An Owner's Guide to the Human Genome

An introduction to human population genetics, variation, and disease

Jonathan K Pritchard
Stanford University
Release 1.0: September 23, 2023

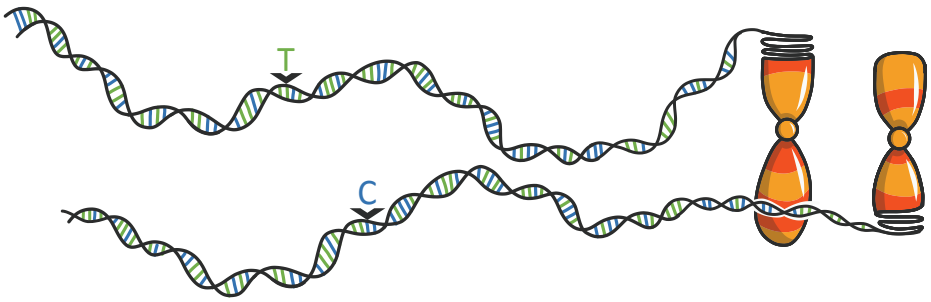


Image Credit: Natalie Telis

Contents

Preface

This book is about the genetic differences between human genomes.

If we sequenced your genome, and my genome, we'd find about five million differences. Most of these differences are SNPs, changing just a single position in the DNA – perhaps you have an A where I have a G. There are also millions of basepairs that are involved in additions, subtractions, or complex rearrangements of chunks of DNA, ranging in size from just a few basepairs to tens of thousands of basepairs.

Human genetics deals with understanding the causes and consequences of all this genetic variation. The story of these genetic variants starts from random mutations – some of which occurred in your parents and grandparents, while others arose more than a million years ago in our ancestral homelands in Africa. These variants have been carried by ancient migrations, and buffeted by the forces of genetic drift, linkage, and natural selection. This pool of genetic variation reflects the history of the human species, and underpins genetic influences on the full range of human traits, behaviors and disease risks.

Key Topics. In this book we seek to understand all the processes above– the forces that govern genetic variation, what they tell us about human history, and the role of genetic variation in human phenotypes and diseases. Specifically, I aim to provide a unified introduction to a set of interconnected topics including:

- The types and distribution of genetic variation;
- Germline and somatic mutation;
- The forces of population genetics: drift, recombination, selection;
- Genetics as a tool for studying human population structure and history;
- The inheritance of human phenotypic variation;
- Large-effect mutations in genetic diseases and cancer;
- The genetic basis of complex traits in health and disease.

Notably, this book combines population genetics, population history, and trait genetics under a single umbrella. While some of these topics are already covered by other introductory texts, there is a tendency elsewhere to separate trait genetics from population genetics. These topics are fundamentally interconnected and cannot be fully understood in isolation.

Human genetics combines modeling, biology, and inference. Aside from the main topics, there are three themes that run through the book: **theoretical models** – usually made precise with math or probability, these provide structure for interpreting complex phenomena; **biology** – everything that we hope to learn about, from the history of human populations, to natural selection, to the molecular mechanisms that link genomic variants to phenotypes; and **inference methods** – often using statistics, these tell us how to extract meaning from complex data.

The role of models. Students of genetics are often surprised by the central role of theoretical models. One remarkable aspect of human genetics, and of population genetics [a], is how often important, deep, principles extend logically from the basic processes of genetics.

[a] ***Population genetics*** *refers to the study of genetic variation in populations, and it will be central to our story here.*

In this regard, I like to think of the founders of population genetics, working in the first half of the 20th Century. They really knew nothing of modern molecular genetics – bear in mind that even the structure of DNA was not known until 1953. And yet they did know some of the most basic rules of inheritance, including:

- the basic rules of Mendelian segregration of alleles and inheritance;
- the existence of chromosomes, linkage, and recombination;
- that mutations can create new alleles;
- that some traits are controlled by a single locus (gene), while others are affected by multiple loci and environmental factors.

Starting from these observations, the founders of population genetics made basic models for the transmission of alleles within families, and within populations. In some models they considered that species have geographic structure, such that individuals are most likely to reproduce with other individuals living nearby. They also considered models of fitness: models where individuals with particular genotypes survive or reproduce better than others, as well as models where the alleles have no effect on survival (so-called *neutral* alleles).

The remarkable thing is that, starting with these very limited observations, they and their successors were able to outline many of the most important processes in population genetics: drift, selection, the role of linkage, and others. Similarly, the genetic principles of plant and animal breeding (known as *quantitative genetics* [b]) were also largely figured out in the 20th Century, again starting from very basic assumptions. Quantitative genetics is also an essential tool for understanding the inheritance of human traits; this will be the focus of Chapter 4.4.

[b] ***Quantitative genetics*** *and* ***statistical genetics*** *study the role of genetic variation in shaping phenotypic variation and will be central themes in Part 4 of the book.*

In other words, these insights from simple models are still fundamentally important today. Much of how we understand aspects of human population genetics is built on top of these basic models. The importance of models here is arguably greater than in any other area of biology.

So does this mean that there's nothing left to learn? Far from it. The early models provided a framework for understanding modern data, but until recently it was impossible to know which aspects of these models would turn out to be most relevant in real life. Moreover, recent discoveries have

motivated important new avenues of theory in many areas.

The role of biology. Again and again, we'll see how modern genomic and functional data illuminate new and unexpected aspects of human history, human evolution, or the genetic basis of human traits and diseases. We'll talk about how genome data has reshaped our understanding of the importance of genetic drift, the modes of adaptation in human population, and the types of genes that have been targets of natural selection in specific environments. We'll discuss the ways in which ancient DNA has transformed our understanding of human prehistory.

We already know far too much to cover everything in one book, but I have aimed to emphasize key concepts that will be useful for understanding the primary literature, as well as interesting examples that illustrate general principles. There's also a huge amount of exciting new work that is adjacent to our main topics, for example in functional genomics and cancer genomics. Unfortunately no single book can cover everything, so my approach will be to give background where needed, while maintaining our focus on genetic variation.

Computer science and statistics. Last but not least, in recent years there has been an enormous growth of new statistical and computational methods for analyzing genome data. Modern experiments often generate terabytes of data, and it takes enormous skill and creativity to extract meaningful biological insight. Data analysis is an essential part of modern genetics. If you're currently a student of genetics, there is perhaps no better single piece of advice than to make sure you become adept at programming and data analysis. (Oh, and I tell students to spend as much time reading science papers as they can!)

My third main goal as we go through each topic is to outline the core concepts that underlie the most important analytical approaches relating to our core topics. This is not a book about statistical methods, *per se*, but these sections should provide a useful jumping-off point for further reading.

Quantitative reasoning in human genetics. In addition to models, another thread that runs through the book is the power of thinking about numbers, scales, and rates, for understanding biology.

I'm reminded of a game in which the quizmaster asks seemingly impossible questions, like: "How many french fries are consumed per day in the US?". The goal is to use logical reasoning to get to a sensible order-of-magnitude answer. For example, you might guess how often an average person eats french fries, and how many they would eat in a typical serving, and then note that there are about 330 million people in the US, to work your way to a reasoned answer.

These types of logic are also very helpful for thinking about genetics and genomics, but rarely taught. I often find in class that students are good at explaining complex molecular mechanisms, yet most have a hard time

thinking about quantitative scales. For example, I might ask "How many heterozygous missense SNPs does a typical person carry?" or "What is the average number of genes spanned by a 1 megabase deletion?". Could you give ballpark estimates?

Throughout the book, I have tried to provide a sense of key numbers and rates to give you intuition for these kinds of questions. I don't want you to memorize every number, but if you can remember rough magnitudes, it will give you a very useful street-sense for thinking about genomics. These can also be very useful for spotting errors when you do data analysis. You'll find a short list of Very Useful Numbers in Chapter 1.1. They may even help you to answer the questions above.

Who is this book for? I hope that this book can be useful for a wide range of readers, from late-stage undergraduates and graduate students, to seasoned experts.

In terms of specific background, I don't assume a great deal of specific technical knowledge of genetics, but certainly the book will be much easier if you are already comfortable with the main concepts of genetics. I expect that a typical reader would already have at least one college-level genetics class, or at least be self-taught to that level. My goal is that students at this stage can use the book as a bridge into the scientific literature. Meanwhile I have tried to write this so that hopefully even most scientists who work in this field will find topics that are new to them.

Similarly, probability, and statistics are also important in human genetics and a basic background in these topics will be invaluable at some points. But to make the book more accessible, I have tried to limit topics that require specific technical knowledge (or sometimes to fence them off into boxes). If you find the mathematical sections difficult, try to use the accompanying text to get the gist of the key points. For readers who want *more* technical detail, you'll find a great deal of useful material and references in the endnotes.

Organization of the material. In the book I have used a mixture of main text and figures, marginal notes and figures, and endnotes to display different types of information. The margin notes contain a mixture of key takeaways, and interesting miscellania (it should be clear from context which is which). Figures that are embedded in the text are usually central to the topic of a paragraph; the marginal figures are usually either for clarification, to show typical data, or for general interest. New terms that you should remember are usually boldfaced and followed by a short definition.

For some topics I have placed either introductory material, or more-complex material – often with math – into Optional Boxes. If you're fairly new to human genetics, you may choose to skip over advanced boxes without losing the general thread; more sophisticated readers should use these for

deeper understanding.

The endnotes contain references to further readings, and sometimes additional comments, caveats and exceptions, or expanded math. Again, readers who are new to the field probably won't need to bother much with the endnotes, but advanced readers can use these as a gateway into relevant literature. The endnotes also serve another purpose. As I write, there is a pair of imaginary readers, one whispering into each ear. On one shoulder, an imaginary student prefers general principles and overall clarity. But on the other shoulder, a specialist grumbles about the many simplifications and exceptions. Some of the endnotes are written to try to cut down on those grumbles.

The material is somewhat cumulative from the first chapter onward, so for example when we're talking about complex traits there are a couple of points where it will be helpful to recall the models of purifying and stabilizing selection. But this overlap is small enough that you'll be ok dipping into particular sections or chapters if you prefer [c].

[c] *Although the material is somewhat cumulative, you will probably be ok if you prefer to focus on specific chapters or sections to learn particular topics.*

Thanks. The structure of this book descends, with modification, from a class for first-year graduate students that I co-taught with Anna Di Rienzo at the University of Chicago from 2002-2013. Since 2013 I have taught elements of this in a variety of classes at Stanford University, at both the undergraduate and graduate levels.

I have learned so much from my many wonderful mentors, colleagues, and trainees over the years, much of which is reflected here. Although there are too many of you to list, I am grateful to you all. I got excellent feedback on rough drafts from many people, including Molly Przeworski and Doc Edge who both read the manuscript twice, additional early advice from John Novembre, Aylwyn Scally, Jenny Tung; also the Edge Lab (Feriel Ouerghi, Shirin Nataneli, Obadiah Mulder, Dandan Peng, Josh Schraiber), the Pritchard Lab (including Alyssa Lyn Fortier, Roshni Patel, Clemens Weiss, Margaret Antonio, and Tami Gjorgjieva), the students of Bio/Gene 247, and Matteo Floris, Guy Sella, Emanuel Goncalves, Vince Buffalo, Ajay Nadig, George Davey-Smith, and Faith Okamoto. I thank the readers for many excellent suggestions, including many that I was not able to implement in this release. And needless to say, any mistakes are my own. Thanks to Lily Leung for getting permissions for use of the copyrighted images. And, of course, I thank my family for their support in all things.

Closing comments. *This is truly an age of discovery in biology, and in human genetics in particular. Every week, fantastic new research papers come online in preprint servers and scientific journals. There is an awesome logic and beauty in genetics, and my greatest hope is that this book will communicate some of this to you.*

Part 1.

An introduction to human genetics

In this first part of the book we cover a collection of topics that will be useful background throughout. This includes

Chapter 1.1: A short list of ***helpful numbers to guide quantitative thinking*** *about many questions in human genetics.*

Chapter 1.2: An overview of ***essential principles in genetics, and how these relate to the human genome****. Some of this will already be familiar to many readers, but the genomic perspective is not usually covered in this way in introductory classes.*

Chapter 1.3: An ***introduction to human genetic variation****: the types of variation, how we quantify variation, and an introduction to how variation affects phenotype. Understanding genetic variation will be the central theme of this book.*

Chapter 1.4: ***DNA sequencing*** *is the fundamental tool for studying genomes, and it's important to understand basic principles about the types of sequencing and the types of data we can collect.*

Chapter 1.5: All genetic variation arose in the past through ***mutation****. We provide an overview of mutational processes and rates, emphasizing topics that are relevant for human genetics.*

1.1 Some Very Useful Numbers for human population genetics.

When you're thinking about genomes and genomic data it's often useful to have a sense of rates and scales[1].

Genome properties.
Genome size: 3.1 Gb (haploid size).
Number of chromosomes: 23 pairs
Number of coding genes: $\sim$20,000
Exons per gene: 8 (median)
Number of genes per megabase: 6.5 (mean)
Total in protein-coding exons: 1% of genome
Total in genes (introns+exons): 40% of genome
Active chromatin (per cell type): 1% of genome
Active chromatin (all cell types): 13% of genome

Length scales. (Orders of magnitude.)
Transcription factor binding site: 10 bp
Enhancer: 100 bp – 1 Kb
Exon (coding): 150 bp
Coding length per gene: 1200 bp (median)
Intron: 1 – 50 Kb
Gene (pre-mRNA): 10 – 100 Kb
Extent of LD: 10 Kb – 1 Mb (varies by locus & population)
Enhancer–promoter interactions: 1 Kb – 1 Mb
Chromatin topological domains (TADs): $\sim$1 Mb
Chromosome lengths: 47 Mb – 250 Mb

Genetic variation.
Heterozygosity: $0.5–1.0 \times 10^{-3}$ (varies by population)
Human-Chimpanzee divergence: 1.4×10^{-2}
Number of common SNPs: $\sim$8 million (at $>$ 5% MAF in global sample)
Number of SNPs for genome-wide SNP tagging: $\sim$1 million
Fst between populations: $\sim$0.10–0.15 between continents

Population genetic parameters.
Mutation rate per generation: 1.3×10^{-8} per basepair (at parental age 30)
Mutation rate per year: 4.0×10^{-10} per basepair
Number of mutations per child: $\sim$70
Recombination rate: 1.2 centiMorgans per Mb (mean, sex-averaged)
Cross-overs per egg: $\sim$42
Cross-overs per sperm: $\sim$26
Effective population size: 10,000–20,000 $(H/4\mu)$

Timescales of population divergence. (Take with grain of salt)
Human to chimpanzee: 6.5 MY
Human to Neanderthal/Denisovan: 600 KY
Deepest human population splits: $\sim$200 KY
Out-of-Africa migration: 65–100 KY
Deepest non-African splits: 65 KY

Notes and References.

[1]Here's an excellent book-length treatment of this topic, with a focus on cell biology, free online [Link]:
Milo R, Phillips R. Cell biology by the numbers. Garland Science; 2015

1.2 A genome owner's starter pack.

A whirlwind introduction to some essentials of the human genome. We emphasize the core function of the human genome as a physical device for storing and replicating biological information.

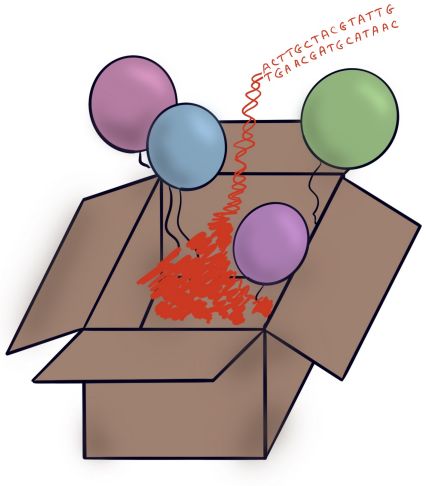


Figure 1.1: *Congratulations on your brand-new human genome! Your custom-made genome has been synthesized for your exclusive use following 4 billion years of evolution.* Lucy Pritchard

A short history. Genetic concepts, and genetic data, are so ubiquitous in modern society that it's easy to forget how recent our understanding of genetics really is.

The origins of modern genetics trace back to the 1850s, when a Moravian monk, Gregor Mendel, used pea plants to learn the most basic rules of genetic inheritance. Mendel's work was published in an obscure journal from Brno (now in the Czech Republic) and was ignored until mainstream scientists rediscovered his manuscript in 1900, thereby kicking off the scientific study of genetics. Thus, genetics had a late start compared to many other scientific fields: for example, cells were first described by Robert Hooke in 1665, and Isaac Newton's theory of universal gravitation was published in 1687.

During the 20th Century, geneticists worked out the basic nuts and bolts of inheritance: phenotypes, mutations, chromosomes and linkage; next, the biochemistry of DNA, RNA and proteins; and ultimately most of the major principles of molecular genetics. Meanwhile, in human genetics, early researchers had learned that certain genetic diseases, like cystic fibrosis or Huntington's disease, are caused by Mendelian mutations in single genes, and by the 1980s they had started to map the genes that are responsible.

At the same time, there was growing realization that most of the ways that humans vary from one another – think of traits like height or weight, diabetes or schizophrenia – are influenced by small contributions from many genes, as well as environmental factors. Last but not least, they had already developed much of the theoretical framework that we use to understand human population genetics today. All of these pieces are central to our story here.

Thus, by the start of the 21st Century, most of the fundamental building blocks of molecular biology and genetics were in place. And yet, at the same time our viewpoint was limited by bottlenecks in measurement: most notably DNA sequencing. The last two decades have seen a revolution in all kinds of biological measurement, but especially of DNA genotyping and sequencing.

Until very recently, it was extravagantly expensive to sequence human genomes: the Human Genome Project completed the DNA sequence for a single genome in 2003, at a cost of $3 billion [2]. With newer, highly efficient technologies, the cost to sequence someone's genome is now well under $1000. These technical advances have ushered in a revolution of human genetics. As of 2022, hundreds of thousands of people have had

their genomes sequenced for research or medical applications. Meanwhile, genome-scale data (SNP genotypes) have been collected for tens of millions of people.

In research settings, high-throughput sequencing is now a universal routine tool. Sequencing has enabled major new insights into the genetic basis of inherited traits, and cancers. Ancient DNA sequencing has revealed important new storylines about the origins and evolution of modern humans. Genome sequencing has also revolutionized our ability to study the *functions* of genomes including our ability to measure which parts of a genome are active in any given cell type. We'll touch on all these topics in later chapters.

The new technologies have also allowed the general public to interface with genetic tools for the first time: millions of people have sent their DNA samples to personal genetics companies that promise insights into customers' family trees, their ancestries, and perhaps even their genetic predispositions. DNA forensics has become an essential part of the criminal justice system, connecting suspects to crime scenes (or exonerating them), and recently using genetic genealogy tools to solve a large number of "cold" cases that had been unsolvable by traditional methods. The use of genetic data in medicine is steadily increasing: millions of mothers have received prenatal genetics screening to provide early detection of chromosomal abnormalities; genome-sequencing is now an important tool in cancer treatment; we are on the cusp of genetic prediction in clinical medicine; techniques like genome-editing with CRISPR and cellular reprogramming promise to transform the role of genetics in medicine. And of course, many aspects of genetic research came together to enable the rapid development and approval of mRNA-vaccines against COVID-19 in 2020.

Genomes, inheritance, and variation. Most of the topics above relate to human genome variation, which is the focus of this book. To understand genetic variation, it's first helpful to think about the genome as a device for storing data, and encoding biological functions. The data stored in your genome (or mine) are inherited from our species' shared ancestors in Africa, via many thousands of generations of mutation, genetic drift, and natural selection. Looking further back into history, your genome is also inherited, albeit with massive modifications, through billions of cell divisions from single-celled ancestors that lived near the beginning of life on earth, some 4 billion years ago.

In the next sections, we look at how DNA stores and encodes biological information [a].

[a] ***Parts of this introductory chapter may be familiar to you already, so feel free to skip over those!***

The DNA molecule. Your genetic data are stored using a molecule called **DNA**, short for deoxyribonucleic acid.

The DNA molecule is shaped like a twisted ladder. Each side of the ladder is called a **strand**, and is made up of four different kinds of chem-

ical building blocks called **bases**: namely **A, C, G, and T** (for adenine, cytosine, thymine and guanine). Along each strand, the bases are linked together by a chemical backbone. A base plus its chunk of backbone is called a **nucleotide**. The distinction between base and nucleotide is not especially important for us here, and we'll use the terms somewhat interchangeably.

The two strands of the DNA molecule fit together with what's called complementary base-pairing: specifically, A on one strand is always matched with T; and C with G. This means that we can have four kinds of rungs: A:T and T:A; C:G and G:C, depending on which base is on which side of the ladder. One rung of the ladder–i.e., 2 bases from opposite strands–is called a **base-pair**.

Another key feature of the strands is that they have a natural direction–analogous to how we always read English from left to right. Each nucleotide is asymmetric, with a so-called 5′ side and a 3′ side (pronounced "5-prime" and "3-prime"). In a DNA strand, all the 5's are oriented in the same direction, so we can label one end of a strand as 5′, and the other end as 3′. Meanwhile, the other strand of the helix is oriented in the opposite direction.

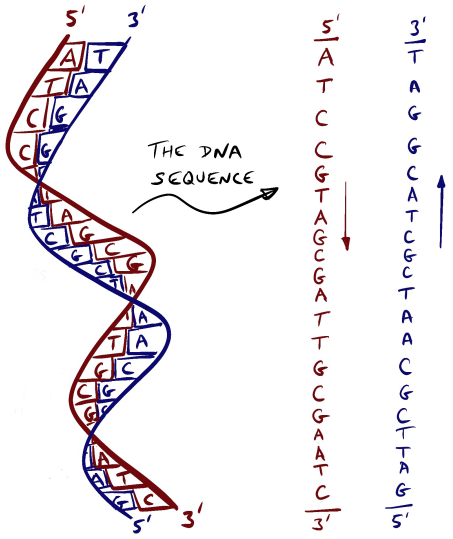


Figure 1.2: *The nucleotides fit together to form complementary strands of DNA, like a twisted ladder. Some of the nucleotides shown in the sequence are obscured at the points where the helix turns perpendicular to the page. Genes can be encoded on either strand, but always in a 5′ → 3′ direction on the relevant strand.*

Again, similar to English which is always written left to right, everything important in genetics happens 5′ to 3′. DNA replication occurs 5′ to 3′. And the copying and decoding of genes – transcription and translation – is from 5′ to 3′. Genes can be encoded on either strand, but since the two strands of a double helix are oriented in opposite directions, genes encoded on one strand are oriented opposite to genes encoded on the other strand.

23 pairs of chromosomes. The DNA in your genome is organized into chromosomes. You have 23 pairs of chromosomes–you got one copy of each chromosome from your mum [b], and one from your dad for a total of 46. That includes 22 regular pairs, creatively named Chromosome 1 to Chromosome 22, roughly in order of size, as well as the sex chromosomes X and Y: biological females have two Xs; males have an X and a Y. The smallest chromosome is actually 21, and not 22, because early studies put them in the wrong order, and the original numbering has been kept.

[b] *In this book, we'll generally use terms relating to sex and parental relationships as a shorthand for referring to biological sex and genetic relationships, while noting that these terms simplify a complicated reality* [3].

Chromosomes 1–22 are referred to as **autosomes** when we want to distinguish them from the sex chromosomes. The two copies of each chromosome pair are referred to as **homologous chromosomes**, or simply **homologs** for short.

Altogether, you have about 3.3 billion base pairs of DNA from each parent: 6.6 billion total in every cell. If we laid out the DNA from a single cell in a straight line, it would be over 2 meters long [c] [4]. Obviously the DNA is not stored in a straight line: instead it's wrapped around small balls of protein called **nucleosomes**, much like spools of thread. The spools themselves are also packaged in an orderly fashion, to fit the whole lot into the cell's nucleus. Together, this highly compact DNA-protein packaging is referred to as **chromatin**, and is the default state for our

[c] *Your body contains ∼40 trillion cells, nearly all of which carry a copy of your genome. If we stretched out all the DNA from all your cells end to end, it would span across the solar system!*

genomes. If you do need to read part of your genome, you'll briefly unspool the bit that you need.

DNA is the world's greatest data storage device. The central role of DNA is to store biological information. Chromosomes are long strings of A, C, G, and T that encode information, sort of the same way a book does, but using 4 letters instead of 26. In a minute, we'll talk about how this information is encoded, but for right now, just pause for a moment to think about the fact that this book of 6 billion base pairs provides complete instructions for all the proteins needed to make each of your cells– and in fact all the proteins you need to assemble a complete person.

As a loose analogy you can think of an organism's genome as being like the computer code (operating system and software) that controls a computer. If you want to make a human or a dog, a worm or a melon, you need to input different strings of DNA. And like computer code, which relies on computer hardware (your phone or laptop) to produce a physical output, a genome relies on elaborate cellular machinery that reads it and interprets it to produce biological functions. Later we'll also talk about the important role of environment in shaping phenotypes.

DNA is an incredibly impressive storage system. If we put this in terms of data storage on a computer, a single human cell carries the equivalent of about 1.5 Gigabytes of computer data [d]! (In computing, a *bit* is a single binary digit that is either zero or one; the basic unit of data storage is a *byte*, which is 8 bits. Since DNA has four possible letters, each base pair is the equivalent of 2 bits, and 4 DNA base pairs carry one byte of data.) So just a few hundred cells carry as much data as your phone – though to be fair, it's the same information repeated in every cell. At the same time, cells are so small that you could actually fit about 100 million cells inside a standard phone.

[d] *Not only is DNA storage physically compact, but it might also seem surprising that the information content of a diploid human genome is also modest compared to modern computing systems, at just 1.5 gigabytes. For comparison, the iPhone operating system is somewhat larger at 2–3 gigabytes, depending on version.*

Indeed, DNA is such an efficient and stable storage system that there is a line of research on how to use DNA to store digital data. DNA is far more compact than computer storage systems, but currently much slower and more expensive for humans to read or write DNA compared to conventional systems [5].

The genomic encoding of biological information. For biological systems, the importance of DNA is that it encodes biological information. One major challenge in genome science is to be able to read the encoded information. What does each of the 3.1 billion base pairs do – if it does anything at all? What would be the impact of a mutation that changes any specific part of the genome sequence?

Soon after the Human Genome Project was completed in 2003, at a cost of 3 billion dollars, one project leader, Eric Lander, famously gave this terse summary of the challenges ahead [6]:

"Genome. Bought the Book. Hard to read."

You can think of the genome as encoding two main types of information: **The first kind of information is contained in genes**. A gene is a stretch of DNA that encodes a protein. (A small fraction of genes encode functional RNAs instead of proteins, but the principles are similar.)

The second kind of information tells each cell how much of each protein to make. This is referred to as **gene regulation**, and is also critically important. For example, the differences between a liver cell, a neuron, or a muscle cell are mainly due to precisely-controlled differences in gene regulation across these cell types.

As we'll see shortly, these two types of information are encoded very differently. Genes encode proteins using a very simple format where each successive block of 3 nucleotides specifies an amino acid.

In contrast, gene regulation is controlled by molecular interactions between DNA sequences and cell-type specific proteins. The language of gene regulation is both highly complex, and highly context-specific: a particular sequence may be interpreted as an important regulatory region in a liver cell, and completely ignored by a neuron. Consequently, while the general principles of gene regulation are fairly well understood, it is still a difficult research problem to predict how a particular DNA sequence will be interpreted in any given cell type. Luckily, it's possible to create accurate maps of regulatory regions using a variety of experimental assays.

Genes and the encoding of proteins. Each gene stores the instructions to make a particular **protein**. If DNA is the information storage device in cells, proteins are the molecules that actually get things done. Much of biology is controlled by different proteins doing different kinds of jobs in cells. (I don't mean to trash-talk the other essential biomolecules, such as lipids – but they are not directly encoded by the genome, and will be a much smaller part of this book's story.)

Even though proteins perform a huge variety of different jobs, they are all made up of the same basic building blocks. These building blocks are small molecules called **amino acids**. Your genome encodes 20 different amino acids, which can be joined together in any order to make a protein. What a protein does is determined by the specific order, and number, of its amino acids. Proteins vary greatly in size, but the average protein in humans is about 400 amino acids long.

Unlike DNA, proteins fold into an enormous diversity of shapes, depending upon their amino acid sequences, and this is part of what determines their biological functions. There is a major field of biology devoted to measuring, and even predicting, 3-dimensional protein folding, and how each protein interacts with other molecules in cells [7].

The genetic code. DNA specifies proteins using a simple code, in which a nucleotide sequence along one strand of the helix encodes a sequence of amino acids.

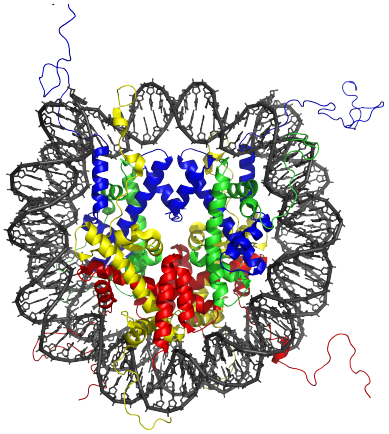


Figure 1.3: **Example of a protein structure.** *Most of your genome (DNA shown here in black) is wrapped around protein complexes called* ***nucleosomes*** *(colors), like thread on spools.* *Credit: Zephyris CC BY-SA 3.0 [Link]*

Remember that DNA is made up of 4 letters: A,C,G,T. So how does DNA encode the 20 amino acids? Just like words in a book, we need more than one letter to encode each amino acid. If we used pairs of adjacent letters (here we mean adjacent on the same strand of the helix), there would be $4^2 = 16$ possibilities: AA, AC, AG, AT, CA, CC, CG, CT, GA, GC, GG, GT, TA, TC, TG, TT. Hmmm. That still isn't enough to code for 20 amino acids. So we're going to need three adjacent letters for each amino acid, as that gets us to $4^3 = 64$ possibility combinations. So for example, AAA in DNA codes for the amino acid Lysine in the corresponding protein.

Now that gives us 64 combinations when we only really need 20. So what does the cell do with the 44 extra triples? Well, three triples, TAA, TAG, TGA, are STOP signs, marking the end of the protein. And beyond that, there is redundancy, so that most amino acids are encoded by multiple triples: eg TGT and TGC both code for the amino acid Cysteine. The other special signal is ATG, which signals as a START sign when it occurs at the beginning of a protein. ATG also encodes the amino acid Methionine. Each block of three nucleotides is called a **codon**.

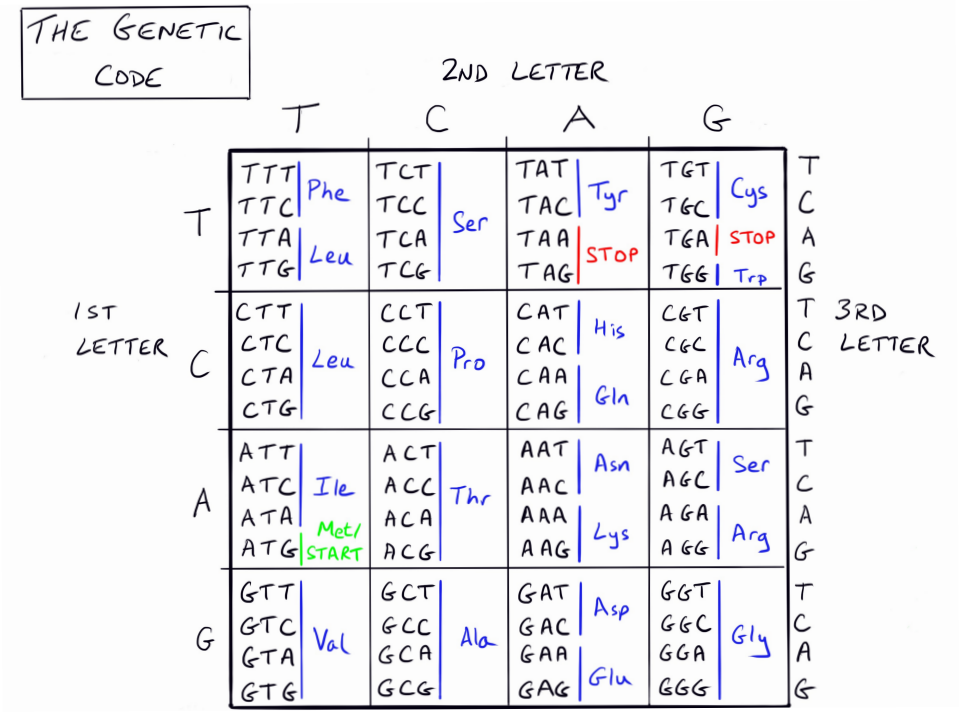


Figure 1.4: *The genetic code: this shows the encoding of DNA triplets for amino acids. The 64 possible DNA codons are shown in black, and the corresponding amino acids are shown in blue using their abbreviations. ATG signals both the protein START and the amino acid Methionine. TAA, TAG, and TGA are protein STOP codes.*

Abbreviations for the amino acids:
Ala: Alanine; Arg: Arginine; Asn: Asparagine; Asp: Aspartic Acid; Cys: Cysteine; Glu: Glutamic Acid; Gln: Glutamine; Gly: Glycine; His: Histidine; Ile: Isoleucine; Leu: Leucine; Lys: Lysine; Met: Methionine; Phe: Phenylalanine; Pro: Proline; Ser: Serine; Thr: Threonine; Trp: Tryptophan; Tyr: Tyrosine; Val: Valine.

This code for translating from DNA to protein is called the **genetic code**. It's interesting that this code is nearly identical in all living things. For example, most bacteria have exactly the same code as humans. There is no fundamental reason why AAA should encode the amino acid Lysine– it just started that way in the first cells to evolve a genetic code, and has been inherited throughout the tree of life ever since, during the last 4 billion years of evolution. Notable exceptions to the "universal" genetic code can be found in the tiny genomes carried by our mitochondria, which encode four of the codons differently [8].

Once we know the genetic code, the encoding from DNA to protein is remarkably simple: it starts with ATG, and then every successive 3-nucleotides encodes a single amino acid until we reach the first STOP. (There's a minor complication, which we'll get to shortly, that blocks of DNA called introns are removed before the protein is decoded.)

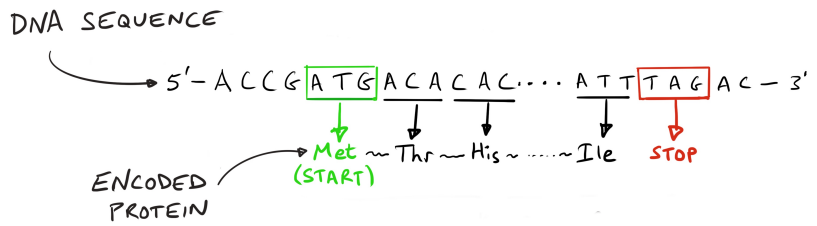


Figure 1.5: **The encoding from DNA to protein.** *The amino acid sequence is interpreted as starting from the first ATG and continuing by threes until the first STOP codon. The DNA sequence shows the coding strand only.*

You can imagine that **this encoding is fragile**, in the sense that just changing a single nucleotide can potentially alter the protein almost completely: for example a mutation that introduces an early stop signal will cause the protein to be immediately terminated; similarly, insertion (or deletion) of a single nucleotide would cause the reading frame of the protein to shift and, from that point on, to encode a completely different amino acid sequence. As we will see in future chapters, both of these types of mutations do occur: they generally cause complete loss-of-function of the affected protein, and depending on the protein, they are often highly deleterious.

DNA → mRNA → Protein. DNA is not interpreted directly into protein, but instead it is first copied into an intermediate called **messenger RNA** (mRNA). RNA is a molecule that is very similar to DNA, but it is usually only one strand of the helix, and is less chemically stable for long-term storage [e]. Note that RNA uses a base called Uracil (U) everywhere that DNA uses Thymine (T). This flow of information from DNA → mRNA → protein is known as the **Central Dogma**.

[e] *Even though RNA is less stable than DNA some viruses, including HIV and the virus that causes COVID-19, use RNA as their main storage molecule instead of DNA.*

Transcription. DNA is stored within the cell's nucleus. In order to make a protein, your cell unwraps the bit of the genome that encodes that gene, and makes mRNA copies of the DNA. This process is known as **transcription**–meaning copying.

Translation. mRNA copies are then transported out of the nucleus into the cell's cytoplasm, where molecular machines called **ribosomes** assemble proteins, using the mRNA sequence as a template. This process converts the biological information from the four-letter alphabets of DNA and RNA into the twenty-amino acid alphabet of proteins. This process is known as **translation**, reflecting the conversion from one type of information (DNA/RNA) into another (protein):

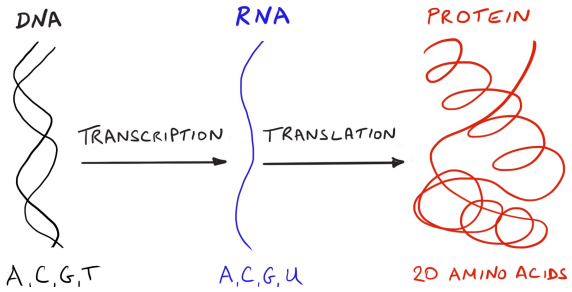


Figure 1.6: **The flow of genetic information.** *DNA provides permanent information storage for cells; mRNA serves mainly as a temporary molecule, used as a template for translation; proteins are highly versatile molecules that perform a wide range of functions. As shown below, the three molecules use different alphabets.*

At this point I should confess that despite the grandiose title of the Central Dogma, a small fraction of genes don't seem to know about this rule, as they produce functional RNAs instead of proteins: for example, some RNA genes encode essential components of the ribosome, and another RNA gene is responsible for inactivating one of the X chromosomes in females [9] [10]. You may be getting the (correct) sense that virtually every rule in biology has exceptions!

Gene structure: UTRs, Exons, Introns, and Splicing. So far, we've been talking about the part of an mRNA that encodes a protein. But this is actually embedded in a much larger transcribed region.

Transcription begins from a location called the **Transcription Start Site**, and terminates at the **Transcription End Site**. The initial immature transcript is referred to as a **pre-mRNA**.

Almost immediately (usually starting during transcription), another key process takes place, in which regions called **introns** are **spliced** (cut) out of the pre-mRNA to produce a shorter, processed mRNA. After being cut out, the introns are trashed, and the nucleotides are recycled. As we'll see shortly, introns are usually much longer than exons, and the final mRNA is usually just a few percent of the initial pre-mRNA [11].

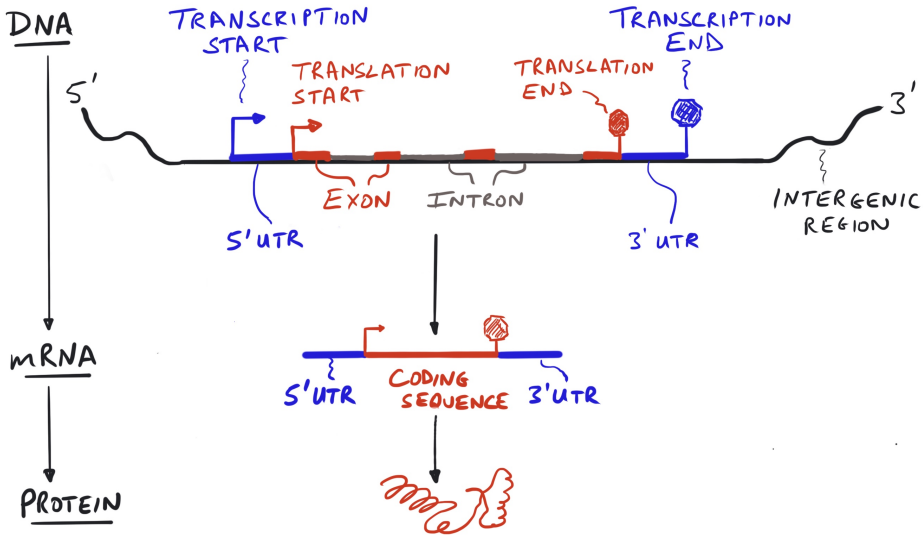


Figure 1.7: **A typical gene structure.** *Transcription initially includes 5′ and 3′ UTRs, coding exons, and introns. The introns are rapidly removed to create the processed mRNA, prior to translation.* ***This is not drawn to scale:*** *typical introns are 10× to 100× larger than exons.*

The final, processed, mRNA is transported from the nucleus into the cytoplasm, where translation takes place. The translation machinery finds the first available start codon (ATG): this is the **Translation Start**. It then

proceeds until it finds the first in-frame STOP signal: the **Translation End**. The regions upstream and downstream of the coding region are known as the **5′** and **3′ Untranslated Regions (UTRs)**. The UTRs often contain information that is used to target the mRNA to particular locations in the cell, or for other forms of regulation.

One advantage of splicing is that it is possible to make different protein products from the same gene, by including or excluding different combinations of exons, or by using different splice sites. This is known as **alternative splicing**, and the different protein products are called **isoforms**. For some genes, distinct isoforms are critical for creating functional diversity of proteins from the same transcripts [12].

Splice site specification. Given that the exons are joined together from a much longer pre-mRNA, this immediately raises another question: how does the splicing machinery know where to cut? What marks the positions of the exon-intron boundaries?

This information is encoded in the DNA (and hence the pre-mRNA, which is what the splicing machinery is actually interacting with) using a variety of signals. First, the splicing code requires a GT at the start, and AG at the end of nearly all human introns (GU and AG in the pre-mRNA). As we'll see in Chapter 1.3, any change in the GT or AG forces splicing to occur at another position – this can dramatically change the encoded protein and can have devastating consequences.

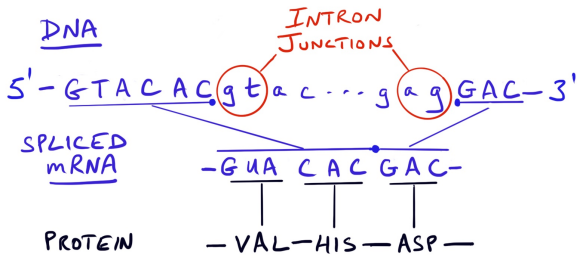


Figure 1.8: **Correct splicing relies on positioning signals encoded in DNA.** *The figure indicates intronic nucleotides with lower-case text. Mutations in the 5′ gt or 3′ ag of the intron disrupt splicing and often result in a nonfunctional protein. T in DNA is U in RNA.*

But of course, there are many GT and AG sites in a typical intron (each of these occurs roughly once every 16 base pairs). So in addition to these required features which help to position the precise splice site, the exon positions are also indicated by a combination of weaker sequence-based signals, where no single nucleotide is fully required for correct splicing: for example the region upstream from the AG usually contains Cs and Ts, as well as an upstream A that is involved in cutting out the intron but can occur at variable distances. These, and other, signals help to position the splice site.

These weaker signals that help position splicing events are very different from the simple and precise algorithmic rules that encode proteins; instead they are more similar to the sequence elements that control gene regulation – which we'll discuss next. The goal of understanding the determinants of splicing is an active research area, using both experiments and machine learning [13].

The encoding of gene regulation. Aside from coding proteins, there's a second kind of information stored in DNA. This information tells a cell which RNAs and proteins to produce, and in what quantities. The production of specific RNAs and proteins is called **gene expression** and the controls of expression are called **gene regulation**. Gene regulation is encoded in the genome, and it turns out that the encoded regulatory information is just as important as the protein-coding sequences themselves.

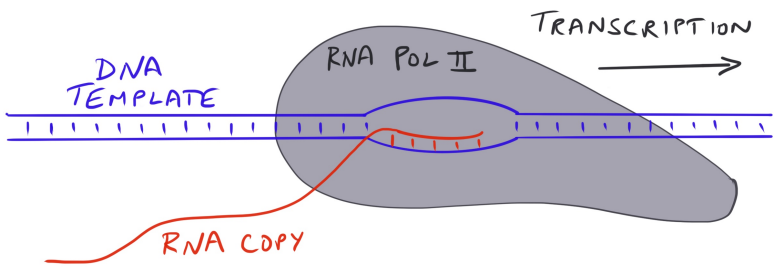
To understand why gene regulation is so important, it's helpful to reflect on the fact that we are immensely complex multicellular organisms. Think about all the different types of cells in a human body: skin cells, heart cells, liver cells, neurons, sperm and eggs, and hundreds of others. These cell types do very different jobs, and look different under a microscope. It turns out that every cell type also expresses a characteristic portfolio of mRNAs and proteins – and this portfolio is a large part of what gives a cell type its identity.

Moreover, cells must produce mRNAs and proteins in very precise proportions (a bit like mixing ingredients for baking). Consistent with this, many genetic diseases are caused by disruptions in the relative proportions of expressed genes [14].

So how is this precise regulatory information encoded in the genome? And, even more strikingly, how does a cell know whether to express the portfolio of genes required for a liver or a neuron or a muscle, even though every cell carries essentially the same genome?

To understand these questions, we first need to detour into some brief details about how gene regulation is encoded in the genome.

The major focus of gene regulation is on controlling transcription. Genes are copied into mRNA – i.e., transcribed – by a protein machine called RNA Polymerase II (**Pol II** to its friends, pronounced "pol-2"). Prior to transcription, Pol II assembles in a region of DNA at the start of the mRNA, known as the **promoter**. The assembly is guided by a set of additional proteins that, along with Pol II form a so-called **Pre-Initiation Complex** within the promoter region. Once Pol II has been assembled at the promoter, it attempts to initiate transcription; if that is successful, Pol II then chugs along the gene, at a speed of ~2 kb/minute [15] to produce an mRNA copy of the DNA [16].



We'll skip over many interesting details about the molecular biology of transcription – but for our story here, the key question is to think about

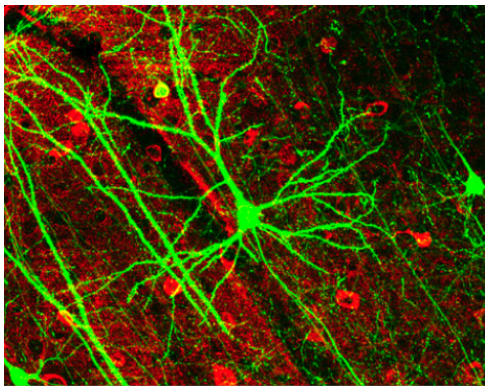


Figure 1.9: **Gene regulation controls the differentiation of cells into many distinct types.** *In this image, two types of neurons in mouse cerebral cortex are stained red and green, depending on whether each cell produces GABA, a key neurotransmitter.* *Credit: Fig. 6F of Wei-Chung Allen Lee et al (2005); [Link] Creative Commons License.*

Figure 1.10: **Transcription:** *RNA Pol II makes an mRNA copy of the DNA. Gene regulatory information is responsible for controlling transcription rates of each gene in specific cell types and conditions.*

the DNA sequences that direct transcription. Crucially, how do DNA sequences position the pre-initiation complex? This determines where transcription will start. And how do DNA sequences control the rate of Pol II assembly and transcription [17]?

These decisions are guided in large part by proteins called **transcription factors (TFs)**. Most TFs have a DNA binding domain that attaches to the genome at specific sequences (**transcription factor binding sites**) [18]. Meanwhile, other parts of the same TF can interact with other proteins to help increase, or sometimes to repress, transcription. As an example, the image below shows the molecular structure of a TF called AP-1, where the purple region of the protein is bound to DNA:

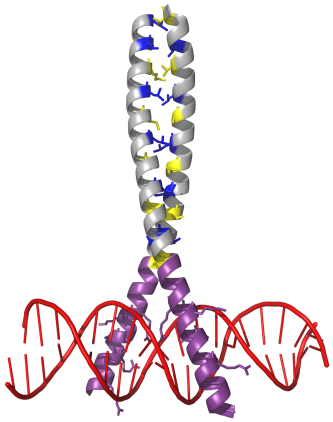


Figure 1.11: **Transcription factor binding to DNA.** *Most TFs have a DNA-binding domain (shown here in purple); other parts of the protein structure can interact with other proteins to control transcription.* *Credit: Houq [Link] CC-BY-SA-3.0*

TF binding usually takes place both within the promoter region itself, as well as at more distant locations called **enhancers**. Enhancers are regions of TF binding that are situated outside the promoter. When TF binding occurs at the enhancers, the DNA can form a loop to bring the enhancer into close physical contact with the promoter. These enhancer-promoter interactions can be essential for assembling the Pre-Initiation Complex which includes Pol II, prior to transcription:

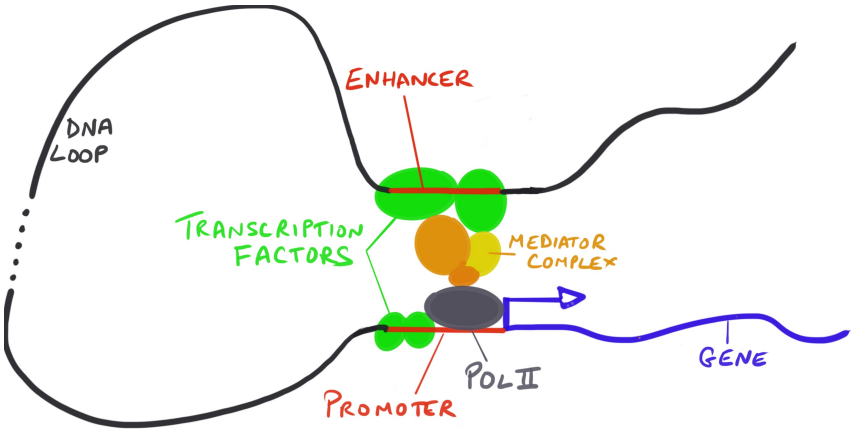


Figure 1.12: **Enhancer-Promoter interactions help drive gene expression.** *Pol II assembles in the core promoter prior to initiating transcription. It is stabilized by protein-protein interactions with TFs bound both within the promoter and at distant enhancers. Promoter-enhancer proteins may attach through protein bridges such as the Mediator Complex.*

Enhancers are often located quite far in DNA distance from the promoters they regulate – usually at distances of tens of thousands up to a million base pairs away, but loop around to create physical proximity.

So what specifies the locations of transcription binding sites, and by extension, of promoters and enhancers? At positions where the protein contacts the DNA directly, each TF has preferred binding sequences that reflect physical interactions in the contact zone between amino acids in the TF, and the nucleotides in the DNA. These preferred sequences are called **binding motifs** – for example, the image below shows the preferred binding sequence, TTTGCAT, for the TF Oct1:

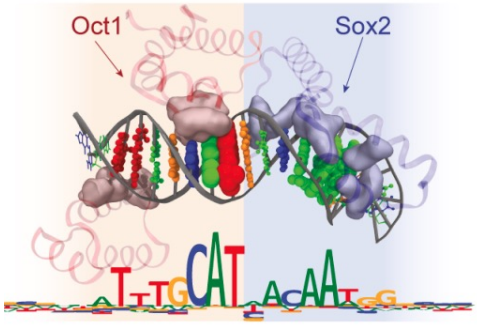


Figure 1.13: **Gene expression is largely controlled by transcription factor (TF) binding of DNA.** *The image shows binding of the Oct1 and Sox2 proteins to DNA (these two factors often bind jointly). The letters below the plot provide a graphical representation of which nucleotides at each position are preferred for binding: larger letters are more important.* *Credit: Fig 3a from Žiga Avsec et al (2020) [Link] CC-BY-NC 4.0.*

However, these binding motifs are neither necessary nor sufficient to predict binding. First, even within TF binding motifs, most nucleotides are not strictly required for binding. In the image above, the sizes of the letters indicate the importance of each position for binding: the largest letters, CAT, are found within most Oct1 binding sites, but the other positions are more variable. Second, since these binding motifs are quite short, they are found found many times in the genome. Most TFs bind only a tiny fraction of all the possible motif matches.

Instead, the specificity of TFs to bind in the correct locations is usually controlled by combinations of factors binding adjacent DNA sequence elements: for example, very often binding is stabilized when ensembles of multiple TFs can bind in a small region [19]. The specific rules that control TF binding are highly complex and vary across cell types; development of computational tools for predicting TF binding sites is an important research area where machine learning techniques have started to make huge progress from around 2015 onward [20].

A related puzzle is that enhancers act by DNA *looping* to create physical interactions with promoters. How do enhancers "decide" where to loop to? While there is a tendency for enhancers to interact with the nearest promoter(s), there are exceptions in which enhancers ignore nearby genes in favor of regulating genes as far as a million base pairs away [21]. The controls of looping are poorly understood at present [22].

Cell type differences in regulation. Lastly, I want to touch briefly on a remarkable feature of genomes. All of your many cell types carry essentially the same genome, and yet they can interpret it differently to produce different portfolios of genes, and these give different cell types their unique identities: for example T cells, or liver cells, or neurons.

The regulatory logic that I've described above starts to hint at how this is possible. Cell type identities are controlled in large part by which en-

hancers are active (and hence which genes are expressed). Enhancer activity, in turn, is controlled by combinations of transcription factors binding. The key point here is that different cell types express different sets of TFs; thereby turning on (or off) different enhancers across the genome.

But how do cells "know" which TFs they should express? In embryonic development, the earliest cells can produce any possible cell type, but as the organisms develops, cells become increasingly specialized. This specialization is controlled in large part by turning off embryonic TFs, and turning on other TFs that are specific to particular cell lineages. The lineage-specific TFs drive programs of gene expression that are appropriate to the corresponding cell types.

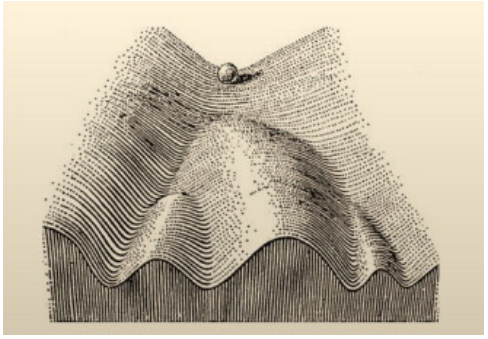


Figure 1.14: **Waddington's landscape metaphor for cellular differentiation (1957).** *Conrad Waddington imagined the increasing specialization of cell types during development as like a ball rolling down a slope. As it rolls it makes random choices that restrict it to increasingly narrow gullies; in cellular terms, we can think of this as turning on lineage-specific transcription factors that drive cell-type relevant programs of gene expression.*

In summary, the encoding of gene regulation works on very different principles than the encoding of proteins. First, gene regulation is analog (i.e., expression level is a continuous variable), unlike protein coding, which is digital (each codon sequence encodes exactly one protein). Secondly, the encoding of expression is controlled by the aggregate effects of many nucleotides and is robust in the sense that single nucleotide changes in the sequence generally have small effects on expression; in contrast, single nucleotide changes such as premature stop codons can completely break protein function.

In the last few pages we have discussed that two major categories of information stored in genomes include the encoding of genes (i.e., mainly proteins); and the encoding of regulatory information (when and how much to make each protein). *How is this information organized in our genomes?*

Bloated genomes: the good, the bad, and the ugly. Remarkably, only about 1% of the genome encodes proteins. A somewhat larger amount codes for regulatory sequences – perhaps around 10% – although the precise amount is uncertain due to the cryptic nature of gene regulatory elements [23]. But most of the remaining ∼90% of the genome sequence shows no clear evidence for biological function. What's there?

To start addressing this question, it'll be useful to have a rough sense of the landscape of genomes and functional elements.

Measurement units of DNA sequence. We'll often need to measure lengths of DNA. The natural units of sequence length are in terms of base pairs but it's convenient to abbreviate different scales with different units (similar to how we use milligrams, grams, and kilograms). So you'll want to remember that:

- 1 bp = 1 basepair
- 1 Kb = 1 kilobase = 10^3 bp
- 1 Mb = 1 megabase = 10^6 bp
- 1 Gb = 1 gigabase = 10^9 bp

Chromosome sizes. The human genome is about 3,100 Mb = 3.1 Gb. Most cells have two copies of the genome (one from mum and one from dad), so that's a total of 6.2 Gb. The chromosomes range in size from 250

Mb (Chromosome 1) down to 47 Mb (Chromosome 21). The mitochondrion has its own genome, contained in a small circular molecule of 16 Kb. For comparison, the genome of SARS-Cov2, the virus that causes COVID-19, is about 29.9 Kb.

Gene numbers and sizes. Meanwhile, the genome contains about 20,000 protein-coding genes [f] (estimates range from about 18,000–22,000 depending on how strict the criteria are that each gene is translated and/or functional). This works out, on average, to about 6.5 genes per Mb, although the distribution of genes is highly uneven.

To give you a sense of scale, the median length of a protein-coding gene, including introns, is 27 Kb [24]. Meanwhile, the coding length is much shorter, with a median length of 1.2 Kb (400 amino acids). A typical gene has 8 exons, and the median size of a coding exon is 122 bp. Introns are more than ten-fold longer, with a median size of 1,600 bp, and a mean of 6,300 bp. Coding exons and UTRs occupy only about 2.5%, each, of the average pre-mRNA before splicing.

You can see an example of a genome region, in a screenshot from the Ensemble Genome Browser. Known protein-coding genes are marked in orange. The vertical bars and boxes are exons or UTRs; horizontal lines are introns. Other possible genes are marked in grey and purple (in practice most of these are likely nonfunctional) [25]. This region is fairly typical, except that the gene density is about twice the genome average.

[f] *Before the human genome was completed, most genome scientists expected that there would be many more genes. In the year 2000, the British scientist Ewan Birney organized a betting pool to guess how many genes there would be. The mean guess was over 60,000; the winner was Lee Rowen who had the lowest guess (24,800) out of more than 460 bets [Link].*

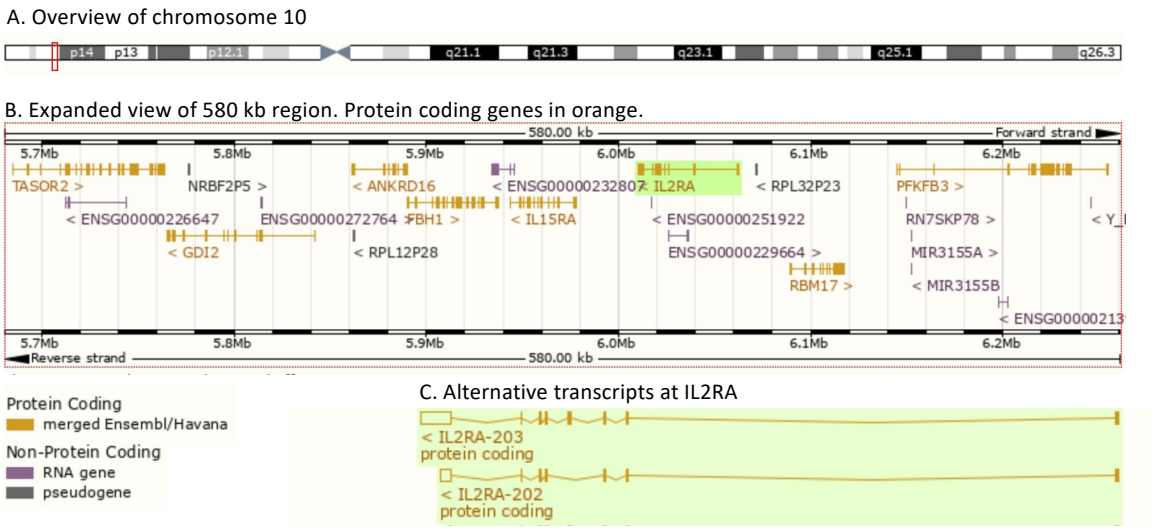


Figure 1.15: Genome browser view of the genome. *Screenshots from the Ensembl Genome Browser show a gene-dense region around IL2RA (an important immune gene).* **A.** *The IL2RA region is marked by the red box at the left end of the chromosome.* **B.** *Coding genes are marked in orange. '>' or '<' mark the direction of transcription of each gene (i.e., whether it is coded on the forward or reverse DNA strand).* **C.** *Expanded view of two possible transcripts at IL2RA. Coding exons and UTRs are marked by filled/open boxes respectively.*

Source: Ensembl browser [Region][Transcripts]

So if only about 1% of the genome is protein coding, and a small fraction of the rest (~10%) encodes regulatory information, then what is all the rest of the genome doing?

Remarkably, most of the genome doesn't have any clear function.

Indeed about two thirds of the genome is made up by **repetitive DNA**: short sequences of hundreds to thousands of base pairs that are repeated many times in the genome. A few of these repetitive elements are involved in gene regulation, but most don't do anything useful for you. In fact they are sometimes referred to derisively as "junk DNA". (These elements really need to hire a better PR team.)

The single most common repetitive element is a 300 base pair sequence called an Alu element, which occurs about 1 million times in the genome! In other words, about 10% of the storage space of the human genome is given up to recording Alu elements – this about 10 times as much space as we devote to storing all genes.

Alu is a type of **transposable element (TE)**. TEs are DNA elements that can copy themselves and reinsert elsewhere in the genome. They are usually considered **selfish DNA**, meaning that they proliferate due to their ability to replicate, while having little or no value to the host genome – in short, they are genome parasites. In the case of Alu, it first infected the genome of our ancestors about 65 million years ago, and has been wildly successful in spreading itself around the genome since then. In some cases Alus and other repetitive elements have evolved new functions, but most are essentially inert elements. These must be copied every time a cell divides, but most do not contribute to genome function.

It is outside our scope but there is fascinating work on mechanisms that have evolved to prevent transposable elements from spreading in the genome, and the ways that TEs evolve to evade those mechanisms [26]. At the same time, there is great work on TEs that have been "domesticated" by host genomes to serve as protein domains or regulatory elements [27].

Genome sizes and TEs. The genome sizes of different organisms vary enormously, as you can see in the table below. Notice the switch from measuring genomes in megabases (Mb) to gigabases (Gb) partway through the table. There's about a 10,000-fold difference in genome size between E coli and Axolotol, even though the numbers of genes varies by less than a factor of 10.

Organism	Genome Size	Number of genes
E. coli (bacterium)	5 Mb	4,000
S. cerevisiae (yeast)	12 Mb	6,000
C. elegans (nematode)	101 Mb	20,000
A. thaliana (flowering plant)	135 Mb	27,000
D. melanogaster (fly)	175 Mb	15,000
human	**3.1 Gb**	**20,000**
Picea abies (spruce tree)	20 Gb	28,000
Axolotl (salamander)	32 Gb	23,000

Table 1.1: Haploid genome sizes for representative organisms. *Notice the enormous range of genome size (by a factor of $\sim 10^4$), while gene numbers vary by less than a factor of* 10*. The largest genomes are cluttered with repetitive DNA. Gene numbers are approximate and are for protein-coding genes.*

Naively, one might perhaps have expected that the genome-sizes of or-

ganisms would reflect their complexity. While it's true that single-celled organisms generally have smaller genomes and fewer genes than multi-celled organisms, there is no very clear pattern of genome-sizes beyond that. (It may not be entirely clear how to measure organismal complexity but, for many of us, it might hurt our feelings to be told that axolotls are ten-fold more complex than we are.) But actually a major determinant of genome-size is how active TEs have been in each evolutionary lineage.

The inheritance of genomes. So far we've been talking about genomes as a device for storing biological information. The next crucial feature is that genomes can be copied to make new cells and new individuals.

In animals, there are two main forms of copying: **mitosis**, in which the genome of one cell is essentially copied to produce two identical genomes; and **meiosis** in which a diploid genome (two of each chromosome) is reduced to haploid (one of each chromosome) to create gametes prior to **fertilization**.

Genome copying: mitosis. Your body started from a single fertilized egg cell. Now that you're reading this, your body contains some 40 trillion cells, each with nearly-identical copies of those original 46 chromosomes. For organisms to increase their number of cells – and to grow in size – the cells need to go through **cell division**.

In cell division, a "parent" cell divides into two "daughter" cells. The parent cell copies each of its 46 chromosomes; then as the cell splits into two, each daughter cell inherits 46 chromosomes to match the genome of the parent. This process of first copying, and then correctly distributing the chromosomes into the daughter cells, is called **mitosis** (pronounced "my-toe-sis") [28].

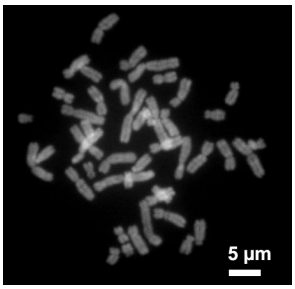


Figure 1.16: **Human chromosomes**, *condensed during mitosis.* *Credit: Steffen Dietzel. CC BY-SA 3.0 [Link]*

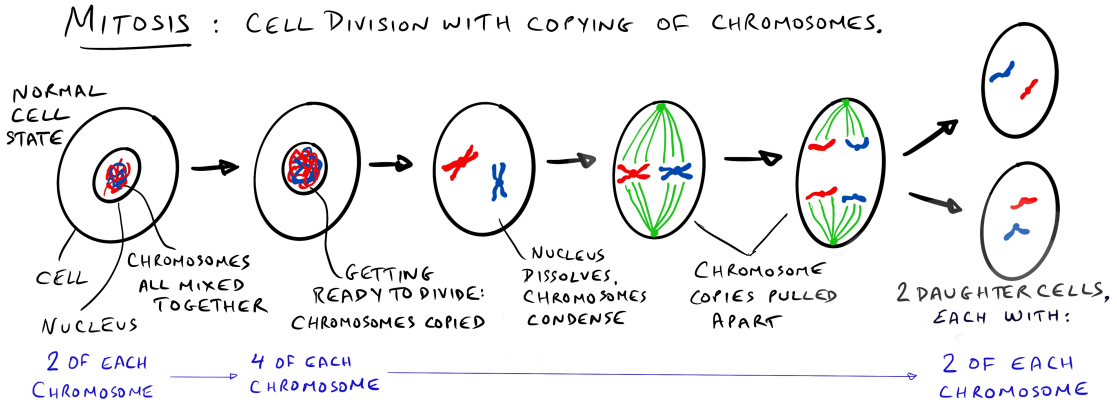


Figure 1.17: Mitosis. *For simplicity, we just show one chromosome; red and blue indicate the two versions of that chromosome carried by each cell (e.g., the chromosome that came from mum in red, and from dad in blue). The x-shaped structures in the middle of the plot show that both the red and the blue versions have been made into pairs of identical copies; one red and blue copy is distributed to each daughter cell.*

Genome reduction and shuffling: Meiosis. In contrast, we need a very different type of cell division to make **gametes** (i.e., sperm and eggs).

While ordinary cells carry 2×23 chromosomes, the gametes only carry 1×23. This is so that when sperm and egg fuse, the fertilized egg have pairs of each chromosome like a regular cell: i.e., 2×23. The process of halving the chromosome numbers to make gametes is called **meiosis** (pronounced "my-oh-sis").

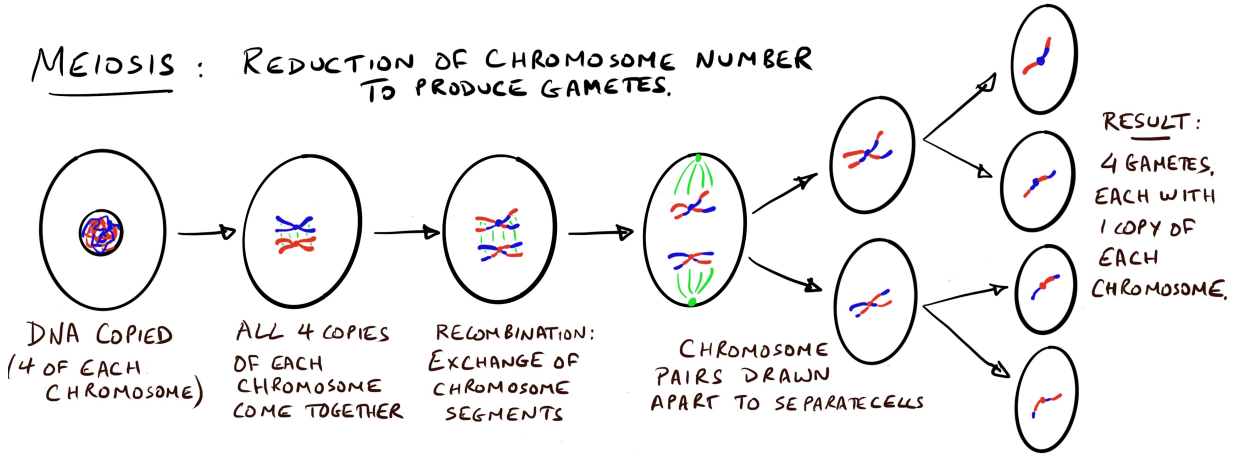


Figure 1.18: Meiosis. *Meiosis starts with DNA copying to make four copies of each chromosome. Next, these come together to exchange pieces: recombination. Then, two stages of cell division result in four gametes, each with one copy of each chromosome. As above, we show one chromosome: red is the version of that chromosome inherited from mum, and blue the version from dad. In females, only one of the four resulting cells develops into an egg.*

Like mitosis, meiosis begins by doubling the amount of DNA in the cells, so that there are 4 copies of every chromosome. It then goes through two rounds of cell division to result in four gametes, each with 1 copy of each chromosome.

Meiosis includes **a crucial process called recombination**, or **crossing-over** [29], which shuffles segments of chromosomes between the maternal and paternal copies. In the figure you can see that the red and blue chromosomes–originally red came from mum and blue from dad–have been shuffled to result in new combinations in each of the 4 gametes.

Meiosis assigns a random 50% of the genome into each gamete. Meiosis is a fundamentally random process that produces a different outcome every time. This is in sharp contrast to mitosis, which is fully predictable: i.e., mitosis produces highly accurate copies of the parent cell every time.

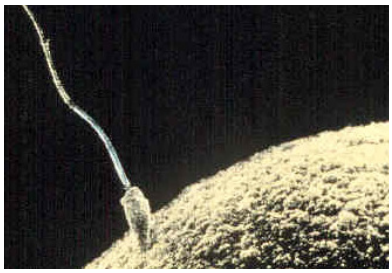


Figure 1.19: **Sperm and egg fusing.** *Unknown author, Public Domain [Link].*

There are two stages of randomness in meiosis: **first, recombination produces a random shuffling of chromosomes**. Later in the book we'll come back to the importance of recombination. **Second, the recombined chromosomes are assigned randomly to gametes**. This combination of two levels of randomness means that every sperm or egg that you produce across your lifetime carries a random, and different, 50% of your genome.

Fertilization. Meiosis is used to create sperm in males, and eggs in females. Each of the sperm and eggs now has a total of 23 chromosomes. Fertilization occurs when a single sperm cell inserts its 23 chromosomes

into the egg to create a fertilized egg that is back to the normal chromosome number of 2×23.

Mutation. The ultimate source of all genetic variation. We'll discuss the types of mutations, mechanisms, and abundance of mutations. Genome replication is extraordinarily accurate, and a typical child carries only about 70 new single nucleotide mutations genomewide. This works out to an average human mutation rate of about 1.2×10^{-8} per base pair per generation. Mutation rates are higher in males than in females, such that a typical child inherits about 3/4 of their new mutations from their dad.

Major data resources for human genetics. In the last part of this chapter we turn our attention to a short description of some of the key resources that are widely used in human genetics. Some combination of these resources were used in virtually all of the modern studies described in this book.

Figure 1.20: **A production line for automated sample preparation**, *built at the Whitehead Institute for use by the HGP. Equivalent work could now be performed on a single benchtop.* *Credit: International Human Genome Sequencing Consortium (2001) [Link]. Used with permission.*

A standard paradigm for research is that if I am interested in a specific research question X, I might collect data relating to X, but I will analyze my new data in the context of other existing data sets. For example, a project that collects any kind of human sequencing data will usually map the sequence reads onto the human Reference Genome, will probably rely on standard gene annotations, and will likely also make use of other more-specialized data sets.

In addition to these large-scale public data sets, researchers also benefit from an enormous number of smaller data sets that analyze specific samples or questions. It's been a huge boon to science that during the last two decades, there has been a strong shift toward making data freely available without preconditions [30]. This is part of a larger movement toward *open science*, which emphasizes the value of making all the results and tools of research publicly available as far as possible [31]. It's now widely recognized that anyone who publishes research in a scientific journal has a responsibility to make the underlying data publicly available [32].

The Human Reference Genome. The central data set that underlies everything practically everything else is the human Reference Genome. For example, in virtually any project that involves sequencing, the first step of data analysis is usually to map reads to the Reference Genome. This genome sequence was the main product of the **Human Genome Project (HGP)**, a huge, $3 Billion international effort that ran from 1990 to 2003, including teams from the US, Britain, Japan, France, Germany, and China.

By the mid-1980s techniques for mapping and sequencing DNA had reached a point where a number of leading scientists started to argue for a "moonshot" type of project to sequence the human genome. Early on, this audacious goal was highly controversial: critics said that it would be purely technical and scientifically uninteresting; that it would divert money from more-focused research; that it is wasteful to sequence the

99% of the genome that is noncoding; that we wouldn't know how to interpret the finished sequence anyway; and that the project would contribute to misconceptions of genetic determinism [33].

Despite the controversy, the genome project was greenlighted by the US Congress in 1990 [g]. During the early years it developed physical and genetic maps of the chromosomes and sequenced several much smaller genomes including the worm *C. elegans* and the fly *D. melanogaster*. During this time frame the costs of sequencing also dropped steadily due to technical advances. But in 1998 a privately funded company named Celera announced a plan to beat the public project to completion using a different strategy, thus spurring the Human Genome Project into a much faster timeline [35].

[g] *This came the year after one of the all-time great presidential malapropisms, when George HW Bush lauded the "**Human Gnome Initiative**" at a White House ceremony on recombinant DNA* [34].

In the end, the public project and Celera battled to a negotiated draw, announcing simultaneous completion of draft genomes in the year 2000. The completion was a major international news event, and the announcement was made by US President Bill Clinton and British Prime Minister Tony Blair in a White House ceremony [Link]. The genome was announced complete three years later [36]. Although the project was controversial at the time, modern biology could not exist in anything like its current form without genomes.

Given that everyone's genome is unique, you might wonder what exactly is in the Reference Genome. In a quirk of history, the sequence is based on a mixture of anonymous donors who were recruited by a newspaper ad in Buffalo, New York, in 1997. At any given position, the reference genome reflects the sequence of a single donor; thus, at many positions, the Reference Genome carries rare, and sometimes even deleterious, alleles [37]. About 70% of the Reference Genome comes from just one of the Buffalo donors, denoted RP-11. Analysis of sequences from RP-11 shows that he had mixed African and European ancestry in roughly equal proportions. Most of the rest of the Reference Genome is derived from ten other donors of East Asian or European ancestry [38].

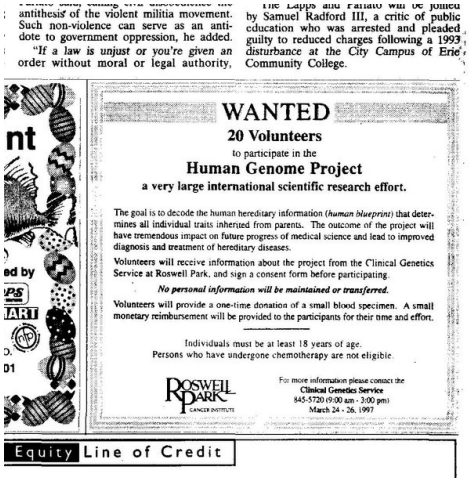


Figure 1.21: **Ad in the Buffalo News, 1997.** *The donors for the Human Genome Project were recruited using this ad, placed in the Buffalo (NY) newspaper on 3/23/1997, by Pieter de Jong of the Roswell Park Cancer Institute.*

Since the end of the Human Genome Project, a group called the Genome Reference Consortium has continued to update the Reference, by fixing assembly errors and providing alternate builds in certain regions with high levels of structural variation. You can download the genome, or browse specific regions using genome browsers at UCSC [Link] or Ensembl [Link]. Genomes for hundreds of other species can also be accessed through the same websites.

While the genome was announced complete in 2003, "complete" was used as a term of art that didn't actually mean *complete*. At that time, existing techniques were unable to span the most repetitive regions of the genome, including the centromeric and telomeric repeat regions, huge arrays of ribosomal DNA genes, and recent segmental duplications. The 2003 genome only covered 2.8 Gb (out of about 3.1 Gb) and included an estimated 341 gaps. The essential problem was that these regions of the genome contain large blocks of highly repetitive DNA that could not be

assembled: imagine trying to assemble a jigsaw puzzle with huge repetitive blocks and a mixture of many pieces randomly sampled from each block. During the next 15 years, even while genome sequencing became massively cheaper (by a factor of 10^5-fold), the sequence reads didn't get longer and most of these regions remained untouchable.

However, by the late 2010s, advances in ultra-long read sequencing using technologies developed by companies called PacBio and Oxford NanoPore enabled extraordinarily long reads that can bridge right across these repetitive elements. Using a mixture of these technologies, the **Telomere-to-Telomere Consortium** announced the first fully assembled human genome in 2021 [39]. We can expect that their work will usher in a new generation of genome sequencing in humans and many other species.

Functional annotation. Of course knowing the genome sequence is only a first step toward understanding the information encoded in it. Since the 1990s there has been a huge amount of work to identify the genes and regulatory elements and to understand their functions. Annotations showing the locations of genes and exons, and their splicing patterns, have been developed by two major projects: **RefSeq** and **GENCODE**. As we discussed above, a more challenging problem is to interpret the regulatory information encoded in genomes. The main approach to this uses a variety of experimental assays; much of this work has been performed and analyzed by the **ENCODE** Consortium. Gene expression profiles for different tissues and cell types are available from **GTEx** and **Human Cell Atlas**, respectively. Information about gene functions can be obtained from many sources, including comprehensive databases from **UniProt**, and **GeneCards**.

Human genetic variation. The Reference Genome only provides a single DNA sequence, and thus doesn't tell us anything about genetic variation across individuals. Thus, as the Human Genome Project was ending, it was recognized that the Genome would be much more powerful if we also had a good catalog of which sites in the genome are variable.

To address this need, from 2002 to 2010, the **International HapMap Consortium** created cell lines from around 100 individuals each from 11 global populations intended to represent some of the world's largest groups. Each individual was genotyped at up to about 3 million known single nucleotide variants across the genome. This landmark work created the first genomewide maps of genetic variation, and paved the way for a huge range of studies.

Subsequently, from 2008 to 2015, the **1000 Genomes Project** performed genome sequencing of a total of 3,200 individuals from 26 human populations. All of the data are freely available for browsing or bulk download [Link]:

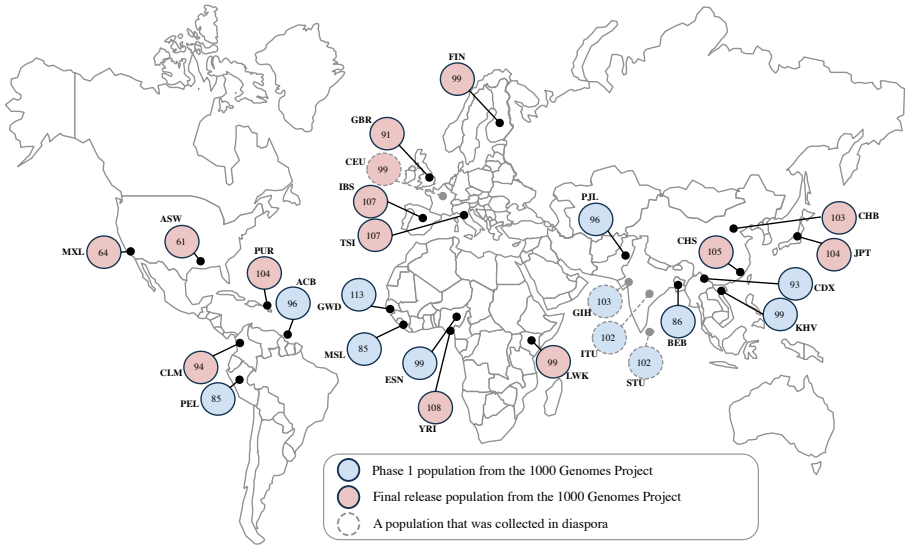


Figure 1.22: **The 1000 Genomes Project** *provides an essential reference set of human genomes from 26 human populations. The blue samples are mainly from HapMap. Some populations (dotted lines) were collected at a different location than their recent ancestral origins, such as CEU (west-Europeans in Utah), and GIH (Gujarati Indians in Houston, Texas).* *Credit: Modified from Fig. 1 of Taras Oleksyk et al (2015) in GigaScience. CC BY 4.*

Four of these populations are often used as example populations for data analysis and figures, so it's worth remembering their sample codes: **YRI** is a group of Yoruba individuals (a west-African ethnic group) sampled in Ibadan, Nigeria; **CEU** is a group of west-European descent individuals sampled in Utah; **CHB** and **JPT** are Chinese Han and Japanese individuals sampled from Beijing and Tokyo, respectively.

While the 1000 Genomes is an essential resource for many purposes, it has poor coverage of some human populations, especially smaller indigenous groups. For example, some groups including southern Africans, Papuans, Pacific Islanders, and Native Americans, are poorly covered by the 1000 Genomes Project. In contrast, the **Human Genome Diversity Panel (HGDP)** and a later extension, the **Simons Genome Diversity Panel (SGDP)**, provide much broader sampling of indigenous populations, albeit with fewer individuals [40]. These panels have helped to reveal wonderful insights into human history that would not have been possible with 1000 Genomes alone; we'll return to these especially in Part 3 of the book.

Figure 1.23: **The Simons Genome Diversity Project** *(shown here) consists of 300 genomes from 142 human populations. The* ***HGDP*** *consists of 1050 genomes from 52 populations.* *Credit: Image courtesy of Simons Foundation [Link].*

Lastly, another important study design involves cataloging genetic variation by sequencing extremely large samples, such as **gnomAD** and **TOPMed** [41].

The genotype-phenotype relationship. Our final category of data sets aim to measure the effects of genotype on phenotype. The most influential of these is an extraordinary dataset collected by the **UK Biobank**, which has collected genome-wide genotypes, and a huge array of phenotypic measures on about 500,000 British residents. Enrollment began in 2006, targeting an age range of 40–69, and continuing to track those individuals through middle and old age. Any qualified researcher can go through an application process to get access to the de-identified data. Due to the relative ease of data access, and the richness of information available, the UK Biobank has had a huge impactful on our understanding of human genetics. It's not a large exaggeration to say that the UK

Biobank has managed to get all of the world's human geneticists studying the British population.

Other very large cohorts have replicated aspects of the UK Biobank, including Biobank Japan, the China Kadoorie Biobank, FinnGen, the Estonian Biobank, the Million Veteran Program and All of Us (the latter are both in the US). These other cohorts either have less data at present, or are less accessible to outside researchers than UK Biobank. There are also disease-specific projects, such as the Psychiatric Genetic Consortium, that aggregate case-control data for focused study of particular diseases. One important concern about current cohorts is that, in aggregate, individuals of recent European descent are over-represented across these studies. This challenges human geneticists to ensure that the future benefits of genetic research can be shared equitably among people from all ancestries [42].

In this first chapter we have given an overview of some important background that will be helpful before we dive more deeply into the main areas of human genetics. We next turn to a more focused description of human genetic variation.

Notes and References.

[2]Although the Human Genome Project was declared complete in 2003, about 10% of the genome was unsequenceable at that time. The first truly complete human genome sequence was reported in 2021 and published the following year:

Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53

[3]As with some other complicated topics, for the sake of brevity we will generally simplify important points relating to sex, gender and familial relationships, except when the complexities are specifically relevant. For example it's convenient to refer to XX and XY individuals as female and male respectively. We do so despite the fact that (i) biological sex is not entirely binary – some individuals have physical characteristics of both sexes due to mutations in sex-determination genes, unusual karyotypes such as XXY, or other causes not all of which are currently understood; (ii) biological sex does not necessarily correspond to gender; gender is actually *more* relevant than biological sex for many aspects of our lived experiences – even though it is generally less connected to the core topics of this book; (iii) familial relationships do not always imply genetic relationships – for example in the case of parents of adopted children.

[4]During our lives, our bodies produce about a light-year of DNA: [Link].

[5]For more on DNA storage systems see eg:

Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. Science. 2017;355(6328):950-4

Kim J, Bae JH, Baym M, Zhang DY. Metastable hybridization-based DNA information storage to allow rapid and permanent erasure. Nature Communications. 2020;11(1):1-8

[6]Improbable Research's video of Eric Lander's 24 second and 7 word descriptions of the human genome: [Link]

[7]In 2021 the AlphaFold team reported huge progress on computational prediction of protein folding, thereby helping to transform this field:

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9

[8] An important set of exceptions to the standard genetic code is found in the mitochondrial genome. The mitochondrion is thought to have evolved from an endosymbiotic prokaryote, and it still retains a very small genome of its own. This genome is so small that rare minor changes in the genetic code have been tolerated by natural selection. Specifically, the genetic code in vertebrate mitochondria differs from the conventional code at four triplets: AGA and AGG are stop codons instead of arginine; TGA codes tryptophan instead of stop; and ATA codes methionine instead of isoleucine.

[9]There are various categories of genes in which the RNA itself is functional. For example, in females one copy of the X chromosome is inactive in each cell; this is achieved in part by transcribing an RNA called Xist off one of the two X chromosomes. The Xist transcript coats that X chromosome and prevents transcription from most other genes. Xist is an example of what is known as a long noncoding RNA (lncRNA). In addition to lncRNAs, other functional RNA genes categories include microRNAs, transfer RNAs, ribosomal RNAs, and piRNAs.

[10]Another important exception to the Central Dogma is that some viruses use RNA as their genetic material, and then use an enzyme called *reverse transcriptase* to make a DNA copy for replication. Reverse transcriptase is also used in the lab to make DNA copies of RNA when we want to sequence RNA.

[11]The fact that the introns are so very long is probably not functionally important in most cases, and instead reflects a tendency for genomes to accumulate noncoding junk, as we will discuss below.

[12]There is some uncertainty about exactly how much alternative splicing is functionally important. One approach that is often used to evaluate functional importance of biological features is whether a feature is maintained (conserved) over evolutionary time, or whether it evolves rapidly, suggesting malleability and (usually) lower functional importance. Curiously, alternative splicing patterns (specifically, exon skipping events) are not very conserved across species – and are less conserved than overall expression levels. However interpretation of this is not entirely clear:

Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338(6114):1587-93

Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012;338(6114):1593-9

Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nature Genetics. 2018;50(1):151-8

[13]There's been quite a bit of interesting work on the sequence controls of splicing; these include both high-throughput experimental approaches as well as machine learning methods to learn highly complex rules from genome sequence data or experiments. See for example:

Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. Learning the sequence determinants of alternative splicing from millions of random sequences. Cell. 2015;163(3):698-711

Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-48

Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biology. 2022;23(1):1-18

[14]For example Down Syndrome occurs in individuals who have an extra copy of Chromosome 21. Chromosome-level changes in copy number change the expression levels of the genes on that chromosome relative to the genes on other chromosomes. It's interesting to note that cells can often tolerate duplication of the entire genome better than duplication of a single chromosome, as whole-genome duplication maintains the relative proportions of genes. Somewhat similarly, many monogenic diseases are due to defects in the core transcriptional machinery, leading to broad transcriptional dysregulation rather than disruption of specific biological pathways; see Table 6.2 of

Calof AL, Santos R, Groves L, Oliver C, Lander AD. Cornelia de Lange syndrome: Insights into neural development from clinical studies and animal models. In: Neurodevelopmental Disorders. Elsevier; 2020. p. 129-57

For example, Cornelia de Lange Syndrome is due to mutations that disrupt the cohesin complex; these cause minor disruptions of many genes leading to diverse developmental disorders.

[15]Muniz L, Nicolas E, Trouche D. RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. The EMBO journal. 2021;40(15):e105740.

[16]Technically, the direct copy of DNA is called a pre-mRNA. This must be spliced to produce the mature mRNA. Most splicing occurs at the same time as transcription.

[17]Expression (i.e., mRNA levels) of any given gene depend on the rate of transcription in the relevant cell type (defined as the number of new mRNAs synthesized per unit time), and the mRNA decay rate. For most genes, control of gene expression acts mainly on synthesis.

[18]An exception is that several proteins called General Transcription Factors are components of the Pre-Initiation Complex and lack DNA binding domains.

[19]Most of the genome is bound by nucleosomes, and TF binding requires nucleosome removal. This can be much more stable if multiple TFs can bind within the same nucleosome-free region.

[20]There's a large, growing body of work using machine learning approaches to predict enhancer regulatory activity, e.g.,

Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods. 2015;12(10):931-4

Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research. 2016;26(7):990-9

Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics. 2021;53(3):354-66

[21]One famous example of long-range looping occurs at the FTO locus:

Sobreira DR, Joslin AC, Zhang Q, Williamson I, Hansen GT, Farris KM, et al. Extensive pleiotropism and allelic heterogeneity mediate metabolic effects of IRX3 and IRX5. Science. 2021;372(6546):1085-91

[22]For empirical work on predicting enhancer-promoter interactions see e.g.,

Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature genetics. 2019;51(12):1664-9

[23]This number is a bit rough because we still don't have a complete accounting of functional regulatory sequences in all cell types. But around 10% of the genome shows signals of evolutionary conservation. This provides an estimate of what fraction of the genome is functional – in the sense that changes in the DNA sequence have consequences for organismal fitness.

Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. 2012;337(6102):1675-8

Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS Genetics. 2014;10(7):e1004525

[24]For statistics about gene sizes see

Piovesan A, Caracausi M, Antonaros F, Pelleri MC, Vitale L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. Database (Oxford). 2016;2016.

[25]Many of these regions are transcribed but not translated; as noted above, these are referred to as long noncoding RNA (lncRNA) genes. Some lncRNA genes play essential roles, but most show limited evolutionary conservation and only

a tiny fraction are currently associated with putative functions, suggesting that most lncRNAs are likely nonfunctional:
Ransohoff JD, Wei Y, Khavari PA. The functions and unique features of long intergenic non-coding RNA. Nature Reviews Molecular cell biology. 2018;19(3):143-57
Ponting CP, Haerty W. Genome-wide analysis of human long noncoding RNAs: a provocative review. Annual Review of Genomics and Human Genetics. 2022;23

[26]See for example L1 silencing mechanisms:
Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, et al. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature. 2018;553(7687):228-32.

[27]For examples in which TEs have been co-opted by their host genomes see
Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351(6277):1083-7
Bartonicek N, Rouet R, Warren J, Loetsch C, Rodriguez GS, Walters S, et al. The retroelement Lx9 puts a brake on the immune response to virus infection. Nature. 2022:1-9

[28]Mitosis and meiosis are complicated and deeply studied processes, and it's impossible to do them justice here. We'll touch on a few of those complexities later in the book as they become relevant.

[29]To be more precise, meiotic recombination events can be resolved either with crossover or non-crossover events. Non-crossovers involve copying of a small region (average 30–40bp in mice) from one chromosome to the other.
Li R, Bitoun E, Altemose N, Davies RW, Davies B, Myers SR. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nature Communications. 2019;10(1):3900
While non-crossovers are very common they are difficult to detect in data. However the term "recombination" is often used in human genetics synonymously with crossovers.

[30]Some of the major resources, such as the human genome and 1000 Genomes Project data sets are freely downloadable. Other data sets such as the UK Biobank contain personal information about research subjects, albeit anonymized, and can only be used by qualified researchers who agree to certain conditions for appropriate use of the data. However in all these cases, researchers have a large amount of flexibility in how they use the data for their own analyses.

[31]Open science: [Link].

[32]For example, the prominent journal Nature writes on their website: "It is a condition of publication that authors deposit their data in an appropriate repository, and agree to make the data publicly available without restriction, excepting reasonable controls related to human privacy or biosafety." [Link], accessed 10/01/2021.

[33]Roberts (2001) wrote "Sydney Brenner of the MRC facetiously suggested that project leaders parcel out the job to prisoners as punishment–the more heinous the crime, the bigger the chromosome they would have to decipher."
Lewontin R. The dream of the human genome: doubts about the Human Genome Project. The New York review of books. 1992;39(10):31-40
Roberts L. The Human Genome. Controversial from the start. Science. 2001;291:1182-8

[34]This was in a White House ceremony in 1989 to award the National Medal of Honor to Stan Cohen and Herbert Boyer who developed recombinant DNA technology; as recalled by Carol Ezzell in Scientific American, July 2000 [Link].

[35]There was a great deal of acrimony between the two groups, not least because Celera's build incorporated data that the Human Genome Project was releasing into the public domain on a daily basis (in part to prevent attempts to patent genes). Some of the back-and-forth can be found here: HGP critique
Waterston RH, Lander ES, Sulston JE. On the sequencing of the human genome. Proceedings of the National Academy of Sciences. 2002;99(6):3712-6;
Celera reply: Myers EW, Sutton GG, Smith HO, Adams MD, Venter JC. On the sequencing and assembly of the human genome. Proceedings of the National Academy of Sciences. 2002;99(7):4145-6

[36]Flagship papers on the Human Genome Sequence:
International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304-51
International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45

[37]There have been occasional calls to change the reference to remove rare alleles, but such large changes to the reference genome would create all kinds of compatibility issues and in this case the medicine may be worse than the disease.
Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biology. 2019;20(1):1-9

[38][Link] and p 146 of the supplementary information of
Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-22

[39]Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53

[40]The HGDP was started at Stanford in the early 1990s by two of my mentors, Luca Cavalli-Sforza and Marc Feldman. This project pioneered the concept of collecting cell lines from diverse human populations as permanent resources for studies of genetic diversity, a concept later adopted by HapMap and 1000 Genomes. The HGDP was used for limited genotyping in the 1990s, genomewide genotyping in the 2000s and, ultimately, whole genome sequencing in the 2010s.
Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Annals of human genetics. 2006;70(6):841-7
Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. science. 2014;343(6172):747-51
Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201-6

[41]Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91
Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43

[42]Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4

1.3 Human genome variation and why it matters.

In the last chapter, we discussed the standard human Reference Genome. But in practice ***everyone's genome is unique, and differs from the Reference at millions of sites****. Here we introduce the concept of genome variation and how it can change the information encoded in genomes.*

SNPs. The most abundant type of genetic variation [a] are **SNPs** (Single Nucleotide Polymorphisms, pronounced *snips*). These are simple sequence differences that affect a single nucleotide: for example in a short stretch of genome your maternal chromosome might read ATCGAAGCC, and your paternal chromosome ATCGGAGCC. Although four nucleotides would be possible at any given position, the vast majority of SNPs only have two alleles [43]:

[a] *In this chapter we'll see many different types of ways that sites or regions of the genome can differ among individuals. We can refer to these genetic differences as* ***variants****.*

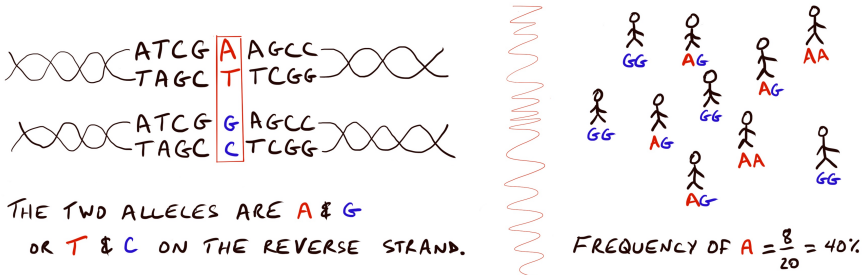


Figure 1.24: **Illustration of an A/G SNP.** *A has a sample frequency of 40%.*

Allele frequency. In the figure above, we show the genotypes for ten individuals at an A/G SNP (i.e., a SNP where the two possible alleles are A and G). Since each person carries two copies of this sequence, they can either be AA, AG, or GG. In this examples there are 8 copies of A out of 20: this gives a frequency of p=0.4 for A, and q=0.6 for G.

We will often use p and q to indicate the frequencies of two alternative alleles, where $p + q = 1$.

If you're analyzing data it's important to be keep track of **which strand of the DNA** the SNP refers to; in the example above we would consider this an A/G SNP if we're looking at one strand, but a T/C SNP on the other strand. This is especially tricky for transition mutations (A/T versus T/A or C/G vs G/C as both alleles are found on both strands). SNPs are usually labeled with respect to the strand used in the Reference Genome, or occasionally with respect to the direction of translation if the focus is on protein-coding variation.

Once we've solved the strand issue, it's also useful to have generic ways of referring to alleles so that we don't have to remember that this particular SNP is A/G and that G is more common. When you read papers, you'll see this done using one of three different naming conventions [44]:
• **Reference/Alternate allele**: The *reference* allele is the allele listed at that position in the Human Reference Genome.
• **Minor/Major allele**: The *minor* allele is the less common allele in a population (frequency $< 50\%$). **MAF** stands for *Minor Allele Frequency*.

• **Ancestral/Derived allele**: The *ancestral* allele is the allele that was present in the common ancestor of humans (this can be inferred if one allele matches the nucleotide found at this position in other great apes), while the *derived* allele is inferred to have arisen by mutation within the human population. **DAF** stands for *Derived Allele Frequency*.

Some authors reserve the term SNP for variants where both alleles are fairly "common" – often defined as MAF>1% – and use the term **SNV (single nucleotide variant)** to include sites with rarer variation. But since the cutoff is arbitrary, here we use the term SNP throughout.

Genotype frequencies and the Hardy-Weinberg model. Aside from allele frequencies, we also need to know about **genotype frequencies**: the fraction of people who have each possible pair of alleles, in this case AA, AG, and GG. People with AA or GG are **homozygotes**, while people with AG are **heterozygotes**.

Part of the power of population genetics is that we can often predict fundamental properties of populations using mathematical models. In most cases, the starting assumptions are quite simple (although sometimes the math is complicated). This brings us to the first important model of this book [b].

[b] *The Hardy-Weinberg model gives us the fundamental relationship between allele frequencies and genotype frequencies.*

To keep things simple, we'll first assume that we have distinct generations, so that there is a population of parents in one generation who mate to produce kids in the next generation. Among the parents, let p represent the frequency of the A allele, and q the frequency of the G allele. We'll also assume that the parents don't choose their reproductive partners based on genotype at this locus.

So for a kid in the next generation, what is the probability that this kid will have an AA genotype?

Answer: There's a probability p that she gets an A from the mum; and similarly from the dad. So the probability that she gets AA from both parents is $p \times p$ or p^2. (The probability of two independent events occurring is the product of the probabilities.)

Using the same logic, the probability that she gets GG from both parents is q^2. Lastly, she could get A from mum and G from dad (probability pq) or G from mum and A from dad (probability qp)–those both result in her being an AG heterozygote, with total probability $2pq$.

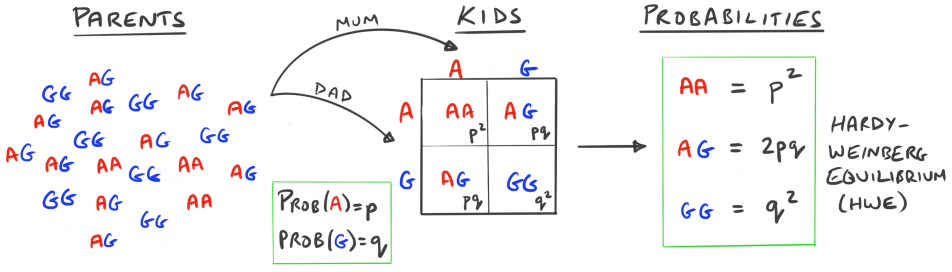


Figure 1.25: **Hardy-Weinberg.** *The genotype of a kid can be modeled as a random draw of two alleles from the population of possible parents. This means that if the allele frequencies in the parents are* $\mathbf{p}$ *and* $\mathbf{q}$*, then the genotype frequencies in the kids are* $\mathbf{p^2 : 2pq : q^2}$.

The key prediction here is that the expected genotype frequencies in the kids' generation is given by the proportions $\mathbf{p^2 : 2pq : q^2}$. This result is known as **Hardy-Weinberg Equilibrium (HWE)**. This result gives us the fundamental relationship between allele frequencies and genotype frequencies.

If you learned about the Hardy-Weinberg rule at school, you might have been taught a whole host of assumptions that this relies upon: for example, random mating, no selection, non-overlapping generations, etc. But in practice **Hardy-Weinberg is usually extremely accurate**, except sometimes with strong population structure. In large part, this is because just **one generation of random mating will restore a population to HW proportions** [45]. Indeed, in data analysis, if a SNP is out of Hardy-Weinberg proportions within a population, this is often taken as an indication that the genotyping may not have worked properly for that SNP [46].

Lastly, the Hardy-Weinberg result has an amusing backstory. It was first published independently in 1908 by GH Hardy, a famous English mathematician, and a German gynecologist Wilhelm Weinberg. Hardy was told about the problem of genotype frequencies by the geneticist Reginald Punnett, with whom he played cricket at Cambridge University. (You may be familiar with "Punnett Squares", used to predict the outcomes of genetic crosses.) Hardy's 1-page paper is written in a bashful tone because he thought the main result was rather beneath him (..."A little mathematics of the multiplication-table type is enough to show...") though it is now seen as the first fundamental result in population genetics. To add further insult, the article was initially rejected by the leading British journal Nature for being "tainted" with Mendelism (which was unpopular in Britain at the time, see Chapter 4.4), and it was eventually published in Science instead [47].

MENDELIAN PROPORTIONS IN A MIXED POPULATION

To the Editor of Science: I am reluctant to intrude in a discussion concerning matters of which I have no expert knowledge, and I should have expected the very simple point which I wish to make to have been familiar to biologists. However, some remarks of Mr. Udny Yule, to which Mr. R. C. Punnett has called my attention, suggest that it may still be worth making.

Figure 1.26: **The start of Hardy's 1908 paper** *on genotype frequencies.*

How many SNPs are there? Popular culture often refers to "the human genome" – but of course in practice everyone's genome is unique (unless you have an identical twin [48]).

Since you inherited one copy of each chromosome from your mum, and one from your dad, one way to measure genetic diversity is to count up how many differences you have between these two genome copies. Any difference between the two genomes – for example, you got an A from mum and a G from dad at a particular position – is a heterozygous SNP. So we can ask, how frequently would you find heterozygous SNPs between the homologous copies of your genome?

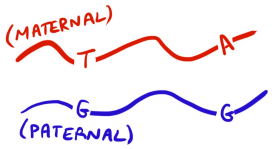


Figure 1.27: **Heterozgygosity** *measures the fraction of sites that differ between the homologous copies of someone's genome.*

The answer is that you can expect to find a heterozygous SNP about once every 1,000–2,000 basepairs, depending on your ancestry. The fraction of heterozygous sites is referred to as **heterozygosity**, and is a useful measure of genetic diversity [49]. Here's a table of heterozygosity estimates from a variety of human populations:

Region	Population	Heterozygosity × 1000
Africa	San	0.95
	Yoruba	0.96
	Maasai	0.93
	Mbuti	0.91
Near East	Palestinian	0.73
	Iranian	0.71
Europe	Spanish	0.69
	Polish	0.67
South Asia	Punjabi	0.71
	Bengali	0.72
East Asia	Thai	0.69
	Japanese	0.67
Oceania	Australian	0.63
	Papuan	0.58
Americas	Inuit	0.63
	Surui	0.50

Table 1.2: Heterozygosity estimates by population, *reported as the mean number of heterozygous sites per 1000 bp. Populations such as 'Australian' refer to indigenous groups.*

Data from Supp. Table 1:AH of Swapan Mallick et al (2016) [Link].

The most striking pattern in these data above is that **heterozygosity is highest in Africa, and decreases with migration distance from Africa.** This reflects the fact that modern humans evolved in Africa; as we will discuss later in the book, modern humans spread out of northeast Africa during the last 100,000 years and eventually colonized most of the world. These population movements caused a steady loss of diversity with distance.

Another way to think about variation is in terms of the total numbers of SNPs. Since your genome is about 3.2 billion basepairs, this table implies that **you have about 1.5–3.0 million heterozygous sites, depending on your ancestry**.

What if we look at SNPs in a larger number of individuals? For example, the 1000 Genomes Project sequenced the genomes of 2500 individuals from a diverse set of global populations. They reported 85 million SNPs, most of which were very rare: 65 million were below frequency 0.5%; 12 million were between 0.5%–5%, and 8 million SNPs were above 5% [50]. In other words, **there is a common SNP with frequency>5% about once per 400 bp**.

It's important to note that, while a study of modest size like 1000 Genomes Project can identify essentially all common SNPs, large sequencing studies continue to find many more novel, rare SNPs [51]. Indeed, we should expect to find that **nearly every possible SNP allele exists somewhere in the world**. The world population is nearly 10^{10} people and, as we'll see in Chapter 1.5, the mutation rate is around 10^{-8} per nucleotide per generation. This implies that nearly every possible single nucleotide variant must occur many times each generation, somewhere in the world. (There are a few exceptions: a tiny fraction of possible mutations would not be

observed because they disrupt biological processes so severely that they prevent embryonic or fetal development.)

Single nucleotide differences between human and chimpanzee. We don't need to limit ourselves to looking at genetic diversity within humans – we can also examine the number of sequence differences between humans and other species [c]. For example, our closest relative is the chimpanzee; our two species evolved from a common ancestor about 6.5 million years ago [52]. The average sequence divergence between the human and chimpanzee genome is about 1.37%. This is only about 15-fold higher than the divergence between the two copies of your own genome. I'll explain a bit later how to think about this 15-fold ratio. We still know very little about specifically which variants are responsible for the major phenotypic differences between humans and chimpanzees, such as our remarkable fondness for cell phones.

[c] *For thinking about human genetic variation,* ***there are a few numbers that are useful to remember****: Heterozygosity in humans is* $0.5–1 \times 10^{-3}$*, depending on population. There are* $\sim$*10M common SNPs. The human and chimpanzee genomes differ by 1.4% – about 40M single nucleotide differences.*

Genotypes and haplotypes. So far we have just been *counting* SNPs, but it will also be important to consider how they are arranged along chromosomes. The identities of alleles along one copy of a chromosome are referred to as a **haplotype**. For example, the first haplotype in the plot below is A-T-A-C-G-A, and the second is A-A-C-C-G-C. As we go through this book, we'll learn a lot from studying the structure of haplotypes:

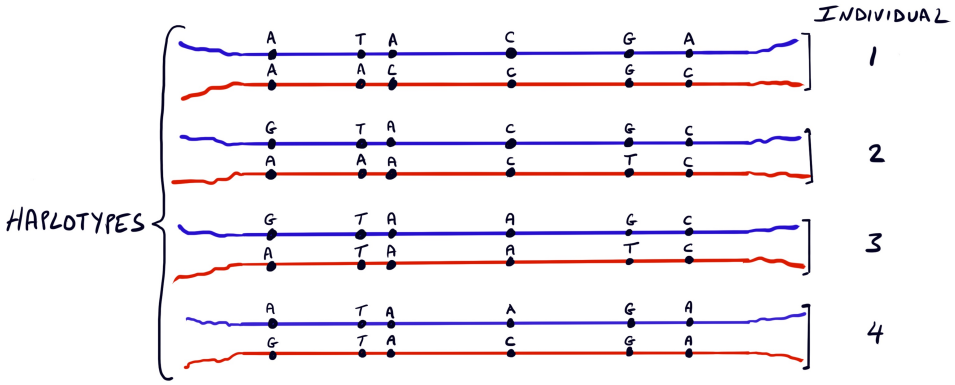


Figure 1.28: **Haplotypes for 4 individuals** *in a small region of the genome. The term* ***haplotype*** *refers to the arrangement of alleles along a homologous chromosome, or sometimes to a pattern seen in a genomic region in multiple individuals.*

Now, one challenge is that standard technologies for genome sequencing are very good at telling us the genotype at any given location, but for a heterozygous site we can't tell which allele is on which haplotype. (We'll cover sequencing in Chapter 1.4, but the issue is that standard DNA sequence reads are much shorter than the usual spacing between heterozygous sites. This is starting to change with the arrival of new long-read sequencing techniques.). So traditional sequencing does not give us the actual haplotypes, but instead genotypes like this:

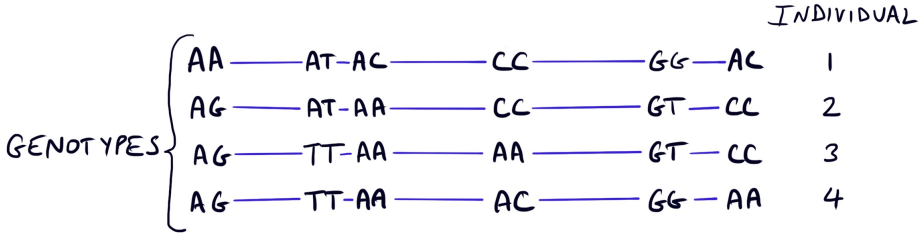


Figure 1.29: **Genotypes for 4 individuals** *in the same region as above. For heterozygous sites we usually do not get a direct measurement of which allele comes from which haplotype so the data for individual must be represented as genotypes.*

The assignment of alleles to haplotypes is referred to as **haplotype phase**; haplotypes can be estimated using statistical techniques that we will cover later.

Lastly, it's often convenient to replace the genotype matrix with a simpler version that recodes the genotypes with the allele counts: 0, 1, 2, depending on the number of minor alleles at each SNP:

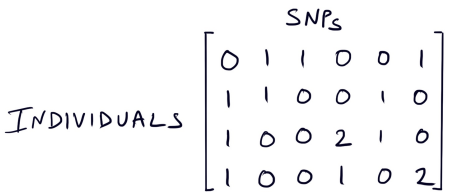


Figure 1.30: **Genotype Matrix representation** *of the data above. The entries in the matrix show the numbers of minor alleles at each position: i.e., the columns show the numbers of G,A,C,A,T,A alleles, respectively).*

Example from real data. One of the first groups to study human sequence variation at large scale was Debbie Nickerson's lab at the University of Washington, in the early 2000s [53]. This plot shows the genotype matrix they obtained for the IGF1 locus on Chromosome 12, using colors to represent the genotype counts 0, 1, and 2:

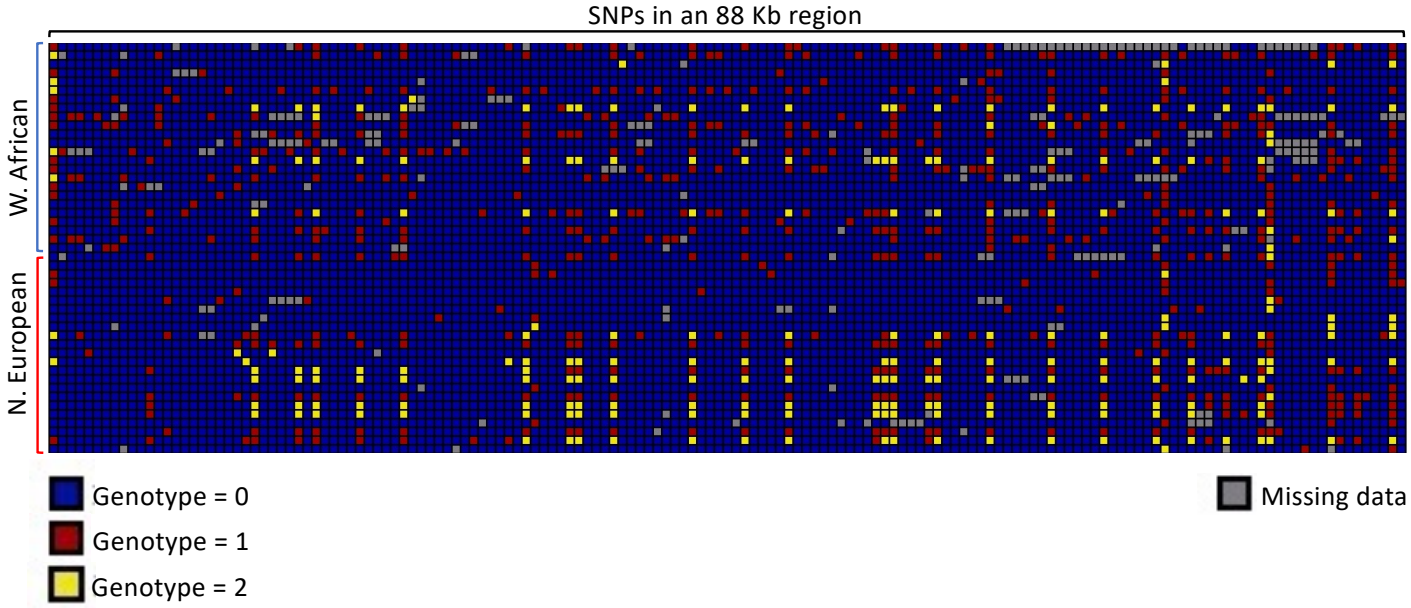


Figure 1.31: Genotype Variation at IGF1. *Each row shows the genotype for a single individual; columns are SNPs within an 88 Kb region containing the IGF1 gene, ordered by position within the region.* *Sequencing was PCR-based and included gaps within the region. Credit: Debbie Nickerson's lab/ Seattle SNPs project [Link].*

This example illustrates several typical features of genomic data:
• Most common variants are shared between human populations, though allele frequencies may differ.
• As expected, many of the SNPs are at low frequency. Indeed this pattern becomes increasingly stronger in larger samples, as nearly all the common SNPs are identified in a small sample like this one, while the new variants discovered with additional individuals would be rare;
• Variants at different sites often co-occur together – for example notice

the block of yellow genotypes on the left-hand side of the region where individuals who are yellow (or red) at one site are usually yellow (or red) at multiple other sites as well. This pattern of genotype correlations across sites is called **linkage disequilibrium (LD)** and will be an important topic for us later.

Figure 1.32: **A gryphon and a greyhound:** *two beasts from an illustrated bestiary (England ~1520). The wide array of possible types of variation – some of them difficult to glimpse with current methods – reminds me of a tableau of mysterious beasts.* *From the Tudor Pattern Book in the Bodleian Digitized Collection.*

Beyond SNPs: Other types of inherited variation. While SNPs are the most common type of variation, there are many other ways that genomes can differ. These include small-scale events such as indels and STRs (short tandem repeats), as well as a diverse variety of larger structural elements. Collectively, we refer to all these different forms of variation as **variants**.

In Medieval books, a bestiary was a compendium of beasts (animals), both real and imagined, with pictures and descriptions; by analogy, here I show a collection of some of the main types of genetic variation. Many of the large repeats that we'll talk about soon remain slightly mysterious: they are very difficult to measure using current DNA sequencing technology, and variation in complex regions is still largely uncharacterized.

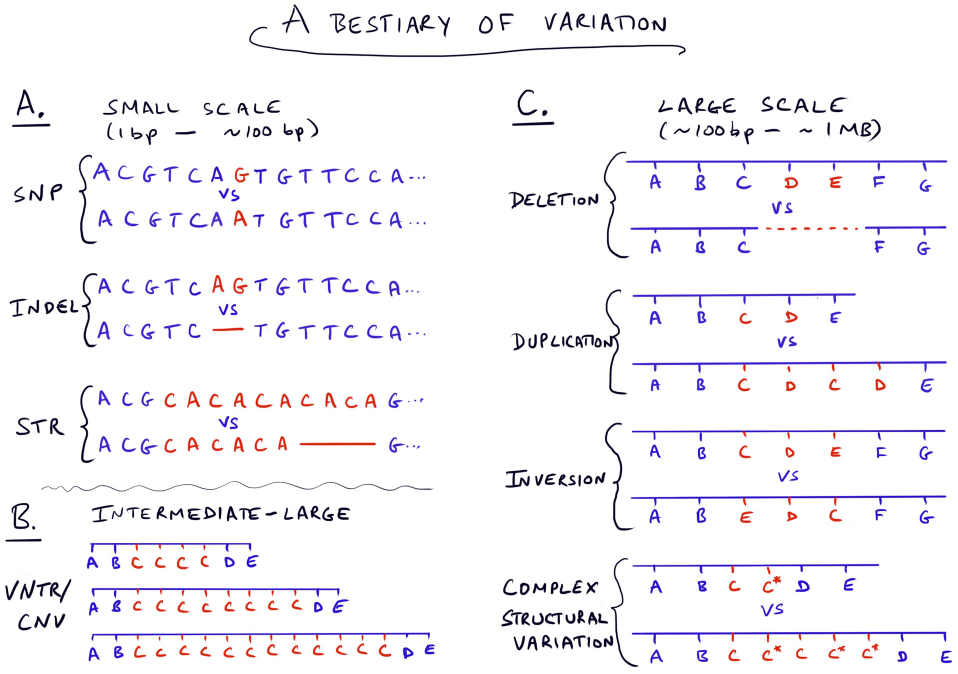


Figure 1.33: **Major types of variation. A.** *These variants affect short stretches of DNA sequence.* **B.** *and* **C.** ***Structural variants:*** *Here* ***the letters represent large blocks of DNA sequences****. These categories are often blurred, and complex structural variants often contain multiple types of events. There are many more SNPs than other kinds of variants, but because they are so large, the structural variants cover more total genome.*

Short-scale variation. The first type of short-length variation are the **indels** as shown in the picture above. This term is a mashup of the phrase **insertions/deletions**, reflecting the fact that without comparing to another genome such as chimpanzee, it is difficult to know whether an indel represents a gain, or loss, of nucleotides relative to the ancestor. These are about one tenth as many indels as SNPs, and most are very short, between 1 and 5 nucleotides in length [54]. As we mentioned above, indels are of special importance in exons, as they result in frameshifts

(unless the indel length is a multiple of 3).

The next important type of short-length variation is the **STR (short tandem repeat)**. STRs are places in the genome where a short DNA sequence (up to around 6bp) is repeated many times (eg., CACACACA) – often dozens of times. During cell division, the copying of STRs is highly error-prone due to a process known as replication slippage. For this reason, STRs are highly variable within populations [d].

[d] *Because STRs are highly variable they are used in paternity testing and **DNA fingerprinting** for forensics.*

Intermediate-scales: VNTRs. Similar to STRs, there are larger blocks of sequences, known as **VNTRs: Variable Numbers of Tandem Repeats**, that can be duplicated many times in the genome [55]. One example is shown at right, where a block of 57bp is repeated between 22–29 times within an exon of the ACAN gene.

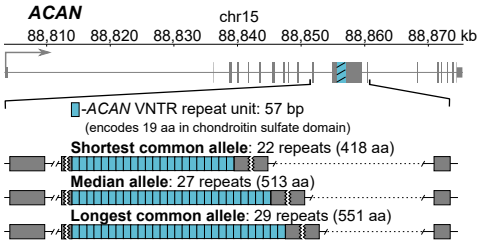


Figure 1.34: **VNTR in the ACAN gene.** *A 57 bp (19 amino acid) repeat is present 22–29 times on different haplotypes. The ACAN protein is part of the extracellular matrix of cartilage, and larger repeat numbers are associated with higher average height.* *Figure 1a from Ronen Mukamel et al (2021) [Link]. Used with permission from the authors.*

Large-scale variation. The last major class of variation are various types of **large-scale structural variation**. Some of these are simple rearrangements of the DNA sequence, including **deletions** and **duplications**. As shown above, deletions are events that cut a segment out of a chromosome, while duplications copy a segment. An individual with a deletion would then carry one copy of any genes that lie inside that deletion (i.e., the copy on the unaffected chromosome); or three copies of those genes in the case of a duplication (two copies on the duplicated chromosome, plus one on the unaffected chromosome). Together, deletions and duplications can be referred to as **copy number variation (CNVs)**.

For reasons we'll discuss shortly, large CNVs – at megabase scale or larger – are usually highly deleterious, and can cause severe genetic syndromes in children. In contrast, smaller CNVs are often benign, especially if they do not overlap genes or key regulatory regions. A typical person carries more than 200 deletions with a median size of 7 Kb, and with 20% deleting more than 25 Kb. Numbers for duplications are similar [56]. While CNVs do not overlap genes, a small fraction – about 10% – of the deletions remove either entire genes or parts of genes (i.e., exons) [e].

[e] *Common deletions are less likely to overlap genes than you would expect if they occurred randomly in the genome. This, and other evidence, implies that natural selection preferentially removes genic deletions.*

Meanwhile, **inversions** take a segment of chromosome and flip it around. Inversions are much less abundant than deletions and duplications, but can be very large. The two best-characterized human inversions include one on Chromosome 8, which is a huge 4.5 Mb at around 50% global frequency, and a very interesting 900 Kb inversion on Chromosome 17 that is at 20% frequency in Europeans and may be associated with fertility and other phenotypes [57].

Lastly, some regions of the genome become crucibles of **complex structural variation**, in which the processes of replication slippage, deletion, duplication, inversion, become layered upon one another, to the extent that there may be huge numbers of different alleles. These regions are very difficult to sequence, difficult to visualize, and generally not well-characterized. However, new technologies for getting very long sequencing reads are starting to open up these regions to study. We'll see more of these topics in Chapters 1.4 and 1.5.

How do SNPs affect the information encoded in genomes? [f] Genomes are a molecular system for storing information; SNPs can alter that encoded information. Roughly speaking, SNPs and other kinds of genetic variation can have two main types of effects: they can change protein coding sequences, or they can change gene regulation.

While SNPs that affect function are of special interest for understanding disease and evolution, it's important to bear in mind that less than ~10% of all possible SNPs have any meaningful impact on the encoded information.

1. Protein-coding variants. The figure below illustrates four important types of variants within part of a coding exon:

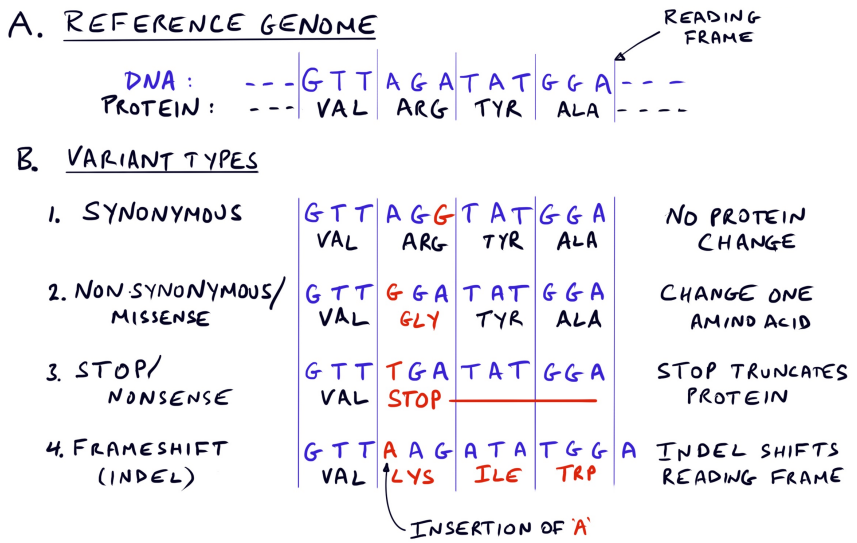


Figure 1.35: **Four types of genic variants** *in a small part of a coding region.* **A.** *Reference sequence: DNA in blue, with corresponding protein sequence in black.* **B.** *Four important categories of exonic mutations. Changes to the DNA and protein sequences are indicated in red. Frameshifts are caused by indels, which we have not covered yet:* ***indels*** *are short insertions or deletions of DNA sequence and they can shift the protein reading frame.*

• **Synonymous.** Remember that DNA encodes proteins using a genetic code in which three consecutive DNA letters correspond to amino acids: 3 DNA triplets (or codons) encode STOP signals and, together, the other 61 possible triplets code for 20 amino acids. This means that about 1/4 of possible one-step mutations simply convert between equivalent triplets. For example, in the illustration, AGA and AGG both encode Arginine. Hence, such a variant is referred to as *synonymous* in the sense that it encodes an identical protein. Synonymous variants generally do not have phenotypic effects [58].

• **Nonsynonymous/Missense.** However, many exonic mutations do change the encoded amino acid: for example AGA → GGA swaps Arginine → Glycine. Such mutations are referred to as *nonsynonymous* or *missense* as they change the meaning of the information encoded in the DNA. The functional impact of missense mutations can range from lethal to no effect, depending on what the gene is, whether the amino acid lies in a key functional domain of the protein, and the chemical properties of the original amino acid and its replacement.

• **Stop/Nonsense.** Three codons (TAA, TAG, TGA) encode the protein Stop signal. Thus for example changing AGA → TGA causes protein

[f] *We'll cover phenotypic effects of variation in Section 4 of the book, but it will be helpful for you to have this preview in mind for the upcoming chapters.*

translation to terminate. Unless a stop mutation is near the 3' end of the coding region, it will usually obliterate protein function. Most mRNAs with premature stop codons are degraded through a process called Nonsense Mediated Decay (NMD) to prevent translation. [g]

• **Frameshift.** So far we have been talking about SNP changes, but as we'll discuss below, it's also possible to have insertions and deletions of DNA sequences. Unless these are in multiples of three nucleotides, they cause the reading frame of the protein to shift. Like stop mutations, unless these are near the end of the protein sequence, these would also generally destroy protein function.

• **Splice Site Disruption.** The next type of variant is a little different. Recall from Chapter 1.2 that genes contain introns within the coding region; these must be spliced from the transcript to produce a functional protein. The positions of exon-intron boundaries are encoded in the DNA: this code includes a required GT at the start, and AG at the end of most human introns, as well as other sequences that help position the splicing machinery. Mutation of either the GT or AG usually prevents splicing or moves the location of splicing. We'll see an example of this shortly.

As a broad generalization, **single nucleotide mutations with large effects, such as in monogenic diseases and cancer are primarily driven by effects on coding sequences**.

2. Effects on gene regulation. As we discussed in Chapter 1.2, the second important function of DNA is to encode gene regulation: how much mRNA (and protein) from any given gene should be produced in a particular cell type, or phase of development.

While the DNA encodes precise patterns of gene regulation, there is no analog to the "genetic code" for proteins. As we discussed in the previous chapter, gene regulation is achieved through a complex dance of DNA-protein and protein-protein interactions that stabilize RNA Pol-II at the promoter and enable transcription. Much of the regulatory information is encoded through sequences that control the ability of transcription factors to bind at particular sites.

Thus, SNP alleles can affect expression by changing the encoded regulatory information – for example by increasing or decreasing the strength of transcription factor binding at a particular site. But, because TF binding is generally controlled by multiple sites, and because expression of any single gene is usually controlled by the interplay of many proteins, **the effects of individual SNPs tend to be quantitative – slightly increasing or decreasing expression – rather than turning expression on or off.**

As a result, it is unusual for individual SNPs in gene regulatory elements to result in single-gene diseases. However, genome-wide there are hundreds of thousands of SNPs with regulatory effects, and **in aggregate regulatory SNPs are the main drivers of most common phenotypic variation, and probably evolutionary change.** [h]

[g] *Variants that destroy the functional protein are called* ***Loss of Function (LOF)****: these would include most stop mutations, splice site disruptions and frameshifts.* ***LOF mutations are usually at very low frequencies*** *due to selection against them.*

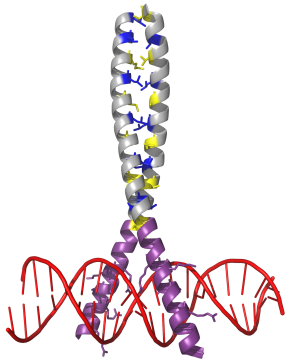


Figure 1.36: **Transcription factor binding to DNA.** *Recall that gene regulation is largely controlled by sequence-specific TF binding to DNA sequences. Thus, sequence changes can increase or decrease TF binding strength, thereby potentially changing expression.* *Credit: Houq [Link] CC-BY-SA-3.0*

[h] *We'll come back to regulatory variation in much more detail in Section 4 of the book.*

We close out this section by reviewing the story of one of the most famous mutations in history, and its interesting mechanism.

Example: hemophilia in the royal families of Europe. One famous example of a SNP mutation comes from the inheritance of hemophilia in the European royal families of the 19th and 20th Centuries.

Hemophilia is a genetic disease, caused by mutations in either of two X chromosome genes that are essential for normal blood clotting. Since males only have one X chromosome, any male with the mutation will have the disease. In contrast, females with one copy of the mutation do not have the disease, but can transmit to their children. Prior to modern treatments, affected individuals often died at young ages, but hemophilia can now be treated using clotting factors.

The 19th century British queen, **Queen Victoria** (1819-1901), is the first person in her family known to have carried the mutation. One of her three sons had the disease; he and two of Victoria's daughters passed it into the royal families of Spain, Germany and Russia. Ultimately, eleven male-line descendants of Victoria had hemophilia, spread across 4 generations. Victoria's last known descendent with hemophilia died in 1945.

Even though Victoria's mutation no longer exists in any living person, recent genetic analysis was able to determine the causal variant.

Figure 1.37: **Tsarevich Alexei of Russia in 1916** *(front right), with his family and cossacks. Anastasia front left.* *Credit: Beinecke Rare Book and Manuscript Library, Yale University; [Link].*

We pick up the story with one of Victoria's great-grandsons, the **Tsarevich Alexei Nikolaevich**, born in 1904 as heir apparent to the Russia throne. Alexei inherited the hemophilia mutation from his mother, the Tsarina Alexandra. He almost died from blood loss at birth, and suffered throughout his life from dangerous hemorrhages resulting from the minor bumps and bruises of childhood. After the Russian Revolution of 1917, Alexei and his family were exiled to Siberia. The following year the Bolsheviks executed the entire family. Much later, amid persistent rumors that Alexei and one of his sisters had escaped, the remains were exhumed and eventually subjected to genetic analysis in the mid-1990s that confirmed their identities [59].

But what about the hemophilia mutation? In 2009, a Russian team sequenced the main hemophilia genes in DNA recovered from Alexei and his sister Anastasia. They identified a causal mutation in the Factor IX gene that was present on Alexei's one X chromosome, and heterozygous in Anastasia [60].

This mutation has a very interesting mechanism. Recall that the 3' ends of introns are indicated, in part, by an AG dinucleotide. In this case, the mutation creates a new AG near the 3' end of the intron, two base pairs upstream from the original wildtype AG splice site. The new AG is preferred by the splicing machinery, and this results in the exon being extended by two base pairs. This in turn creates a frameshift in the reading frame, leading to a nonfunctional protein:

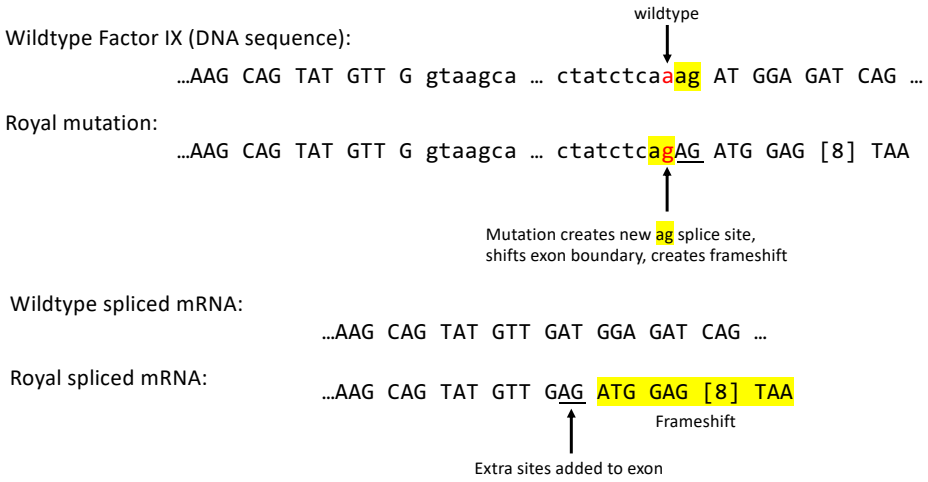


Figure 1.38: **Mechanism of the royal hemophilia mutation.** *Exonic nucleotides are in upper case letters and intronic in lowercase. The a→g mutation is marked in red; it creates a new 3′ ag splice site, which shifts the position of the second exon to the left by two base pairs. Thus the underlined G becomes part of the exon in the mutated gene. The lower panel shows the spliced mRNA with coding triplets indicated. Addition of G to the exon creates a frameshift (highlighted), which extends another 10 amino acids before terminating in TAA (STOP).* *Credit: Redrawn from Rogaev et al 2009 [Link]*

The royal mutation is just one of many different mutations that can cause hemophilia: the global prevalence of hemophilia is about 1 per 5,000 male births, caused by more than 1,000 different mutations in the two main hemophilia genes. A database of mutations in the Factor IX gene provides a sense of the relative importance of different disease mechanisms at this gene:

Mutation type	Number	% severe
missense	558	39%
frameshift	130	78%
nonsense	77	75%
splice site	83	41%
promoter	18	17%

Table 1.3: Mutation types in the Hemophilia B disease database *(Factor IX). Notice that most frameshift and nonsense mutations cause severe disease (unless they are near the end of the transcript), while other mutation types are less often severe.* *Simplified data from Table 2 of Tengguo Li et al (2013) [Link]. For brevity, minor categories including structural variants are not shown.*

As you can see above, most of the Hemophilia mutations affect either the protein coding sequence (missense) or entirely rewrite the protein sequence (frameshifts, nonsense, and splice site mutations).

Just a tiny fraction of the mutations are located in the promoter; these presumably change gene expression. This distribution of mutation mechanisms is typical of monogenic diseases. In contrast, we'll see later that regulatory variants are the major drivers for complex traits.

How does structural variation affect the information in genomes? As we discussed for SNPs, structural variants can affect both protein coding sequences and expression, but with a wide diversity of possible outcomes.

Changes in protein-coding sequences. Some smaller-scale variants such as STRs and VNTRs sit inside protein coding sequences; hence changes in copy number alter the protein coding sequence, as we described above for the ACAN gene which affects bone growth (and therefore height). Another famous example, Huntington's disease, is caused by an STR re-

peat (CAG, coding the amino acid glutamine). Alleles with large numbers of CAG produce long glutamine tracks; these form aberrant protein clumps in neurons, leading to a severe neurological symptoms. We'll come back to Huntington's Disease in Chapter 1.5.

But while this type of mechanism is important in a handful of genes, it impacts relatively few genes across the genome [61].

Copy number changes. In contrast, changes in copy number usually act through a completely different mechanism: they **alter the expression levels** of any genes contained within the affected segments: typically to 50% of wildtype for a heterozygous deletion, or 150% for a duplication.

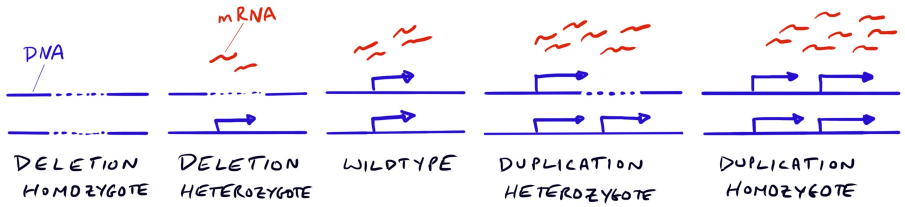


Figure 1.39: **Expression reflects copy number.** *The cartoon shows what's known as an* ***allelic series*** *in which the copy number of a particular gene (marked by the arrow) ranges from* 0 *to* 4 *copies in different individuals. mRNA (and protein) expression is usually roughly proportional to the gene's copy number.*

Does this matter?

It turns out that cellular systems are often sensitive to the precise expression levels of genes. For example, cellular differentiation is controlled by transcriptional regulatory networks, and small changes in expression of key genes can lead to widespread changes in expression. Secondly, many proteins act as components of multi-protein complexes that must be formed in precise ratios. Under- or over-expression of any component of a complex can have deleterious consequences.

Reflecting the importance of expression levels, some genes are described as **haploinsufficient**, meaning that a single functioning copy of that gene would not be sufficient for normal development or health. There are several hundred genes with known haploinsufficiency, resulting in major phenotypic effects.

Meanwhile, a much larger number of genes are **copy-number sensitive**: i.e., copy-number changes have measurable effects on survival or reproduction: it's estimated that, for about 20% of genes, loss of one copy results in a 10% loss of fitness [62]. (Here, "fitness" is a combined measure of survival and reproduction.) Thus, most large deletions or duplications (1 Mb or more, say) are extremely likely to overlap one or more copy-sensitive genes. Such large events often cause severe genetic syndromes in heterozygotes and are usually very rare in the population.

Adaptation through copy-number changes. However, copy number changes are not universally negative. Very occasionally copy number expansions evolve as a mechanism for increasing expression. For example, the amylase gene, AMY1, which is responsible for breaking down dietary starch, is present in our genome with a variable number of copies, ranging from around 2-16 copies per person [63]. These copies appear as **tandem repeats** within a single genomic region ("tandem" here meaning adjacent, rather

than scattered around the genome), as you can see in the FISH image below [64]. (FISH is a technique in which DNA probes with fluorescent tags are hybridized to specific DNA sequences so that they can be imaged with microscopy.)

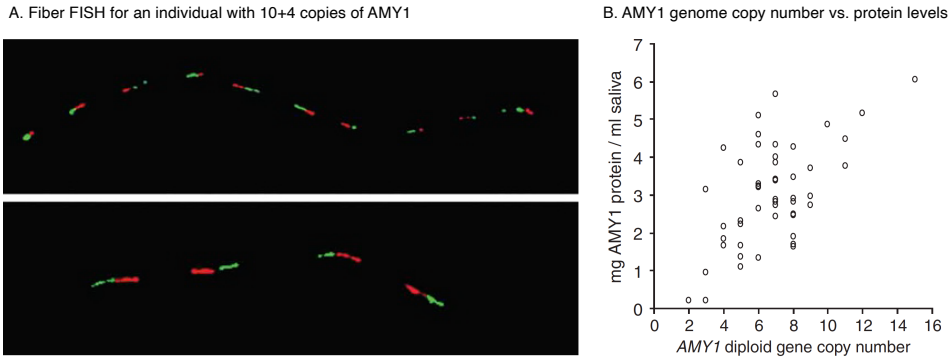


Figure 1.40: **Genome and protein variation of salivary amylase. A.** *Fiber FISH has been used to label individual copies of the AMY1 locus in this microscopic image. The two images show the two homologous copies of this region; each AMY1 copy is marked by one green and one red block, showing* 10 *copies on one homolog and* 4 *on the other.* **B.** *Genomic copy number of AMY1 is highly correlated with Amylase protein levels in the saliva, in a sample of* 50 *individuals.* *Credit: Modified from Figure 3a and 1c of George Perry et al (2007) [Link] Used with permission.*

As you can see above, variation in amylase copy number also has functional consequences: higher copy number is correlated with higher protein levels in the saliva, which may enhance starch digestion.

There's one more fascinating aspect to the amylase story: The copy number of amylase in humans is greatly expanded relative to other great apes (e.g., the copy number in chimpanzees is around 3 per haploid genome). This has led to the hypothesis that the copy number expansion is an evolutionary adaptation to the higher levels of starch in human diets compared with our evolutionary ancestors. Remarkably, some other species that are associated with humans (and may also have high-starch diets) also have expansions of the amylase locus, including dogs compared to wolves, and mice and rats compared to other rodents [65].

Chromosome inheritance errors: aneuploidy. Lastly, we'll talk briefly about **a completely different kind of genetic variation that is usually not inherited**: the cases where the fertilized egg inherits too many, or too few chromosomes, a situation known as **aneuploidy**. In most cases, aneuploidy has major phenotypic consequences, and is generally not transmitted to subsequent generations [i].

Recall from our last chapter that we discussed the process of **meiosis**. Most human cells contain two sets of 23 chromosomes, but eggs and sperm each carry half the usual number: 1 set of 23 chromosomes each. Meiosis is the reduction process in which the number of chromosomes is cut from 46 to 23. A fertilized egg then receives 23 chromosomes from each parent to bring it back to the correct number of 46 total chromosomes.

Sometimes there are errors in meiosis where two homologs stick together during cell division and both wind up in the same cell. When this happens, the fertilized egg ends up with either an extra chromosome (3 copies of one of the chromosomes for a total of 47) or missing a chromosome (1 copy of that chromosome for a total of 45). These errors occur most of-

[i] *Like the discussion of copy number variation above, aneuploidy illustrates a fundamental principle:* ***organisms are very sensitive to changes in the precise ratios of expression across genes****. Small variations in expression, spread across many genes, are major drivers of human phenotypic variation, evolution, and disease [66].*

ten in older mothers, for reasons we'll explain in the next chapter. (It's outside our scope in this chapter, but aneuploidy can also arise during mitosis and is a common feature of cancer genomes.)

Aneuploidy where a chromosome is present in single copy is referred to as **monosomy**, while aneuploidy with three copies of a chromosome is a **trisomy**. Most of the possible aneuploidies have very severe effects on global gene regulation, and the embryos do not survive to birth.

One exception is that embryos with an extra copy of the smallest chromosome, Chromosome 21, do sometimes survive to birth. These children have Down Syndrome, and usually have developmental delays and may suffer from health issues including heart problems. Children with trisomies of two other chromosomes (18 and 13) can also survive to birth, but they have severe disabilities and low survival.

Additionally, embryos with extra, or missing copies of the X and Y chromosomes also often survive to birth. Although these individuals often exhibit developmental problems, the X/Y aneuploidies are generally much less severe than autosomal aneuploidies for reasons we'll describe below. Individuals with at least one Y are usually assigned male sex at birth, regardless of the total number of Xs and Ys. This is because the SRY gene, which encodes a transcription factor that turns on male developmental programs, is located on the Y chromosome. The side image shows the **karyotype** – the number and identities of the chromosomes – in a boy with Klinefelter Syndrome (two X and one Y).

The most frequent types of aneuploidy in humans that survive to birth are listed below. As we'll discuss shortly, these include the three chromosomes with the fewest genes, and unusual combinations of the X and Y chromosomes.

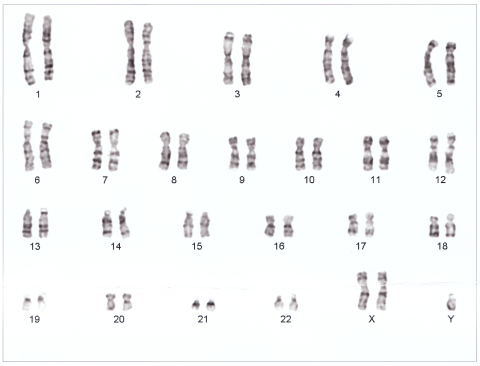


Figure 1.41: **Karyotype of a boy with Klinefelter Syndrome**: *he has two X and one Y chromosome. The chromosomes were imaged while condensed during mitosis, and then positioned in order.* *Credit: Nami-ja [Link] Public Domain*

Syndrome	Frequency	Notes
Trisomy 13 (Patau Syndrome)	1/5,000	Severe developmental issues; 5-10% survival at first year
Trisomy 18 (Edwards' Syndrome)	1/5,000	Severe developmental issues; 5-10% survival at first year
Trisomy 21 (Down Syndrome)	1/1,000	Mild to moderate intellectual disability, low fertility
X (Turner Syndrome)	1/2,000	Female, extensive developmental issues, infertile
XXX (Triple X Syndrome)	1/1,000	Female, frequent physical/learning issues, often fertile
XXY (Klinefelter Syndrome)	1/500	Male, may have some feminized features, low fertility
XYY	1/1,000	Male, symptoms usually absent, normal fertility

Table 1.4: *The main types of aneuploidy that can survive to birth. Frequency estimates reflect rates among live births; birth rates for some of these conditions are declining due to prenatal screening.*

Why does aneuploidy affect development? An individual with a monosomy or trisomy can still make the full complement of proteins. But as we discussed above, cells depend on having a precise balance of all their proteins. Individuals with an extra (or a missing) chromosome produce

either too much, or too little of all the proteins on that chromosome, and these imbalances lead to major developmental problems. By and large, these are not due to the impact of a few genes that are particularly sensitive to copy number, but instead the accumulated effects of hundreds to thousands of genes with one extra copy each.

One indication of this is that the severity of trisomies reflects the number of genes on each chromosome: it's no coincidence that Trisomy 21 (Down Syndrome) is the least severe trisomy, and that 18, and 13 are next in line, as these are the three chromosomes with the fewest genes (among the autosomes). A similar result can be seen in mice, where there is a tight inverse correlation between chromosome length and how long the corresponding trisomies can survive [67]:

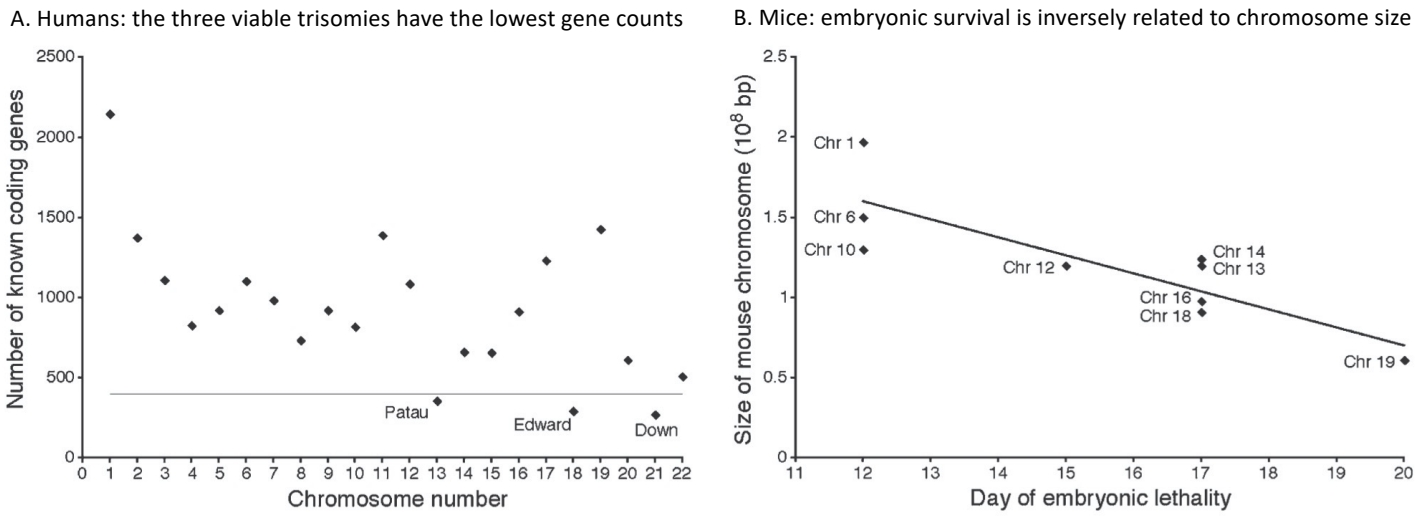


Figure 1.42: Trisomies of chromosomes with fewer genes are more viable. A. *The plot shows gene number for each autosome (y-axis), ordered by chromosome number* 1–22. *The three autosomes with viable trisomies are those with the fewest genes.* **B.** *In a study of trisomies of different mouse chromosomes, there was a strong relationship between chromosome size and how long the mice could survive.* *Credit: Modified from Figure 2 of Eduardo Torres et al (2008) [Link] Used with permission.*

The situation for the sex chromosomes is a little different, but reflects the same principles.

Remember that usually females have 2 X chromosomes, and males have an X and a Y. Ordinarily in females one X chromosome in each cell is silenced by a process known as **X-inactivation**: in other words, only one chromosome is used for gene expression. This is important so that there is no major mismatch between the gene expression levels of X chromosome genes in males and females. This same mechanism rescues people who have three or more X chromosomes, because X-inactivation ensures that there is only one active X, regardless of the actual number of Xs.

Meanwhile, there are only about 50 protein-coding genes on the Y chromosome, and many of these are involved in development of the male reproductive system and of sperm, so individuals with an extra Y chromosome (XYY) are generally healthy.

The fact that individuals with unusual X/Y karyotypes often do have some degree of symptoms is because around 100 genes escape X inactivation [68]. Consequently, individuals with unusual XY karyotypes do not maintain the correct dosages for these genes: this can lead to developmental and health issues, especially for Turner's Syndrome (X-) patients, as outlined in Table 1.4 [69].

Karyotypes evolve rapidly over evolutionary time! Given the strong constraints within the human population in maintaining correct chromosome numbers, it may come as some surprise to learn that closely-related species often evolve quite different karyotypes [70] [j].

[j] *For an interesting table of chromosome numbers in different species see: Link.*

For example, humans have 23 pairs of chromosomes, while other great apes have 24 pairs: this is because human chromosome 2 is a fusion of two ancestral ape chromosomes [71].

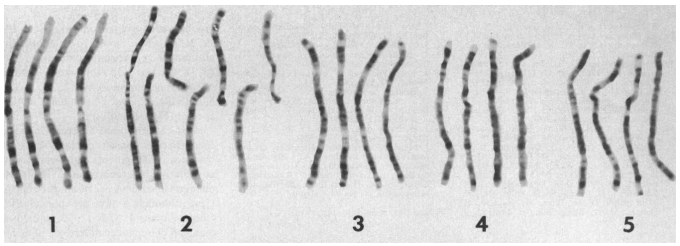


Figure 1.43: **Partial karyotypes of the great apes.** *The image shows human chromosomes 1–5, pictured alongside the corresponding great ape chromosomes. Left to Right: human, chimpanzee, gorilla, orangutan. Human Chromosome 2 is a fusion of two chromosomes that are separate in the other apes.* *Credit: From Figure 1 of Jorge Yunis and Om Prakash (1982) [Link]Used with permission.*

While the overall structure of the great ape genomes are largely similar aside from this fusion event, our next-nearest relatives, the gibbons, have undergone extraordinarily rapid chromosome evolution. For example, the genome of the northern white-cheeked gibbon has been dramatically reorganized: it has 26 pairs of chromosomes, but even more strikingly it differs from humans by 96 different rearrangements of large chromosomal blocks [72]!

Ordinarily, one might expect major chromosomal rearrangements to have deleterious effects, and these are almost vanishingly rare within human populations [73]. Thus it's surprising that closely-related species can evolve dramatically different karyotypes. One potential explanation is that this may be driven in part by the evolution of new centromeres that can hijack meiosis to gain a selective advantage and spread through populations – but this is an exciting area that is not yet well understood [74].

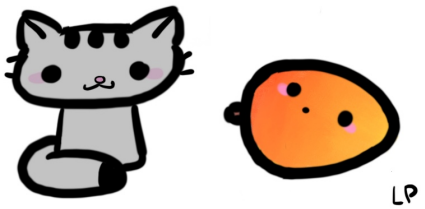


Figure 1.44: *We like to tease our pet cats that they are one chromosome short of a mango! (Cats have 19 pairs of chromosomes – and mangoes have 20.)* *Credit: Lucy Pritchard*

In summary, the genomes of any two people differ at millions of positions, including SNPs, as well as a variety of more complicated types of sequence differences. In the next chapter we will discuss how these differences can be detected by DNA sequencing.

Notes and References.

[43]To be more precise, the vast majority of SNPs only have two alleles at any appreciable frequency. However, as we discuss below, virtually every possible allele that is one step away from the reference genome exists somewhere in the world (excluding alleles that would be incompatible with life).

[44]You can imagine that there are pros and cons to each naming system. The *reference allele* is rather arbitrary, because it depends on whether the allele happens to match the individual who was sequenced at that position for the Human Genome (and sometimes that individual had a super rare allele). The *minor allele* label is particularly useful for rare alleles, but it can lead to inconsistent labeling across different samples if the allele frequency is near 0.5. The *derived allele* label is attractive in having a clearer evolutionary interpretation, but it involves an inference about which allele is ancestral that may be uncertain or even incorrect for some SNPs.

[45]For autosomal loci, one generation of random mating (i.e., random with respect to the SNP in question) immediately restores HW proportions regardless of the starting allele frequencies. This means that a process like selection must be implausibly strong to drive meaningful departures from HWE. Note that X-linked loci do not reach HWE immediately (but do converge within a few generations).

[46]Genotyping issues that lead to departures from HWE can occur for various reasons, and the details depend a bit on the specific technology. One common reason for errors is that the sequence surrounding a putative SNP is duplicated elsewhere in the genome and so the sequencing reads or genotyping assay contain a mixture of DNA fragments from two different locations. Suppose that these two duplicated versions of this region differ at exactly one position, and this position has been inferred incorrectly as a SNP. Then all individuals would appear to be heterozygous.

[47]Edwards A. Anecdotal, Historical and Critical Commentaries on Genetics: GH Hardy (1908) and Hardy–Weinberg Equilibrium. Genetics. 2008;179(3):1143.

[48]Genomes of "identical" (monozygous) twins are in fact nearly identical: the genomes of a monozygous pair differ by only ∼5 early developmental mutations in non-repetitive sequences, as well as presumably additional STRs and other more-mutable sequences that are more difficult to measure:

Jonsson H, Magnusdottir E, Eggertsson HP, Stefansson OA, Arnadottir GA, Eiriksson O, et al. Differences between germline genomes of monozygotic twins. Nature Genetics. 2021;53(1):27-34

[49]We can generalize the concept of heterozygosity to consider the expected heterozygosity under random mating. The expected heterozygosity is useful if we don't have access to individual-level genomes, and the estimator also has lower variance. For example, if we know the allele frequency p_s at every SNP s in a region of size L, then we can compute the expected heterozygosity as

$$\frac{1}{L}\sum_s 2p_s(1-p_s).$$

(Note that in practice the formula above is slightly biased since we only have estimates of p_s rather than true values; an unbiased formula can be derived by computing the heterozygosity summed over all pairwise comparisons.)

[50]1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68

[51]Large sequencing studies continue to find many more novel, rare SNPs: for example the gnomAD Project identified 230M high confidence variants – nearly one every 10 bp – by sequencing about 16,000 genomes. Note that the gnomAD Project had higher sequencing depth than 1000 Genomes, and this accounts for why they detected more new variants per individual. gnomAD Project:

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43

[52]We'll return to questions about divergence among the great apes in Chapter 2.2.

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, et al. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483(7388):169-75

[53]This was laborious work that relied on PCR amplifying regions of interest, followed by Sanger sequencing. Anna Di Rienzo's lab, at the University of Chicago, also did important work in this area at around the same time.

Frisse L, Hudson R, Bartoszewicz A, Wall J, Donfack J, Di Rienzo A. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. The American Journal of Human Genetics. 2001;69(4):831-43

Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. nature Genetics. 2003;33(4):518-21

[54]Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. Human Molecular Genetics. 2005;14(1):59-69;

Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional

impact of short insertion–deletion variants identified in 179 human genomes. Genome Research. 2013;23(5):749-61

[55]VNTRs are also sometimes known as minisatellites, while STRs are also microsatellites.

[56]Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253):aab3761

[57]Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, et al. A common inversion under selection in Europeans. Nature Genetics. 2005;37(2):129-37

Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Research. 2012;22(6):1144-53.

One effect of inversions is that they disrupt local recombination in heterozygotes. In some species this enables the evolution of co-adapted gene clusters, but there are no clear examples in humans: Inversion coadapted complexes

Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. Trends in Ecology & Evolution. 2018;33(6):427-40.

[58]The main exceptions where a synonymous variant has a phenotypic effect are usually due to some regulatory function that overlaps with the same positions – for example that the variant is contained with a transcription factor binding site or exonic splicing enhancer.

[59]For a good account of the genetic testing, with quite a bit of historical and forensic context see

Coble MD, Loreille OM, Wadhams MJ, Edson SM, Maynard K, Meyer CE, et al. Mystery solved: the identification of the two missing Romanov children using DNA analysis. PloS One. 2009;4(3):e4838.

[60]Rogaev EI, Grigorenko AP, Faskhutdinova G, Kittler EL, Moliaka YK. Genotype analysis identifies the cause of the "royal disease". Science. 2009;326(5954):817-7

[61]Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Zheng Y, McCarroll SA, et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science. 2021;373(6562):1499-505

[62]The most relevant studies test for a depletion of LOF mutations compared with a neutral background. If this is detected it implies that there is at least some degree of selection against heterozygous LOFs. The effects of haploid gene deletions should be roughly functionally similar to haploid LOFs.

Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91

Agarwal I, Fuller ZL, Myers S, Przeworski M. Relating pathogenic loss-of function mutations in humans to their evolutionary fitness costs. bioRxiv. 2022

[63]Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nature Genetics. 2007;39(10):1256-60 CITE NOVEMBRE TOO

[64]While the main form of variation at Amylase1 is variation in copy number, it turns out that there is also additional complex structure within the region, as the gene copies appear in several slightly different forms that are variable across individuals:

Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nature Genetics. 2015;47(8):921-5

[65]Pajic P, Pavlidis P, Dean K, Neznanova L, Romano RA, Garneau D, et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. Elife. 2019;8:e44628

[66]It's interesting to note that polyploidy (usually 3 or 4 copies of *all* chromosomes) can be less deleterious than aneuploidy of a single chromosome. Many species, across the tree of life, have evolved polyploid genomes, and it's believed that our own ancestors went through two rounds of whole genome doubling in early tetrapod evolution. Moreover, some human tissues, including liver, placenta, and heart are polyploid. This indicates that problem with aneuploidy is that changes the relative proportions of genes (stoichiometry) relative to one another, not the absolute changes in expression of specific genes.

[67]Torres EM, Williams BR, Amon A. Aneuploidy: cells losing their balance. Genetics. 2008;179(2):737-46

[68]These mainly fall into three categories: (1) There is a pair of *pseudo-autosomal regions*, containing a total of 3 Mb of DNA and 20 genes, that are shared between the X and Y chromosomes and are important for proper chromosomal pairing during meiosis and mitosis; (2) Secondly, there are about 25 genes with essential roles in gene and protein regulation, that have homologs on the X and Y chromosome. These genes have evolved to escape X-inactivation because both XX and XY individuals have two functional copies; (3) genes that are not particularly dosage-sensitive. For estimates of the number of genes that escape X inactivation see Balaton 2015 [Link]

[69]For a more detailed discussion of this see

Posynick BJ, Brown CJ. Escape from X-chromosome inactivation: an evolutionary perspective. Frontiers in Cell and Developmental Biology. 2019;7:241

For analysis of X-Y homologs and their functions see:

Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature. 2014;508(7497):494-9

[70] Ferguson-Smith MA, Trifonov V. Mammalian karyotype evolution. Nature Reviews Genetics. 2007;8(12):950-62

[71] Yunis JJ, Prakash O. The origin of man: a chromosomal pictorial legacy. Science. 1982;215(4539):1525-30

Ventura M, Catacchio CR, Sajjadian S, Vives L, Sudmant PH, Marques-Bonet T, et al. The evolution of African great ape subtelomeric heterochromatin and the fusion of human chromosome 2. Genome Research. 2012;22(6):1036-49

[72] Carbone L, Harris RA, Vessere GM, Mootnick AR, Humphray S, Rogers J, et al. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. PLoS Genetics. 2009;5(6):e1000538

Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature. 2014;513(7517):195-201

[73] There are rare examples of balanced translocations that are inherited within families, but I'm not aware of any chromosomes fusions or fissions.

[74] Chmátal L, Gabriel SI, Mitsainas GP, Martínez-Vargas J, Ventura J, Searle JB, et al. Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. Current Biology. 2014;24(19):2295-300

1.4 DNA sequencing: a fundamental tool for studying biology.

In which we take a detour to discuss DNA sequencing and genotyping.

Microscopes were invented in the early 1600s, and opened up a new unimagined world. Robert Hooke is credited with the discovery of plant cells in 1665; shortly after Antonie van Leeuwenhoek observed microbes including bacteria and protozoa, as well as human cells including spermatozoa, blood, and muscle cells. Despite important recent advances in microscopy, DNA and many of the processes that we focus on in genetics, are too small to see clearly by microscopy. In particular, DNA molecules are much too small to read the nucleotides directly.

However, starting from the 1970s, there has been a rise of new technologies that allow us to read the nucleotide sequences of DNA molecules: this is **DNA sequencing**. In modern biology, **DNA sequencing has become a truly transformative tool**, opening up new avenues of exploration that were unimagined prior to the sequencing era.

First, and probably most obvious, it's now relatively cheap and easy to sequence genomes. We now have genome sequences for thousands of different species. Hundreds of thousands of different humans have now been genome sequenced.

But beyond this, there's a dizzying array of different things that we can now measure using DNA sequencing technology [a]. For example if a patient has cancer, we can sequence the genome of the cancer cells, to understand what genetic changes enable uncontrolled growth (which may indicate the use of particular treatments). When a woman is pregnant, we can sequence her baby's DNA, using free-floating DNA fragments in the mother's bloodstream to predict genetic diseases before birth. We can use DNA sequencing to measure the microbial population that lives in everyone's guts (the microbiome), or in agricultural or wild soil samples. We can use sequencing to detect the presence of viruses in patients, or in the environment: on surfaces, in the air, or in wastewater. Sequencing has been an essential tool for tracking the evolution of the SARS-CoV-2 virus, and to identify outbreaks of novel strains.

In the lab, DNA sequencing has also transformed genomics research. We use DNA sequencing to measure many different aspects of how a genome functions in different cell types: for example which parts of the genome are bound by a particular protein; which parts of the genome are actively involved in regulating genes; which genes are being expressed, and how much; what cell types are present in a tissue sample; sequencing is being used to transform developmental biology. Almost any lab experiment involving genomes or cellular functions can now be set up to end with data collection by DNA sequencing.

In short, DNA sequencing is like a new microscope for the 21st century[75].

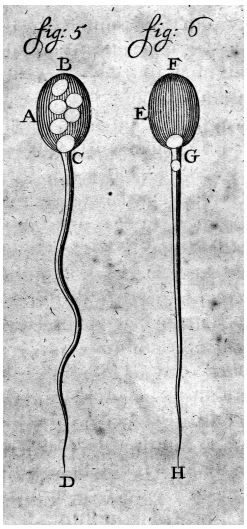


Figure 1.45: **van Leeuwenhoek's drawings of spermatozoa (1719)**. *The early microscopists revealed for the first time a world of cells, fine structures of tissues, and microbial life.* ***DNA sequencing is, similarly, now revealing new worlds.*** *Credit: Opera Omnia (1719). [Link] CC BY 4.*

[a] *DNA sequencing is a fundamental technology that allows highly efficient detection, counting and sequencing of biological molecules. As such it is transforming a vast array of different applications in biological research and medicine, not just limited to the determination of genome sequences.*

A short history of sequencing. The first practical DNA sequencing was achieved in the 1970s. Two scientists (Fred Sanger and Walter Gilbert) won the Nobel prize in 1980 for developing techniques that could sequence up to around a hundred basepairs of DNA at a time. During the half-century since this work, sequencing has become millions of times faster and cheaper, now enabling rapid, affordable whole genome sequencing.

First generation sequencing. The technique introduced in 1977 by Fred Sanger – now known as **Sanger sequencing** – provided the first practical sequencing, and was the basis for nearly all sequencing projects until about 2005 [76]. Sanger sequencing is still used in lab-work for small-scale applications [77].

In Sanger sequencing, DNA polymerase is used to copy a single-stranded DNA template. The reactions are run in a soup of ordinary DNA nucleotides and a small fraction of so-called dideoxy nucleotides, which block further extension of the sequence. In the original Sanger sequencing, template copying was performed in four distinct reactions, one for each termination nucleotide: with dideoxy-A, dideoxy-C, and so on. Ultimately, the dideoxy-A reaction would contain a collection of DNA fragments of different sizes, corresponding to all the fragment sizes that end in A. The fragments were run out on an electrophoretic gel that separates molecules by size. DNA fragments were labeled with radioactive atoms so that they could be detected on x-ray films. You can see an example in the image at the right.

In later iterations of Sanger sequencing, each dideoxy nucleotide was instead labeled with a different fluorescent dye. This means that sequencing could be performed in a single reaction, and the bands can be recognized by color using a scanning laser. This in turn enabled further miniaturization, as each reaction could be run through a thin capillary tube, instead of in a large slab gel.

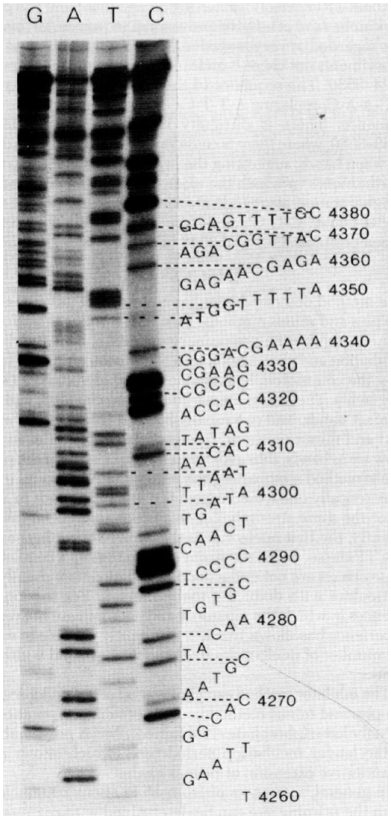


Figure 1.46: **Sequencing autoradiograph** *from the 1977 paper that introduced Sanger sequencing. The image displays a short sequence from the virus ϕX174. Each band corresponds to fragments of a specific nucleotide length (shortest at the bottom); each lane contains fragments that terminate in the nucleotide shown at top. Though blurry at points, this allowed the sequence to be read from the image, as shown at right.* *Credit: Figure 2 from Fred Sanger et al, 1977 [Link].*

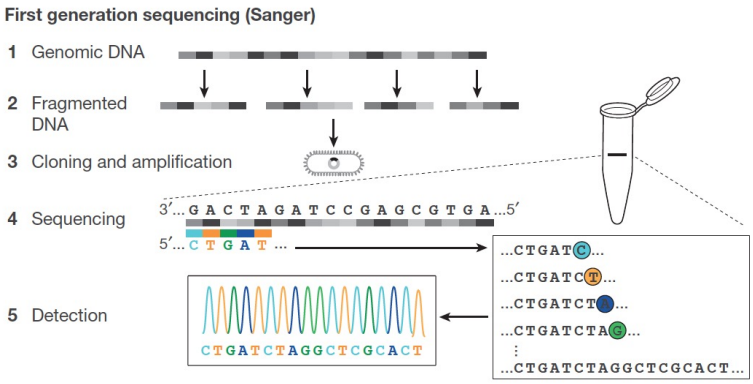


Figure 1.47: **Sanger Sequencing.** *In modern Sanger sequencing, DNA fragments of different sizes are labeled with fluorescent dyes according to the 3′ nucleotide on each fragment. The fragments are size-separated using gel electrophoresis, and the colors are recorded by a scanning laser as they migrate through the gel, thereby providing the DNA sequence.* *Credit: From Figure 1, Jay Shendure et al. 2017. [Link]Used with permission.*

These improvements to Sanger sequencing enabled the first sequencing of eukaryotic genomes in the 1990s, leading to the draft human genome in 2000 – completed at a cost of $2.7 Billion, calculated in 1991 dollars [78].

Second generation sequencing. But at these prices, genome-scale work was extremely expensive, and limited to "genome centers", which were

essentially dedicated sequencing factories.

The early 2000s saw a major paradigm shift, with a group of new sequencing technologies that achieved enormous advances over Sanger sequencing. These new methods – also called **next-generation** or **massively parallel sequencing** – were dramatically faster, and required smaller amounts of expensive reagents. These became commercially available by around 2006 and greatly reduced sequencing costs, enabling individual labs to perform genome-scale sequencing projects for the first time. Over the next few years one technology, owned by the company Illumina, gained a dominant position in the DNA sequencing market and currently enjoys a near-monopoly in high throughput sequencing [79].

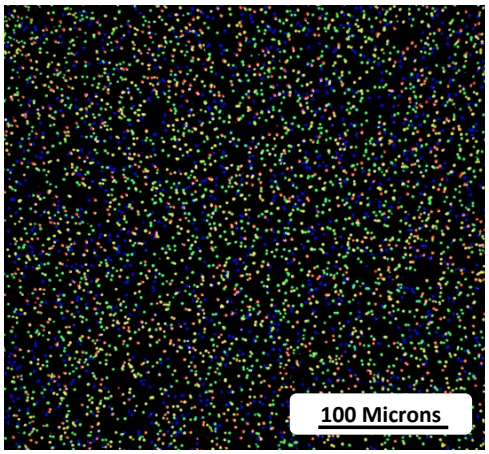


Figure 1.48: **Illumina flowcell.** *The image shows a tiny part of a flowcell. Each dot represents a DNA cluster, and the colors indicate the nucleotide added in the current cycle.* *Original image source unknown.*

Illumina's approach [80] starts by attaching billions of DNA fragments to a solid surface called a **flow cell**, which is similar to a microscope slide. These are used to create colonies of identical single-stranded DNA fragments. Sequencing proceeds using a **sequencing by synthesis** approach, in which fluorescent nucleotides are added to the complementary strand one-at-a-time. At each cycle of the experiment, one of the four possible nucleotides is added to each colony, depending on the sequence on the template strand, and the corresponding colors are recorded. The sequence of colors for each colony indicates the correct sequence.

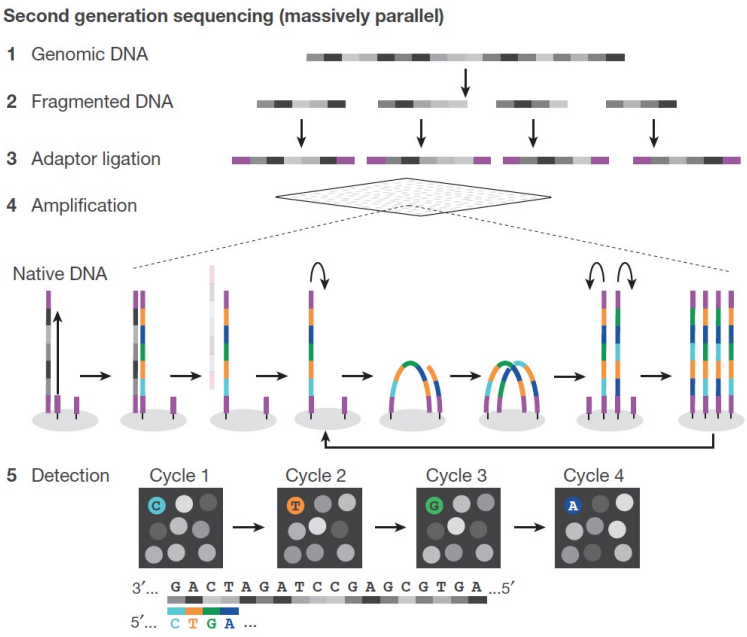


Figure 1.49: **Massively parallel short-read sequencing**, *e.g., Illumina. Colonies of identical single-stranded DNA fragments are attached to a solid surface. Sequencing occurs through DNA synthesis, as colored nucleotides are added one at a time and imaged.* *Credit: From Figure 1, Jay Shendure et al. 2017. [Link]Used with permission.*

Illumina sequencing is vastly more efficient than Sanger sequencing. In Sanger sequencing each reaction must run through a separate electrophoretic channel (a capillary, or lane on a gel); in contrast, Illumina sequencing is only limited by the number of DNA colonies that can be placed and imaged on a flow cell without overlapping.

However, an important limitation of Illumina sequencing is that the sequence reads are relatively short. The current main platform sequences 150 bp from both ends of a larger molecule (typically one might input

DNA molecules of perhaps 600 bp, and then sequence 150 bp from each end). Modern Sanger sequencing reads are slightly longer, reaching up to ~800 bp. In sharp contrast, Third Generation methods, which we discuss next, are routinely tens of kilobases and can reach megabase lengths.

Figure 1.50: **Paired-end reads** *are a standard sequencing format on the popular Illumina platform. This involves adhering both ends of a larger DNA fragment to the flow cell, and sequencing from both ends. The precise read lengths and fragment sizes vary across applications.*

Third generation sequencing. The 2010s have seen the emergence of a third paradigm for sequencing, which for the first time involves direct sensing of individual DNA molecules. At the time of writing, these are lower throughput and with higher error rates than second-generation sequencing, but they can provide extremely long sequence reads of individual molecules, potentially even to megabase-length reads.

At present, the two leading commercial technologies are from Oxford Nanopore Technologies and Pacific Biosciences. Oxford Nanopore's approach measures electrical conductance of a DNA molecule as it passes through a biologically-derived membrane channel (a nanopore). The different nucleotides can be recognized as producing different electrical signatures. PacBio positions individual DNA polymerases inside a measurement well that can accommodate a single DNA molecule. The well detects light emitted by fluorescent nucleotides that are incorporated, one-at-a-time, into a single growing DNA strand.

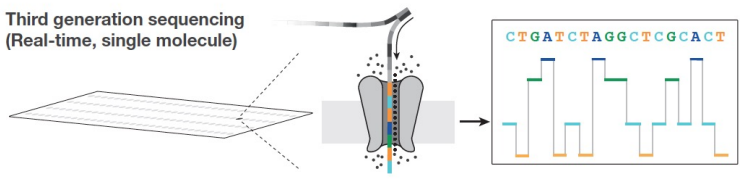


Figure 1.51: **Third generation single-molecule sequencing.** *Technologies including Oxford Nanopore and PacBio pass single molecules through molecular sensing devices. These can provide read lengths up to about* 1 *Mb.* *Credit: From Figure 1, Jay Shendure et al. 2017. [Link]Used with permission.*

At present, the long-read technologies are lower throughput and more error-prone than Illumina's short-read platforms; a 2020 paper estimated that the error rate per base pair for a single sequencing read can be as high as 10% versus around 0.1% for Illumina [81]. However, by sequencing to high depth it is possible to achieve comparable overall accuracy, albeit at higher cost per sample (about $5000 on the Nanopore platform for a clinical-grade genome in 2022 [82]). We can expect error rates and costs for 3rd-Gen sequencing to continue to drop.

Moreover, 3rd-Gen long reads enable some applications that are difficult, if not impossible, with short reads: these include sequencing of complex regions of the genome, and haplotype phasing. **3rd-Gen long reads were essential for the first truly complete human genome sequence, published in 2022** [83]. Furthermore, we can anticipate new advances in this space: for example, because these methods sequence individual molecules without amplification, it's possible to study DNA or RNA modifications such as methylation. Lastly, Oxford Nanopore sequencing is performed on portable USB devices that plug directly into a laptop – this makes it practical for field applications such as infectious disease surveillance in developing countries.

We can summarize the three main sequencing approaches as follows:

Sequencing Type	Read Length	Throughput	Error Rates	Single molecule
1st Gen (Sanger)	~800 bp	low	low	no
2nd Gen (e.g., Illumina)	~2×150 bp	very high	low	no
3rd Gen (e.g., Nanopore, PacBio)	up to ~1 Mb	medium	high	yes

Moore's Law and the dropping costs of DNA sequencing. During the last three decades, the increases in speed, and decreases in cost, of DNA sequencing have been absolutely gobsmacking [b]. It's interesting to compare the enormous improvements in the DNA sequencing industry to gains in another industry, computing, which has famously benefited from extreme miniaturization. **Moore's Law** is an observation from the computer industry that the number of transistors on a circuit chip doubled roughly once every two years; this remarkable rate of progress fueled the rise of computing.

DNA sequencing improved even more rapidly than Moore's Law from around 2007-2012, driven in large part by the transition to massively parallel sequencing. Subsequent stagnation in costs partly reflects that a single company currently controls the vast majority of the short-read sequencing market [84].

[b] *"Back in 1990, sequencing 1 million nucleotides cost the equivalent of 15 tons of gold (adjusted to 1990 price). At that time, this amount of material was equivalent to the output of all United States gold mines combined over two weeks. Fast-forwarding to the present, sequencing 1 million nucleotides is equivalent to the value of ~30 g of aluminum. This is approximately the amount of material needed to wrap five breakfast sandwiches at a New York City food cart." –Yaniv Ehrlich (2015).*

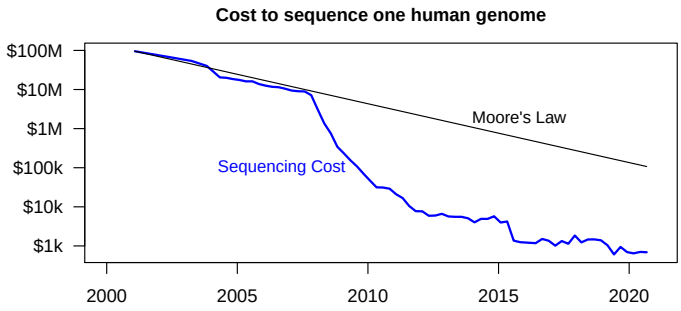


Figure 1.52: **The rapidly declining cost of DNA sequencing.** *The blue line shows the estimated cost to sequence one genome, from $95M in 2001 to $700 in 2020. The black line shows the Moore's Law prediction, projected forward from 2001.* *Credit: Redrawn from a figure and data by the US National Human Genome Research Institute (NHGRI) [Link].*

In the remainder of the chapter we will discuss the applications of sequencing in more detail, with an emphasis on "resequencing".

Sequencing applications in human genomics. We can classify most sequencing applications into three broad categories:

- **Genome resequencing and polymorphism discovery.** If we sequence your genome, and mine, what are all the places where we differ from each other? The analysis is usually performed by identifying differences from the Reference Genome. These applications are referred to as **resequencing** when analysis is based on comparison to a reference.
- **De novo genome sequencing and assembly.** How can we use sequencing to determine the genome of an unstudied species, or to determine the human genome in regions of high structural complexity and variability? Affordable 2nd- and 3rd-generation se-

quencing have now enabled genome assemblies for many thousands of different species, spread widely across the tree of life [85].

- **Sequencing as a molecular counting tool.** Most of this is outside our scope here, but since around 2005, there has been a huge shift toward using DNA sequencing as the readout for a huge array of molecular experiments: What are the expressed or regulatory regions of the human genome in any given cell type? What regions of the genome are amplified in a cancer cell? Which CRISPR guides lead to better cell survival in a functional screen?

In the remainder of this chapter we focus on the first of these goals, which is most essential for the topics of this book.

Genome resequencing and polymorphism discovery. Suppose we want to characterize the genetic variation in a sample of individuals; or perhaps we want to search for potential causal mutations in a child with severe developmental delays; perhaps we wish to find driver mutations in a cancer genome. How should we tackle these problems using current technologies?

Ideally you might imagine DNA sequencing providing a fully accurate end-to-end read-out of each chromosome. But no current technology can provide this directly. Instead, in practice we must balance a desire for high accuracy and completeness against considerations of cost and speed. At present (writing this in 2022), most resequencing projects are using Illumina short-read sequencing because Illumina reads are relatively cheap and accurate [c]. We discuss limitations below.

[c] *In this section we focus on **whole genome sequencing (WGS)**. At the end of this chapter we describe two alternatives: **exome sequencing**, and **genotyping**.*

Resequencing with short reads. Remember that human chromosomes are 50-250 Mb long, but what we get are billions of short reads of ∼150bp. To make matters worse, there is no easy way to indicate where in the genome each read comes from. Indeed, most genome sequencing uses what is called **shotgun sequencing**, in which we break the genome into many small fragments, sequence them, and rely on our ability to make sense of the sequence reads when we have them. One further challenge is that all DNA sequencing comes with occasional errors (e.g., about 1 per thousand nucleotides per sequencing read on Illumina [86]) and the data analysis must be robust to this [d].

[d] *As we describe below, we usually have many sequence reads spanning every position in the genome. This allows us to correct errors so that the final error rate in a finished sequence is much lower than the raw rate per read.*

A common pipeline for sequencing human genomes is as follows:
• **Extract DNA** from a tissue sample, e.g., from blood cells. The initial sample contains millions of cells (each with its own copy of the genome), and so the DNA fragments that we sequence are a mixture of many different copies of the same genome;
• **Smash up the genome** into ∼600 bp fragments of DNA for shotgun sequencing;
• **Map reads** to a standard reference human genome (see below);
• **Infer genotype differences** from the reference (e.g., SNPs and structural variants)

• **Interpretation of variation** – for example, identifying disease-associated variants.

The output from the sequencing machine consists of billions of short DNA sequence reads. Since we don't know in advance where each read comes from, the first step of analysis is to **map the reads** to a reference human genome. Conceptually you can think of read-mapping as being like taking a string of nucleotides and sliding it along the genome sequence until you find a location where it matches – very much like fitting a piece into a jigsaw puzzle. There are efficient computational algorithms to do this. The matching process must allow for modest levels of difference from the reference, as there may be SNPs, indels or sequencing errors.

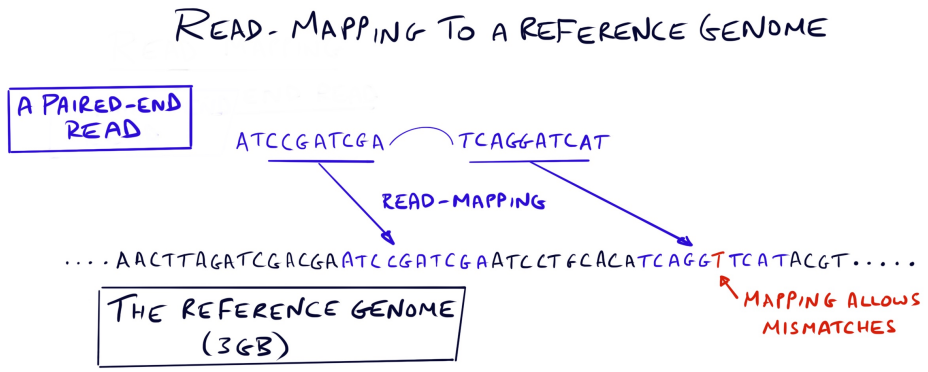


Figure 1.53: **Read mapping.** *Each paired-end read is compared to the reference genome to figure out where in the genome it came from – much like fitting a piece into a jigsaw puzzle. For a paired-end read we do not know the sequence in the internal part of the DNA fragment, but we do know the approximate size: for a correct match both sequence ends should fit into the reference genome with a gap of a few hundred base pairs between them. Low levels of mismatches (in red) are allowed as these may reflect SNPs or other types of variation.*

As described above, read mapping assumes that a read only matches a single location in the genome. But many sequences in the genome are repeated two or more times, so that it is ambiguous where a read comes from. This is especially problematic in the most complex regions of the genome, including centromeres, subtelomeric regions, ribosomal DNA clusters, and other locations where large blocks of DNA are repeated many times. Reads from transposable elements can also be difficult to map (although this depends on size – smaller elements such as the 300 bp Alu repeat, are generally mappable as long as any part of a paired-end read hangs off into unique sequence outside the element) [87]. Collectively, the most ambiguous regions are referred to as **unmappable**, and cover about 10% of the genome [88]. Third generation long-read sequencing is starting to resolve these regions that are inaccessible to short reads.

Genome coverage. A key experimental parameter for genome sequencing is referred to as **read depth** or **genome coverage**. These terms refer to the average sequencing depth in mappable regions of the genome.

For example, if we sequence a DNA library to a depth of "30X coverage", this means that an average (mappable) position in the genome is covered by 30 reads. 30X coverage is a commonly-used standard for high quality genomes: this relatively high read depth ensures that with high probability we have good coverage of both chromosomes at every position in the genome – as discussed next, this allows high-quality genotype calls

Figure 1.54: **Reads tiled across a genome.** *The number of reads (solid lines) that span any given position is the coverage. In practice, the reads are much longer relative to the reference genome than shown here.*

throughout the mappable genome. Since the human genome is about 3.1 GB this implies that we need ~100 GB of sequence data per genome.

SNP calling. Our next goal is to identify SNPs and other types of variation. When we see a mismatch between the sequence read and the reference genome this might indicate one of several possibilities: a homozygous difference from the reference; a heterozygous difference; a sequencing error [e]. By getting deep sequence coverage of the genome we can distinguish these three scenarios:

[e] *Recall that an allele that matches the reference genome is known as the **reference allele**. The allele that differs from the reference is the **alternate allele**.*

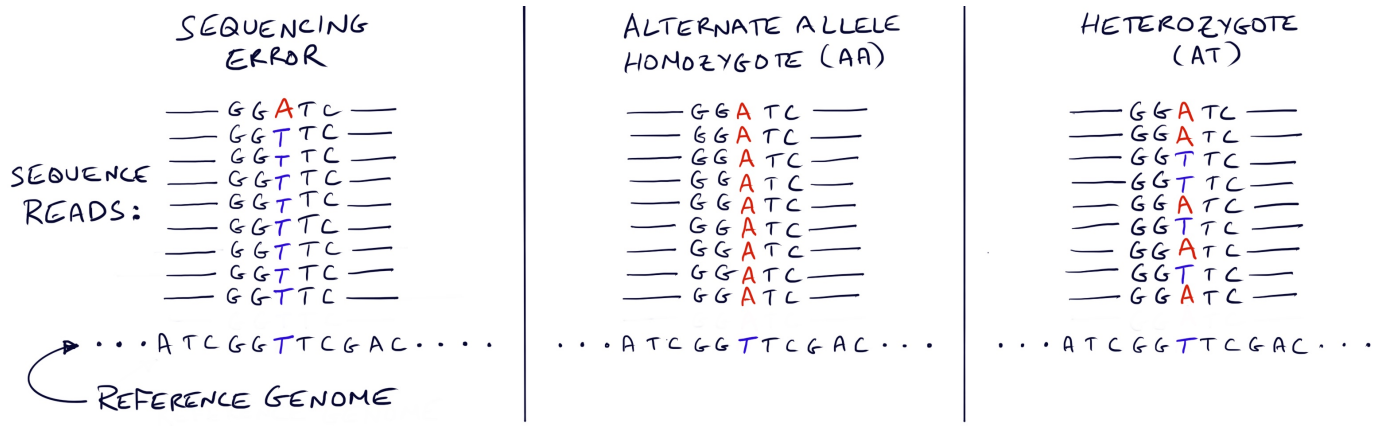


Figure 1.55: SNP calling from sequence data. *Sequencing errors occur at a rate of about 0.1% of nucleotides in current short-read data, but most of these mismatch the reference on only a single read. In contrast, homozyotes for the alternate allele differ from the reference on all reads (or nearly all reads, again remembering that there is a low error rate). Lastly, heterozygotes match the reference on about 50% of reads.*

An error rate of one per thousand nucleotides might sound pretty good, but it actually means that we get a lot more sequencing errors than actual SNPs. So for example with 30X coverage, we'll get a sequencing error roughly once every 30 bp, while true differences from the reference occur less frequently – around once per 500-1000 bp. This issue is particularly acute when we want to detect new mutations – as we'll see in Chapter 1.5, these are extremely rare in the genome, occurring about once per 100 million base pairs. This means that we need multiple supporting reads to confidently detect novel variants.

Lastly, one important point here is that while the read mapping tells us

where each read comes from in the genome, it cannot distinguish between the two homologous copies you have of each chromosome (the one from mum and the one from dad). This means that at a heterozygous site, we don't know which allele comes from which chromosome. This is another situation where long-read technologies can help out, by linking together heterozygous alleles that lie on the same chromosome.

Larger structural variants. So that gives you a sense of how SNP detection works. How can we detect larger structural variants like large deletions, using short reads?

Simplifying somewhat, we can think of two different kinds of information to detect events that are much larger than the scale of a single read-pair. First, most structural variants change the average depth of sequence reads: for example a heterozygous deletion cuts the average read depth to 50% of the genome average.

Second, structural variants may be recognized by the presence of read-pairs that span unexpectedly large distances along a chromosome, inconsistent with the DNA fragment sizes.

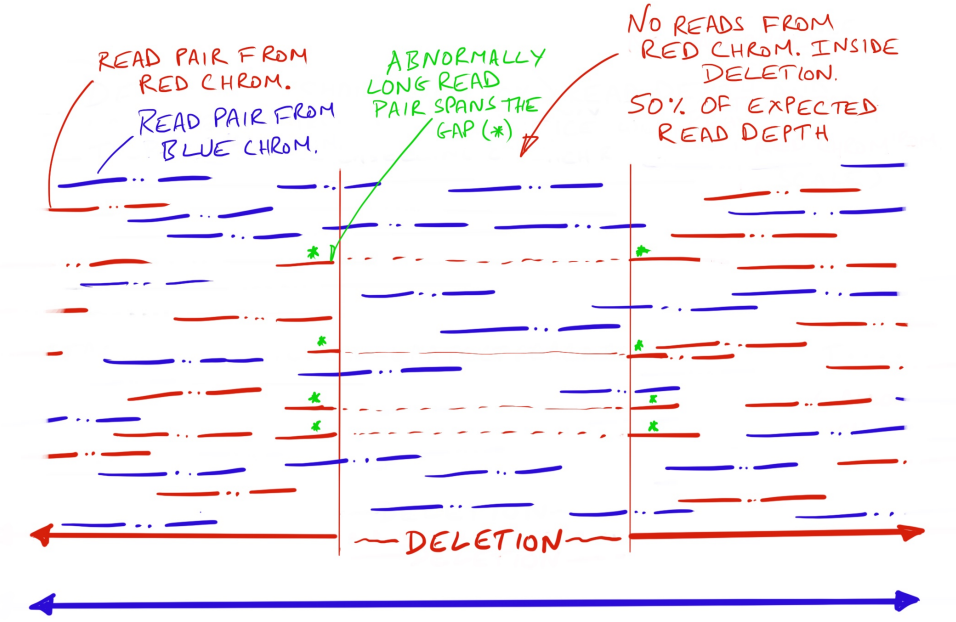


Figure 1.56: **Detection of a deletion from short-read data.** *Two homologous chromosomes (maternal and paternal copies) are shown at the bottom of the figure. The sequence read-pairs that come from each are indicated by red and blue lines above. Within the deletion the average coverage is about 50% of the average, because there are no red reads; we also see some read pairs that span across the deletion. When we map those back to the reference genome they seem extraordinarily long.* ***Note: In real-life analyses we usually do not know which homolog (the red and blue colors) each read comes from.***

However, in practice, many structural variants can be difficult, if not impossible to detect using short-read data. One important reason is that they are often associated with repetitive regions of the genome where read mapping is extremely difficult; we'll cover this further in Chapter 1.5. This is one area where the extremely long reads from 3rd-Gen sequencing are a real game changer compared to short reads. Long reads can span right across high complexity regions and make it far easier to detect the number and orientation of repeated elements and structural variants.

Low-budget approaches to studying genome variation. So far we have focused on **whole genome sequencing (WGS)** applications. But recall that only a small fraction of the genome is involved in coding for genes or controlling gene regulation. Given this, one might greatly reduce sequencing costs by sequencing only the functional regions. One approach to doing this is called **exome sequencing**. Exome (from the words *exons* + *genome*) sequencing uses a lab technique to preselect all the DNA fragments that span gene exons. Since exons only span about 1% of the genome, this greatly reduces the necessary amount of sequencing.

The single biggest advantage of exome sequencing is reduced cost compared to whole genome sequencing. The disadvantage is that obviously it misses all the functional variation outside exons. Later in the book we'll discuss how severe disease mutations tend to be concentrated in exons, but lots of other important types of variation are outside exons, and would be missed by exome sequencing. We may expect exome sequencing to become less relevant as sequencing costs continue to drop.

Genotyping. Last, I'll mention a completely different approach to measuring a limited subset of the genome. The term **genotyping** refers to a variety of different experimental methods that can **determine a person's genotype at a specific set of pre-selected SNP positions** (and nowhere else in the genome).

Current commercial genotyping platforms measure between 500,000 and 2 million SNPs. Genotyping provides less information than a full genome sequence–for example it cannot tell you if carry a rare mutation in a disease gene, as that mutation is unlikely to be included on the genotyping array.

While exome sequencing and genotyping are both used to get a cheaper look at a fraction of the genome, they have very different pros and cons. Exome sequencing provides complete DNA sequence for arguably the most important 1% of the genome. It is useful for identifying rare protein-coding mutations. In contrast, genotyping gives a truly genome-wide look at genetic variation but does not detect rare mutations.

However, genotyping is widely used because it can be applied to very large numbers of samples, and is accurate and relatively cheap (less than $100 per sample). If you have sent a DNA sample to a personal genetics company such as Ancestry or 23andMe, they probably did not sequence your genome, but instead used genotyping. Genotyping is also used in many large-scale research studies. At the time of writing (2022), tens of millions of people have been genotyped, either commercially or for academic research–far more than have been genome sequenced.

In summary, DNA sequencing technology has improved by more than 1 million-fold in the last 30 years. This has enabled cheap resequencing of human genomes for research and clinical applications; genome sequencing of thousands of diverse species; and widespread use of sequencing as a molecular counting tool for many applications. This continues to be a fast-moving area of innovation.

Notes and References.

[75]This phrasing is borrowed from Shendure et al (2017); that paper is a great source for history and technology of sequencing:

Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345-53. Another useful review is:

Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;17(6):333-51

[76][Link]

[77]Sanger sequencing is convenient for quick-turnaround applications in lab-work like checking that a plasmid has been constructed correctly, checking genome edits, or confirming that a PCR product contains the expected sequence.

[78]Cost of the Human Genome Project: [Link]

[79]One potential competitor is Beijing's BGI Genomics which has acquired and refined a technology called nanoball sequencing, originally from Complete Genomics.

[80]Background on Illumina technology, see eg [Link].

[81]Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. NAR Genomics and Bioinformatics. 2020;2(2):lqaa037. Note that PacBio's HiFi approach reads the same molecule multiple times, thereby lowering error rates to be competitive with Illumina.

[82]A 2022 paper considered the application of ultra-rapid genome sequencing in critical settings. They showed that it's possible to obtain extremely rapid (same-day) clinical-grade genome sequences on the Nanopore platform at a cost of about $5000 per sample.

Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, et al. Ultrarapid nanopore genome sequencing in a critical care setting. New England Journal of Medicine. 2022;386(7):700-2

[83]Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53

[84]Illumina has achieved near-monopoly status in the US in genome sequencing. In general monopolies lead to higher prices and lower rates of innovation in industries dominated by a single player: [Link].

[85]For one ambitious current effort in this direction see [Link].

[86]A 2018 paper estimated Illumina error rates at 0.24% per base pair

Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports. 2018;8(1):1-14

[87]Teissandier A, Servant N, Barillot E, Bourc'his D. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. Mobile DNA. 2019;10(1):1-12.

[88]Lee H, Schatz MC. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics. 2012;28(16):2097-105

1.5 Mutation: The ultimate source of genetic variation.

DNA is an exquisitely robust data storage system: a typical baby is born with just mutation one per ∼100 million base pairs (that's about 70 genome-wide). Nonetheless, mutations are central to our story, as they are the source of all genetic variation, for good and bad, enabling evolution and causing disease.

Figure 1.57: **A mutation** *is a change to the genome sequence: in this example a C→T mutation (also G→A on the other strand). A mutation at a single position, like this, is referred to as a* ***single nucleotide mutation*** *or* ***point mutation****.*

The existential challenge of DNA storage and replication. Each cell in your body carries a single precious copy of your genome. Errors in the genomes of your germline cells (the cells that produce gametes) can cause genetic diseases in your children; errors in somatic cells (cells of the body) can lead to cancer or other diseases of aging. Thus, safeguarding the integrity of the genome is a fundamental requirement for all cells.

And yet, every genome copy suffers a constant barrage of **DNA damage**: i.e., events that create molecular alterations, or **lesions**, in the DNA. But as we shall see, the vast majority of these lesions are repaired by DNA repair pathways. Only a tiny fraction of these result in **mutations** – i.e., events in which DNA repair or proofreading fails, resulting in permanent (uncorrectable) changes to the genome sequence of a cell.

It's estimated that a typical cell suffers 70,000 lesions per day[89]! The metabolic processes playing out continuously within each cell produce a variety of small nasty molecules such as reactive oxygen species that can cause DNA damage [a]. Meanwhile, hydrolysis reactions can cleave chemical bonds in DNA [90]. External mutagens including x-rays and gamma-rays, UV radiation (in exposed skin), and mutagenic chemicals such as nicotine, alcohol, and asbestos cause further damage.

The resulting lesions include many possible nucleotide modifications including addition and removal of methyl groups, deamination and depurination (in which a base is released), chemical modifications such as pyrimidine dimers in which adjacent thymines or cytosines form inappropriate covalent bonds parallel to the DNA helix.

Other damage events can cause breaks in the DNA molecule, including single strand breaks (one strand of the double helix breaks, while the other strand stays intact) or – much worse – double strand breaks, in which the helix breaks apart completely. Double strand breaks must be repaired rapidly to maintain cell viability. It's been estimated that a mammalian cell suffers 55,000 single strand breaks and 25 double strand breaks per day [91].

Moreover, the genome of 6 billion base pairs must be copied at every cell division. **DNA replication** provides further opportunity for errors, either when copying damaged sites that have not yet been corrected, or by errors introduced in the copying process itself. A typical cell in your body is descended through tens to hundreds of cell divisions – each involving

[a] *There is a large literature on mechanisms of DNA damage and repair, but most of this will be outside our scope in this book, except where it intersects with our main themes.*

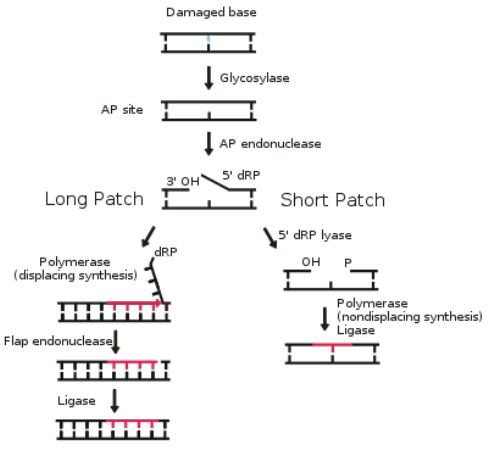


Figure 1.58: **Example of a DNA repair pathway:** *Base Excision Repair. Here, a damaged base (blue) is removed, and patched (red bases), using the other strand as template.* ***Cells suffer thousands of DNA lesions per day, but nearly all are repaired using pathways including this.*** *Credit: Amazinglarry [Link] Public Domain.*

genome copying – since you were a single fertilized egg.

Clearly, protecting the genome against decades of spontaneous chemical damage and mutagens, and accurate DNA replication through trillions of cell divisions during one's lifetime, is an existential challenge. Multicelled organisms couldn't survive if all this DNA damage resulted in permanent changes in genomes.

Consequently, cells have evolved an exquisitely complex molecular machinery of proteins responsible for detection and repair of spontaneous DNA damage, and for highly accurate DNA replication and proofreading. And when DNA damage is so severe that it cannot be repaired – as can happen with double strand breaks – there are alternate pathways for programmed cell death.

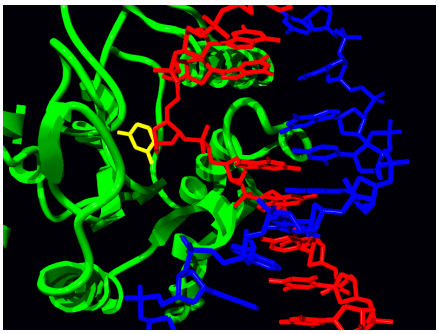


Figure 1.59: **Molecular structure showing repair of a damaged base**. *The DNA strands are red and blue. The repair enzyme uracil glycosylase is in green; it has flipped a nonstandard base (uracil, in yellow) out of the red strand prior to removal and correction by Base Excision Repair.* *Credit: TimVickers [Link] Public Domain.*

As we shall see, the repair and proofreading pathways are absolutely gob-smackingly effective, with germline mutation rates on the order of one per billion base pairs per year.

Mutations and evolution. *That's very impressive...* – you say – *But aren't some mutations good? Don't we also need mutations to enable adaptation?* Yes indeed, this is true. A tiny fraction of mutations are advantageous and, over thousands of years, these are the drivers of evolutionary change. Mutation enables what Darwin called "descent with modification": if there were no mutation, there would be no modification – and no evolution.

This suggests a paradox: On average, mutation is bad for individuals, but in the long-term mutations are necessary for species to adapt and survive. As we shall discuss later, natural selection acts mainly on short-term effects – in this case, the direct fitness cost of mutation – and lacks the foresight to consider possible future benefits to the species (Chapter 2.6). Thus, selection generally favors mutation rates to evolve as low as reasonably possible; fortuitously these rates are still high enough to enable adaptation [92] [93].

In the remainder of the chapter we discuss the rates and mechanisms of mutations.

Germline mutation rates. In animals, there is a strict separation between cells of the **germline** (which produce gametes–eggs or sperm), and the **soma** or **somatic cells** (which produce the body of the organism).

Mutations arise in both types of tissues, but they have very different implications: germline mutations can be passed on to future generations and, as such, they touch nearly every topic in this book; somatic mutations are not passed on, but can lead to cancer and potentially other diseases of aging.

Detecting germline mutations. We can detect ***de novo*** (new) germline mutations by sequencing genomes of families. The example below shows the sequencing of a **family trio**: both parents and a child. If the child has

an allele that is not present in either parent, this must have arisen by mutation, most likely in the germline of a parent [94]:

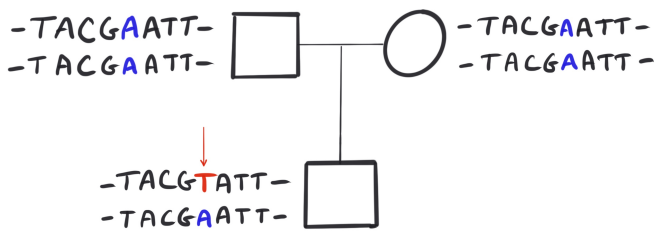


Figure 1.60: **Single nucleotide mutation in a child.** *De novo mutations can be detected by genome sequencing of family trios: here, both parents are homozygous for A, while the child is a T/A heterozygote.*

Starting around 2010, with access to 2nd-generation sequencing, there has been a series of studies characterizing mutation rates in a variety of populations [95]. Several of the most extensive studies have come from the genetics company DeCODE, based in Iceland [96]. The plot below, from DeCODE, shows a histogram of the number of single nucleotide mutations per child, across a large sample of families:

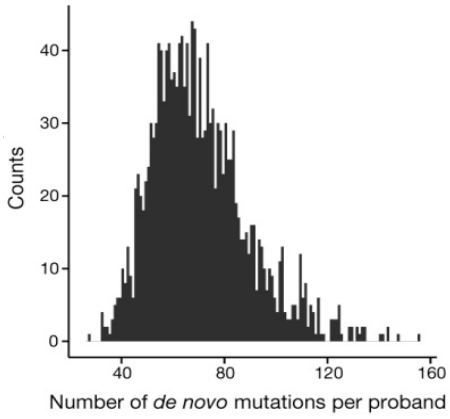


Figure 1.61: **Number of new mutations per child.** *The plot shows the distribution across children in many families.*

Credit: Figure 1d of Hákon Jónsson et al (2017). [Link] Used with permission.

As you can see, a typical child inherits about 70 single nucleotide mutations. Hmmm... does that seem like an awful lot of mutations to you? Well, bear in mind that only about 1% of the genome is protein coding, so a typical child will have about 0–1 mutations in protein coding regions, and perhaps a couple more in regulatory regions. Most of these will not have detectable effects. It's been estimated that about 1.5% of children are born with a loss-of-function mutation, such as premature stop, in a highly constrained gene. Such mutations are a major cause of childhood developmental disorders [97].

We are now ready to estimate the genome-wide mutation rate. The human genome is about 3.1 GB, but in the study above they could only get high-quality sequence data for about 2.68 GB (i.e., excluding repetitive regions). Remembering that each child gets two genome copies (one from each parent) we can estimate the average mutation rate as the average number of mutations divided by the sequenceable genome size [b]:

$$\frac{70 \text{ muts}}{2 \cdot (2.68 \times 10^9) \text{ bp}} = 1.3 \times 10^{-8} \text{ muts/bp} \qquad (1.1)$$

[b] *The human mutation rate is about* 1.3×10^{-8} *mutations per base pair per generation, or just slightly more than one mutation per 100 Mb. This is a fundamental parameter, and useful to remember.*

We'll see shortly that mutation rate increases linearly with the age of the parents; this estimate is for an average parental age of 30. Equivalently this corresponds to a mutation rate of about 4.0×10^{-10} per base pair per year of the parent's ages.

DNA replication is remarkably accurate. At this point I like to emphasize that DNA storage and replication is just remarkably, astonishingly, accurate. The DNA in your parent's germ cells was stored for 2–4 decades or more, and replicated hundreds of times, with an aggregate error rate of just one point mutation per 100 million base pairs [98]!

To put this in perspective, compare this to the process of copying books. Before the invention of the printing press, medieval scribes used to make hand copies of the Bible and other texts. The Bible contains about 700,000 words, or about 3.5 million letters. So to be as accurate as DNA replication, a scribe would have to copy almost 30 Bibles with just a single letter mistake. (In truth, hand-copying of texts was notoriously error-prone and medieval scholars were known to grumble about the "foolish" mistakes of their scribes [99].)

Figure 1.62: **Medieval book copying (1148).** *A scribe would have needed to copy 30 Bibles with just a single mistake to be as accurate as the transmission of human DNA from one generation to the next.* *Credit: British Library article [Link]; Digitized Worms Bible [Link]; Public Domain.*

More mutations in older parents; more mutations in dads. In 1912, the German doctor Wilhelm Weinberg (of Hardy-Weinberg fame) reported that children with a skeletal defect called achondroplasia had older-than-average fathers. During the following 60 years, similar patterns were seen for several severe dominant diseases: namely, that the risk of disease increases with parental age, and especially with the ages of fathers.

Although it was not yet possible to sequence the mutations directly, these disease cases were interpreted as arising from *de novo* mutations in the parents. Here's an example from a 1987 paper, before the genes for achondroplasia and most other diseases were known:

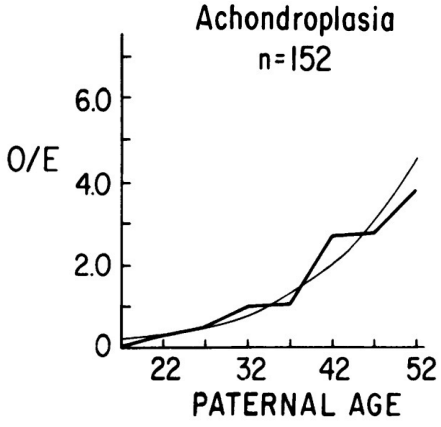


Figure 1.63: **Pre-genome era evidence that mutation rates increase with age (1987).** *The plot shows that rates of achondroplasia increase with paternal age; the y-axis is prevalence in each age-bin, divided by mean prevalence.* *Credit: Modified from Fig. 1 of Neil Risch et al (1987). [Link]*

These, and other, observations were taken as indirect evidence that the *de novo* mutation rate increases with age, and is higher in fathers [100] [101].

A century after Weinberg's work, this hypothesis was confirmed, with sequencing studies showing that

- The number of mutations per child increases roughly linearly with

parental age, with a much higher slope in dads;
• Dads have higher mutation rates at all ages (by a ratio of 3:1 if dad and mum are the same age [102]), as you can see below:

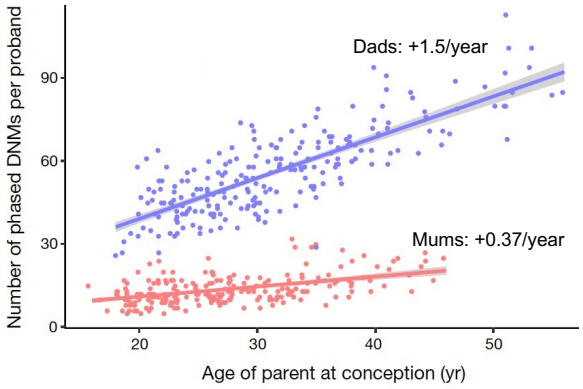


Figure 1.64: **Numbers of new mutations from each parent, as a function of parental age.** *Each data point shows counts for a single child. The increase in mutation counts as a function of parental age is statistically significant in both sexes.* *Credit: Modified Figure 1e of Hákon Jónsson et al (2017). [Link]*

This plot emphasizes that the kids of older parents (especially older dads) inherit a lot more mutations. In fact, if you look back at the histogram of the number of mutations inherited per child – ranging from around 40–120 – most of this variation can be explained by differences in the ages of the parents (plus random sampling variation) [103].

Mutation rates for other types of variation, including STRs and structural variants, are also higher in males than in females [104]. But there's one important exception to this rule: **chromosomal segregation errors** – such as Down Syndrome, caused by transmission of three copies of Chromosome 21 – **are mainly from meiosis errors in mums**. We'll cover this at the end of the chapter.

For decades, the higher mutation rates in males were believed to reflect the fact that there are more cell divisions in the male germline than in the female germline. But as we'll discuss shortly, recent work suggests that most mutations are due to spontaneous damage rather than cell divisions.

Mutation rates in somatic tissues. So far we have been talking about the rates of inherited (i.e., germline) mutations. Mutations also occur within the tissues of our bodies; these are important as drivers of cancer, and also likely contribute to some diseases of aging [105]. How do somatic mutation rates compare to germline mutation rates? How do they vary with age, and across tissues?

It turns out that it is much more difficult to study somatic mutations, than germline mutations. Since mutations occur very rarely in the genome (as low as 1 mutation per 10^8 base pairs), the sequencing error rates have to be extremely low, otherwise errors will outnumber true mutations detected. For studying inherited mutations, we can generally assume that all cells in a tissue sample from a child will carry the same mutations. In contrast, for somatic mutations, mutations that occur early in development may be carried by most or all of the cells in a tissue, but mutations

that occurred recently may be carried by only one or a few cells. Thus, to get accurate somatic mutation rate estimates we need to be able to detect mutations that are present in a single DNA molecule.

Recent techniques based on so-called *duplex sequencing* make this possible [106]. In short, both strands of the same DNA helix are used as independent templates for PCR amplification and sequencing. Variant nucleotides are only confirmed as mutations if they are observed from both strands.

Using these methods, recent work has provided the first direct measurements of somatic mutation rates [107]. As you see below, these tend to accumulate roughly linearly with age, similar to germline mutations. Overall rates are roughly 20–50 times higher than for germline, though still exceedingly low in absolute terms [108].

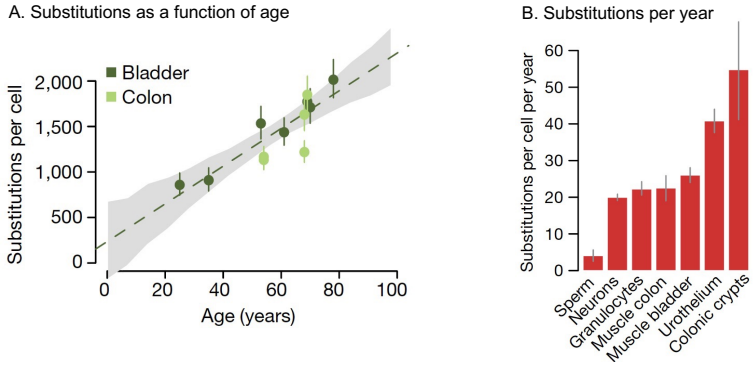


Figure 1.65: **Mutation accumulation in somatic tissues. A.** *Average numbers of mutations per cell in individuals of different ages for bladder and colon.* **B.** *Rates of mutation accumulation per year in different tissue types.*
Credit: From Figure 3 of Federico Abascal et al (2021). [Link] Used with permission.

One other key observation is that the mutation rates in different tissues are not strongly related to rates of cell division, suggesting that a large fraction of mutations arise from spontaneous damage instead of errors in DNA replication. For example, in the plot above, cortical neurons and urothelial cells undergo little or no cell division, but nonetheless have fairly typical mutation rates [109]. If you recall that a typical cell is estimated to suffer $\sim$70,000 genome lesions per day, this implies that only around one lesion per million actually results in mutation.

Mutation rates in cancer. There's one important exception to the rule that human mutation rates are very low: cancer.

Cancer refers to a collection of diseases in which somatic cells start to replicate in uncontrolled manner. In healthy tissues, cell replication and cell death are tightly constrained. As we shall see in Chapter 4.3, cancers are evolving systems that gain the ability to expand without the usual constraints. Typically, the transition into a full-blown cancer state involves multiple mutations across a collection of genes that suppress or enhance cell division. Mutations that enable faster cell division or metastasis are selectively favored within a developing cancer, even though they are severely detrimental to survival of the patient.

Consequently, some cancers arise cells with particularly high rates of exogenous damage; for example, skin cells that are exposed to UV light suf-

fer high rates of DNA damage (this is why you should wear sunscreen!). Secondly, many cancers actually evolve high mutation rates, by gaining mutations in DNA repair or proofreading genes. Cancer cells with higher mutation rates are more likely to accumulate additional key mutations that increase rates of cell division or metastasis. This leads to indirect positive selection on so-called *mutator* genotypes. The plot below shows numbers of mutations (per megabase) across a broad range of cancer types. At the high end, these numbers are 100–1000-fold higher than in healthy somatic tissue.

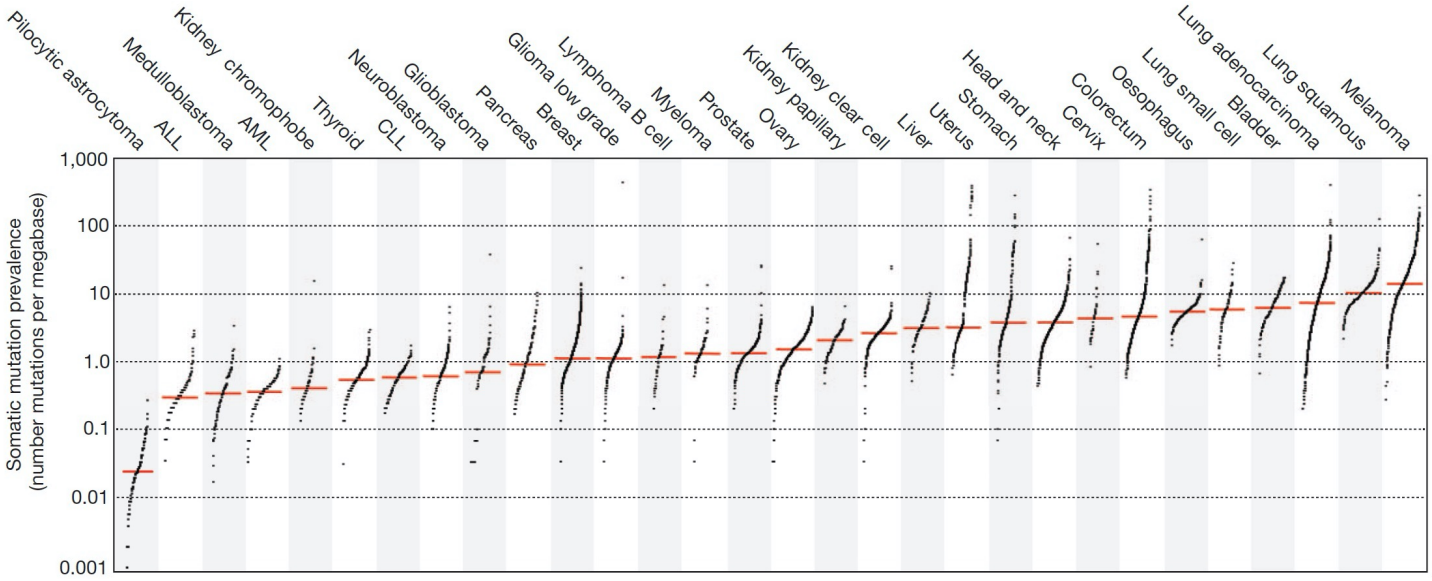


Figure 1.66: High mutation prevalence in cancer. *Mutation rates per megabase for different cancer types. Each data point shows the rate in a different patient; horizontal red lines show the median for each cancer. Notice that for many cancers the numbers are in the range of* 1–100 *mutations per megabase; higher than mutation numbers in healthy somatic tissue (*∼0.1–1 *per Mb).* *Credit: Figure 1 of Ludmil Alexandrov et al (2013). [Link] Used with permission.*

Types and mechanisms of mutations. Up to now, we've been focusing on single nucleotide mutations, and ignoring distinctions between different types of mutations. But you'll remember that the genome contains many different types of variation – including indels, STRs, and structural variants. These different types of mutations occur at widely varying rates, and this fact greatly influences the distribution of genetic variation and disease.

The table below shows estimates of germline rates for important subtypes of single nucleotide mutations, as well as a range of other events [110]. As we will explain shortly, the molecular mechanisms vary widely across different types of mutations, leading to widely varying rates.

As you can see below, single nucleotide mutations make up the majority of all mutations, but some other types of errors – notably STRs – occur at very high rates in particular sequence contexts. Meanwhile, although

structural variants occur at low rates, they are important because they can affect large genomic regions within a single mutational event:

Mutation Type	Number	Rate per site
Single Nucleotide	70	1.2×10^{-8}
Transition (non-CpG)	35	6.2×10^{-9}
Transversion (non-CpG)	21	3.8×10^{-9}
CpG Transition	12	1.1×10^{-7}
CpG Transversion	1	9.6×10^{-9}
Mitochondrial DNA	0.01	6×10^{-7}
Small Indel (70% deletion; 30% insertion)	~5	8.2×10^{-10}
Short Tandem Repeat* (30% contraction; 70% expansion)	>85	5×10^{-5} (average)
Structural Changes* (60% deletion; 30% duplication†)	>0.16	
Aneuploidy (live birth#)	~1/160	

Table 1.5: Genome-wide mutation mutation counts and rates. *Estimated average numbers of mutations per child, genomewide; all numbers are approximate and assume an average parental age of 30. Note that although structural mutations are relatively rare, they often affect tens of kilobases or more of DNA sequence.* *Estimates are from short-read data and detect a restricted subset of mutations, especially for structural variants. STR rates vary widely across motif lengths and types. †Other structural variants not listed include TE insertions and more-complex events. #Aneuploidy rates at fertilization are much higher. Credit: modified from an unpublished table by Ziyue Gao.

We'll now give a brief overview of types of mutations and mechanisms; this is a large and complex area, so my goal here is to give you an introduction to some of the key points, and not to be comprehensive.

Single nucleotide mutations: Transitions, transversions, and CpGs. Starting with single nucleotide ("point") mutations, the first key classification are transitions and transversions. To understand these, recall the chemical structure of DNA. Each rung in the DNA ladder contains a purine (A or G) paired with a pyrimidine (C or T). Purines have two rings, and pyrimidines have one ring. A **transition** switches one purine for another (on the other strand that's switching between pyrimidines); a **transversion** switches from purine to pyrimidine, or vice versa.

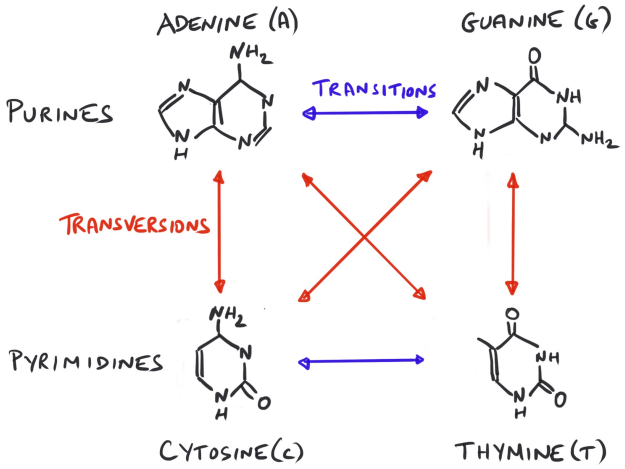


Figure 1.67: **Transitions and Transversions.** *Transition mutations switch between purines (2 rings) or between pyrimidines (1 ring); transversions switch between types. Transitions switch between similar molecules and occur at higher rates than transversions.*

The reason this matters is that because the purines (and similarly the

pyrimidines) resemble each other, the most frequent errors are transitions: i.e., they swap between purines, or between pyrimidines. If all possible point mutations occurred at equal rates, we would expect only 1/3 of mutations to be transitions (count the blue arrows versus red arrows, above). But transitions occur at nearly twice the rate of transversions, so that around 2/3 of point mutations are transitions.

There's one special type of point mutation that is particularly important: **CpG mutation**. In many organisms, including mammals, cytosine can optionally carry a methylation side group. In mammals this occurs almost exclusively when C and G occur at successive nucleotides: i.e., 5'–C–G–3', known as a "CpG". (The 'p' in CpG represents the phosphate that connects successive nucleotides on the same strand, and distinguishes this from the base pairing of G and C on opposite strands.) CpG methylation plays a critical role in preventing undesirable gene expression; consequently most CpGs in the genome are methylated except near the transcription start sites of expressed genes.

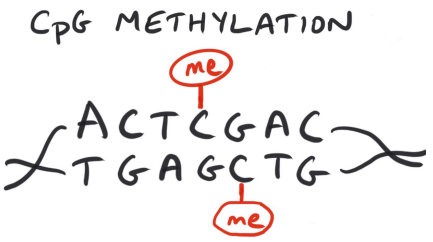


Figure 1.68: **CpG Methylation**. *Most cytosines in a CpG context carry an extra methyl group; cytosine methylation plays an essential role in gene silencing in mammals. Methylated Cs are highly mutable.*

This is relevant here because methylated cytosines can spontaneously convert to thymine. If these are not properly repaired, they cause C→T mutations. These mutations occur at a very high rate, ∼20-fold higher than other transition mutations.

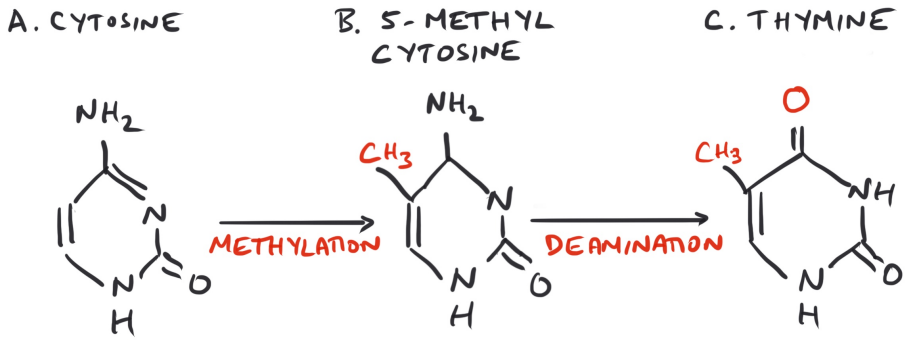


Figure 1.69: **Chemical structure of cytosine methylation and mutation.** *Most cytosines (**A**) become methylated (**B**) when they are in a CpG context. Methylated-C is prone to spontaneous deamination, which results in thymine (**C**). The thymine would be opposite a G on the other strand which tells the cell it must be repaired. Rare failures to repair the T result in C→T mutations.*

Another important special category is for **mitochondrial DNA**, in which mutation rates are even higher than at CpGs. Mitochondria evolved from bacterial symbionts early in the evolution of eukaryotes; they still maintain a small circular genome of 16 kb with 37 genes. DNA repair pathways in mitochondria are more limited than in the nuclear genome, and one important pathway (mismatch repair) may be absent [111]. Consequently the point mutation rate for mitochondrial DNA is about 50-fold higher than in the main genome, at about 6×10^{-7} per base pair per generation [112]. As we shall see later, the high mutation rate of mitochondrial DNA made it an important target of study for early work on human origins, when DNA sequencing was technically challenging and expensive (Chapters 3.2 and 3.3).

Short Tandem Repeats. Some of the highest error rates in the genome occur at short tandem repeats (STRs). Recall that STRs consist of long strings of a repeated motif such as CACACACA.... It turns out that it's difficult for cells to copy these long strings accurately. The main type of error consists of adding or subtracting one repeat. STR mutation rates

have been estimated at a rate of 3×10^{-4} per STR per generation for two-nucleotide repeats, and 1×10^{-3} for four-nucleotide repeats–making these mutation rates as much as a hundred thousand-fold higher than for single nucleotides[113]. Due to their extremely high mutation rate, STRs are highly variable from person-to-person. For this reason STRs are the most commonly-used genetic marker for "DNA fingerprinting" in forensics.

The high rates of STR mutation are due to a process known as **replication slippage**, in which one strand loops out during DNA replication, leaving one or more repeats unpaired:

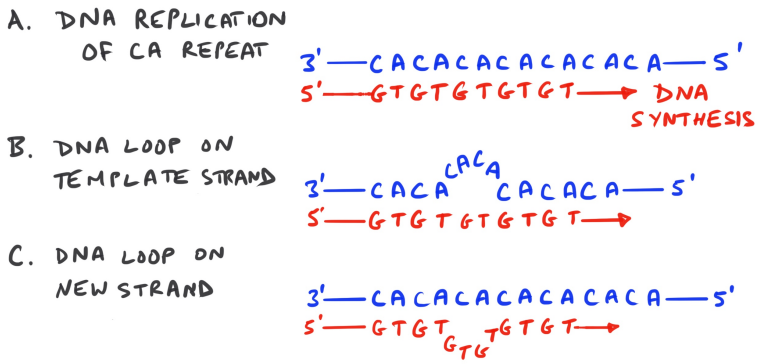


Figure 1.70: **Replication slippage model of STR mutation.** *During DNA replication one strand bubbles out to form a short single-stranded loop, with standard DNA base-pairing on either side of the loop. This causes either loss (B) or gain (C) of repeats, depending on whether the loop is on the template or replicating strand.*

Some STRs play important roles in functional variation and disease: for example in Chapter 1.3 we discussed an STR of CAG repeats within the coding sequence of the Huntingtin gene. The number of repeats is highly mutable; that's ok as long as the number of repeats stays within the normal range – up to 35 in this gene – but longer STRs cause Huntington's disease. Similarly, noncoding STRs sometimes affect gene regulation and contribute to complex traits [114].

Structural variants. Lastly, structural variants–including deletions, duplications, inversions, and more-complex changes in copy number – are another major feature of genetic variation and disease risk, and of genome evolution over longer timescales.

Broadly speaking, most structural mutations are likely due to a few main processes including **recombination errors** [115] and **DNA replication errors** [116]. For both of these processes, repetitive sequences play important roles in confusing the cellular machinery, leading to structural mutations. Alternatively, other events may arise from erroneous **double strand break repair** of damaged DNA, which does not require large-scale sequence homology and hence is not strongly clustered in repetitive regions [117].

The first of these mechanisms, i.e., recombination errors, occurs through a process called **NAHR (non-allelic homologous recombination)**. In NAHR, a DNA sequence that is repeated within a genomic region causes misalignment of homologous sequences during meiosis (i.e., the sequences are *homologous*, meaning that they are (nearly) identical copies of a single original sequence, but *non-allelic*, meaning that they are from distinct chromosomal locations. If a recombination event occurs within the mis-

aligned sequences, this leads to structural changes.

As can be seen here, deletions and duplications can be viewed as alternative products of NAHR, when the repeats are oriented in the same direction. Alternatively, NAHR between inverted repeats leads to inversions.

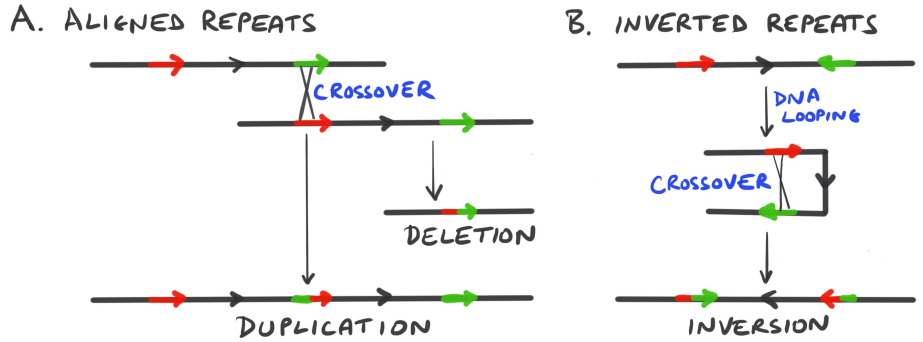


Figure 1.71: **NAHR model.** *A genomic region contains a pair of repeated elements, in green and red. If these misalign during meiosis, cross-over events lead to rearrangements: either deletion/duplication products if the elements are oriented in the same direction, or inversions if they are oriented in reverse orientations. In B, a loop has formed allowing incorrect cross-over between sequences within the same chromosome.* *Redrawn from Figure 1 of Lee and Lupski (2006) [Link]*

As a rule, deletions and duplications are more likely to have functional consequences, because they change **gene dosage** (i.e., the number of copies of genes contained within the region). Since expression of a gene is, usually, roughly proportional to its dosage, this can have functional consequences including possibly genetic diseases, if the affected region contains so-called **dosage-sensitive** genes. In contrast, inversions do not change copy number, and are less likely to cause major effects.

One genomic region that is susceptible to NAHR is at a locus known as 17p11.2 that is responsible for a pair of neurological disorders [c]. The cartoon below shows that the DNA sequence marked in white (CMT1a-REP) appears twice in the region, separated by 1.4 Mb. As discussed above, the two Chromosome 17 homologs can misalign at the repeated region during meiosis; if recombination takes place this, produces both a duplication and a deletion product. This event occurs at a rate[118] of about 10^{-4} to 10^{-5}, which is low in absolute terms, but far more frequent than specific point mutations.

[c] *The 17p11.2 notation uses a classical naming system for chromosome regions that were visible by microscopy prior to the genome sequencing era. This indicates a locus on the p-arm of Chromosome 17 at cytological band 11.2.*

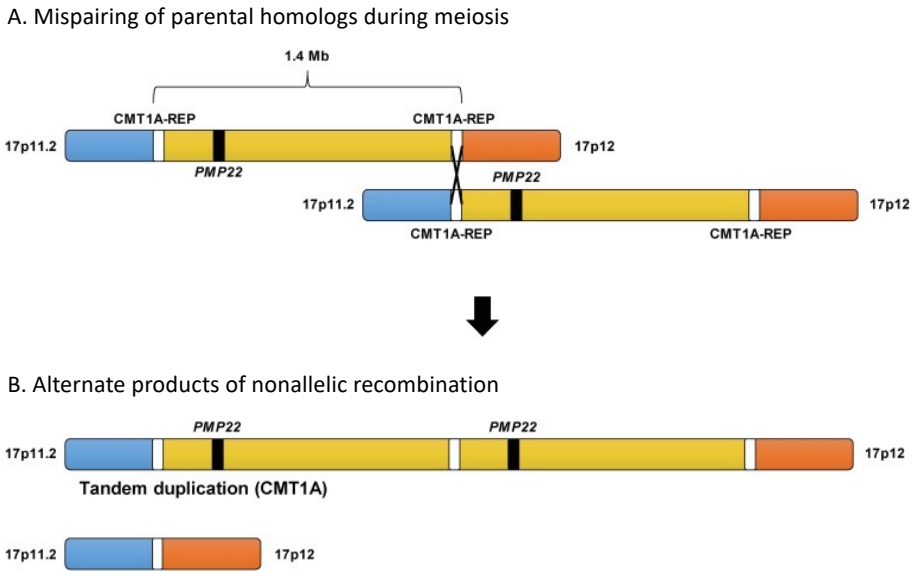


Figure 1.72: **NAHR mechanism at the Charcot-Marie-Tooth (17p11.2) locus.** *The CMT locus is shown as a series of colored blocks for each of the two parental homologs. A duplicated sequence (in white, labeled CMT1a-REP) is present twice, 1.4 Mb apart.* **A.** *During meiosis the duplicated region can mispair, potentially leading to NAHR.* **B.** *NAHR can produce two products: either the entire region is duplicated, or deleted.* *Credit: Modified Figure 1 from Harrison Pantera et al. 2020 [Link] Used with permission.*

The affected region contains a dosage-sensitive gene named PMP22 which

encodes a peripheral nerve myelin protein. Individuals with the duplication (leading to over-expression of PMP22) suffer from a peripheral neuropathy called Charcot-Marie-Tooth, while individuals with the deletion (and under-expression of PMP22) have a different neuropathy with distinct symptoms [119] [120].

The example above illustrates a common mechanism in which large low-copy repeats surround a dosage-sensitive gene, driving recurrent genetic disorders. That is a relatively simple example, but because repeats can drive structural mutations, repeat-dense regions can become crucibles of repeated structural mutations. In some genomic regions, the different haplotypes vary greatly in terms of overall structure, repeat content and orientation [121]. One such example is shown below:

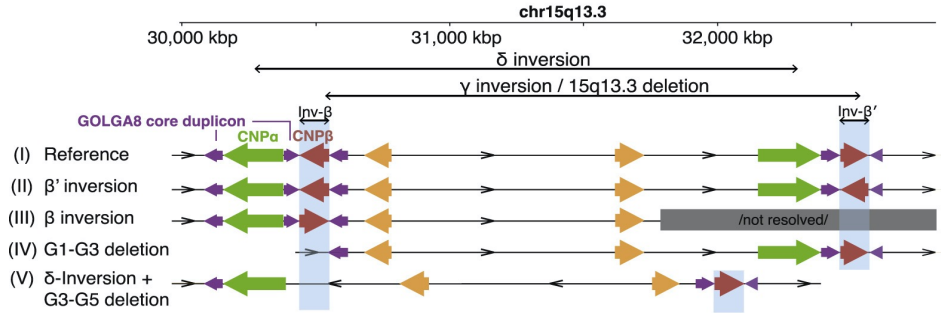


Figure 1.73: **Complex repeat structure at the Prader-Willi/Angelman Syndrome (15q13.3) locus.** *Repeated sequences are shown as directed colored arrows. Based on sequencing of healthy individuals, authors identified five common haplotypes that differ in content or orientation of repeat units.* ***Notice here the diversity of repeat structures and orientations across common haplotypes, typical of repeat-rich regions of the genome.*** *Credit: From Figure 5c of David Porubsky et al. 2022 [Link] CC-BY-NC)*

This region is also noteworthy because it is home to a pair of deletion syndromes called Prader-Willi and Angelman Syndromes [d]. These are caused by deletions that occur between the two maroon arrows, labeled CNPβ. Based on the NAHR model that we discussed above, the authors propose that this deletion may be restricted to haplotypes where the CNPβ arrows point in the same direction, as in Haplotype II.

In the last part of the chapter, we return to some broader topics about the overall patterns and distributions of mutations.

The puzzle of male-driven mutation. Early in this chapter we discussed the point that most mutations (around 75–80% in humans) come from fathers [122]? Why?

For decades there was a standard explanation for this. The key idea was that DNA replication during cell division is the main driver of mutation, and there are many more cell divisions in the male germline than in the female germline, as follows.

Both males and females go through about 30 rounds of cell division early on, as the embryo is developing. If the developing embryo is a girl, they develop into nearly-mature egg cells, and then stop development. Later, when she is an adult, each egg cell completes development right before it is released in ovulation. In contrast, in males, the germ cells that produce sperm stop dividing until shortly before puberty, but after puberty they continue to divide throughout life.

[d] *It's outside our main focus here but the 15q13.3 deletion syndromes have a remarkable inheritance pattern. When the deletion is inherited from the mother, it causes Angelman syndrome (developmental disabilities and motor defects); when it comes from the father, it causes Prader-Willi Syndrome (chronic overeating and related health issues). The difference arises because the region is imprinted: the gene UBE3 is only expressed from the maternal copy and loss of UBE3 causes Angelman; in contrast, SNRPN is only expressed from the paternal copy and its loss causes Prader-Willi [Link].*

This argument helps to explain why older dads transmit more mutations than younger dads. It's also tempting to explain the excess of mutations from dads compared to mums simply as a result of the far greater number of cell divisions in males.

But this calculation also suggests another prediction, namely that the *fraction* of paternal mutations should increase as the parents get older, because the male germline accumulates more and more cell divisions with age, while the female germline does not. Using modern data we can test this. Do we see this pattern in the data? Unfortunately for the number-of-cell-divisions model, we do not:

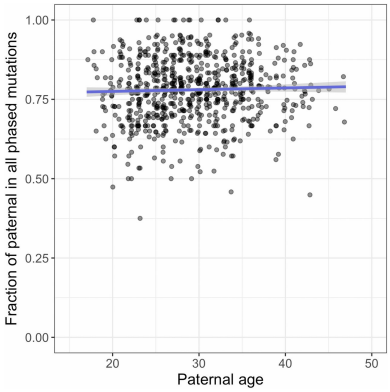


Figure 1.74: **Fraction of paternal mutations as a function of dad's age.** *Each point shows data for one child; the blue line is the regression fit. Under the model where this is controlled by number of cell divisions we would expect a strong positive slope; the fact that the slope is flat argues against this model.* *Note that the parental ages are matched in this analysis. Credit: Figure 1 from Ziyue Gao et al, 2019 [Link].*

As you see above, the fraction of mutations from dads is around 80% at all ages. Moreover, we now have data from many different mammals, with a wide range of generation times. In all these species males have more germline cell divisions than females, but the precise ratio varies widely depending on the specific details of development, age at puberty and reproduction. But, oddly enough, the fraction of paternal mutations is remarkably similar across all these species. These show only a very weak increase with generation time, across species whose generation times range from weeks to decades.

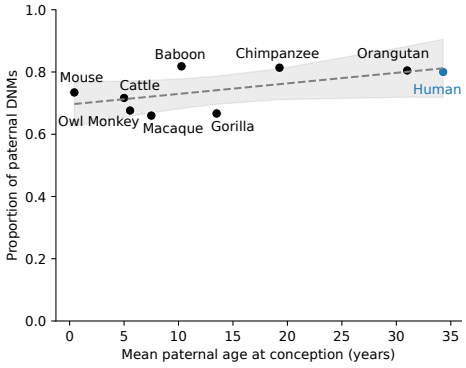


Figure 1.75: **Proportion of paternal mutations in different mammals as a function of generation time.** *This proportion is surprisingly consistent around 75% even though these species vary greatly in terms of generation time, and the ratio of male:female germline cell divisions.* *Credit: Kindly modified by Marc de Manuel Montero and Felix Wu, based on Figure 3B from Felix Wu et al, 2020 [Link] CC BY 4.0*

These results, as well as other analyses [123], suggest that the standard story based on number of cell divisions is not correct. Instead they point to a model where most mutations are not due to DNA replication, but are caused by DNA damage, accumulating steadily with age. (Remember that cells suffer thousands of lesions a day, and if only a tiny fraction of

that is not properly repaired it results in mutations.) We know that there is at least some contribution from non-replicative mutations (i.e., caused by damage), because the mutation rate increases with age even in mothers (albeit slower than in dads), even though mother's germ cells are not dividing.

In summary, the current data suggest that non-replicative mutations are the main driver of germline mutation, but we have to assume that these rates are about 3x higher in testes than in ovaries. It's not known yet why the rate is so much higher in testes, although this does seem to be a broadly conserved feature across at least mammals, birds and reptiles [124].

The puzzle of chromosome segregation errors. There's one huge exception to the rule that genome errors are rare, and male-biased, and that's for **aneuploidy** – i.e., cases where a cell does not carry the correct set of chromosomes: i.e., 23 pairs for a diploid human cell [e] [f].

In sharp contrast to mutations, aneuploidy is inherited mainly from mothers, especially older mothers. For example, around 93% of Down Syndrome cases (3 copies of Chromosome 21) come from chromosomal errors in the egg [125]. Furthermore, the rate of Down Syndrome increases dramatically with the age of the mum: from less than 0.1% in 20-year old mothers to around 1% at age 40 and 3% at age 45.

[e] *Aneuploidy is not a mutational process, but we cover it in this chapter under the broad umbrella of the types of genome alterations that can be transmitted to a zygote.*

[f] *See also Chapter 1.3 for more about aneuploidy.*

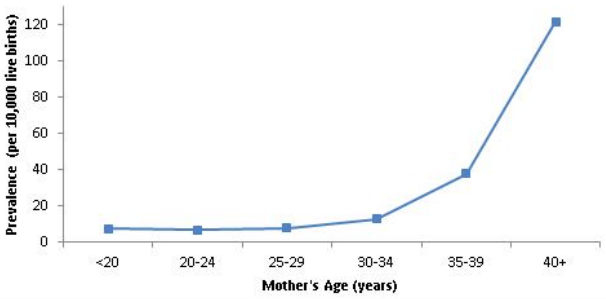


Figure 1.76: **The prevalence of Down Syndrome (Trisomy 21) increases rapidly with mother's age.** *Credit: CDC educational materials. [Link]. Public Domain.*

And Down Syndrome is really just the tip of the iceberg: it's possible for oocytes to carry gains or losses of any of the chromosomes. However, for most other possible aneuploidies, the resulting embryos fail to develop properly, let alone to survive to full term pregnancy.

It turns out that in older women a strikingly large fraction of oocytes carry at least one aneuploidy: by about age 45, more than 50% of oocytes in a typical woman carry chromosomal defects [126] [127]. In contrast, aneuploidy rates in sperm are around 1–4% [128]. As well as causing chromosomal disorders including Down Syndrome, these high rates of aneuploidy are a lead cause of infertility among older women:

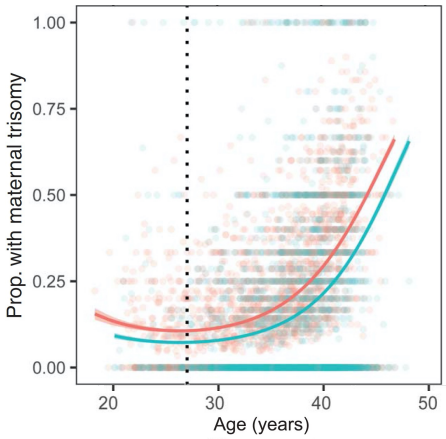


Figure 1.77: **High rates of trisomic oocytes in older women.** *The fitted curves show total trisomy rates across all chromosomes, at two developmental timepoints. Plotted points are for individual patients. As you can see, trisomy increases rapidly after age 35.* *Credit: From Figure 1e of Jennifer Gruhn et al (2019) [Link] Used with permission.*

Earlier in this chapter I emphasized how extraordinarily accurate DNA storage and replication are. Thus, by contrast, female meiosis is remarkably error-prone.

The molecular mechanisms for this are currently an active research area [129], but in broad strokes they are related to a very curious aspect of how egg cells develop in mammals. During fetal development, female germ cells migrate to the ovaries, where they undergo several rounds of mitotic cell division. A subset of the cells then enter meiosis to produce mature oocytes. Recall that meiosis is a process involving two rounds of cell division that produce haploid gametes.

Oddly enough, in normal female development, egg maturation halts in the middle of the first round of cell division, known as Meiosis 1. The oocytes must then wait, *for decades (!)*, until they are re-activated prior to ovulation. At that point, the homologous chromosomes are pulled apart to complete Meiosis 1. Meiosis 2 is completed later, upon fertilization.

While they are waiting to complete Meiosis 1, the homologous chromatids are tethered together by a protein complex called a kinetochore, as well as at crossover sites (which result in recombination). The chromatids sit in this tethered configuration for up to 40+ years until they are pulled apart by the meiotic spindle to complete cell division. It seems that multiple components of the meiotic machinery may deteriorate with age, including the kinetochore, and the assembly of the meiotic spindle [130] [131].

So, from an evolutionary point of view, why are aneuploidy rates in female meiosis so high? Curiously, it does not seem that female meiosis has evolved to minimize the rates of aneuploidy. A first line of evidence comes from analysis of crossover points. Crossovers in females are set up during fetal development and play an essential role in stabilizing the homologous chromatids for the completion of Meiosis 1. In human females (but not in males), about 25% of crossover sites are not fully assembled, and these incomplete cross-overs are a major driver of trisomy 21 [132].

Secondly, the meiotic spindles (which pull the chromatids apart) are actually less stable in human oocytes than in other mammals [133]. Somewhat perplexingly, this is because human oocytes do not express a key spindle-stabilizing protein, KIFC1, used by other mammals (and also used in

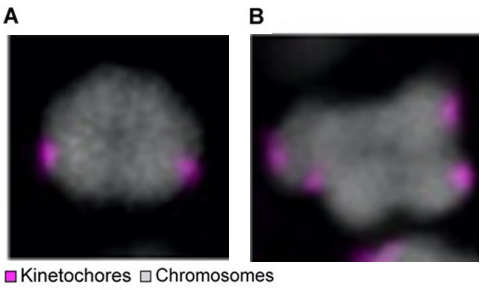


Figure 1.78: **Kinetochores (purple) can drift apart in human oocytes.** *In Meiosis 1 each chromosome pair consists of two pairs of sister chromatids; each of the four chromatids has its own kinetochore (purple).* **(A)** *When pairs of chromatids are tightly bound, there are two purple dots, one for each pair of sister chromatids;* **(B)** *when the sister chromatids drift apart, all four kinetochores can be seen. Separation of kinetochores increases with age and is thought to contribute to aneuploidy.* *Credit: Figure 2 from Agata Zielinska et al, 2015. [Link]*

human mitotic cells). This hints that human spindles have specifically evolved to be unstable.

It's not yet clear why meiosis may have evolved to be more error-prone than strictly necessary. One intriguing type of explanation is that oocytes are known to be susceptible to the evolution of "selfish" centromeres that hijack the process of meiosis to increase their chances of transmission (known as centromeric drive). Error-prone meiosis may evolve as either a consequence of centromeric drive, or as an antidote to it. For more on this, see [134]. A second type of explanation notes that maximizing fertility may not always lead to higher female fitness, especially in humans and other primates, which makes high investments in each offspring. In this hypothesis, since most aneuploidy leads to failure of implantation, aneuploidy serves to lower female fertility in an age-dependent fashion [135].

In summary, the genome is astonishingly well-protected against mutations; most inherited mutations come from fathers, and mutation rates increase with parental age in both sexes. In contrast, most aneuploidy comes from meiosis errors in older mothers, for reasons that are still not entirely clear.

In the next section of the book we will talk about the inheritance of mutations within families and within populations. Some mutations are inherited within populations for thousands of generations, or even eventually spread throughout an entire species.

Notes and References.

[89] Tubbs A, Nussenzweig A. Endogenous DNA damage as a source of genomic instability in cancer. Cell. 2017;168(4):644-56

[90] Gates KS. An overview of chemical processes that damage cellular DNA: spontaneous hydrolysis, alkylation, and reactions with radicals. Chemical Research in Toxicology. 2009;22(11):1747-60

[91] See Tubbs et al (2017), above.

[92] This paragraph touches on several complex topics. In most cases, natural selection pushes mutation rates to be as low as possible; exceptions include so-called 'mutator strains' in bacteria, as well as cancers, which generally evolve high mutation rates. There is presumably some molecular or physiological limit to how low mutation rates can be (it's also been argued that there may be a metabolic cost to having arbitrarily accurate DNA repair). However, Michael Lynch has argued that multi-celled organisms are generally not close to any fundamental limit because natural selection becomes ineffective when the mutation rate is low-enough. For reasons we'll explain in Chapter 2.6, this means that mutation rates are mainly determined through an interaction between selection and effective population size.

Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proceedings of the National Academy of Sciences. 2012;109(45):18488-92

[93] I should also point out that it's an over-simplification to say that evolution does not act on long-term effects. As a thought experiment, imagine a species with a magical repair pathway that lowers the mutation rate to zero. In the short term, this new repair pathway would presumably be favored, as there would be no fitness cost due to mutations. But in the long term, this species could not adapt to changing environments, and would likely eventually go extinct.

[94] In practice, when we do genome sequencing, we're actually sequencing from a somatic tissue (usually blood). So this study-design potentially over-estimates the *de novo* mutation rate by including somatic mutations in the child. We can get a more accurate estimate by sequencing 3-generation pedigrees: we know that 50% of germline mutations should be transmitted to a grandchild in the third generation. It turns out that the 2- and 3-generation estimates are quite similar as few mutations occur early enough in somatic development to appear as heterozygous sites in sequencing of bulk tissue while not contributing to the germline.

[95] Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010;328(5978):636-9;

Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet. 2014;15(1):47-70

[96] Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488(7412):471-5

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017;549(7673):519-22

[97] Agarwal I, Fuller ZL, Myers S, Przeworski M. Relating pathogenic loss-of function mutations in humans to their evolutionary fitness costs. bioRxiv. 2022

[98] Great thread about how amazing DNA replication is: [Link].

[99] E.g., Amos van Baalen writes about medieval copying errors; in one cited example: *In his Latin poem 'On Scribes', the English scholar Alcuin of York (c. 740–804) admonishes scribes to "take care not to insert their silly remarks" and that "their hands not make mistakes through foolishness".* [Link].

[100] Weinberg W. Zur vererbung des zwergwuchses. Arch Rassen-u Gesel Biolog. 1912;9:710-8

Crow JF, Denniston C. Mutation in human populations. Advances in Human Genetics 14. 1985:59-123

Risch N, Reich E, Wishnick M, McCarthy J. Spontaneous mutation and parental age in humans. American Journal of Human Genetics. 1987;41(2):218

[101] It was also inferred from studies of sequence evolution of the X, Y and autosomes, that mutation rates are higher in males; eg

Shimmin LC, Chang BHJ, Li WH. Male-driven evolution of DNA sequences. Nature. 1993;362(6422):745-7

[102] Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. Proceedings of the National Academy of Sciences. 2019;116(19):9491-500

[103] About 70% of the variance in de novo mutation count is explained by parental age

Kaplanis J, Ide B, Sanghvi R, Neville M, Danecek P, Coorens T, et al. Genetic and chemotherapeutic influences on germline hypermutation. Nature. 2022;605(7910):503-8.

[104]Structural variation: Belyeu JR, Brand H, Wang H, Zhao X, Pedersen BS, Feusier J, et al. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. The American Journal of Human Genetics. 2021;108(4):597-607. STRs: Mitra I, Huang B, Mousavi N, Ma N, Lamkin M, Yanicky R, et al. Patterns of de novo tandem repeat mutations and their role in autism. Nature. 2021;589(7841):246-50

[105]One emerging theme in cancer biology is that most aging tissues are susceptible to clonal expansions of specific cell lineages with proliferative advantages. An example where this contributes to aging is through clonal expansions in immune cells and their link to CAD:

Jaiswal S, Natarajan P, Silver AJ, Gibson CJ, Bick AG, Shvartz E, et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. New England Journal of Medicine. 2017;377(2):111-21

[106]Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nature Protocols. 2014;9(11):2586-606

[107]Abascal F, Harvey LM, Mitchell E, Lawson AR, Lensing SV, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021;593(7859):405-10

[108]To put this in context, the highest mutation rate of nearly 60 per year implies around 1 mutation per 100 million base pairs.

[109]See again Abascal et al (2021)

[110]Single nucleotide variation: Kong et al (2012), Jonsson et al (2017); Indels: Jonsson et al (2017); Structural variation: Belyeu et al (2021); STRs: Sun et al (2012), Mitra et al (2021), Steely et al (2021), Kristmundsdottir et al (2023). Mitochondrial DNA: Fu et al (2013)–converted from rate per year assuming a generation time of 30 years. References not given previously:

Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nature Genetics. 2012;44(10):1161-5

Steely CJ, Watkins S, Baird L, Jorde L. The Mutational Dynamics of Short Tandem Repeats in Large, Multigenerational Families. bioRxiv. 2021

Kristmundsdottir S, Jonsson H, Hardarson MT, Palsson G, Beyter D, Eggertsson HP, et al. Sequence variants affecting the genome-wide rate of germline microsatellite mutations. Nature Communications. 2023;14(1):3855

Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Current Biology. 2013;23(7):553-9

[111]Fontana GA, Gahlon HL. Mechanisms of replication and repair in mitochondrial DNA deletion formation. Nucleic Acids Research. 2020;48(20):11244-58

[112]Fu et al (2013), cited above.

[113]Sun et al (2014), cited above

[114]Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nature Genetics. 2016;48(1):22-9

[115]Carvalho C, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics. 2016;17(4):224-38

[116]The second major class of mechanisms is due to errors in DNA replication and repair. These are much more complicated than NAHR, and involve a variety of different pathways. These include mis-templating of repetitive regions during DNA replication, or during repair of replication errors. See Carvalho and Lupski (2016) and see:

Ottaviani D, LeCain M, Sheer D. The role of microhomology in genomic structural variation. Trends in Genetics. 2014;30(3):85-94

[117]These mechanisms involve non-homologous end joining or micro-homology mediated end joining. See:

Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell. 2010;143(5):837-47

[118]I cannot find a rate estimate, but the prevalence of CMT is about 1/2500 births, and the 17p11.2 locus is reported to be responsible for nearly half of cases.

[119]Hereditary Neuropathy with Liability to Pressure Palsies

[120]The Charcot-Marie Tooth locus was the first genetic disorder to be found that is usually due to structural variation, in 1992:

Roa BB, Garcia CA, Pentao L, Killian JM, Trask BJ, Suter U, et al. Evidence for a recessive PMP22 point mutation in Charcot–Marie–Tooth disease type 1A. Nature Genetics. 1993;5(2):189-94

An interesting footnote to the story is that the PMP22 gene was discovered by a team led by James Lupski. Lupski, a pioneer in studies of structural variation, is himself affected by Charcot-Marie Tooth syndrome; however Lupski's genome sequence showed that his own symptoms are due to mutations in a different gene: described here: [Link], and here:

Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. New England Journal of Medicine. 2010;362(13):1181-91

[121]Porubsky D, Höps W, Ashraf H, Hsieh P, Rodriguez-Martin B, Yilmaz F, et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. Cell. 2022;185(11):1986-2005

[122]Key recent work on this problem comes from Molly Przeworski's lab: Gao et al (2019), cited above, and:

Gao Z, Wyman MJ, Sella G, Przeworski M. Interpreting the dependence of mutation rates on age and time. PLoS biology. 2016;14(1):e1002355

Wu FL, Strand AI, Cox LA, Ober C, Wall JD, Moorjani P, et al. A comparison of humans and baboons suggests germline mutation rates do not track cell divisions. PLoS Biology. 2020;18(8):e3000838,

de Manuel M, Wu FL, Przeworski M. A paternal bias in germline mutation is widespread across amniotes and can arise independently of cell divisions. bioRxiv. 2022

[123]Wu et al (2020) and de Manuel et al (2022), cited above.

[124]The ratio is around 3:1 in mammals and 2:1 in birds and reptiles: de Manuel et al (2022) [Link]

[125]Vraneković J, Božović IB, Grubić Z, Wagner J, Pavlinić D, Dahoun S, et al. Down syndrome: parental origin, recombination, and maternal age. Genetic Testing and Molecular Biomarkers. 2012;16(1):70-3

[126]Kuliev A, Zlatopolsky Z, Kirillova I, Spivakova J, Janzen JC. Meiosis errors in over 20,000 oocytes studied in the practice of preimplantation aneuploidy testing. Reproductive biomedicine online. 2011;22(1):2-8

[127]Gruhn et al (2019), from which the figure is taken, proposes that the small uptick at younger ages is a real effect, and is due to a distinct signature of Meiosis 1 errors that declines with age; however this a very weak signal compared to the primary signature of increased aneuploidy at older ages.

Gruhn JR, Zielinska AP, Shukla V, Blanshard R, Capalbo A, Cimadomo D, et al. Chromosome errors in human eggs shape natural fertility over reproductive life span. Science. 2019;365(6460):1466-9

[128]Greaney J, Wei Z, Homer H. Regulation of chromosome segregation in oocytes and the cellular basis for female meiotic errors. Human Reproduction Update. 2018;24(2):135-61

[129]This section greatly simplifies a complex field. For more on this, you can start with: Greaney et al (2018), cited above;

Nagaoka SI, Hassold TJ, Hunt PA. Human aneuploidy: mechanisms and new insights into an age-old problem. Nature Reviews Genetics. 2012;13(7):493-504

Webster A, Schuh M. Mechanisms of aneuploidy in human eggs. Trends in cell biology. 2017;27(1):55-68

[130]Zielinska AP, Holubcova Z, Blayney M, Elder K, Schuh M. Sister kinetochore splitting and precocious disintegration of bivalents could explain the maternal age effect. Elife. 2015;4:e11389

Patel J, Tan SL, Hartshorne GM, McAinsh AD. Unique geometry of sister kinetochores in human oocytes during meiosis I may explain maternal age-associated increases in chromosomal abnormalities. Biology Open. 2016;5(2):178-84

[131]One interesting aspect of this is that cross-overs play an important role in tethering the sister chromatids. Even though the crossovers (i.e., recombination events) are set up during fetal development, it turns out that children of older mothers have more maternal crossovers. This suggests that oocytes with more cross-overs are more likely to be non-aneuploid, and thus to produce successful pregnancies.

Wang S, Hassold T, Hunt P, White MA, Zickler D, Kleckner N, et al. Inefficient crossover maturation underlies elevated aneuploidy in human female meiosis. Cell. 2017;168(6):977-89

[132]Wang et al (2017), cited above.

[133]So C, Menelaou K, Uraji J, Harasimov K, Steyer AM, Seres KB, et al. Mechanism of spindle pole organization and instability in human oocytes. Science. 2022;375(6581):eabj3944

Bennabi I, Terret ME, Verlhac MH. Meiotic spindle assembly and chromosome segregation in oocytes. Journal of Cell Biology. 2016;215(5):611-9

[134]Centromeric drive:

Zwick ME, Salstrom JL, Langley CH. Genetic variation in rates of nondisjunction: association of two naturally occurring polymorphisms in the chromokinesin nod with increased rates of nondisjunction in Drosophila melanogaster. Genetics. 1999;152(4):1605-14

Malik HS. The centromere-drive hypothesis: a simple basis for centromere complexity. Centromere. 2009:33-52

Kursel LE, Malik HS. The cellular mechanisms and consequences of centromere drive. Current opinion in cell biology. 2018;52:58-65

Lampson MA, Black BE. Cellular and molecular mechanisms of centromere drive. In: Cold Spring Harbor symposia on quantitative biology. vol. 82. Cold Spring Harbor Laboratory Press; 2017. p. 249-57

Hurst LD. Selfish centromeres and the wastefulness of human reproduction. PLoS Biology. 2022;20(7):e3001671

[135]This model notes that aneuploidy can increase the gap between successive children to allow greater maternal care for each child, and to reduce fertility in older women who might otherwise care for their existing children or grandchildren. In this view, incomplete crossovers are a feature, not a bug of the system. It's hard to rule out this type of explanation, but it strikes me as a rather clumsy physiological mechanism to regulate fertility. Wang et al (2017), cited above.

Part 2.

Population genetics: the forces that shape genetic variation

In Part 2 we turn our attention to ***Population Genetics****: the study of the processes that shape genetic variation. The key forces are mutation (Chapter 1.5), drift, recombination, and natural selection.*

Population genetics is unusual in biology in having a powerful theoretical framework that allows us to understand many different phenomena in terms of specific, formal models. Here we go through the key models, and show how they shed light on aspects of human variation. We start with an overview of neutral models, and then show how these are extended to study natural selection.

Specifically, we will cover the following:

Chapter 2.1: ***Genetic drift*** *and the most fundamental model in population genetics:* ***the Wright-Fisher model****.*

Chapter 2.2: ***The Coalescent****, which models the inheritance of genetic material backwards in time and is a hugely powerful tool for many problems.*

Chapter 2.3: Shared inheritance at linked sites in the genome causes genotype correlations called ***linkage disequilibrium****; linkage is broken down by* ***recombination****.*

Chapter 2.4: Population structure, and models to understand ***why allele frequencies differ across populations****.*

Chapter 2.5–2.7: ***Natural selection****, including* ***models and data for diverse forms of selection****.*

2.1 Genetic Drift: What happens to alleles over time?

The two copies of your genome (one inherited from your mum and one from your dad) differ at about 3 million SNPs. Each of these arose as a point mutation some time in the past: about 70 are new mutations from your parents, while most of them are inherited from very distant ancestors. In fact, most SNPs that you carry arose as mutations in distant ancestors, hundreds of thousands of years ago, living in sub-Saharan Africa. In this chapter, and the next one, I'll explain why.

Every generation, new mutations are introduced into the population (around 70 per child). You can imagine tracking what happens to these mutations over time. Most mutations are lost from the population within a few generations, but sometimes a mutation can increase in frequency by chance alone. The random changes in allele frequencies over time are known as **genetic drift** [a].

[a] *Until Chapter 2.5 we'll assume that all variation is **neutral**: i.e., that there is no advantage to having one allele or the other. This is a good assumption for the vast majority of point mutations.*

You can think about the spread of a new mutation as being like what would happen if you walked into a casino with a dollar. You decide that you are going to keep playing until you either go bust or you beat the house. Most likely, you go bust pretty quickly, but if you have some early luck, you might be able to build up your cash reserves and play for a while. Very very rarely (theoretically at least) you might be able to play long enough to bankrupt the casino [136].

This is how it is for a new mutation. Most mutations are **lost** from the population within just a few generations (that's like you going bust in the casino). But a tiny fraction spread by chance to be common. And a very few, eventually, spread throughout the entire population, so that the newer allele reaches frequency 1 (that's like you bankrupting the casino). We refer to this situation as **fixation**; or we say that the new variant has **fixed**.

Figure 2.1: **The casino analogy** *is not entirely accurate because real casinos have a built-in advantage which means that you have slightly less than fair odds of winning each bet; that's not to mention the big burly guys who come over when you start to beat the house and discuss loudly how much they enjoy busting kneecaps.*
Credit: Lucy Pritchard

Before we go on, I need to remind you of some jargon: At the position of a mutation, we'll refer to the *original* allele as the **ancestral allele**, and the *new* allele will be the **derived allele**.

A new derived allele (i.e., a new mutation) starts out with 1 copy in the population. If use **N to denote the number of individuals** in the population, then the starting **allele frequency p** of a derived allele is

$$p = \frac{1}{2N}.$$

The factor of 2 in the denominator is to account for the fact that chromosomes come in pairs so everyone has two copies of each locus (for the autosomes [137]).

Over time, due to genetic drift, the allele frequency p either drifts down to zero (the derived allele is lost) or eventually drifts up to 1 (the derived allele is fixed). As I will explain, a new allele is usually lost within a few generations, but fixation takes tens of thousands of generations.

Thought experiment: random changes in allele frequencies over time. The tiny and remote island of Pitcairn is situated in the south Pacific, roughly halfway between New Zealand and Peru. It is currently home to about 50 inhabitants. They are descended from a party of 29 founders who landed there in 1790: nine mutineers from the British ship Bounty, along with 20 Polynesians they had kidnapped. The population has never exceeded 250 inhabitants [138].

Figure 2.2: **Pitcairn Island.** *NOAA, Public Domain [Link]*

Imagine an A/G SNP that was present in the founding population of the island. Suppose that the derived allele G was at a frequency of 55% in the founding group in 1790. Assuming there's no advantage to having either A or G in terms of either survival or reproduction....should we expect that G would stay at a constant 55% frequency over time?

Answer: Probably not. In each generation, the kids get a random sample of the alleles from the previous generation. Since there are so few people in each generation, the number of G alleles will vary by chance from one generation to the next, as illustrated below. This random change is genetic drift.

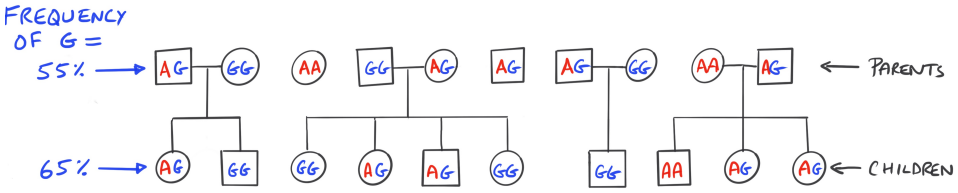


Figure 2.3: **Random sampling of alleles.** *Allele frequencies change from one generation to the next due to random processes: how many children each person has and which alleles they pass on.*

Although most human populations may seem very different from the population of Pitcairn Island, **genetic drift occurs in all populations, though usually much more slowly**.

The Wright-Fisher (WF) model of genetic drift. The Wright Fisher model provides a framework for modeling how allele frequencies change over time [b].

If you wanted to model genetic drift on Pitcairn Island, you might try to get a pedigree for the population over time, and then try to understand how allele frequencies might change through this pedigree. But in practice we don't have pedigrees like this in most populations and, in any event, the complexities of real-world pedigrees tend to obscure the general principles of how frequencies change over time [140].

Instead, the Wright-Fisher model proposes some simplifying assumptions that allow us to understand the fundamentals of population genetics within a very basic model for how allele frequencies change over time. As we will show in the next few chapters, this model is naturally extendable to cover all the other main processes in population genetics, including recombination, natural selection, and population size changes and population structure. While the structure of the model is relatively simple, it provides a powerful framework for understanding genetic variation in real populations.

[b] *The **WF model** [139] is named after two early-20th century founders of population genetics, Sewall Wright and Ronald Fisher.*

We start by assuming a population with N individuals ($2N$ copies of each locus). We assume that there are discrete generations, and that the N individuals mate at random to generate N individuals who form the next generation, ignoring constraints on the sexes of parents [c].

To make the model as simple as possible, you can think of all $2N$ alleles being thrown into a giant bag. Then, we generate the genotypes in the next generation as follows: reach into the bag, draw out an allele at random, write it down, and throw the allele back into the bag. (In a probability class, this process is referred to as *sampling with replacement*.) This is illustrated here:

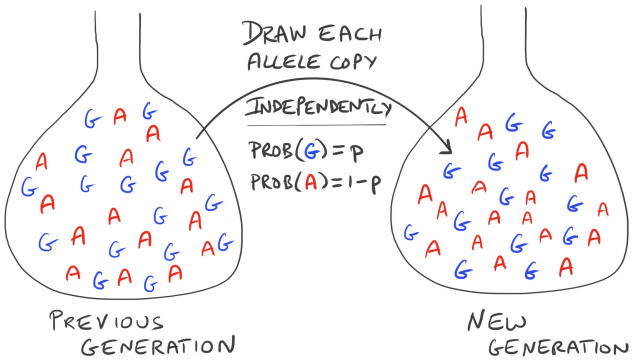


Figure 2.4: **WF sampling.** *Imagine that all* $2N$ *copies of a site are thrown into a big bag (left). We draw alleles out of this bag to make the new generation (right). After we draw out an allele and record it on the right, we drop the original back into the bag on the left. We do this* $2N$ *times to make the new generation.*

This process gives rise to a probability distribution called the **binomial distribution**, which we'll describe shortly.

Before we get to that, here is what this looks in practice. Here I'm assuming a starting allele frequency of $p = 0.55$ as before. Let's suppose that we do a single generation of Wright-Fisher sampling: what is the range of possible outcomes?

The histograms below show the distributions of possible outcomes from repeating this experiment many times in populations of two different sizes.

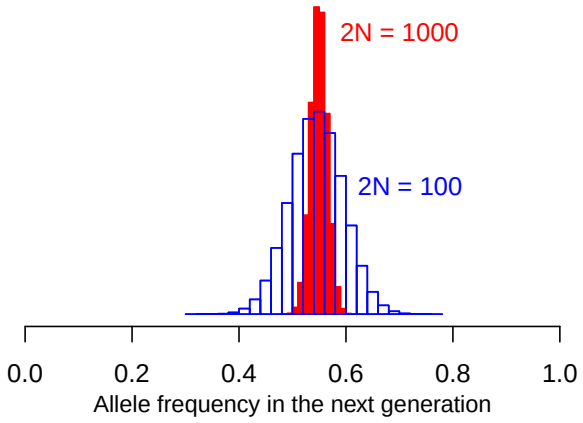


Figure 2.5: **Genetic drift in a single generation.** *Histograms of binomial sampling outcomes for* p_1 *given* $2N$=1000 *(red) vs.* $2N$=100 *(overlaid in blue).*

As you see, both populations are centered on the previous allele frequency (0.55) [d], but the range of outcomes is much wider in the smaller population. This is intuitive, because there is a greater amount of randomness from sampling in the smaller population.

[c] *Here we focus on the numbers of each allele, and ignore the pairing of alleles in diploid genotypes. When we need to think about genotypes in Chapter 2.5, we can predict the proportions using Hardy Weinberg.*

[d] *To be more precise, the* ***expected value*** *of the new distribution equals the frequency in the previous generation. In statistics,* ***an expected value indicates the average (mean) of a distribution****.*

Binomial sampling.

Binomial sampling comes up in many contexts where we make a series of random, independent draws, and each time there is a probability p of one outcome, and $q = 1 - p$ of the other outcome [141].

In WF sampling, we make $2N$ independent draws to create the next generation. Suppose that k is the number of times we draw the derived allele. In this case, the allele frequency after one generation of sampling is $k/2N$, which we will refer to as p_1. The expected value of p_1 (denoted $\mathrm{E}(p_1)$) is simply

$$\mathrm{E}(p_1) = \mathrm{E}(\frac{k}{2N}) = p, \tag{2.2}$$

meaning that *on average* the frequency in the next generation is centered on the current frequency. This doesn't tell the whole story, however, because as shown in the histograms above, the actual, observed p_1 can vary around p. We can measure how much p_1 varies using the **variance** of p_1. By definition, variance measures the average squared difference from the mean:

$$\mathrm{Var}(p_1) = \mathrm{E}[(p_1 - p)^2] \tag{2.3}$$

Using standard properties of the binomial distribution we can show that:

$$\mathrm{Var}(p_1) = \frac{p(1-p)}{2N} \tag{2.4}$$

Notice that the variance in p_1 is inversely proportional to the population size N. This makes sense: a larger population size means that you're getting a bigger sample of the allele frequency from the previous generation. Another important quantity is the **standard deviation** (SD), which is the square root of the variance, namely:

$$\mathrm{SD}(p_1) = \sqrt{\mathrm{Var}[p_1]} \tag{2.5}$$

$$= \sqrt{\frac{p(1-p)}{2N}} \tag{2.6}$$

A very useful rule of thumb is that 95% of the time p_1 will be within two standard deviations of p.

Election polling also follows a binomial distribution. We can get some intuition for binomial sampling by thinking about a completely different context where it comes up: election polling. Suppose that Dumbledore and Voldemort are running against each other for President.

In a particular state, 55% of voters plan to vote for Dumbledore, and 45% for Voldemort. To get a pre-election poll, we phone 100 people, chosen at random, to ask whom they plan to vote for. Assuming that we can get a representative sample of the voting population, the binomial distribution tells us that there is a 95% chance our estimate will be within two standard deviations of the true value: i.e., between 45.1% and 64.9% for Dumbledore [142] [143].

However, suppose that we phone 1,000 people instead of 100. Now we expect a much more accurate estimate: in the range 51.8–58.1%.

These examples illustrate two properties: first, each time we do the survey we get a random estimate centered around the true value. Second, the random error is reduced with a larger sample compared to a smaller sample. Both properties are relevant for allele frequency sampling.

Binomial sampling over successive generations produces genetic drift.

So far, we have talked about genetic drift for a single generation. Now let's think about what happens over the course of many generations.

The crucial thing now is that the result of binomial sampling in one generation gives you the starting point for binomial sampling in the next generation. This will produce a series of allele frequency changes over time called a Markov chain, or more colorfully, a **random walk** [e].

Let's go back to Pitcairn Island. In our hypothetical example, the derived allele started at a frequency p_0=0.55 in the founders (the subscript 0 on p_0 is to indicate that this is generation 0, before any kids have been born on the island). Let's suppose that in the next generation, due to random sampling, the frequency of A goes up to 60% (p_1=0.60). Now, we repeat the random sampling to create generation 2–but this time, the input frequency of A is 0.60, so the expected distribution is centered around 0.60. This process repeats, with the frequency of A drifting up or down by chance depending on the previous frequency. So for example, we might get a sequence of allele frequencies like this:

p_0=.55, p_1=.60, p_2=.57, p_3=.60, p_4=.52, ...

but another SNP with the same starting frequencies might go

p_0=.55, p_1=.45, p_2=.42, p_3=.30, p_4=.35, ...

This is illustrated here:

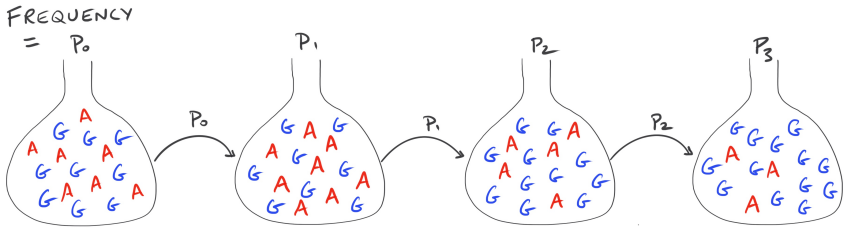


Figure 2.6: **WF sampling over multiple generations.** *The allele frequency in each generation is a binomial sample centered on the allele frequency in the previous generation. Over many generations, this randomness allows the frequencies to drift away from the initial starting point.*

Drift works the same way in larger populations, but the *rate* of drift is slower, simply because the binomial variation is smaller in each generation. The next plots show simulations of this process in populations of different sizes. Each line plots an independent random outcome:

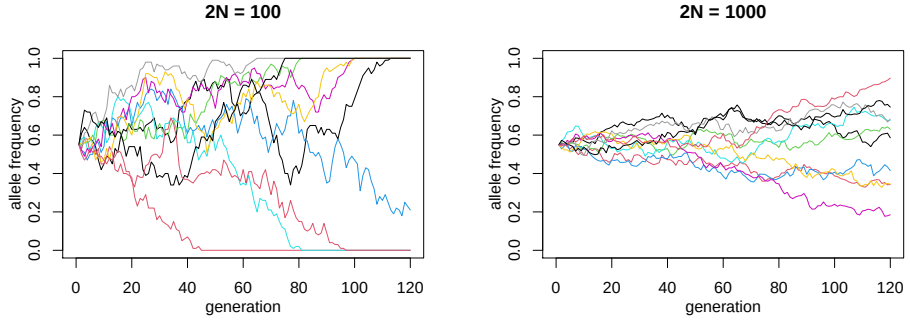


Figure 2.7: **Simulations of genetic drift** *from a starting allele frequency of 0.55. Each plot shows ten independent simulations. Notice that the range of possible outcomes diverges much faster in the smaller population size, with some simulations reaching fixation (frequency=1) or loss (frequency=0) within the timescale of the simulation.*

Eventually, the G allele will either reach 100% frequency, in which case we say that it has *fixed*, or 0% in which case we say it has been *lost*. In random walk theory, 0 and 1 are referred to as *absorbing states*: meaning that the random walk ends if it reaches those values.

[e] *Genetic drift is an example of a mathematical model known as a random walk. You can imagine a drunkard stumbling backwards and forwards along a number line with walls at 0 and 1 until he bumps into either wall and stops.*

It's outside the scope of this book, but infinite random walks in 2 and 3 dimensions have very interesting properties. I have always enjoyed the aphorism from mathematician Shizuo Kakutani that "A drunk man will find his way home, but a drunk bird may get lost forever." [Link].

Mutation and drift. So far we have been talking about drift of alleles that are already common. But in practice, each SNP starts life as a new mutation. Some mutations drift up to become common. How can we model this?

We'll assume that each mutation creates a new allele that did not exist previously in the population. This is known as an **infinite sites** assumption (this simplifies the math and is usually a good approximation [144]). Under this assumption, each new mutation has a starting allele frequency of one copy in the population: i.e., $p_0 = 1/2N$. The new allele now drifts until it either reaches loss or fixation:

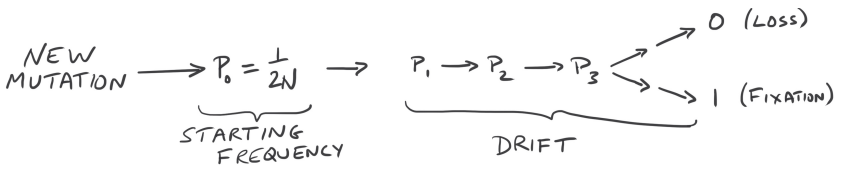


Figure 2.8: **The life-cycle of a SNP.** *A mutation generates a new variant. This is initially at frequency $1/2N$. Its frequency drifts until it eventually reaches loss or fixation.*

The next figure illustrates this process for 200 mutations introduced at different times in a population of 100 individuals. As you can see, most of the mutations stay rare and are quickly lost; however a few drift up to become more common and, in this example, one eventually reaches fixation.

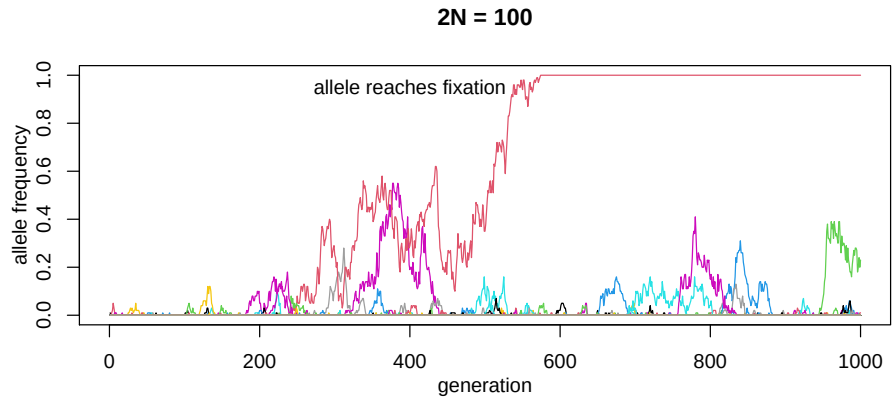


Figure 2.9: **Genetic drift of new mutations.** *Each line shows the simulated trajectory of a different mutation, starting at a random generation number, and drifting independently of the other mutations. This simulation included 200 mutations, most of which stayed rare and are hard to see on this plot.*

Most alleles in a population are very rare. Every new mutation starts out rare in the population (at a frequency of $1/2N$), and most are quickly lost, while a very small fraction drift up to become common. You can see this in the simulation plot above, where only a few of the 200 mutations drifted above 10% frequency [f].

What is the probability that a new mutation reaches fixation?

One very useful fact that we'll derive in the next chapter is that **the probability that a derived allele currently at frequency p will eventually fix is also p. For example, the probability that a new mutation will eventually fix is 1/2N.** Since human populations number in the tens-of-thousands to millions, the fraction of new mutations that eventually fix is very very small.

[f] *You can think about new mutations within our casino analogy. Think about what happens when a crowd of punters all walk into a casino with one dollar each and start gambling – most never get much money and go broke very quickly, but a very few lucky players build up a sizable purse and play for a while before going broke.*

Mutation, drift and the amount of genetic variation. If we put these concepts together, we're now ready to think about genetic variation in populations.

In Chapter 1.3 we discussed how to quantify the *amount* of genetic variation in different populations. One important measure of genetic diversity is **expected heterozygosity**. We define this as follows. Suppose that you sequence a genomic region on one homolog of one random individual, and on one homolog of a different individual. Expected heterozygosity is the average fraction of sites at which these two haploid sequences differ.

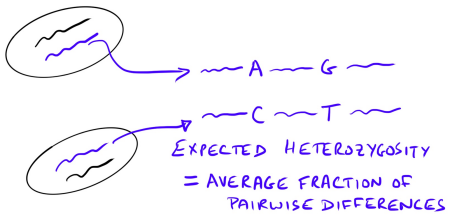


Figure 2.10: **Expected heterozygosity** *is the average fraction of sequence differences between a random pair of allele copies.*

In modern human populations, **expected heterozygosity is ∼0.5–1 heterozygous sites per kilobase**, depending on the population. (See Table 1.2, Chapter 1.3.)

What determines expected heterozygosity? First, mutation plays a critical role of creating new variation in the population. Secondly, the average effect of drift is to remove variation. (Of course, drift sometimes allows rare alleles to become common, but this is always transitory, and in the absence of new mutations, all variants eventually drift to fixation or loss.)

Thus, we can **understand expected heterozygosity as a balance between two forces: mutation, which inputs new variation, and drift, which tends to remove it.** The next box shows a derivation of expected heterozygosity under the WF model. You can skip this if you prefer.

Optional derivation: Computing expected heterozygosity.

Imagine picking two random copies of a locus from the population. We want to write down how the probability that a single nucleotide is heterozygous in the next generation depends on population size and mutation. We set H_1 to be the initial heterozygosity, and H_2 the heterozygosity in the next generation. We also define μ (pronounced mu) as the mutation rate per base pair per generation.

Imagine we pick two random alleles in generation 2. Under our model, there is a probability $1/2N$ that they are descended from the same parent allele in generation 1. This has an effect of reducing heterozygosity by a factor $1/2N$.

On the other hand, alleles that were identical in the parents (with probability $1 - H_1$) might have mutated in either parent: i.e., with probability $\sim 2\mu$. (For simplicity I'm ignoring some extra terms that relate to rare double events, such as two mutations or both mutation AND inbreeding; I'm also making a standard simplifying assumption that mutations always create new alleles.)

We can now write a simple recursion for the expected heterozygosity in generation 2, given what it is in generation 1:

$$H_2 = \underbrace{H_1 \times (1 - \frac{1}{2N})}_{\text{Het goes down by } 1/2N} + \underbrace{(1 - H_1) \times 2\mu}_{\text{Het goes up due to mutation}} . \quad (2.7)$$

This last equation tells us how heterozygosity changes from one generation to the next. Let's suppose we're at a steady state between loss of heterozygosity (from drift) and gain of heterozygosity (from mutation). In that case, we can consider an equilibrium value H that is the same on the left and right hand

sides of the equation, and solve for this:

$$H = H \times (1 - \frac{1}{2N}) \; + \; (1 - H) \times 2\mu \tag{2.8}$$

After some algebraic rearrangement we get

$$H = \frac{4N\mu}{1 + 4N\mu} \tag{2.9}$$

Since $4N\mu$ is usually very small ($\sim$0.1% in humans) it's customary to simplify this last expression to

$$H \; \approx \; 4N\mu \tag{2.10}$$

which matches the value we will derive in the next chapter using a very different technique called the coalescent.

To summarize the math, we just showed that the expected heterozygosity is $4N\mu$. In other words, heterozygosity is proportional to both population size N (because larger population size lowers the rate at which alleles are lost to drift) and mutation rate μ (because higher mutation rate increases the influx of new variation).

Moreover, it turns out that $4N\mu$ is a fundamental parameter in population genetics, that controls the amount of neutral genetic variation. We won't use this notation here, but it's so fundamental that it's sometimes given a special name, θ. We'll come back to interpreting $4N\mu$ on the next page, but we have to introduce *effective population size* first.

Effective population size. It's time for me to confess that the binomial sampling model requires several assumptions that you might think are silly: for example the population size is constant over time; mating occurs completely at random; we don't differentiate between males and females; generations don't overlap; and that there is no inherent tendency for some individuals to have more kids than others (in technical terms, we say that the number of kids is Poisson-distributed).

But it turns out that the simple model actually works well in place of more complicated scenarios, provided that we use a fudge factor called **effective population size** (N_e). Basically the idea is that we will use this number N_e in place of N, where N_e reflects the actual rate of drift under a more realistic scenario.

For example, if a few males have very many offspring, as happens in some species (and in historical examples like Genghis Khan and his sons who fathered huge numbers of offspring across central Asia [145]), this can greatly reduce N_e. Similarly, when the population size fluctuates over time, N_e is strongly shifted toward the times when the population size is smallest, because drift happens much faster in those generations [146] [g].

In future, when we're talking about theoretical models, it's most conve-

Figure 2.11: **Male elephant seals fighting for dominance.** *The winners can control harems of up to 50 females. High variation in reproductive success reduces N_e relative to the actual population size N.* *Credit: Hullwarren CC BY-SA 3.0 [Link]*

[g] *Most of the ways in which real populations differ from the idealized WF model tend to reduce N_e.*

nient to describe the models in terms of N, for idealized populations. But when we talk about data from real populations, we almost always need to think about the data in terms of N_e instead of N. As we'll see below, N_e is usually much smaller than a census estimate of the population size.

Estimating N_e from data. Remember from above that in the idealized model, the expected heterozygosity per site is given by $4N\mu$. When we look at real populations, we replace this with $4N_e\mu$ to reflect that the actual rate of drift is controlled by *effective* population size.

As we noted above, the effective heterozygosity in humans ranges from 5×10^{-4} to 1×10^{-3} per base pair, depending on the population. The mutation rate is about 1.3×10^{-8} per base pair, per generation. If we use N_e in place of N then for the higher end of this range we have

$$4N_e\mu = 1 \times 10^{-3} \tag{2.11}$$

$$N_e \approx 20,000 \tag{2.12}$$

and $N_e \approx 10,000$ for populations at the lower end of the range [147]. In summary, the long-term effective population sizes of humans are around 10,000–20,000 individuals.

These estimates may seem absurdly low, given that the current world population is almost 8 billion. In part N_e is so small because it's a type of average [148] over roughly the last million years and the human population was far smaller for most of that time than it is now; the effective size is also made smaller by the various other ways that real human populations differ from the ideal model. But it's difficult to fully interpret exactly why effective population sizes are what they are [149]. Nonetheless, N_e provides a powerful tool for modeling patterns of genetic variation, especially if we allow it to vary over time–as we will in the next chapter.

The WF model with haplotypes. So far we have been discussing mutation and drift for individual SNPs. But of course, each SNP is contained within a DNA sequence, which may contain multiple variant sites (also known as a *haplotype* [h]). How should we think about mutation and drift in the context of haplotypes?

[h] Genetic variation is inherited within **haplotypes** and we'll talk a lot about these in later chapters.

Here we're going to introduce a basic haplotype model. Importantly, this type of model can be extended in many ways – for example with recombination, selection, or population structure – and we'll use this basic model as a scaffold again in later chapters.

First, let's assume we want to model mutation and drift for a genomic region of L basepairs in length. Now, each generation will comprise two steps: (1) Mutations can arise anywhere in the sequence, at a rate μ per base pair, per haploid sequence. (2) In the sampling process, each haploid sequence in the next generation is drawn at random from the previous generation, sampling with replacement. (Similar to before, this is like putting all $2N$ haplotypes in a bag, and drawing out the next generation one at a time, always writing down the new haplotype, and throwing the old one back in the bag.) This is shown here:

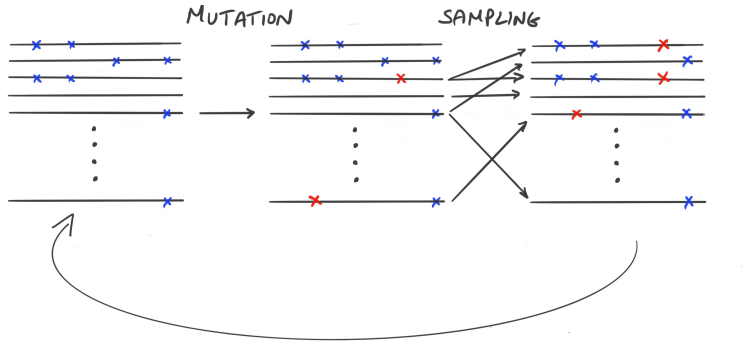


Figure 2.12: **A WF model for haplotypes.** *Each line is a haploid sequence (there are 2N of these to represent the entire population). Blue crosses indicate derived alleles at SNPs. New mutations (red crosses) are placed at random locations within the sequence. WF sampling acts on the haplotypes instead of on alleles.*

WF simulation of haplotype variation. One powerful feature of the WF model is that we can use it to simulate data under a wide variety of evolutionary models [i].

To close the chapter, we'll take the intuition suggested above, and turn this into pseudocode (a kind of recipe) that we could use to simulate haplotypes. If you have experience with programming you could try this out. Even if you don't have experience with programming, think about the steps here, and how they relate to the underlying model.

One key idea here is that we will start with an arbitrary genotype matrix, and then iterate through many generations of mutation and WF sampling until the simulation reaches equilibrium levels of genetic variation (and the starting point doesn't matter any more). For reasons we'll cover in the next chapter this takes at least about $4N$ generations [150]. The genotype matrix at the end of the simulation is a random draw from this mutation-drift equilibrium.

Here's some basic pseudocode for a Wright Fisher model with mutation:

- *Genotype matrix:* Create a genotype matrix G, that contains $2N$ rows (each row is a haplotype) and L columns (each column is a site in the sequence). We'll designate four possible nucleotides using the integers 0, 1, 2, 3 [151].
- *Initialization:* Set every entry in the genotype matrix G to 0.
- *for* `generation` *in* 1 *to* `Max-Generations` *do:*
  {
  - *Mutation:* For each site in each row of G, mutate the existing allele with probability μ.
  - *WF sampling:* Create a new temporary genotype matrix, named G'. For each row of G', pick a random integer u between 1 and $2N$, inclusive. Copy row u from G into G' (this simulates WF sampling with replacement). When all $2N$ entries of G' are filled, copy G' back into G before starting the next generation.

  }

[i] *This approach to simulation is called* ***forward simulation*** *to distinguish it from the backward-in-time approaches we will encounter in the next chapter.*

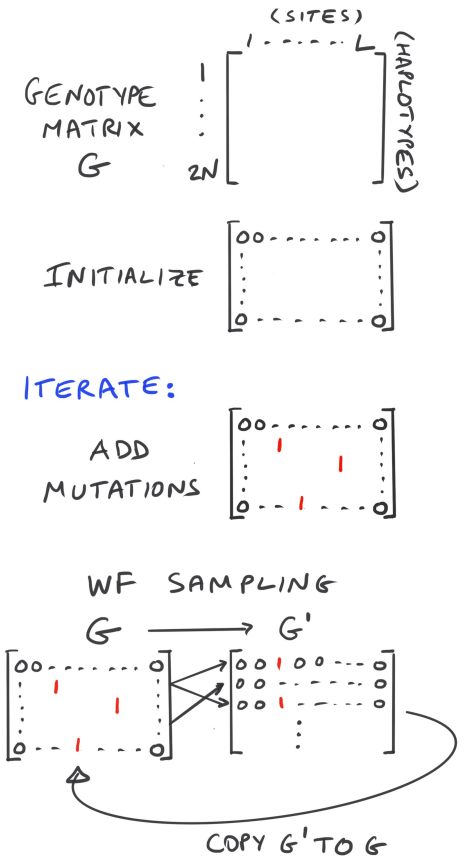


Figure 2.13: **Illustration of WF pseudocode.**

Possible improvements [Optional]. The pseudocode above is ok, but it's slower than necessary and it wastes memory because (i) many haplotypes are identical, and (ii) most sites are not variable. We could make this much more efficient if we just keep track of the distinct haplotypes, and how many there are of each haplotype. We also don't need to store all the nonvariable sites – instead we can just store the positions of derived variants. Lastly, when we add mutations, we can generate the total number of new mutations each generation using a single Poisson random variable with mean $2N\mu L$, and then modify the existing haplotypes at random positions [152].

Simulation software. Forward simulations provide a flexible approach for modeling population genetic data, and can be applied to a wide range of possible models. They are usually computationally slower than backward simulations (next chapter) but aside from very simple models they are easier to implement and far more flexible. A popular software package called **SLiM** provides a powerful toolkit for simulating a wide range of interesting models [Link] [153].

In this chapter we introduced the Wright Fisher model and the fundamental concept of genetic drift (and the interplay of mutation and drift). Nearly everything else in population genetics depends on the basic processes of mutation and drift. In the next chapter, we'll introduce the coalescent, which gives us a very different way to understand drift.

Notes and References.

[136]In practice the size of your cash holdings over time when gambling in a casino is more analogous to the drift of a deleterious variant, since casino betting is set up to favor the house. We'll describe drift of deleterious alleles in Chapter 2.5.

[137]The counts would be different for sex chromosomes: there are $N/2$ Y chromosomes, and $3N/2$ X chromosomes, assuming equal numbers of males and females.

[138]You can read more about Pitcairn Islands here: [Link] and specifically about the mutiny here [Link]. The peak population size was 250 inhabitants in 1936.

Another example of an extremely isolated population is Tristan da Cunha. This is a tiny island in the south Atlantic– at 1700 miles west of Cape Town in South Africa it is the most remote inhabited island in the world. Tristan da Cunha is currently home to about 270 people who descend mainly from 8 men and 7 women from Europe and the US who settled the island in 1816:

Soodyall H, Nebel A, Morar B, Jenkins T. Genealogy and genes: tracing the founding fathers of Tristan da Cunha. European Journal of Human Genetics. 2003;11(9):705-9

[139]Sewall Wright, RA Fisher, and a third scientist JBS Haldane, are often credited as developing many of the key ideas of modern population genetics, mainly in the first half of the 20th Century. This formed a key component of the so-called Modern Synthesis, which united Darwin's theory of evolution with the growing understanding of heredity started by Mendel.

[140]It's outside our scope here, but techniques for studying frequency changes in known pedigrees are referred to as *gene dropping*. For an excellent example see

Chen N, Juric I, Cosgrove EJ, Bowman R, Fitzpatrick JW, Schoech SJ, et al. Allele frequency dynamics in a pedigreed natural population. Proceedings of the National Academy of Sciences. 2019;116(6):2158-64

[141]Binomial sampling. The probability of getting k successes is

$$\frac{n!}{k!(n-k)!}p^k q^{n-k}, \tag{2.13}$$

where the function $n!$ is pronounced "n factorial" and calculated as $n \times (n-1) \times (n-2) \times \cdots 3 \times 2$. For more on the binomial see [Link].

[142]Here we approximate the sampling distribution as binomial, assuming that the size of the poll is much smaller than the number of voters. The standard deviation of the binomial proportion is $\sqrt{p(1-p)/n}$ where p is the true proportion and n is the number of voters that we phoned (instead of $2N$ for number of allele). The true estimate will lie within $+/-$ two standard deviations about 95% of the time.

[143]These example are meant as illustrations, but in practice, the biggest challenge in election polling is not binomial sampling error but getting a representative sample of the voting population. In particular, it may be more difficult to reach some types of likely voters than others. For this reason, analysis of polling data usually involves techniques to reweight the samples to better reflect the expected demographic and political composition of likely voters.

[144]Remember that only about 0.1% of sites are common SNPs so this is a very useful approximation for most applications within species. However the assumption breaks down in analyses of very large sample sizes, especially at hypermutable CpG sites. It also doesn't work well for phylogenetic models of distantly related species as over longer timescales a larger fraction of the sites have accumulated substitutions.

Harpak A, Bhaskar A, Pritchard JK. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genetics. 2016;12(12):e1006489.

[145]About 8% of the men in central Asia carry a single Y chromosome haplotype that is estimated to descend from a common ancestral haplotype 1000 years ago. The age and geographic distribution of the haplotype suggest that it was likely spread by Genghis Khan and his male relatives:

Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, et al. The genetic legacy of the Mongols. The American Journal of Human Genetics. 2003;72(3):717-21

Balaresque P, Poulet N, Cussat-Blanc S, Gerard P, Quintana-Murci L, Heyer E, et al. Y-chromosome descent clusters and male differential reproductive success: young lineage expansions dominate Asian pastoral nomadic populations. European Journal of Human Genetics. 2015;23(10):1413-22

[146]When population size fluctuates rapidly over generations, the effective population size is given by the harmonic mean. Long-term changes in N are less-well modeled by a simple change in N_e.

[147]I'm rounding here since all the other numbers are somewhat rounded (and in any event heterozygosity varies across the genome and across populations). Given these particular numbers, the precise value of N_e would be 19,230.

[148]The harmonic mean.

[149]It's difficult to fully interpret effective population size estimates. Humans have extremely low heterozygosity (and hence N_e) compared to a wide range of other species. Although chimpanzees and gorillas now have very small populations, they actually have higher long-term N_e than humans. Meanwhile, Neanderthals were even less diverse than modern humans, as are a few contemporary species with very small populations, such as lynx and wolverines. Although N_e can be difficult to interpret, it still provides a powerful tool for modeling patterns of genetic variation, especially if we allow N_e to vary over time as is typical in more advanced models.

Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species? PLOS Biology. 2012;10(9):e1001388

[150]We want to run the simulation long enough to ensure that the simulation can reach a stationary distribution with respect to the amount of genetic variation (and so the starting point is no longer relevant). One way to think about this is that the population MRCA in the final generation (see the next chapter) should exist within the simulation. On average, the time to the MRCA is $4N$ generations, so we would want to run this for at least $4N$, and probably more like $10N$ generations to be safe.

[151]The way I'm writing this it's actually finite sites mutation, instead of the infinite sites model alluded to earlier. The finite sites model is a bit more intuitive here.

[152]We can also convert this into an infinite sites model by representing the mutated position using a real number on the interval [0,1]. Derived alleles will be represented by 1.

[153]Messer PW. SLiM: simulating evolution with selection and linkage. Genetics. 2013;194(4):1037-9

Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. Molecular Biology and Evolution. 2019;36(3):632-7

2.2 More on genetic drift: The coalescent.

Here we introduce a different way of understanding the Wright-Fisher model, called the coalescent, but now looking backward in time. The coalescent may seem confusing at first but is incredibly powerful for understanding genetic variation and for data analysis [a].

[a] *The 19th Century Danish philosopher Søren Kierkegaard quipped that "Life can only be understood backwards, but it must be lived forwards." This quote encapsulates the difference between coalescent models (backward-in-time) and the Wright-Fisher model (forward-in-time)* [154].

A short history. In the early 20th Century, when people first started studying population genetics, it was natural to think about evolutionary models forward in time, and these ideas were developed into the Wright-Fisher model during the 1920s. For 50 years forward-in-time models were the main tools for understanding evolutionary processes.

But it turns out that forward models are not easily adapted for use in data analysis. When the first molecular data started to arrive at the end of the 1960s, this drove the development of new questions and models in population genetics [155]. One huge innovation was the coalescent, developed independently by three scientists in 1982 and 1983: John Kingman, Richard Hudson, and Fumio Tajima [156].

Like many breakthroughs in science, the coalescent stands the conventional thinking on its head. Instead of thinking about evolution forward in time to reach the present day, we look backward at the ancestors of modern samples. Many problems in population genetics, especially for neutral models, suddenly become far easier [157].

Inheritance of genetic material from a shared ancestor. The central concept of the coalescent is that the DNA sequences carried by present-day individuals – you or me, for example – are copies of DNA sequences carried by individuals in the distant past. Your genome and my genome are descended from many shared ancestors that lived hundreds of thousands of years ago.

To train your intuition, we start by thinking about inheritance of DNA within families. Imagine comparing your own genome with that of a second cousin (second cousins share great-grandparents). In some parts of the genome you, and that second cousin, inherited the exact same chunk of chromosome from one of your great-grandparents (marked in red, below). On average you share 1/32nd (about 3%) of your genome with that second cousin:

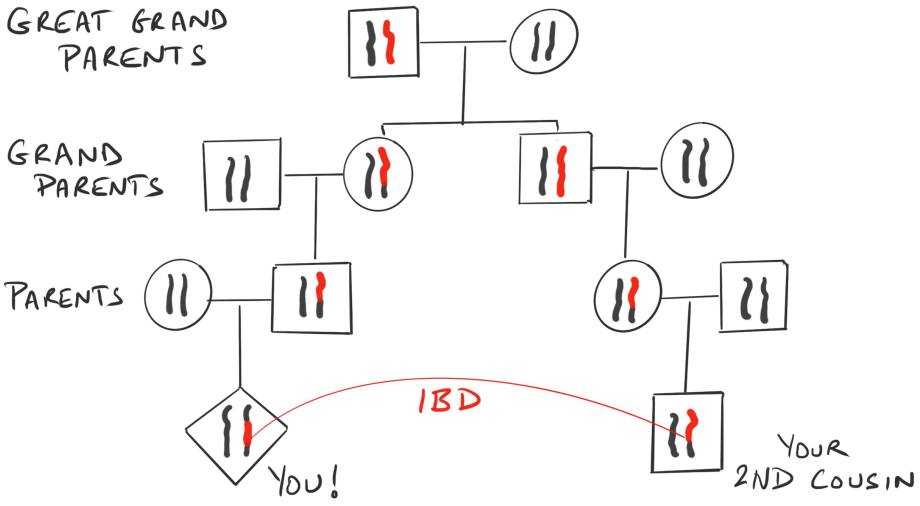


Figure 2.14: **Shared ancestry between second cousins.** *Inheritance of one chromosome (i.e., one homolog) from a great grandparent shown in red. For the two cousins the overlapping segment is said to have* ***coalesced*** *in their great-grandfather. (With such recent ancestry, the overlapping part of the red segments in the two cousins is also said to be* ***identical by descent (IBD)****.)*

We say that this part of your genome **coalesced** with the corresponding part of your cousin's genome 3 generations ago; the great-grandparent is your **common ancestor** at this locus. Coalescence means that this part of both your, and your cousin's genomes, are descended as copies of this ancestral genome [b].

Here we're focusing on coalescence within a family pedigree, between two people who are "related" in the usual sense of the word. But as I shall explain next, in fact *everyone* in a population is related in the same way, although coalescence is usually far more ancient [c].

[b] *With such recent shared ancestry we expect the two copies of this region to be identical, aside from any new mutations. Regions shared within ~10 generations are referred to as* ***identical by descent (IBD)****.*

[c] *There is an important distinction between* ***pedigree ancestors*** *(e.g., you have 8 great-grandparents) and* ***genetic ancestors, which are the focus here****. As in the picture above, you have two copies of any small region of your genome,* ***each of which comes from just a single parent, grandparent, great-grandparent and so on.***

The coalescent refers to ancient shared ancestry within populations.

Let's pick an arbitrary location in the human genome. You have two homologous copies of this locus. Pick one of those two copies at random. You inherited this copy from one of your parents – your mum, say – who got it from one of her parents, and so on backwards in time.

Now do the same thing for one of your friends. Pick one copy of this locus in your friend. Do these two copies have a common ancestor? Perhaps surprisingly, the answer is yes, although that common ancestor probably lived hundreds of thousands of years ago.

To see this, we're going to use the Wright-Fisher model again. Remember that going forward in time, the WF model generates each generation by random sampling with replacement from the generation before. We can think of this in terms of drawing colored balls out of bags. Each time we pull out a ball we write down its color and toss it back into the bag. Here, two red balls in the present generation are both copies of the same "ancestor" red ball two generations ago:

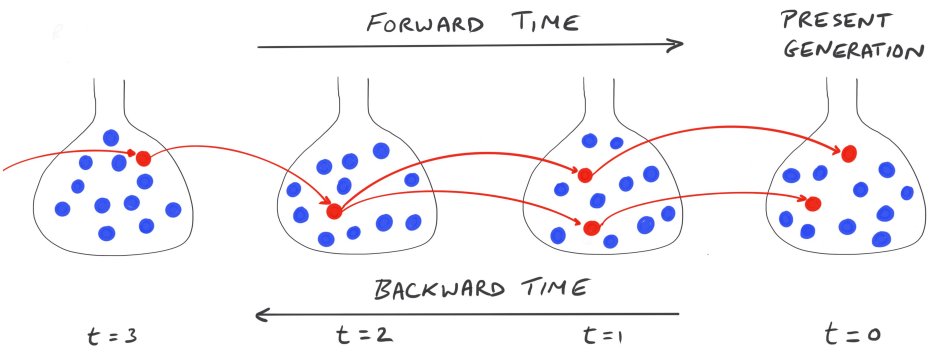


Figure 2.15: **Coalescence in the WF model.** *Two copies of this locus in the present generation are marked by red balls. These descend from a common ancestor (i.e., they coalesce) two generations ago.* ***In coalescent models it is most natural to measure time backward from the present.***

Measuring time forward or backward. The illustration above also shows that we can measure time either forward or backward. For the WF model it's natural to count generations forward from some arbitrary starting point, as we did in the previous chapter. But in the coalescent we will define the present day as $t = 0$ and count generations backward in time.

The genealogy of a sample. So far we have been talking about the ancestry for a pair of copies of this locus. Can we extend this to think about the ancestry of m copies of this locus? For example, we could sequence this locus in $m/2$ diploid individuals – how should we think about the ancestry of these m sequences?

You can think of the ancestry of the samples as coming from a **coalescent genealogy** (or just **genealogy**, or **tree**) that represents the relationships of all m sequences. This genealogy is embedded within the forward WF process:

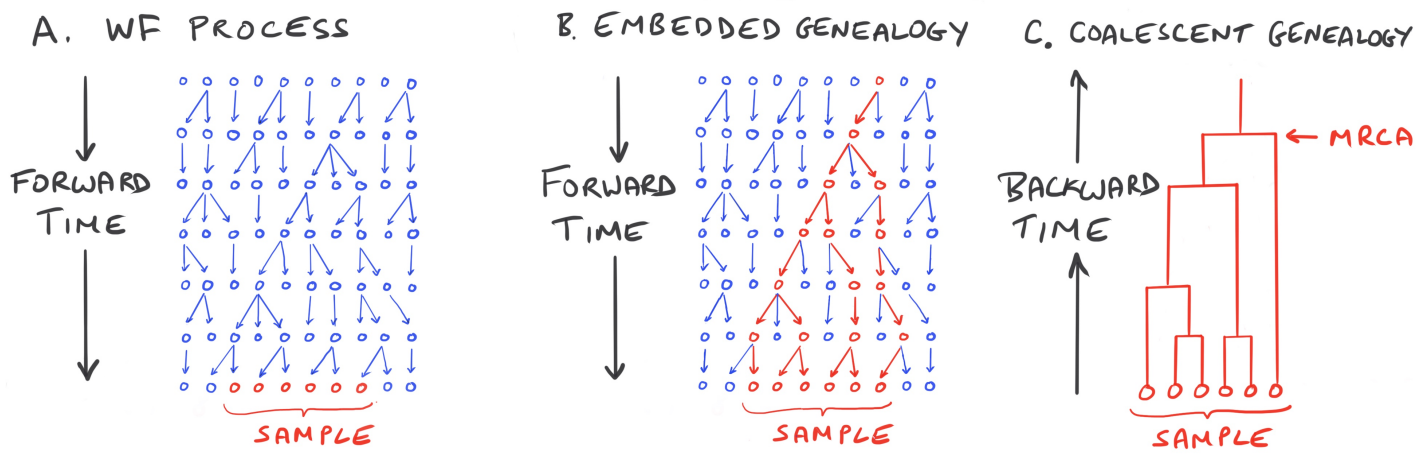


Figure 2.16: The WF history contains an embedded coalescent genealogy. A. *WF genealogy for a small population. This includes six chromosomes sampled at the present day, in red.* **B.** *Red circles and arrows indicate the ancestors of the sampled chromosomes, embedded within the WF process.* **C.** *The coalescent genealogy abstracts away all irrelevant details of the WF process, showing only the ancestral relationships of the 6 samples and the coalescent times.*

Notice that although the WF process runs forward in time, we can only reconstruct the genealogy backward in time, after we are told which six present-day samples are relevant. At that point we can find the genealogy by tracing backward through the ancestors of the sample. Looking

forward in time, there is nothing particularly remarkable about the chromosomes that wind up being ancestors, versus those that do not, and their relevance to the present-day sample only becomes clear in retrospect.

Eventually, we reach a single common ancestor of the entire sample, known as the **most recent common ancestor (MRCA)**, which we will return to shortly.

Time to coalescence. For the sake of simplicity, the pictures above show coalescence within a few generations. But how long would coalescence take in real populations. In fact, how sure can we be that any two copies of this locus ever find a common ancestor – i.e., that they ever coalesce?

Looking backwards in time, each copy of this locus has a random parent from among the $2N$ possible chromosomes in the previous generation. So the probability that they both descend from the *same* parent is $1/2N$.

Conversely, the probability that they do *not* have a common ancestor in the last generation is $1 - 1/2N$. What is the probability that we go back at least t generations without a common ancestor? Assuming this process is independent from one generation to the next, we can multiply the probabilities, giving us

$$(1 - \frac{1}{2N})^t. \tag{2.14}$$

Now the important thing here is that $(1 - 1/2N)$ is < 1, so if we multiply it by itself many times, this number steadily approaches zero. This means that **if we go far enough back in time we can guarantee that any pair of copies of this locus have a common ancestor**.

Ok, so any two copies are guaranteed to eventually coalesce, but how long will this take? To answer this we need to take a short detour:

Understanding waiting-time distributions: the geometric distribution. To understand coalescent models you should know a bit about mathematical models of waiting times. To make this more concrete, suppose that I have a 20-sided die. I keep rolling the die until it lands with the '20' face up (and then stop). How many times do I need to roll the die?

Obviously, the waiting time is random: there is a 1 in 20 chance that the '20' comes up on the first roll – or I might need to roll many times. But we can calculate the average number of rolls, and we can also write down what is called the **probability distribution** which in this case is a general formula for the probability that the '20' first comes up on any specific roll.

First of all, we consider the probability of getting a '20' on any particular roll. We'll call this probability p, and it is simply $1/20$ since we have a 20-side die. Then the probability of NOT getting a 20, is 1-p, or 1-(1/20). One important property of probabilities is that the probability of multiple independent events is the product of the probability of observing each separately, so the probability of NOT

getting a '20' in the first t rolls is

$$(1-p)^t. \tag{2.15}$$

The probability of getting a '20' on the next roll is p, so the total probability that the first '20' occurs on roll number $t+1$ is

$$p \times (1-p)^t \tag{2.16}$$

This function describes the waiting times for events and it is called the **geometric distribution** [Link]. We can get a sense of how long you have to wait to roll a '20' by computing Equation 2.16 for different values of t. For example, there is a 0.4 probability (i.e., 40%) of rolling a 20 within the first ten rolls:

$$1-(1-\frac{1}{20})^{10}. \tag{2.17}$$

Can you be confident that you will eventually roll at '20' if you are patient enough? Yes. Using this formula, we find that there is a 64% chance of getting a '20' within 20 rolls, 99.4% probability of getting a '20' within 100 rolls, and 99.996% within 200 rolls. The probability of eventually getting a '20' converges to 1 as you roll infinitely long.

Lastly, an important property of the geometric distribution is that **the average waiting time to the first success is simply $1/p$**: so in this example, 20 rolls.

Understanding waiting-time distributions: the exponential distribution. The geometric distribution measures time in terms of a discrete number of events or trials. But for our purposes we can approximate the geometric with a continuous distribution called the **exponential distribution** [Link]. For our setting, the two distributions are virtually equivalent [158], but the exponential distribution is much easier to work with.

Like the geometric, the exponential distribution is also used to model waiting times, but in settings where time is measured in continuous units. For example, I might ask: "How long will it be until the next earthquake on the Stanford campus?". Let λ be the rate of earthquakes per day [159]. Then, according to the definition of the exponential distribution, the probability that the next earthquake will occur exactly t days from now is

$$\lambda e^{-\lambda t} \tag{2.18}$$

and the total probability of having an earthquake any time within the next t days is

$$1-e^{-\lambda t}. \tag{2.19}$$

An example of this function is plotted below. Finally, the average waiting time [160] to the next earthquake is $1/\lambda$. Notice this has the same form as the average waiting time in the geometric distribution, $1/p$.

In our models, we are interested in waiting times until coalescent events. We measure time in generations, and set λ to be the rate of coalescence per generation, namely $(1/2N)$; hence the average coalescence time will be $2N$ generations.

The time distribution for two samples. We're now ready to model the distribution of coalescent times for two copies of a locus.

Remember that each generation there is a probability $1/2N$ that the two copies will coalesce. As described above, we'll model the waiting time to coalescence using the exponential distribution, which an excellent ap-

proximation to the geometric (and easier to work with, mathematically).

The next plot shows what is called a "cumulative distribution" of coalescence times under this model (Equation 2.19 with $\lambda = 2N$). As we showed in the last chapter, for human populations, the longterm (effective) population size N is around $20,000$ [d].

The way to interpret this plot is that the y-axis shows the probability that two samples coalesce within the most recent t generations (plotted on the x-axis):

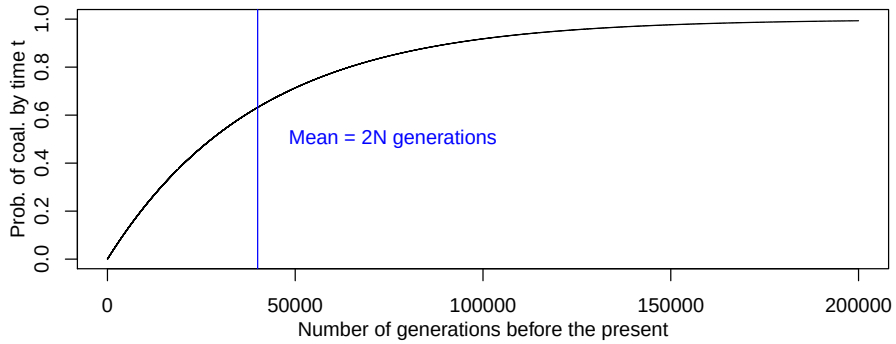


Figure 2.17: **Cumulative distribution for coalescence times.** *This shows the probability that two samples coalesce within the past t generations, assuming $N = 20,000$.*

As this plots shows, there is a 50% chance that coalescence occurs within the last $1.4N = 28,000$ generations (slightly less than the mean of $2N$ generations). And it's almost certain that coalescence occurs with $10N = 200,000$ generations. Note that if we assumed a different population size, this would change the numerical scale on the x-axis, but not the shape of the plot, which is simply proportional to N_e.

The last important point here is that *these timescales are really long* in terms of human evolution. Let's assume that the average generation time is about 25 years [161]... then the average coalescence time of $2N$ generations is 1 million years ago, before the appearance of anatomically modern humans.

The coalescent for larger samples. So far we have been talking about the coalescent for a pair of samples. Suppose instead that we sequence a particular locus in $m/2$ individuals, giving us a sample of m copies of the locus. Remember that the genealogy is embedded within the WF process.

How can we model the genealogy without having to bother with the WF process?

Imagine that we trace the ancestry of these m copies back in time. Going backward in time, *we will pick two of these lineages random* to coalesce into a common ancestor. Now (always looking backward in time) there are $m - 1$ copies. This process repeats until we get down to 2, and then finally to one common ancestor.

The process looks like this:

[d] *Remember that when we need to allow for the complexities of real world populations, the rate of coalescence depends on the* ***effective population size*** N_e*, rather than true population size N. For simplicity we'll discuss the models in terms of N, but you can think of subbing in N_e for real-life situations.*

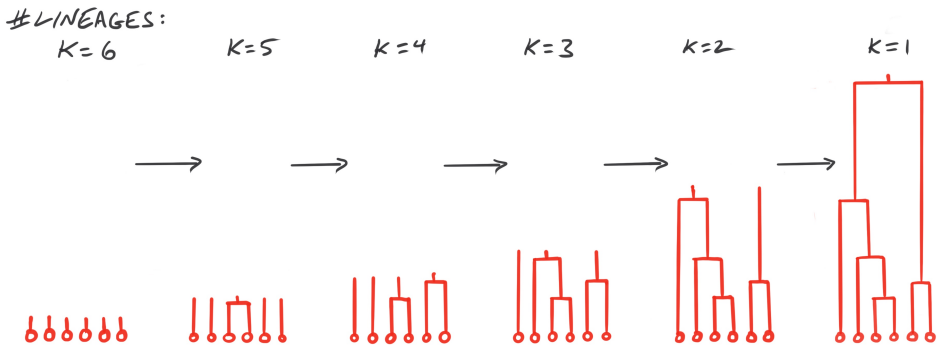


Figure 2.18: **Stepwise construction of a genealogy.** *At each step we randomly join two lineages. This results in a random* ***topology*** *– i.e., branching structure – that relates the m samples.*

Next we need to model the waiting times between coalescent events. We'll use T_k to be the number of generations when there are k lineages on the tree. (Here I use m as the number of samples in the present and k, ranging from 1 to m, as the number of distinct lineages at times in the past.) We showed above that T_2 has an exponential distribution, with a mean of $2N$ generations. What about larger values of k?

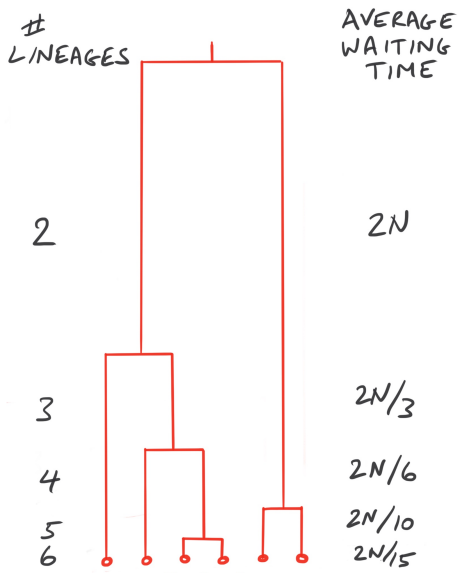


Figure 2.19: **Expected times in the genealogy.** *Here T_k labels the time during which there are k lineages. T_k is a random draw from an exponential distribution with mean $4N/k(k-1)$.*

How long does it take to go from k lineages to $k-1$?

To get this, we need to compute how long it takes for *any two* of the k lineages to merge into a single ancestor. The key thing here is that there are a lot of possible pairs that we could make out of k lineages. Specifically, there are [e]

$$\frac{k(k-1)}{2} \text{ possible pairs.} \tag{2.20}$$

Since there are $k(k-1)/2$ ways to get a possible coalescent event, this means that the waiting time to the first coalescence is reduced by a factor $2/k(k-1)$ compared to the waiting time when there are only two samples. Specifically, the waiting time when there are k lineages is exponen-

[e] *You can compute the number of possible pairs as follows. List the k lineages in some arbitrary order. The first lineage can pair with $k-1$ other lineages; the second can form $k-2$ pairs not counting the pair with the first lineage... and so on. The sum $(k-1)+(k-2)+(k-3)+...+2+1$ equals $k(k-1)/2$.*

tially distributed with mean:

$$\mathrm{E}[T_k] = \frac{4N}{k(k-1)} \tag{2.21}$$

For example when $k = 2$, the average waiting time is $2N$ generations. When $k = 10$, the average waiting time is 45-fold shorter: $2N/45$ generations [f]. When $k = 100$, the average waiting time is nearly 5000 times shorter: $2N/4950$.

In other words, **the most recent coalescent events – when there are many lineages – occur within a few generations, while the oldest coalescent events can easily take a million years.**

One question of particular interest is: *How long ago was the MRCA of a sample (or even of the entire population)?*

[f] *To get some intuition for this, imagine k cars with blindfolded drivers, driving erratically around a large parking lot. The time until the first crash is much shorter when there are many cars, and hence many possible pairs that could crash. E.g., with two cars the time until the first crash would be 45 times longer than for ten cars.*

Optional math on time to the MRCA. To compute the time to the MRCA, we add together the waiting times between each node. Here $T_{\mathrm{MRCA(m)}}$ is the random time to the MRCA for a sample of size m.

$$T_{\mathrm{MRCA(m)}} = T_2 + T_3 + T_4 ... + T_{m-1} + T_m, \tag{2.22}$$

where T_k represents the random waiting time during which there are k lineages, and is an exponential random variable with mean $4N/[k(k-1)]$. So the average time to the MRCA is:

$$\mathrm{E}[T_{\mathrm{MRCA(m)}}] = \sum_{k=2}^{m} E(T_k) \tag{2.23}$$

$$= \sum_{k=2}^{m} \frac{4N}{k(k-1)} \tag{2.24}$$

As the sample size gets large, this sum converges to a fixed value (the derivation requires techniques on infinite series):

$$\lim_{m\to\infty} \mathrm{E}[T_{\mathrm{MRCA(m)}}] = 4N. \tag{2.25}$$

In other words, as the sample size goes to infinity (or in practical terms, the entire population), what we see is that on average the most recent common ancestor for the entire population is $4N$ generations in the past.

The key result here is that for an average location in the genome, **the common ancestor for the entire population is 4N generations ago** ($\sim$2 million years, for humans). On average, **half of the total time back to the common ancestor is spent waiting for the last two lineages to coalesce.**

The genealogy has both random topology and random times. Before moving on, I want to emphasize one last important point about the coalescent: although we have been focusing on average properties, genealogies are inherently random, and vary in two important ways: *both the*

topology *(i.e., branching patterns) and coalescent times are random draws from the coalescent process*. This is illustrated below for genealogies with $m = 4$:

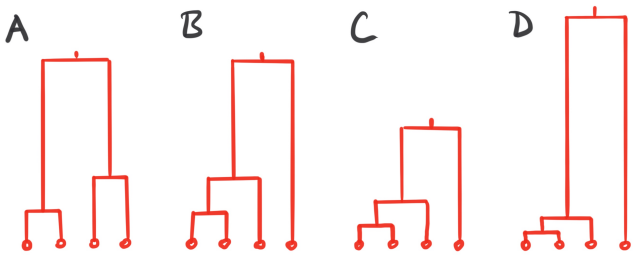


Figure 2.20: **Random outcomes of the coalescent process. A** *and* **B** *differ in their branching patterns (topologies), while* **B**, **C**, *and* **D** *have different coalescent times.*

In practice, the genealogies in different regions of the genome vary widely, and as we shall see next, this influences the allele frequencies and numbers of SNPs at any given locus.

Coalescent with mutation. So far, we have been talking about the genealogy. The genealogy reflects the inheritance of a DNA segment through time. Conceptually you can think of this as tracking the copying of DNA molecules through thousands of meioses, and reflecting the fact that a particular stretch of DNA in different people is a copy of the same ancestral DNA sequence in some distant ancestor.

Now we need to add mutations into the model. Patterns of genetic variation in modern samples reflect the combination of coalescence and mutation. As you get used to the structure of the coalescent, it provides a powerful tool for understanding patterns of genetic variation. We'll come back to this theme repeatedly in the upcoming chapters.

To make this concrete, let's suppose that we sequence a stretch of L base pairs ($L = 5$ kb, for example) in m samples (without recombination). We assume that new mutations arise at a rate μ per base pair per generation. It's going to be helpful now to label the lengths of branches on the tree (in generations); we'll do this using b_i for branch i:

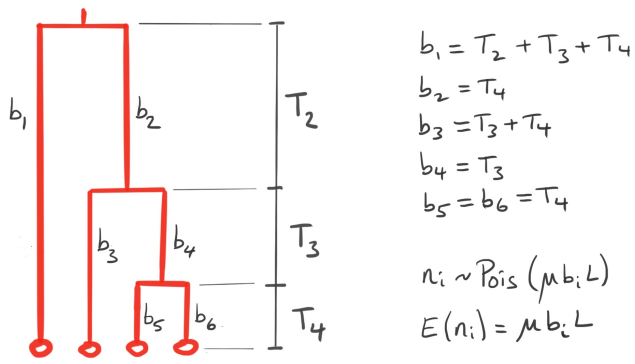


Figure 2.21: **Example of branch lengths in a genealogy.** *The branch lengths b_i show the length (in generations) of each branch. The numbering is arbitrary. Note that the specific branching patterns depend on the random tree topology.*

Notice above that we can write the branch lengths b_i in terms of the times between coalescent events (remember that T_k denotes the time when there are k lineages), although the specific branches and their lengths depend on the random topology.

Now, let n_i be the number of mutations on branch i. What is the **expected number of mutations $\mathbf{n_i}$**? This is the product of the mutation rate μ, branch length b_i, and sequence length L:

$$\mathrm{E}[n_i] = \mu b_i L \tag{2.26}$$

That is the *expected* number, but the actual number of mutations on any particular branch is random; this is modeled using the **Poisson distribution**. For more about the Poisson see [162].

Here's an example of what this might look like for a sample tree. Mutations are shown on each branch in blue; the tips of the tree (A-F) show six samples collected in the present day. *Mutations occur in ancestors along each branch, and are inherited by all the samples that lie below them.* So for example on the tree below, mutation 1 is inherited by sample A only, while mutation 2 is inherited by samples A-D. This means that we can go from the tree on the left, to the haplotypes on the right:

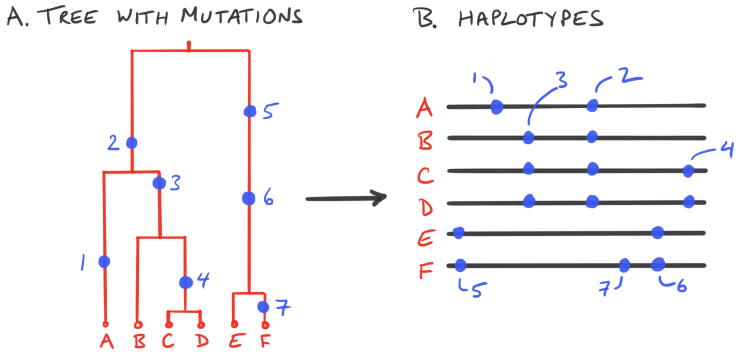


Figure 2.22: **Coalescent tree, mutations, and haplotypes. (A)** *Example genealogy with 7 mutations.* **(B)** *The corresponding haplotypes, with blue circles indicating derived alleles (at arbitrary locations).* *The labeling A-F corresponds to the sample labels in the tree. The assignment of alleles to haplotypes is entirely determined from panel (A). However the sequence positions of the mutations were assigned randomly while drawing panel (B).*

Trees, branches, and derived allele frequencies. The picture above hints at a key connection between the tree topology and the allele frequencies in a sample. If a branch is above j samples, then any mutation on that branch will occur exactly j times within the sample. This is shown below for two example topologies:

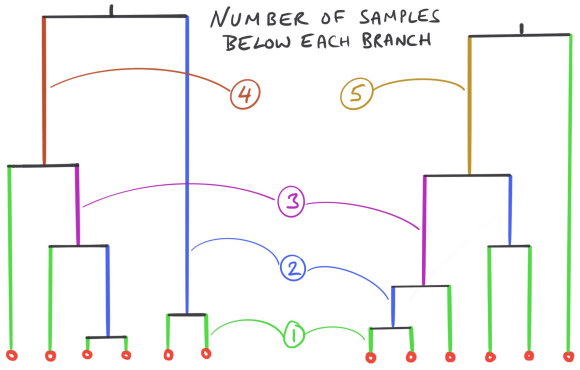


Figure 2.23: **Tree topologies and allele counts.** *Branches are colored according to the number of samples (tips of the tree) below each branch. For example, mutations that occur on blue branches will be present exactly twice in the sample. Notice that the branch lengths and possible allele counts differ between the random tree topologies.*

The branches labeled in green, above, are of particular interest as they lead to just a single sample. These are often referred to as **terminal branches**, and mutations that occur on them are referred to as **singletons** as they are found in only a single sample.

Quantitative aspects of variation in the coalescent. Thus far, I have described the coalescent at a conceptual level, as a way of understanding the structure of genetic variation. But we can also use it as a tool for making quantitative predictions about variation.

To start: *How many sequence differences can I expect between two samples, in a region of L basepairs?*

Recall that the coalescent time for two samples, T_2, is exponentially distributed with mean $2N$ generations, with an average μL mutations per generation along each branch. It follows that the expected number of mutations between each sample and the common ancestor is $T_2 \mu L$, and twice that for the total number of differences between the two modern day samples [163]:

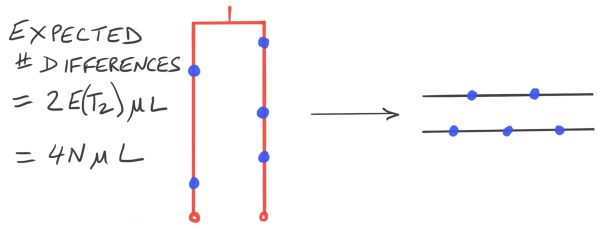


Figure 2.24: **Number of differences between two samples.** *The expected number of differences between two samples (equivalent to heterozygosity) is the product of their average coalescent time ($2N$) times the mutation rate along both branches, $2\mu L$.*

It's convenient to divide this by L, which gives us the expected number of differences per base pair. That is equivalent to **heterozygosity per site**, H, which we computed in the last chapter using the WF forward model:

$$H = 4N\mu. \tag{2.27}$$

Happily, the forward and backward approaches gives us the same result.

Most heterozygous SNPs are very old. I mentioned before that you have about 3 million heterozygous SNPs in your genome. How old are the mutations that produced these heterozygous alleles?

Using this model we know that for any random part of your genome, the average time to the common ancestor of two homologous copies is $2N$ generations (or about 1 million years). On average, a mutation occurs halfway along the branch to the common ancestor... this tells us that **the average variant in your genome is due to a mutation that happened 500,000 years ago(!)** and many are much older [164].

To put this into perspective, modern humans evolved in sub-Saharan Africa. About 70,000 years ago, some populations started spreading out of Africa into the Middle East, and then went on to colonize nearly all of the world's landmasses [g].

The number of SNPs found in a sample. The calculation above tells us the expected number of mutations in a sample of 2. How many mutations should we expect in a sample of size m?

[g] *I like to think there is some beauty in the fact that most of the heterozygous sites in your genome, or in mine, arose as mutations in distant ancestors in Africa half a million years ago.*

Optional: number of SNPs in a sample (the math). Suppose that you sequence a region of L basepairs in m samples. What is the expected number of variable sites (i.e., SNPs) that you detect? In addition to the result itself, this box illustrates the kinds of calculations that are (relatively) easy to do using the coalescent.

To get this, notice that we can break the problem down into two parts: (1) What is the **total branch length** – i.e., the sum of all the branch lengths; and (2) How many mutations do we expect to have occurred given the tree length?

To get the tree length, you might want to start by thinking about computing the length of every branch, and then adding all those together. But this is complicated because it depends on the branching structure of the tree, which is random. Instead, we can make the calculation easier by adding together a contribution from the time between each coalescent event. Specifically, what is the total branch length during the time when there are k lineages? Well the expected time is $4N/k(k-1)$, and there are k branches; multiplying these together gives $4N/(k-1)$ total branch length in this epoch. Next, adding together all the epochs, the expected total tree length is

$$\sum_{k=2}^{m} \frac{4N}{k-1}. \tag{2.28}$$

Then we can multiply by the mutation rate to get the expected number of variable sites (denoted S) in a sample of size m, in a region of L base pairs. After minor rearrangement and a shift in the sum index we get:

$$S = 4N\mu L \sum_{k=1}^{m-1} \frac{1}{k}. \tag{2.29}$$

When $m = 2$ this agrees with the result we got before for heterozygosity.

Equation 2.29 in the box provides an important result: the expected number of SNPs in a sample of size m.

One key point is that **as the sample size grows the MRCA time converges to 4N, while the number of segregating sites grows indefinitely at a rate proportional to the log of the sample size, $\ln(\mathbf{m})$** [h]. This is because as you increase the sample size, new samples usually add additional short branches near the bottom of the tree – slightly increasing the total branch length but not changing the MRCA time.

[h] *In human populations the number of rare variants actually grows a bit faster than ln(m), for reasons we'll explain shortly.*

The site frequency spectrum (SFS). Suppose that we collect genome sequence data from m samples. Let s_i be the number of SNPs at which the derived allele is present exactly i times. For example, s_1 gives us the number of singletons, s_2 the number of doubletons, and so on. The total number of SNPs, S, is related to s_i simply by summing over all the possible allele frequencies from 1 to $m-1$:

$$S = \sum_{i=1}^{m-1} s_i. \tag{2.30}$$

The vector of allele frequencies s_1, s_2, s_3, ..., is referred to as the **site frequency spectrum (SFS)**, and is a simple but important description of genetic variation [i].

[i] *As we'll discuss later, some types of natural selection, as well as other departures from the basic model such as recent population growth, can be detected because they distort the SFS away from this baseline model.*

What determines the SFS? Take a look back at Figure 2.23. The expected value of s_i depends on the amount of branch length that sits above exactly i samples: for example, s_2 depends on the amount of branch length that sits above pairs of samples. If we focus on genome-wide data, this has the effect that we will sample many different trees (in different parts of the genome) so that we can average over the randomness of the coalescent process.

The derivation of the SFS is beyond the scope of this book, but you can read about it here: [165]. Although the math is a little tricky, it produces the pleasingly simple result [j] that the expected number of variants with a derived allele frequency i is proportional to $1/i$:

$$\mathrm{E}[s_i] = \frac{1}{i} \times 4N\mu L. \tag{2.31}$$

[j] *This result also implies that the expected tree length above exactly i samples is $4N/i$.*

This distribution is plotted here:

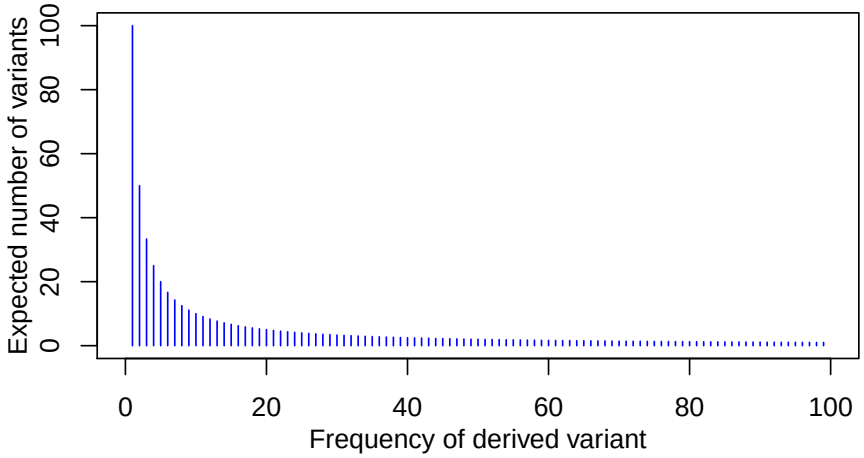


Figure 2.25: **The Site Frequency Spectrum (SFS).** *Here the expected SFS is plotted for $m = 100$ and $4N\mu L = 100$. Notice that most variants are rare. Here, 55% of the variants are below 10% frequency. This pattern is even more dramatic in large samples: in a sample of $m = 10,000$, 76% of the variants are at $< 10\%$ frequency.*

One key thing to notice is that most variants are rare. A useful rule of thumb is that allele frequencies are uniform on a log scale: in very large samples there are as many variants with derived frequencies between 0.1 and 1 as there are between 0.01 and 0.1, or between 10^{-5} and 10^{-4}.

Lastly, we can also get intuition for this from the WF model. In the last chapter I pointed out that most derived alleles are very rare, and only a small fraction are common: every new mutation starts out rare (i.e., at $1/2N$ frequency). Most are lost quickly, while only a few are lucky enough to drift up to become common. Thus, the WF model gives us a different conceptual tool to reach a similar conclusion.

The coalescent with population size changes. I have been describing the coalescent under the simplest possible population model: constant size and no population structure. This basic model is referred to as the **vanilla coalescent**.

But real populations often differ from this simple model, and it's important to think how this might affect the coalescent. In this section I'll de-

scribe how to think about two types of changing population size that are important for humans: bottlenecks and population growth.

Population bottlenecks. In population genetics, a bottleneck refers to a reduction in population size, often but not always followed by a return to the original population size. Bottlenecks are important because they greatly increase the rate of genetic drift.

Bottlenecks have been important features of human evolution, including during the spread of populations as they left Africa and colonized the globe during the past $\sim 80,000$ years [166]. This is why non-African populations have less genetic variation than Africans.

Bottlenecks have also reshaped patterns of variation in some populations within much more recent timescales – for example the ancestors of modern Jews went through a tight population bottleneck $\sim$1000 years ago [167].

In the WF model, we can think of the bottleneck as increasing the variance in allele frequencies: some alleles increase dramatically, while others decrease:

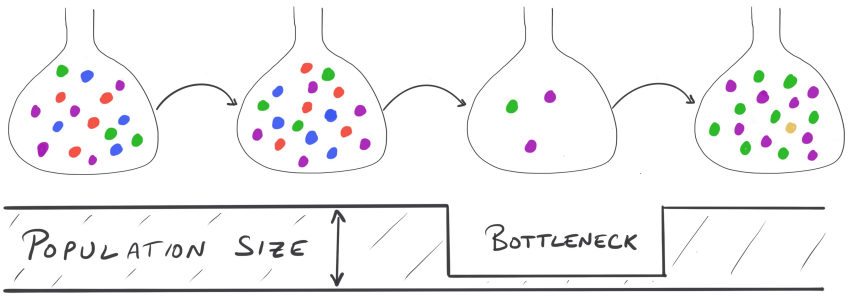


Figure 2.26: **WF drift through a bottleneck.** *Bottlenecks greatly increase the rate of drift due to low N_e.*

Of course, we can also think of this in terms of the coalescent. Remember that the rate of coalescence is $k(k-1)/4N$ per generation. If our model allows N to vary with time then, when N decreases, the rate of coalescence will increase at an inverse rate.

This means that we will get an increased rate of coalescence within the bottleneck, and fewer ancient lineages. The few lineages that predate the bottleneck are likely to have many descendants:

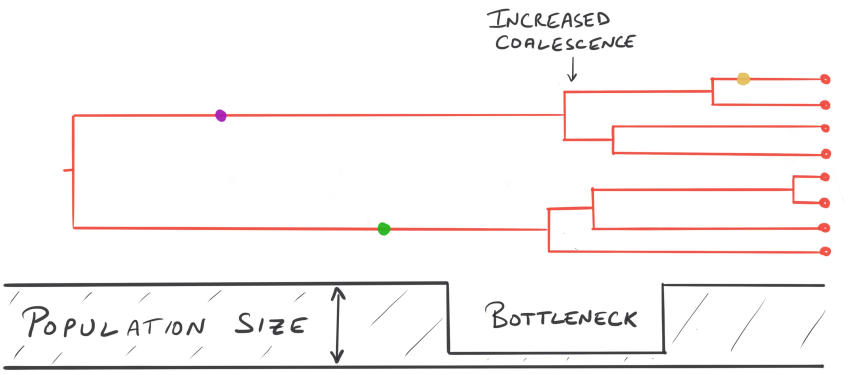


Figure 2.27: **Coalescent through a bottleneck.** *The rate of coalescence during the bottleneck is greatly increased due to low N_e. The purple and green mutations occurred on lineages that survived the bottleneck and are at high frequency in the final sample (at right). The yellow mutation postdates the bottleneck and is at low frequency. The tree here is tipped on its side to emphasize similarity to the WF picture above.*

This example also helps to **illustrate the intimate connection between coalescence and drift**: in a sense, drift in the WF model occurs *because* lineages are coalescing.

Population growth. Another key feature of real human populations is dramatic population growth, from ∼1 million in 10,000 BCE to ∼8 billion today. How did this affect the coalescent process, and genetic variation?

Here, the logic is opposite the bottleneck situation: a very large population size slows down the rate of coalescence at very recent times. As a result, **recent growth hugely increased the number of very rare variants**.

To understand this, it would be most natural to model population growth as following an exponential increase over time [168]. But the math for coalescence with exponential growth is a bit clunky and obscures the main points, so we'll consider a simpler model:

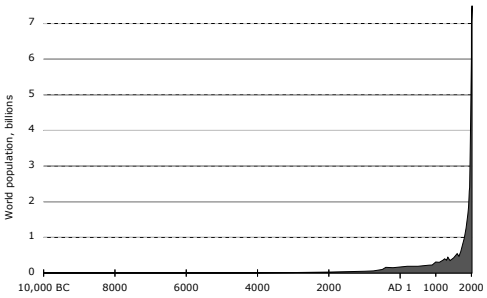


Figure 2.28: **Exponential human population growth.** *Estimates of total world population during the past 12,000 years.* *Credit: El T [Link], Public Domain. Data: [Link].*

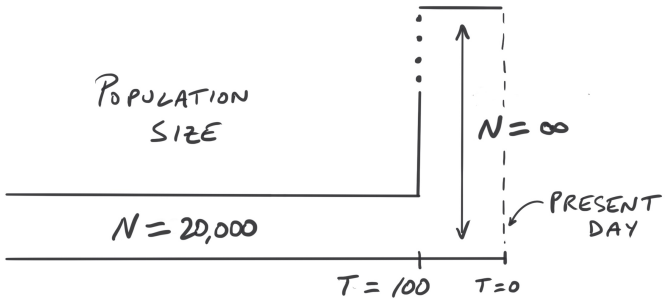


Figure 2.29: **Instantaneous growth.** *In this simplified model, the ancestral population size is* 20,000*, followed by instantaneous growth to infinite size* 100 *generations ago.*

In the model above, we consider a population that grew instantaneously to infinite size, 100 generations ago. How would this extreme model change the properties of trees, compared to a model of constant $N = 20,000$?

Recall that in the vanilla (constant size) model, for large samples the first coalescent events occur very quickly. But in the infinite growth model, there is no coalescence in the most recent time period, thus greatly extending the terminal branches:

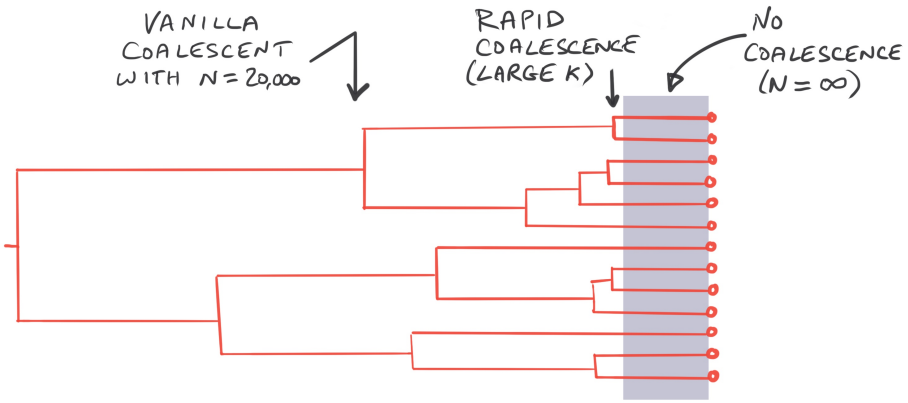


Figure 2.30: **Coalescent tree in the instantaneous growth model.** *There is no coalescence for the first* 100 *generations (grey region) due to infinite population size. Note: the picture is not drawn to scale.*

The longer terminal branches produce many more singleton mutations. Recall for the vanilla model that the expected number of singletons is

$4N \times \mu L$ (Equation 2.31).

But in the infinite growth model, every tip is extended by 100 generations. Since there are m tips, the expected number of singletons is now $(4N + 100m) \times \mu L$. So for example, in a sample of $m = 1000$, the number of singletons is more than doubled! Meanwhile, the deeper structure of the tree is unaffected, aside from slightly pushing back all the expected times.

While the infinite growth model is unrealistic it still provides valuable insight. Under a more realistic model of continuous exponential growth there is a strong reduction in the rate of recent coalescence relative to the vanilla model, thereby increasing the lengths of recent branches. In summary, **recent exponential growth leads to a dramatic increase in low frequency variants.**

Footprints of population history in real data. In a 2012 paper, Tennessen et al described genome-wide (exome) sequencing data from about 1100 individuals of African-American, and 1300 individuals of European-American ancestry [169] [170]. They found over 500,000 SNPs, of which 86% were at less than 0.5% frequency and 57% were singletons.

The plot below illustrates the SFS for these two samples: Each line plots the proportion of alleles (Y-axis) in bins of allele frequencies (X-axis). Both axes are on log-scales; on these axes the theoretical null (constant N) is approximately a straight line [171].

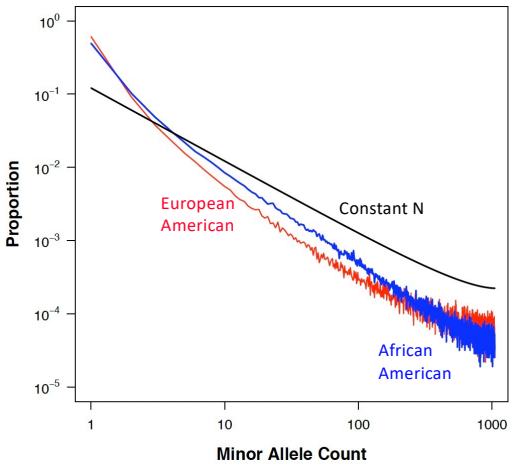


Figure 2.31: **The SFS in human populations: huge excess of rare variants.** *Notice that the real data (colored lines) are well above the theoretical prediction (black line) in the upper-right hand part of the plot.* *Credit: Modified Figure S9D from Jacob Tennessen et al 2012 [Link] Used with permission.*

As you can see, the real data from both populations show a much higher fraction of rare variants (higher in the upper right) compared to the null. This is direct evidence for rapid recent population growth.

The authors then fit a model of historical population sizes (often called a **demographic model**) that can fit the full SFS data. The model is shown below, including a tight European bottleneck, and extreme recent population growth to reflect the huge excess of rare variants relative to the vanilla coalescent model:

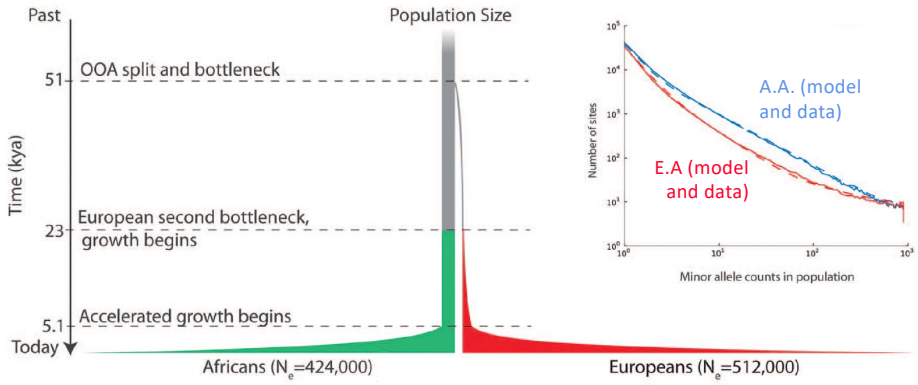


Figure 2.32: **Fitted demographic model.** *This model is designed to fit the SFS data for African- and European Americans (see upper-right panel).* ***The inferred model illustrates extreme recent growth in both populations, and a strong European bottleneck.*** *Note that more recent estimates include times and population sizes that are roughly doubled due to updates in mutation rate estimates since 2012. The image was simplified by not showing European mixing into African Americans in the last 400 years. Credit: Modified Figure 2B from Jacob Tennessen et al 2012 [Link] Used with permission.*

As you can see in the inset in the upper right of the plot above, the proposed model provides a good fit to the observed SFS. While the precise parameter estimates vary among papers in this area, all of them agree on the presence of a tight bottleneck for non-African populations, and extreme recent population growth.

The coalescent and the fixation process. Thus far we've been using the coalescent to understand genetic variation. But it also provides a useful intuition for understanding how alleles fix. *Crucially: A variant is fixed in the present day if and only if it was present in the population MRCA.*

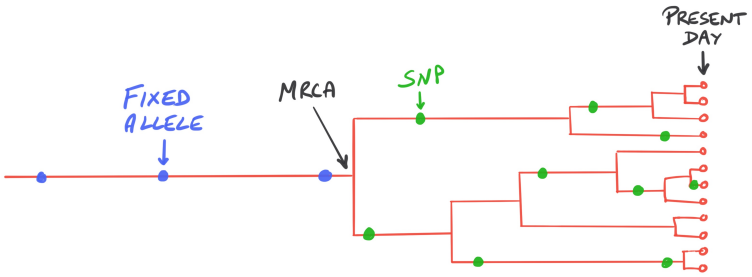


Figure 2.33: **Fixed variants are present in the population MRCA.** *The blue variants are fixed in the present day population because they were carried by the MRCA; the green variants are SNPs.* *Assume that the MRCA shown for this sample is in fact the MRCA of the entire population.*

This now provides intuition for two important results that I stated in the last chapter:

The probability of fixation for an allele now at frequency p, is simply p. You now know that any present-day sample has a common ancestor sometime in the past. Flipping this around, if we imagine going far enough forward in time (on the order of $4N$ generations), we know that exactly one copy of this locus will eventually be a common ancestor of the entire current population. So for a SNP now segregating at frequency p, there is a probability p that in the future a lineage carrying it will become the ancestor of everyone.

The average time to fixation for a new mutation is $4N$ generations. The logic here is similar: If a new mutation eventually fixes, this means that it

is destined to become the common ancestor of everyone in a future population. We know that the expected time back to a common ancestor is $4N$ generations. Forward in time, it takes approximately $4N$ generations until the first time that the mutation is the common ancestor of everyone [172]:

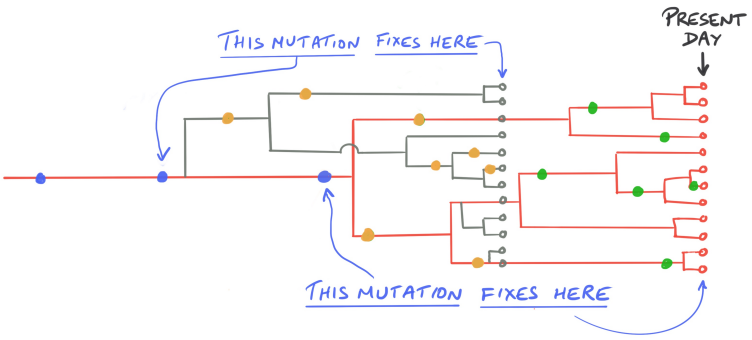


Figure 2.34: **Mutations fix at different times** *depending on when their lineages first become the population MRCA. All three blue mutations are fixed in the final population at the far right, but the first time at which they fix depends on the structure of the coalescent.* *Also of interest: the yellow variants are SNPs in the gray sample, but many of them are lost by the time of the final (red) sample.*

Coalescent simulation of haplotype variation [Optional]. As in the previous chapter we end with a basic outline for how to simulate haplotypes, this time using the vanilla coalescent model. If you're good at programming you may wish to try this [173].

Data storage: It's useful to create a *data structure that represents nodes of the tree*. There are m of these to represent each of the present day samples, and $m-1$ for the ancestral tree nodes. Each node stores the time, as well as a pointer to the parent node, and to each child. (The child nodes are null for the present day samples, and the parent pointer is null for the MRCA.) It also stores a list of the derived variants present at this node.

You'll also want:
- a list of the *locations of mutations* within the sequenced region;
- the *current time* before present, for use while constructing the tree;
- a *list of current active lineages (nodes)*, for use while constructing the tree.

Construct tree:
Initialize the current time at 0.
The initial active lineages are the m present-day samples. Set the times for these to 0 and all their children to null.

for (k starting at $k = m$, down to $k = 2$) do:
{

- *Coalesce lineages:* Pick two of the active lineages at random to coalesce. Create a new node, with these two lineages as children. Drop those from the active list and replace with the new node;
- *Generate node time:* Update the current-time by adding a random time $\sim$ $Exponential(4N/k/(k-1))$. Set the time at the new node equal to the new current-time;
- *Update lineage counter:* $k = k - 1$

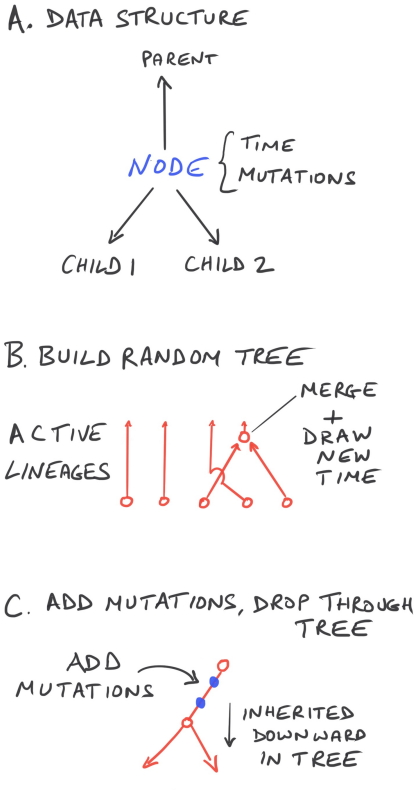


Figure 2.35: **Coalescent simulations.**

}

Add mutations: Starting from the top of the tree, visit each branch i in turn and do:

{

- *Calculate the branch length* b_i as the elapsed time between the parent node and the child node;
- *Simulate the number* of mutations as $\sim Poisson(b_i \mu L)$;
- *Simulate the position* of each mutation as $\sim Uniform(0, L)$;
- *Drop the mutations* down through every node below the mutated branch to the present-day samples.

}

Comments. This is conceptually a bit more complicated than the Wright-Fisher pseudocode, but it's far more computationally efficient, as we don't need to track a huge number of ancestors that are not relevant to variation in the present-day sample.

This type of algorithm provides an extremely efficient tool for simulating genetic data. As a rule, coalescent simulations are much faster than WF simulations, but they can be less versatile, and more difficult to modify to new situations. There are numerous **free software packages** for coalescent simulations, including **msprime** [Link].

Well done! In the last two chapters you have learned the two most fundamental tools for understanding patterns of genetic variation. In the next two chapters we'll discuss how to fold recombination and population structure into these basic models.

Notes and References.

[154]Credit for finding this quote goes to the late Paul Joyce: [Link].

[155]We'll talk more about these early data in Chapter 2.7, along with the other major conceptual development of the 1970s and 80s, the Neutral Theory.

[156]Inspiration for the coalescent was motivated in part by developments in population genetics during the 1970s. John Kingman (later Sir John Kingman) was a mathematician at the University of Oxford with particular interest in stochastic processes. He came to this problem after conversations with a group of Australian population geneticists: Pat Moran, Warren Ewens, and Geoff Watterston. In a trio of papers published in 1982, Kingman framed the process in highly mathematical terms and published in mathematical journals; in one of these he coined the term "coalescent" (hence the occasional name "Kingman Coalescent" for this model). Kingman only worked in population genetics for a couple of years. Despite the huge impact of the coalescent work, Kingman commented to me many years later (2022) that "Coalescent theory is very far from the thing I am most proud of", preferring instead his contributions in queuing theory (which later became important in the development of the internet [Link]), and perhaps his role as a university administrator, including as head of the University of Bristol (England) starting in 1985.

Meanwhile, Richard (Dick) Hudson was a PhD student at the University of Pennsylvania and at UC Davis. He published a pair of papers a year after Kingman (but unaware of Kingman's work) that describe–almost as an afterthought– the nuts and bolts of the basic coalescent model, as well as important extensions to handle the coalescent with recombination, all for the purpose of performing highly efficient simulations. He later went on to develop extensive tools for coalescent simulation.

The third key person, Fumio Tajima, a Japanese scientist then at the University of Texas Houston, published a 1983 paper that outlines the structure of genealogies and the coalescent and showed how this can be used to derive important sample statistics in population genetics. Published in the same year as Hudson's work, in some ways Tajima's presentation is the most modern in flavor (and is the paper in which I first encountered the coalescent as a graduate student, some ten years later).

Kingman JFC. The coalescent. Stochastic processes and their applications. 1982;13(3):235-48,
Kingman JF. Origins of the coalescent: 1974-1982. Genetics. 2000;156(4):1461-3,
Hudson RR. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 1983:203-17,
Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 1983;23(2):183-201,
Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105(2):437-60

[157]Early, highly readable reviews of the coalescent were written by Dick Hudson and Magnus Nordborg. (You can find online versions of the book chapters via Google Scholar: for Hudson 1990 see [Link]; for Nordborg 2000 see [Link])

Hudson RR. Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology. 1990;7(1):44
Hudson R. The how and why of generating gene genealogies. Mechanisms of molecular evolution. 1993:23-36
Nordborg M. Coalescent theory. Handbook of Statistical Genomics: Two Volume Set. 2019:145-30 .

[158]Differences between the geometric and exponential only arise in very special settings: for example when the sample size is large compared to the total population, and also in problems looking at coalescence within relatives.

[159]At the time of writing there have been two major earthquakes at Stanford (in 1906 and 1989) since its founding in 1885. So a simple-minded estimate of λ for major earthquakes would be $\sim 4 \times 10^{-5}$ per day. For an entirely gratuitous picture of a smashed car outside Stanford's Old Chem Building in 1989 see [Link]. USGS data: [Link].

[160]The mean of the exponential distribution with rate parameter λ is given by

$$\int_{t=0}^{\infty} t \cdot \lambda e^{-\lambda t}\, dt = \lambda^{-1}. \tag{2.32}$$

[161]Estimates for long-term average generation times are in the 25-30 year range. I chose 25 here to make round numbers, and that's roughly balanced by using a population size on the high end for human populations.

[162]The **Poisson Distribution** is a widely used model for the (random) number of rare events that occur in a specified time – for example the random number of earthquakes in a 100-year period. It depends on a single parameter, which gives the expected number of events. To read more see [Link].

$$\text{number of mutations} \sim Poisson(\mu L b_i) \tag{2.33}$$

[163] We want to compute the expected number of pairwise differences, m, between two samples under a constant population size model. Note that m is distributed as Poisson$(2\mu LT)$, where μ is the mutation rate per base pair per generation, L is the length of the region in base pairs, and T is the realized coalescent time of the two samples. We use $\Pr[T]$

to denote the probability density function for T (i.e., the exponential distribution with mean $2N$). Then we have:

$$\begin{aligned} \mathrm{E}[m] &= \int_0^\infty \mathrm{E}[m|T]\ \Pr[T]\mathrm{d}t && (2.34)\\ &= \int_0^\infty (2\mu LT)\ \Pr[T]\mathrm{d}t && (2.35)\\ &= 2\mu L \int_0^\infty T\ \Pr[t]\mathrm{d}t && (2.36)\\ &= 2\mu L\ \mathrm{E}[T] && (2.37)\\ &= 2\mu L 2N = 4N\mu L && (2.38) \end{aligned}$$

or simply $4N\mu$ per base pair.

[164]The mean is actually a bit older than this even, because there's an additional ascertainment effect in which the distribution of coalescent times at sites with variation is older than the unconditional mean.

[165]For a proof of the θ/i result, by Richard Hudson, see
Hudson RR. A new proof of the expected frequency spectrum under the standard neutral model. Plos One. 2015;10(7):e0118087

[166]Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences. 2005;102(44):15942-7

[167]Waldman S, Backenroth D, Harney É, Flohr S, Neff NC, Buckley GM, et al. Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14th century. Cell. 2022;185(25):4703-16

[168]The classic paper on exponential growth is
Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991;129(2):555-62

[169]Tennessen JA, Bigham AW, O'connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-9

[170]I'm highlighting this work because it illustrates our major points. There is a long history of papers in this area, with sample sizes and genome coverage generally increasing over time.

[171]The slight uptick at the right occurs because the data are plotted in terms of the minor allele frequency instead of derived allele frequency.

[172]This argument is not entirely rigorous, and the classic results on this use forward-in-time diffusion theory.

[173]Here is a link to some similar sample code by Goncalo Abecasis [Link]. When I get time I expect to post a file that follows this code more closely.

2.3 Linkage, recombination, and LD.

Within small linked regions of the genome, the coalescent process generates correlations between the genotypes at different SNPs. This is known as linkage disequilibrium (LD). Meanwhile, at larger distances, recombination breaks down LD by shuffling genotypes. Here we discuss how the opposing forces of linkage and recombination shape genetic variation.

The concepts of linkage, recombination, and LD appear in almost every topic in human genetics, including natural selection, population history, population admixture and introgression, and the genetics of complex traits.

A first look at haplotype structure. The first time anyone sequenced the same locus in multiple individuals was in 1983. In a landmark study, Marty Kreitman, who was a graduate student at the time, sequenced the ADH gene in 11 lines of the fly *Drosophila melanogaster* [174]. The figure below shows a simplified version of the complete data set from that paper:

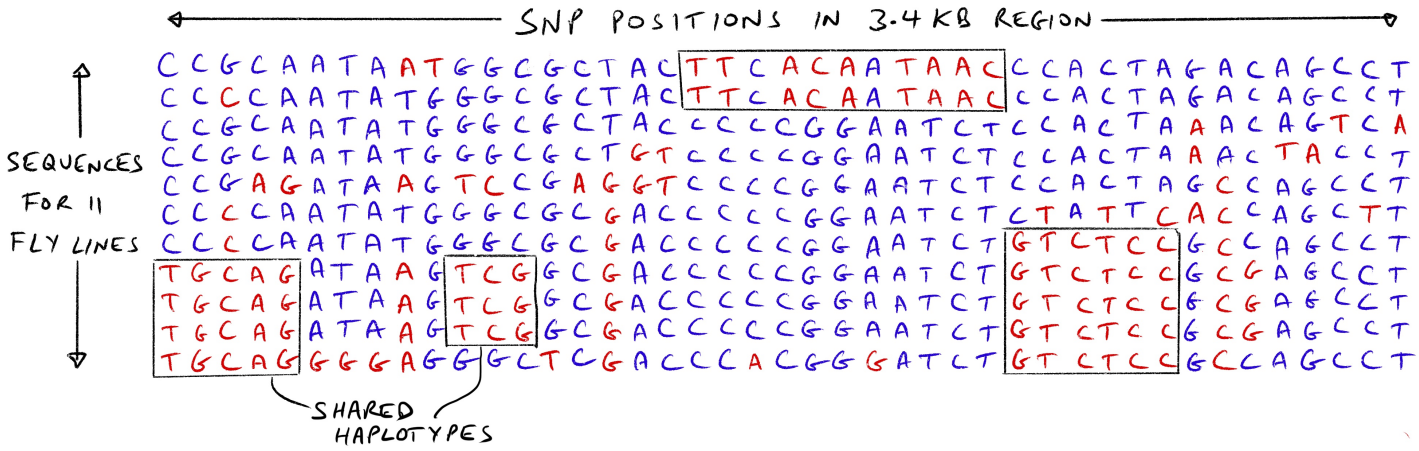


Figure 2.36: Haplotype structure. *Each row shows the genotype for a single fly line, and the columns show genotypes at SNP positions (most sites are not shown as they were identical in all 11 lines). The major allele at each position is shown in blue. Examples of blocks of shared haplotypes are indicated.* *Note that each line was constructed to carry only a single haplotype. For simplicity, a few indels are not shown. Data: Martin Kreitman (1983) [Link].*

Each row shows the sequence of alleles found on a particular chromosome copy in the population. We refer to the set of alleles found at variant positions within a linked region as a **haplotype** [a].

[a] *For another early example, this time from human data, see Figure 1.31.*

Looking at these haplotypes, one feature may jump out at you: **particular combinations of alleles at different SNPs frequently appear together**. For example, on the left, a block of alleles TGCAG is shared among four lines, all of which (and one other) later carry another block: GTCTCC.

This is a very typical feature of genetic data: particular alleles at nearby SNPs often appear together more often than expected by chance. This nonrandom assortment of alleles at different sites is referred to as

linkage disequilibrium (LD). *How do we understand this?*

In the next couple of pages, we'll talk about how **linkage generates LD, while recombination tends to break down LD**. As before, both the backward-in-time models (the coalescent) and forward-in-time models (Wright-Fisher) provide complementary kinds of intuition, and we'll use both approaches.

Linkage generates haplotype structure (or equivalently, LD). Sites that are close together in the genome are usually inherited together. This is called **linkage**.

When we introduced the coalescent in the last chapter, we ignored the possibility of recombination, focusing on sequences that are **completely linked**. In this setting, there is a clear relationship between the branching structure of the tree, and the corresponding haplotypes:

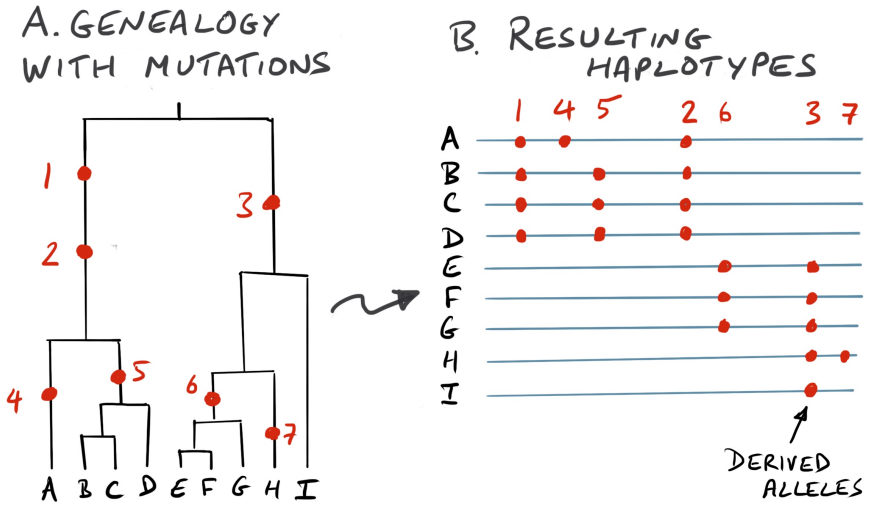


Figure 2.37: **The coalescent process generates haplotype structure. A** *shows a coalescent tree without recombination. The red circles indicate mutations.* **B** *shows the corresponding haplotypes; the red circles indicate derived alleles using the same numbering as in panel A.* *The positions of the mutations within the sequenced region are random.*

For example in the tree above, mutations 1 and 2 occurred on the same branch, and hence those two derived alleles always appear together. Mutations 1 and 5 are on adjacent branches, and so derived alleles 1 and 5 *usually* appear together (although haplotype A has the derived allele at 1 but not at 5).

For a tree without recombination, there are very strong constraints on the possible configurations of the derived alleles across haplotypes. For example, if we focus on two SNPs at a time, you might expect that there could be four possible haplotypes. If we label the ancestral and derived alleles as A/a at the first SNP, and B/b at the second SNP, then in principle the haplotypes could be A-B, A-b, a-B, a-b.

But in the absence of recombination we can only get either two or three of the four possible haplotypes, depending on where the mutations occur. As you can see in the examples below, we can get either 2 or 3 of the possible combinations, but not all 4:

Furthermore, looking across all SNPs together, there are additional con-

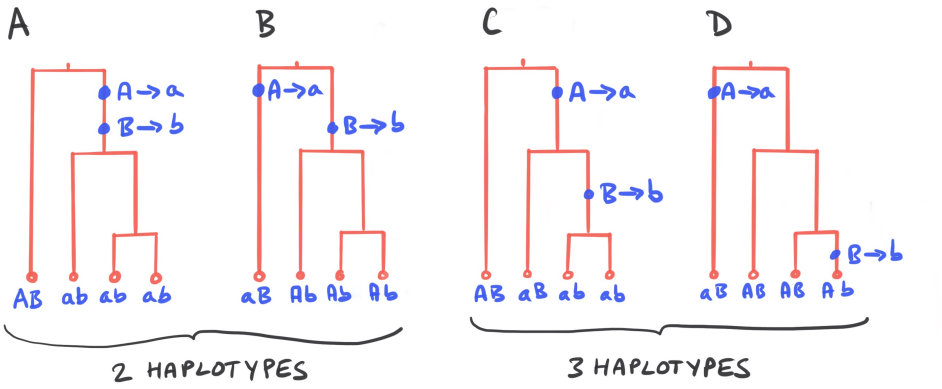


Figure 2.38: **Pairwise LD in the absence of recombination.** *For any pair of SNPs we can observe either 2 or 3 out of the 4 possible haplotypes (depending on where the mutations lie on the tree). While this is illustrated here for 4 samples, it is true regardless of sample size.*

straints: the alleles must be nested in a way that is consistent with existence of a single tree. Haplotypes that are consistent with a single tree are said to form a **perfect phylogeny** [Link]). I suggest that you draw some examples of trees with mutations, to see what configurations are possible.

Recombination. But in practice, most regions of the genome are subject to recombination. Recombination plays a crucial role in shuffling haplotypes, and producing combinations that would be impossible in the absence of recombination.

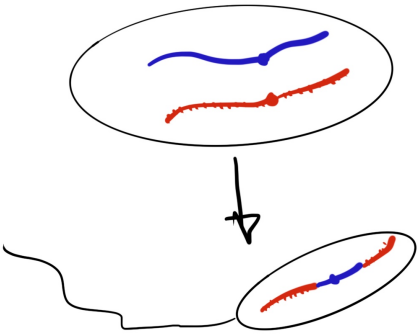


Figure 2.39: **Recombination.** *Sperm and eggs carry recombined mixtures of the parental chromosomes; typically there are around 1-2 switches per chromosome, known as* ***crossovers****, positioned randomly along each chromosome.*

A quick refresher on recombination. Recall that during the production of eggs and sperm, the chromosomes go through meiosis. In humans, this reduces the number of chromosomes from 46 to 23. During this process, the maternal and paternal chromosomes are broken and then joined back together so that chromosomes in the resulting gametes are mixtures of the parental chromosomes. This is called **recombination**, or **crossover** [175]. Crossover events are positioned more-or-less randomly across the genome with an average of 26 crossovers per sperm and 42 per egg.

Genetic distance. It will be helpful to talk about genetic distance, which measures the rate of crossover, between different positions along a chromosome. Genetic distance is measured in terms of **centiMorgans (cM)**.

We define the genetic distance x, between two points on a chromosome to be x cM if the average number of crossovers between those two points is x/100 per meiosis. For example, if two points are 10 cM apart, then we expect 0.1 crossovers per meiosis.

Figure 2.40: **Crossover observed in laboratory whiteboard pens.**

Furthermore, we'll be most interested in short genetic distances, for which **we can also interpret genetic distance in centiMorgans as the percent probability of a crossover** in the specified interval [176]. For example, if the genetic distance between two sites is 1 cM, then there is about 1% probability of a crossover per meiosis in that interval.

Lastly, it's helpful to define relate genetic distances to base pair distance in the DNA sequence. For this purpose we define the **recombination rate**. This is commonly measured in cM/Mb: that's 100 times the expected number of cross overs per megabase. The average recombination rate in the human genome is about 1.2 cM per Mb. In other words, there

is about a 1.2% probability of a crossover event per megabase [b] [177].

So, why do we care about this here? Recombination is central to our story because it shuffles haplotypes:

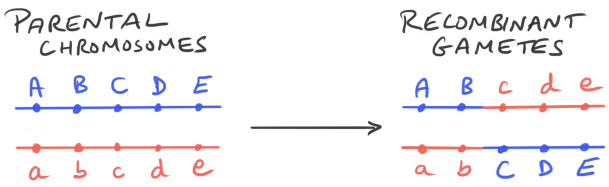


Figure 2.41: **Recombination mixes haplotypes,** *often creating new combinations that did not exist previously.*

In this way, **recombination generates new combinations of alleles that would not be possible with complete linkage**. For example, when there is recombination we *do* expect to see all four possible haplotypes for a pair of SNPs, unlike what I showed you above. However, the rates depend crucially on genetic distance.

As we shall see, for SNPs that are close together in the genome (less than ∼0.01–0.1 cM, or about 10–100 Kb) linkage is a stronger force than recombination and there tends to be strong haplotype structure. At larger distances (more than ∼0.1 cM), recombination is highly effective at shuffling genotypes, and LD is generally weak. In the next sections we'll see why this is.

Measuring LD between pairs of SNPs. To make this discussion more precise, it's helpful to define some measures of LD that we can study in models, and in real data.

Imagine two SNPs. One has alleles A and a, with allele frequencies p_A and p_a; the other SNP has alleles B and b, with frequencies p_B and p_b. Then there are four possible haplotypes: AB, Ab, aB, and ab, with frequencies p_{AB}, P_{Ab}, and so on:

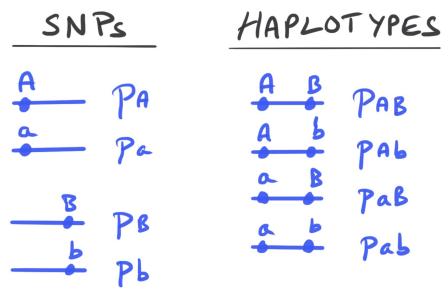


Figure 2.42: **Notation for allele and haplotype frequencies at two SNPs.** *Here p denotes a frequency.*

If I didn't tell you anything about these SNPs in advance, what would you guess for the haplotype frequency p_{AB}? The most natural thing would be to guess that the alleles are independent of each other, in which case $p_{AB} = p_A p_B$.

This intuition is captured by a measure called D, which is the difference between the observed and expected frequency of the AB haplotype:

$$D = p_{AB} - p_A p_B. \tag{2.39}$$

If genotypes at the two SNPs are independent (i.e., the SNPs are in linkage equilibrium), then $D = 0$. [c]

It may seem arbitrary to define D in terms of just the AB haplotype, but a little algebra will show that if we redefined D in terms of a different haplotype (e.g., with respect to Ab), then the only thing that would happen is that D would switch signs to become $-D$. Since the allele labeling is usually arbitrary, in practice we'll only pay attention to the absolute value $|D|$.

The second important measure of LD is known as D'. A weakness of D

[b] *It's useful to remember that the human recombination rate is about 1% per megabase.*

[c] *Note: If the alleles are labeled 0 and 1 at each SNP, then D can be interpreted as the statistical covariance between alleles at the two SNPs.*

as a measure of LD is that its possible range depends on the allele frequencies of the two SNPs, so it doesn't immediately tell us if LD between two SNPs is weak or strong.

To solve this limitation, D' is adjusted to range between -1 and 1 regardless of the allele frequencies:

$$D' = \frac{D}{D_{\max}} = \frac{D}{\min(p_A p_b, p_a p_B)} \quad \text{for } D > 0 \tag{2.40}$$

$$= \frac{D}{\min(p_A p_B, p_a p_b)} \quad \text{for } D < 0 \tag{2.41}$$

As with D, the sign of D' depends on the labeling of the alleles, so most papers use the absolute value, $|D'|$.

The formula for $|D'|$ is a bit messy, but a little algebra reveals a crucial property: **$|D'| < 1$ if and only if all four possible haplotypes are present. In other words, $|D'| < 1$ implies that there must have been recombination between the two sites**.

The third important measure of LD is called r^2, and again builds off D:

$$r^2 = \frac{D^2}{p_A p_a p_B p_b}. \tag{2.42}$$

The value of r^2 ranges from 0 to 1, where 0 means that the SNPs are completely independent. **A value of $r^2 = 1$ is referred to as perfect LD** [d], **and occurs if and only if there are just two of the four possible haplotypes: i.e., only AB/ab or only Ab/aB.**

[d] *Note: r^2 can also be interpreted as the squared correlation coefficient: the statistical correlation between genotypes at two SNPs is $r = D/\sqrt{p_A p_a p_B p_b}$.*

As we'll see later in the book, **r^2 is the natural parameter for measuring the contribution of LD to genetic associations in complex trait genetics** [178].

Strong recombination breaks down LD. We can now show a key result for how LD behaves in a model with strong recombination (and no drift).

Suppose we create an artificial population where two SNPs start out in strong LD, with an initial value of $D{=}D_0$ in generation 0. Let c be the probability of crossover, per generation, between these two SNPs. (See the Box below for a precise definition of c and a derivation.)

In the next generation, the LD (denoted D_1) is predicted to be:

$$D_1 = (1 - c)D_0, \tag{2.43}$$

and over successive generations the decay of D simply multiplies, so that in generation t we have:

$$D_t = (1 - c)^t D_0. \tag{2.44}$$

This implies that unless the recombination rate is very small, LD decays very quickly. For two SNPs that are 20 Mb apart, say, we expect that $c \sim 0.2$. After ten generations, $(1 - 0.2)^{10}{=}0.1$, meaning that within ten

generations D decays to just 10% of its starting value. In this time, LD between unlinked parts of the genome ($c = 0.5$) would essentially disappear. Here's a plot showing decay of D over time:

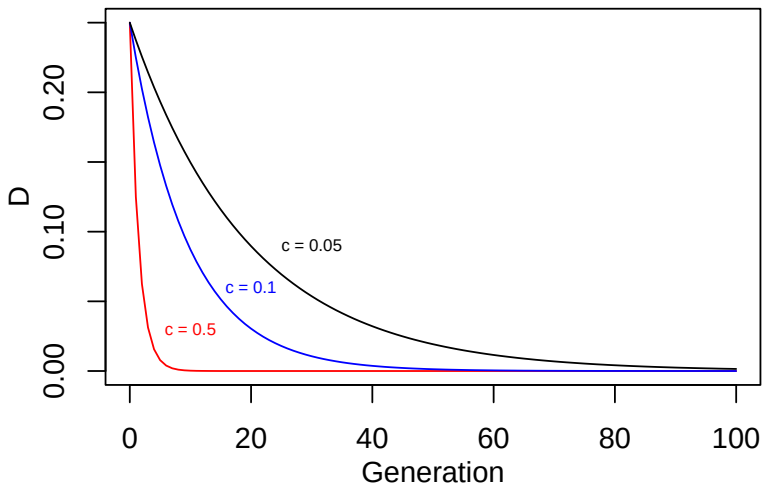


Figure 2.43: **D decays within a few generations for large recombination rates** *(Equation 2.44, assuming $D_0 = 0.25$). Timescales of tens of generations are very short compared to timescales of drift – which takes place over tens of thousands of generations.*

This result also holds for D' since, without drift, the denominator is unaffected by recombination.

Optional: Decay of LD due to recombination. Here we sketch out the argument for how D decays over time (Equation 2.44).

First, my definition of c above was a bit sloppy. To be more precise, c will be the probability that the two alleles passed into a gamete both came from the same parent (i.e., both from the mother, or both from the father). At short genetic distances, c is closely approximated by the probability of at least one cross-over between the two SNPs, but this definition implies that the maximum value of c is 0.5, which corresponds to random assortment of alleles on different chromosomes [179].

Our derivation assumes that recombination is happening much faster than drift. Specifically we assume the allele frequencies p_A and p_B stay constant, while the haplotype frequencies, p_{AB}, etc, change due to recombination. (To be more precise, this approximation will be accurate when the change in D due to recombination, cD, is much larger than the rate of drift, which is $\mathcal{O}(1/N)$.)

Let p_{AB} be the frequency of the AB haplotype in the current generation. What do we expect for p^*_{AB}, the frequency in the next generation? Here I use the notation * to indicate the value of a parameter in the next generation.

The haplotype frequency p^*_{AB} depends on two effects. First, recombination breaks apart AB haplotypes at a rate $c \times p_{AB}$. Second, recombination creates AB haplotypes at a rate $c \times p_A p_B$: this is the probability of randomly assembling an AB haplotype as a result of recombination. So we have

$$p^*_{AB} = p_{AB} - cp_{AB} + cp_A p_B \tag{2.45}$$
$$= (1 - c)p_{AB} + cp_A p_B \tag{2.46}$$

Subtracting $p_A p_B$ from both sides we write this in terms of D:

$$p^*_{AB} - p_A p_B = (1-c)p_{AB} + c p_A p_B - p_A p_B \tag{2.47}$$
$$= (1-c)p_{AB} - (1-c)p_A p_B \tag{2.48}$$
$$D^* = (1-c)D. \tag{2.49}$$

Then, using the same logic over multiple generations, the decay of LD follows

$$D_t = (1-c)^t D_0. \tag{2.50}$$

To summarize, so far we have seen that:
• if there is no recombination, the basic properties of the coalescent genealogy tell us to expect strong LD;
• at large distances (for SNPs more than a few centiMorgans apart, say, and certainly for SNPs on different chromosomes), recombination rapidly eliminates LD.

We now need to explore what happens at intermediate distance scales, between ∼1 kb to 100 kb, where recombination and coalescence compete against each other.

The coalescent with recombination: the ARG. To understand these models, we can incorporate recombination into the coalescent. This produces a more complex structure called an **ancestral recombination graph (ARG)** [180].

To begin, we'll look at two positions in a sequence, labeled L (left) and R (right). Going backward in time, we now have *two* kinds of events: both coalescence and recombination. **As before, coalescence joins lineages, but now recombination can split sequences apart so that each side of a breakpoint becomes a separate lineage.** This is visualized here:

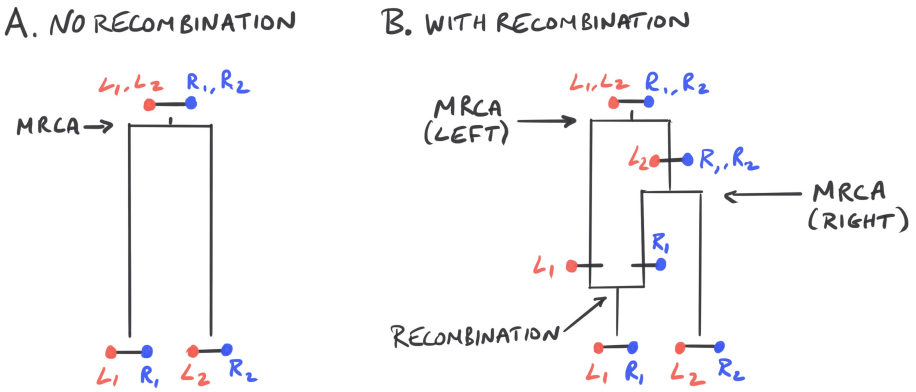


Figure 2.44: **An ARG for two samples.** *L and R indicate the left and right-hand ends of haplotypes 1 and 2.* **A.** *Coalescence without recombination.* **B.** *Going backwards in time Lineage 1 splits due to recombination. L_1 takes the left-hand path, while R_1 takes the right. The two ends of the locus have different MRCAs as indicated.*

In the figure above, you can see in panel B that, going backwards in time, Lineage 1 is split apart by a recombination event so that we have a different coalescent time for the left-hand side of the region (red) versus the right-hand side (blue).

We can extend this idea to consider more samples and more recombina-

tion events. Instead of considering just two sites, we consider a sequence of length c (in units of genetic distance), and recombination can occur anywhere within this region. We indicate the position of a recombination event by a fraction: for example, recombination at 0.3 occurs 30% of the way along the sequence.

In this figure, the full ARG is shown at the left. This contains a series of four so-called "marginal" trees at different positions across the sequence, shown in red, each with a different branching pattern. As you can see, the ARGs can become very complicated:

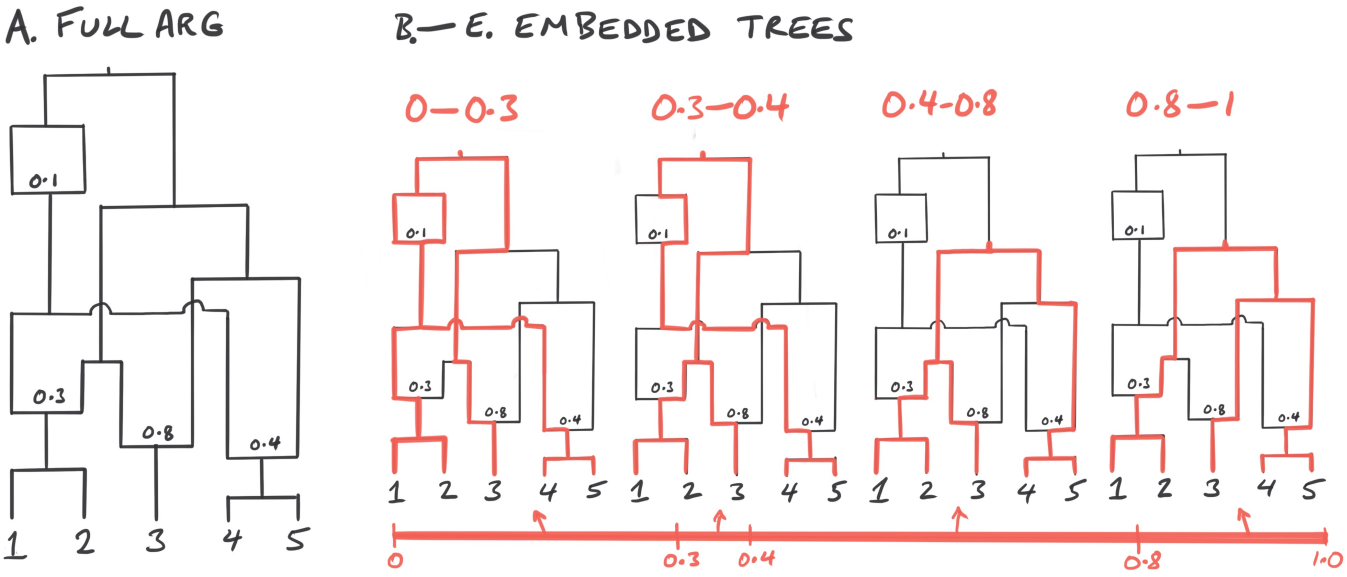


Figure 2.45: ARG and embedded "marginal" trees. *Split points represent recombination events; we use the convention that the left-hand side of the sequence follows the left-hand branch, and the right-hand side follows the right branch. The red numbers at top show which part of the sequence is relevant to each marginal tree: e.g., the first tree covers positions 0–0.3. Note that the recombination at 0.1 does not affect the marginal trees.*

The graph above contains 4 recombination events. These split the region into four distinct blocks, each with a different coalescent tree. However, the trees don't change entirely: haplotypes 1 and 2, as well as 4 and 5, are closely related across the entire region.

I've described all this in terms of the full ARG process, but it's worth noting that the sequence data only depend on the marginal trees at each position (the red trees), and it can be easier to think about the process just in terms of these trees, and the fact that they change as you move along the sequence [e].

Breakdown of LD within the ARG. How does recombination affect haplotype variation? Crucially, **recombination can create mixtures of haplotypes. This is illustrated in the example below, where addition of a single recombination event produces all four possible haplotypes** – remember that this would not possible in the absence of recombination:

[e] *We won't cover inference methods in detail, but in practice the modern inference methods focus on estimating marginal trees rather than the full graph including all recombination events.*

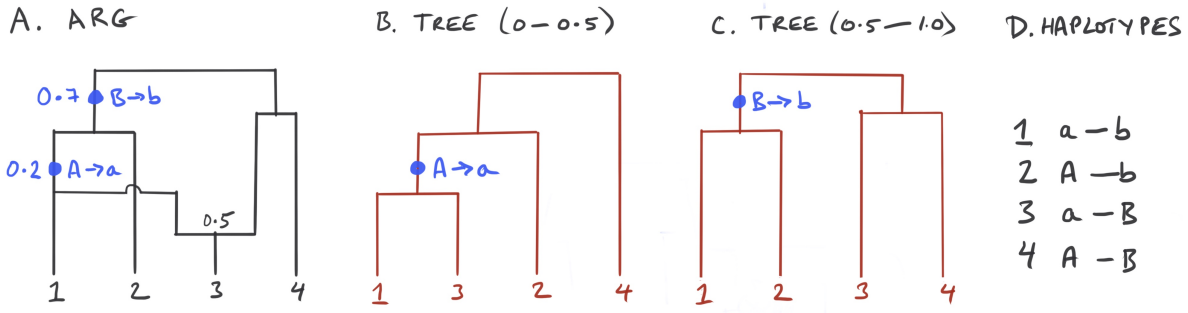


Figure 2.46: Recombination can generate all four haplotypes for two SNPs. *The mutations are at positions 0.2 and 0.7 along the sequence, as indicated; the recombination is at 0.5.*

The tug-of-war between coalescence and recombination. It's difficult to get really deep intuition for properties of the ARG. But I think it's helpful to think about it as a competition between two key processes: the tree-structure of the coalescent creates haplotype structure, while recombination tends to break it apart.

The outcome of this competition is determined by a compound parameter, $4Nc$: the ratio of the rate of recombination (c) to the rate of coalescence ($1/2N$) (and an extra factor of 2, see below). A large value of $4Nc$ basically means there is a high rate of recombination per unit rate of coalescence, so recombination tends to be the winner (and conversely for small $4Nc$).

The next box explains why $4Nc$ is a natural parameter.

Optional: Timescales in the ARG. Let's start with two samples, and a region of length c. What's the probability that these two samples coalesce without recombination?

Going backwards in time, as usual, coalescence occurs at a rate $1/2N$ per generation. Meanwhile, recombination occurs at a rate c per generation (in either lineage), so $2c$ in total. So the probability of at least one recombination before coalescence is

$$\frac{2c}{2c+1/2N} = \frac{4Nc}{4Nc+1}. \tag{2.51}$$

(This result uses a method for "competing exponentials"; don't worry if it's not familiar.)

More generally, take a slice through the ARG at any time. Suppose that we have k lineages. What are the waiting times to the next event (backwards in time, as usual)? Coalescence decreases the number of lineages from k to $k-1$; this occurs at a rate $k(k-1)/4N$ per generation as before. Meanwhile, recombination increases the number of lineages from k to $k+1$; this occurs at a rate ck, where c is the total recombination rate across the segment of interest [181].

So the probability that the next event is a recombination event is

$$\frac{kc}{kc+k(k-1)/4N} = \frac{4Nc}{4Nc+k-1}. \tag{2.52}$$

This formula suggests the following:

- $4Nc$ is the natural parameter to describe the role of recombination in an ARG.

• In large samples, coalescence predominates at recent timescales (when k is large), while recombination is more effective at scrambling the lineages further back in time (when k is small).

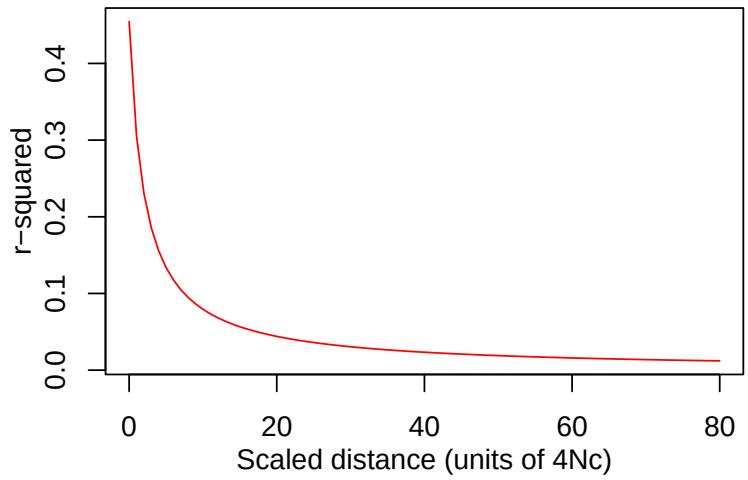
To summarize, consideration of the ARG highlights a few key points:

- sites that are close together tend to share the same genealogy, hence SNPs are in high LD;
- genealogies become less and less correlated with increasing genetic distance, thus reducing LD;
- the scale of LD depends on the product of N and c (usually written as $4Nc$);
- in large samples, the most recent coalescent events occur faster than recombination, so closely related haplotypes can be shared over large recombination distances, even at distances where overall LD is low.

Decay of r^2 with distance. So we have a qualitative prediction that LD should decay with genetic distance. Can we predict this more precisely?

It turns out that the expected r^2 can be approximated as a ratio of covariances in coalescence times among sequences at different distances [f]. I won't present the math for this (it's a bit fiddly) but you can read about it here [Link] [182].

And here's how average r^2 decays as a function of distance:



To give you a sense of scale, on average 100 kb in humans is around $4Nc = 80$; this model predicts that LD should decay to be low within around $10 - 100$kb, which is fairly typical in practice.

Recombination and LD in human data. Most of this basic theory was already understood by the 1980s and 1990s. But for a long time we didn't have the tools to measure this in real data [183].

[f] *Note: these results focus on the typical levels of LD at very short distances with recombination and coalescence; as such it differs from the earlier results predicting rapid decay of D starting from an initial condition of unnaturally high LD.*

Figure 2.47: **Predicted decay of mean r^2** *between pairs of SNPs, as a function of distance. To interpret the x-axis, note that* ***for humans 4Nc=80 corresponds to c=0.1 cM or ∼100 kb*** *at the genome-average recombination rate.* *The function plotted here is $(10+4Nc)/(22+13(4Nc)+(4Nc)^2)$, which approximates the mean of r^2 between common SNPs. See [Link].*

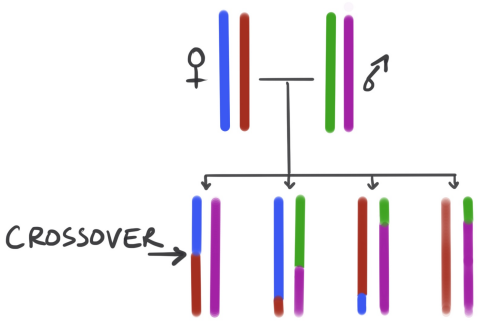


Figure 2.48: **Pedigree studies of recombination.** *Traditional genetic mapping studies used a scaffold of genetic markers to count crossover events within pedigrees – shown here for a single chromosome in parents and four kids.*

This started to change in the 1990s, alongside the Human Genome Project. At that time, one goal was to create genetic and physical maps of the genome. One main approach was to genotype a genome-wide scaffold of genetic markers (STRs) in families, and count recombination events directly. With these data it was possible to estimate recombination rates along each chromosome, as shown here in this recombination map of Chromosome 1, made in 2002:

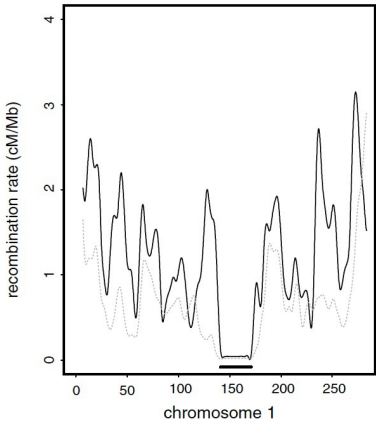


Figure 2.49: **Pedigree-based recombination map for Chromosome 1**. *These data, from a 2002 pedigree study, show estimated* ***recombination rates at megabase scale*** *(females, solid line; males, dotted line).* *The region without recombination at around 150 Mb marks the centromere. Credit: From Figure 2 of Augustine Kong et al (2002) [Link] Used with permission.*

As you can see here, at this scale recombination rates vary from about 1–3 cM/Mb (except for the centromere), and are higher near the telomeres, which is pretty typical of the genome. Female rates (solid line) are generally higher than male rates (dotted), consistent with the fact that genome-wide rates in females are 1.6× the male rates.

But the resolution of this type of map was limited by the number of available STRs (about 2 per Mb), which meant that they could not study fine-scale variation in rates [184].

So when the first high-resolution SNP data came along, it was a big surprise to find that the LD data revealed something much more striking!

But before we get to this, I need to explain a little about how to visualize LD. Let's suppose that we genotype a bunch of SNPs across a region. One thing we could do is to show colored haplotypes in the style of Figure 1.31, but it's hard to get a quantitative sense of the data from this. Instead, a commonly-used approach computes a matrix of r^2 or D' between all pairs of SNPs, and displays it with a color scale, like this:

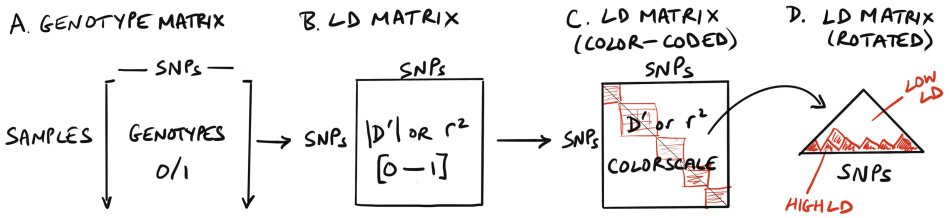


Figure 2.50: **Visualization of LD patterns.** *Panel* **A** *displays the full haplotype data but is difficult to interpret quantitatively.* **B** *Instead it's common to display the data as a matrix of pairwise LD; and often color-coded* **C**. **D** *Finally, the matrix is rotated, and only the top half is shown.*

In panel D above, the SNP pairs that are close together are plotted near the base of the triangle, and SNPs that are far apart are higher up. Thus,

we expect that LD will usually be high (red) near the base, and decrease going up.

By the early 2000s, as it became possible to collect SNP data at higher density, some very interesting patterns started to emerge [185]. Our model above suggests that LD might be expected to decay smoothly with distance, but this is not the case at all. Instead, LD structure forms striking blocks of high LD (so-called **haplotype blocks**), separated with lower LD between blocks. Here's a typical example from a 500 Kb region of the genome:

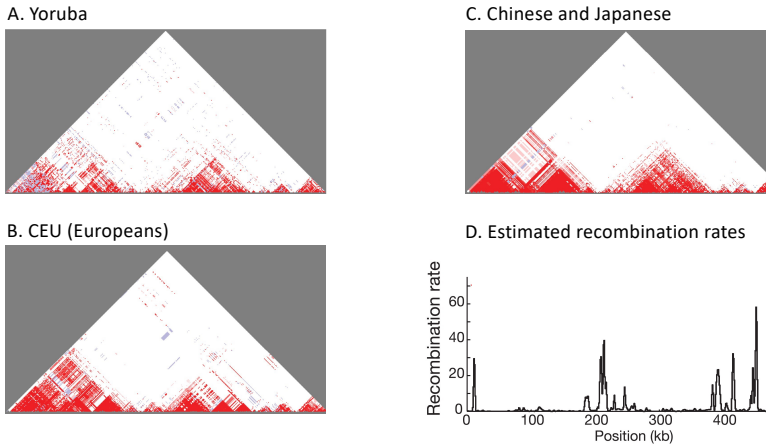


Figure 2.51: **Fine-scale patterns of LD and recombination** *for a 500 Kb region on Chromosome 2, in three population samples (***A–C***). Red entries indicate pairs of sites with* $|D'| = 1$*; white sites indicate* $|D'| < 1$*. Notice that the overall structure is largely shared across populations, but the extent of LD is lowest in the African population (Yoruba) and highest in east Asian population.* **D.** *Estimated recombination rates in cM/Mb across the same region. The peak recombination estimates are much higher than in the pedigree map.* *Credit: Modified from Figure 8 of HapMap (2005) [Link] Used with permission.*

In the plot above, red indicates $|D'| = 1$. Remember that $|D'| < 1$ *(white) indicates that all 4 possible haplotypes are present, and that past recombination must have shuffled genotypes between the two sites*. The blocky structure of the D′ matrices suggests that most recombination is taking place at the boundary points between adjacent blocks.

It's possible to use the LD structure to estimate a fine-scale recombination map (panel D) [186] – this supports the visual impression that most recombination is concentrated into narrow regions with extremely high recombination rates. These locations are referred to as **recombination hotspots**.

These early results have proved to be typical of the genome overall: **the structure of LD tends to form blocks, with generally high LD inside blocks, and lower LD between blocks. This reflects the structure of recombination, which is mainly concentrated into narrow hotspots.**

Genome-wide, more than 30,000 hotspots have been identified, with additional recombination spread among weaker hotspots [187]. **This helps to set the scale of LD, which typically extends around 10–100 Kb, depending on the genomic region** [g].

PRDM9 and the hotspot paradox. The discovery of tens of thousands of recombination hotspots immediately suggested a new question: What controls the locations of hotspots? Work on this question led to a fascinating saga spanning molecular genetics, human genetics, and evolutionary biology.

[g] *The figure above also illustrates another typical pattern, namely that* ***LD is lowest in African populations due to their larger long-term effective population size****.*

The first major progress came in a 2005 paper by Simon Myers and colleagues, which reported that a certain 7-nucleotide sequence motif is highly enriched within hotspots [188]. The presence of this short DNA motif at many hotspots suggested that the locations of hotspots are, at least in part, directed by local DNA sequences [h]. This situation is reminiscent of binding sequences for transcription factors, and suggested that recombination events might be directed by an unknown DNA-binding protein that recognizes this motif.

[h] *Terminology: DNA sequences that control local activity are said to act **in cis**, while external factors such as a DNA-binding protein that recognizes those sequences are said to act **in trans**.*

But this exciting observation immediately raised a theoretical problem known as the **hotspot paradox** [189]. The hotspot paradox argues that, due to the molecular details of recombination, evolution should tend to remove cis-acting hotspot motifs.

To explain this, I need to say a bit about what happens during recombination. During meiosis, the homologous chromosomes pair up. Crossovers are initiated by one of the two homologs (the blue one, in the example below). The initiating chromosome undergoes a double-strand break, and part of the chromosome is chewed back in both directions around the break. Eventually, this damaged region is repaired using the other (red) chromosome as a template, in a process known as **gene conversion**:

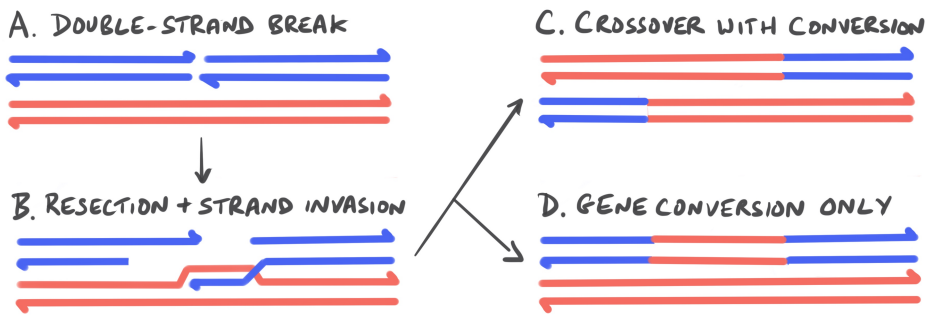


Figure 2.52: **Simplified model of recombination with gene conversion**. **A.** *Recombination is initiated with a double-strand break in the blue chromosome.* **B.** *Several hundred bp around the DSB are resected (chewed away) on the blue chromosome; then one strand invades the red chromosome. This is resolved in one of two endpoints:* **C.** *crossover with gene conversion to repair the damaged section using the red chromosome as template, or* **D.** *gene conversion without crossover. Note: this is a simplified account of a complex process. Figure modified from [Link].*

The key point here is that, within the gene conversion region, it is the initiating chromosome (blue) that is copied over by its partner (red). Both resulting chromosomes end up with the red sequence inside the converted region.

Now, let's suppose that one chromosome carries a hotspot motif but the other does not (for example there could be a SNP for which one allele breaks the motif). Then, the chromosome with the motif can initiate the crossover. But that sequence would then be replaced by gene conversion from the other non-motif chromosome. This is known as **biased gene conversion**:

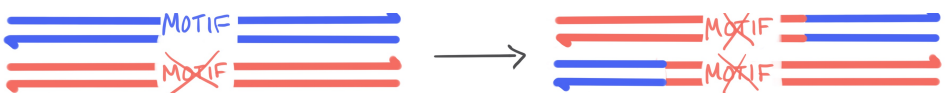


Figure 2.53: **Biased gene conversion**. *An allele that encodes the hotspot motif (blue) will tend to be replaced by an alternative allele that breaks the motif (red).*

In other words, *biased gene conversion tends to remove hotpots!* We haven't

covered selection yet, but *this is mathematically equivalent to a form of selection in favor of alleles that remove hotspots* [190]!

Based on this logic, the hotspot paradox argues that any time a SNP arises inside a hotspot motif, it will tend to spread through the population as if it were positively selected. *Over time, this should tend to eliminate all hotspots. So why are there any hotspots left?*

Around the same time as discovery of the hotspot motif, another intriguing observation was made by comparing LD in humans and chimpanzees.

Recall that chimpanzees are our closest living relatives, and that our genome sequences are extremely similar, differing at only about 1.4% of sites. Given this, you might naively expect most hotspots to be shared if they are controlled by cis-acting motifs. But, remarkably, studies of LD in chimpanzees found no meaningful overlap of hotspot locations between humans and chimps beyond random expectations [191]:

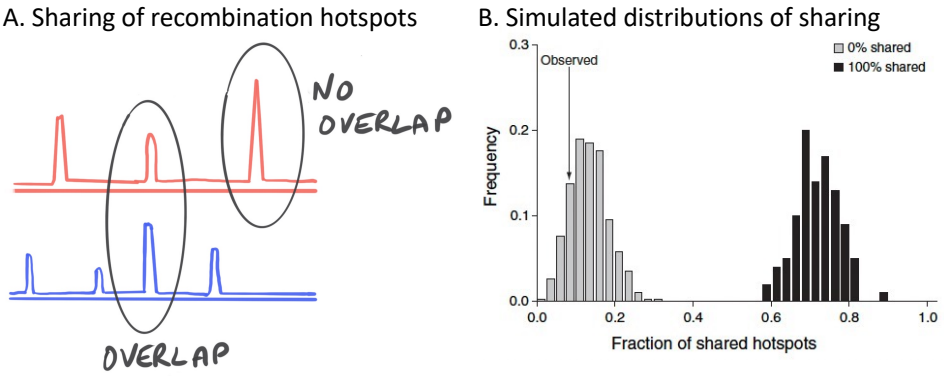


Figure 2.54: **No sharing of hotspots between humans and chimpanzees**. **A.** *One study estimated that just 8% of hotspots overlap between humans and chimpanzees.* **B.** *Simulations showed that this is consistent with no true sharing of hotspots (since hotspot regions are estimated imprecisely, some overlap is expected even by chance).* *Credit: Panel B is Figure 2 from Susan Ptak et al (2005), [Link] Used with permission.*

And an independent study of recombination events in pedigrees in a European-American population showed that, even within humans, not everyone uses the same hotspots at the same rates [192]. All this was very intriguing. If hotspot locations are controlled by local sequences, then shouldn't most hotspots be shared?

Many of these questions started to be resolved by a set of papers in 2010 that identified a gene called PRDM9 as the missing, central, player in this entire saga [193]. PRDM9 encodes a protein with a so-called "zinc finger" domain that is responsible for DNA binding.

The zinc finger domain has a specific affinity to – you guessed it – the previously-discovered hotspot motif. The plot below shows DNA binding predictions from a 2010 paper, based on the protein sequence of the most common European PRDM9 allele. This substantially matches the DNA sequence enriched within hotspots:

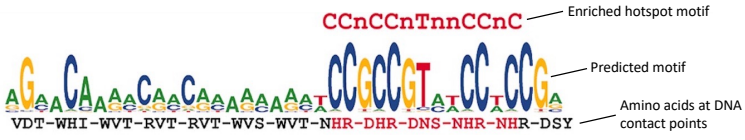


Figure 2.55: **Predicted binding preferences of the PRM9 'A' allele.** *The sequence motif in red at top represents a consensus of the motif enriched at hotspots ('n' indicates no clear consensus). The DNA 'logo' plot shows predicted binding preferences of the PRDM9 protein based on the corresponding zinc fingers; the sizes of the letters reflect the predicted strength of preference for each nucleotide.* *Modified from Figure 2 of Baudat et al (2010) [Link] Used with permission.*

Once PRDM9 binds to the DNA, it recruits additional machinery to

initiate double-strand breaks (which can result in crossovers).

Of particular interest, the zinc finger DNA-binding domain is encoded within a minisatellite repeat section of the gene [i]. Each repeat (or "finger") consists of 28 amino acids (84 bp), of which 4 amino acids touch the DNA and provide binding specificity. Most of the other 24 amino acids are identical across repeats.

Recall that the copy number in such regions is often highly variable due to mispairing during DNA replication. In fact, this is the case at PRDM9, where dozens of alleles have been found in humans. Furthermore, the alleles frequently differ specifically in the amino acids that contact the DNA. The image below shows differences between two of the most common human alleles, A and C, as well as allelic variation across nine different human alleles:

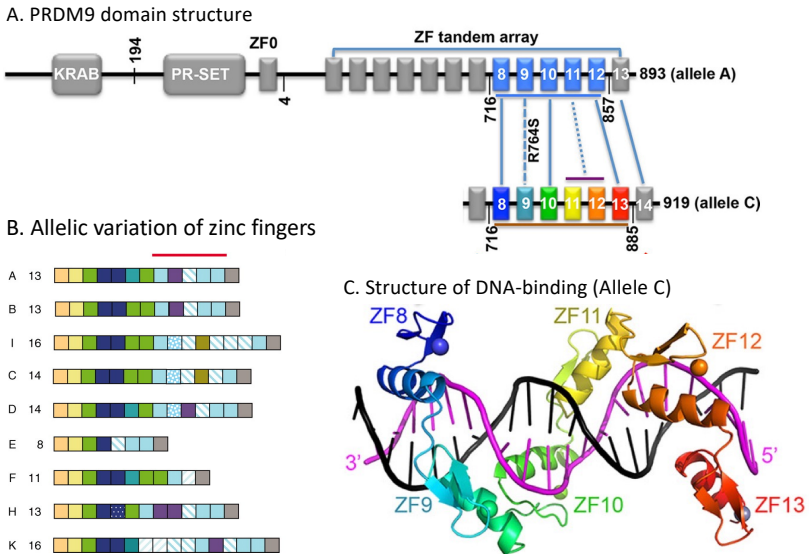


Figure 2.56: **Structure of PRDM9**. **A.** *Domains of the PRDM9 gene. Zinc fingers 8–13 are responsible for DNA binding and differ between Alleles A and C.* **B.** *Diversity of zinc finger structure across nine human PRDM9 alleles. Each box represents a single zinc finger, and colors indicate distinct sequences. Notice that alleles differ both in the numbers of fingers as well as their sequences, especially within the main DNA binding region (red bar).* **C.** *DNA binding structure of the C allele. Fingers 8-13 are responsible for DNA sequence recognition.* *Panel A and C, modified Figure 1 of Patel et al (2017) [Link], CC BY 4; Panel B part of Figure 2 of Baudat et al (2010) [Link] Used with permission.*

[i] *Minisatellites are similar to STRs, but with longer repeat units. They tend to be highly variable due to replication slippage, and are also often referred to as VNTRs: Chapter 1.3; Figure 1.34.*

Importantly, in many cases, the different alleles have different DNA binding preferences. For example, allele C, which is common (36%) in west Africa but rare outside Africa, uses completely different hotspots than allele A (plot at right) [194].

Meanwhile, chimpanzees also have completely different PRDM9 alleles from humans – thus neatly explaining the complete lack of overlap between the human and chimpanzee hotspot maps.

Lastly, the rapid evolution of PRDM9 neatly resolves the hotspot paradox. There has indeed been systematic loss of human hotspots during recent human evolution [195], but this is counteracted by regular jumps in PRDM9 binding preferences due to the evolution of new alleles [196].

As a consequence of all this, the selective pressure imposed by hotspot evolution has made PRDM9 one of the most rapidly evolving vertebrate genes – and a fascinating story involving molecular biology, population genetics, and evolutionary biology.

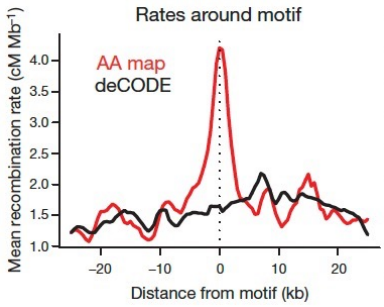


Figure 2.57: **Population differences in hotspot usage at C allele binding motifs.** *Recombination rates averaged across all instances of the C allele binding motif in African Americans (red) and Europeans (black). (Very few other genes exhibit strong functional differences across human populations; in this case it reflects the unique evolutionary pressures acting on PRDM9.)* *From Figure 3 of Anjali Hinch et al (2011) [Link] Used with permission.*

In the last part of this chapter we return to models of LD, but now with a different flavor. The new model is slightly heuristic, but much easier to work with in data analysis – and hopefully more intuitive.

Haplotype copying models. While the ARG can be considered an "exact" representation of chromosome ancestry [197], its complexity makes it extremely difficult to use in statistical analysis [198].

But in a landmark 2003 paper, Na Li and Matthew Stephens introduced an alternative framework known as a **haplotype copying model** [199] that approximates key elements of the ARG process, while being much simpler and far more computationally tractable [200]. This model has inspired many methods for a variety of important problems [201].

The central concept of the copying model is to define a *conditional sampling probability* for the "next" haplotype in a sample. Suppose that you have already observed K haplotypes in some region of the genome (by either sequencing or genotyping). You then sequence one more haplotype: before you look at the data, what would you expect this next haplotype to look like?

Intuitively, within a small region of DNA sequence *we should expect the next haplotype to be similar to (i.e., "copy") one we have already seen, but might not be identical due to occasional mutations*.

Secondly, over a larger region, we might expect that *the next haplotype will first copy one haplotype, and then switch to copy a different one, reflecting past recombination events*.

A third key point is that if we have a very large reference panel (large K), then it's more likely that the next haplotype will be similar to something we have already seen, compared to if we were using a small reference panel. So *the rate of both switches and mutations should decrease with K*.

One possible outcome from this process is illustrated here:

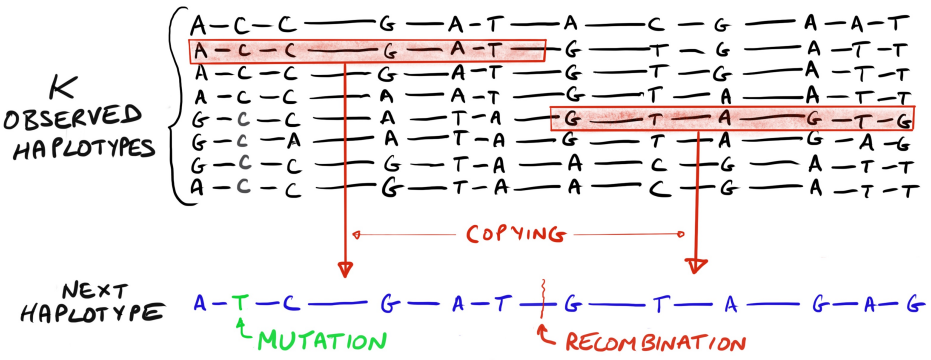


Figure 2.58: **The Haplotype Copying Model** *defines a probability distribution for the next haplotype, modeling it as a mosaic of haplotypes that have already been observed. It allows for occasional differences due to mutation or errors, as well as switches due to past recombination events.*

More formally, these ideas suggest that we could define a **conditional sampling probability**. This allows us to compute the probability of observing any specific sequence of variants as the next haplotype. Under this model, haplotypes that can be generated as simple mosaics of the previous haplotypes are more likely [202].

This model can easily handle recombination hotspots, simply by allowing a higher switch rate any time the copying process passes over a hotspot. Conversely, the model can be used to detect hotspots, as locations where haplotypes often switch parents.

The upcoming box provides some technical detail on the copying process:

Optional: The Conditional Sampling Probability for Haplotype Copying. We define the conditional sampling probability for the next haplotype as follows, assuming a reference panel of $K \geq 1$ haplotypes observed so far. For motivation and details see Li and Stephens (2003).

We focus on genotype data at a set of S SNPs, where the SNP number s ranges from 1 and S. This process defines what is known as a Markov process for the $k+1$ haplotype, conditional on the K haplotypes observed so far:

Initial parent. At the first variant position, s=1, pick a reference panel haplotype k between 1 and K, at random. (We will start copying from this haplotype.)

Next, repeat the following until $s = S$:

- **Determining the allele value.** When copying from haplotype k, we usually copy the allele in haplotype k but we allow for a low probability of single nucleotide mismatches due to mutation (or other events such as genotyping errors or gene conversion). Specifically, at site s, with probability $1-[\theta/(K+\theta)]$ the allele in the new haplotype is set to equal the allele at site s in haplotype k; otherwise we set it to an alternate allele. Here θ reflects the rate of mutations or mismatches in the data [203].
- **Recombination.** When we move from SNP s to SNP $s+1$, we decide whether to switch to copying a different haplotype [204]. Let c_s be the expected number of crossovers between these two SNPs, per generation. With probability $e^{-4Nc_s/K}$ we continue copying from the current haplotype. Otherwise, with probability $1-e^{-4Nc_s/K}$ we introduce a recombination event: in that case we select a new random haplotype parent k' between 1 and K.
- **Increment SNP position.** Set s to $s+1$.

The expression for the switch rate is motivated by noting that the average coalescence time for a new haplotype into an existing panel of K samples is $\sim 2N/K$, so the expected number of recombination events along either branch between the two SNPs is $\sim 4Nc/K$, and the probability of zero recombinations is $e^{-4Nc/K}$.

Notice that here c is measured in units of genetic distance, so it naturally allows for a higher jump rate across hotspots.

One huge advantage of copying models is that they are highly tractable for computational analysis. For example, unlike the ARG, they are amenable to efficient tools for data analysis called Hidden Markov Models (HMMs). The details of HMMs are outside our scope [205], but I'll briefly outline one major application of copying models:

Phasing and imputation. Recall that one of the main ways of collecting genome data on individuals is by genotyping. In genotyping, we measure the genotype of an individual at a pre-specified set of SNPs (com-

monly ~1 million SNPs). Until recently, genotype data has been much cheaper than genome sequencing [j]. However, these data are incomplete in two key ways:

- We do not know the genotype at SNPs that were not on the array;
- We do not know **haplotype phase**: i.e., for heterozygous SNPs, we do not know which allele goes on which homolog.

However, *we can use the concepts of LD and haplotype structure to fill in the missing data*. This is referred to as **phasing** – inferring which heterozygous alleles are from the same homologs; and **imputation** – for inferring genotypes at SNPs that were not on the array. Imputed data are valuable for many purposes as they allow us to approximate whole genome sequencing data at a fraction of the cost. Phased data are needed any time we want to analyze haplotypes and, in any event, most imputation algorithms work by phasing the data simultaneously, as I'll discuss below.

Most applications of phasing/imputation build off a panel of known haplotypes, such as data from the 1000 Genomes Project [206] to enable phasing and imputation in a new sample, as shown here:

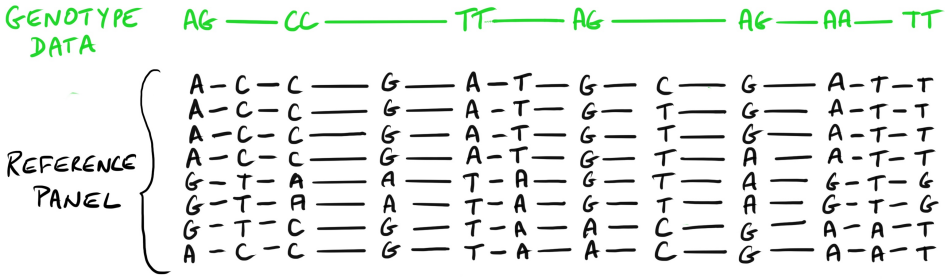


Figure 2.59: **Imputation and Phasing.** *It is common to collect genotype data on a subset of SNPs (green). With the help of a reference panel of known haplotypes (black) we wish to infer haplotype phase and impute the missing genotypes at unmeasured SNPs.*

Under the copying model, we can view the data in a diploid individual as coming from two unknown paths threading through the reference panel. The HMM machinery allows us to identify likely paths and, from this, to infer phase and missing genotypes [207]:

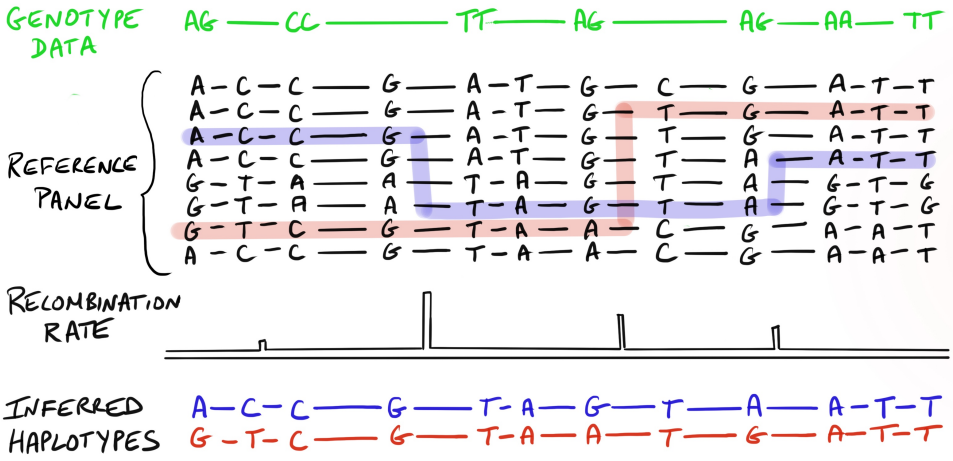


Figure 2.60: **Likely haplotypes inferred from genotype data.** *For a diploid individual, the genotype data result from two independent copying paths through the reference panel. The algorithm finds likely pairs of paths (red and blue) consistent with the genotype data; switch rates are higher at recombination hotspots. There may be multiple likely paths. Once likely paths have been identified we can infer phase and impute variants at ungenotyped SNPs (bottom).*

This type of approach provides the basic structure for how we can phase and impute data from genotypes. While more advanced techniques in-

[j] *These issues also come up for sequencing: in particular, traditional short-read sequencing does not determine haplotype phase.*

clude various bells and whistles and speedups, this type of idea has been used to analyze data from tens of millions of people.

In this chapter we've talked about how linkage, recombination and drift shape patterns of genetic variation in the genome, including LD. These processes are fundamental to understanding other aspects of human variation including natural selection and disease genetics.

Notes and References.

[174]For a short but fascinating history of Kreitman's seminal paper, see Casey Bergman's blogpost here: [Link]. The paper itself is:

Kreitman M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature. 1983;304(5925):412-7

[175]The terms recombination and crossover are often used interchangeably in the human genetics literature; however many recombination events result in local exchange of material (known as gene conversion) without crossing over. The non-crossover events are difficult to detect from genetic variation data.

[176]Genetic distances (cM) are defined in terms of the expected number of crossovers. This is a sensible way to define the distances so that they add together in a sensible way. However in a lot of practical contexts we actually want the probability of ≥ 1 crossovers. Luckily for short distances – up to about 10 cM, say – these are almost exactly the same (since double crossovers are unlikely) and we can ignore the distinction.

[177]Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 2019;363(6425):eaau1043

[178]Measures of LD and significance of r^2 for tag SNPs:

Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics. 2001;69(1):1-14;

LD scores and LD score regression:

Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, of the Psychiatric Genomics Consortium SWG, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. 2015;47(3):291-5.

[179]We define c as the probability that the two alleles passed into a gamete both came from the same parent (i.e., both from the mother, or both from the father). This has the result that the maximum of c is 0.5 (and not 1 as might seem intuitive). Suppose that two SNPs are on different chromosomes, then they are transmitted independently, as predicted from Mendel's laws. In these cases the pairing of alleles is like a coin toss, so c reaches its maximum, $c = 0.5$. This is also true for SNPs on opposite ends of the same chromosome, though it is less obvious as it depends on the mechanics of chromatid pairing in meiosis.

[180]The ARG was first developed (but not really described as such) by Richard Hudson

Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 1983;23(2):183-201

A short but clear description of the ARG is presented by Nordborg 2001 [Link].

[181]Thus the number of lineages, k, forms a Markov chain over time. Since the rate of increases in linear in k, and the rate of decreases is quadratic in k, this will eventually converge to a single ancestor, known at the Ultimate Ancestor (UA). Since the UA likely predates the marginal MRCAs everywhere in the sequence, this is of mathematical but not practical interest.

[182]McVean GA. A genealogical interpretation of linkage disequilibrium. Genetics. 2002;162(2):987-91

[183]For a review of the state of the art in 2001 see Pritchard and Przeworski 2001, cited above.

[184]Pedigree studies are also greatly limited by the number of families analyzed. In this case, the authors measured recombination in 1257 meioses, or in other words, an average of 12 recombination events per cM. This means that they could get adequate estimates at Mb scale, but even with more markers they would not have been able to get a higher resolution map. In general, LD-based maps have higher resolution because they average over many more meioses (i.e., past meioses in the history of population) compared to pedigree-based maps.

[185]I'm slightly oversimplifying the historical narrative here. A few early papers suggested the presence of specific recombination hotspots based on LD data, starting as early as 1984:

Chakravarti A, Buetow K, Antonarakis S, Waber P, Boehm C, Kazazian H. Nonuniform recombination within the human beta-globin gene cluster. American Journal of Human Genetics. 1984;36(6):1239. Meanwhile, Alec Jeffreys (most famous for inventing DNA fingerprinting) and colleagues provided compelling experimental evidence for a small number of hotspots in a series of papers around 2000:

Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics. 2001;29(2):217-22

But the fact that LD patterns are mostly dictated by hotspot locations was not fully evident until a series of papers in 2001-2005.

[186]Later in the chapter I'll give some intuition for one method to estimate this, based on the Li and Stephens model.

These plots used a different approach based on McVean 2002 (cited above)

[187] McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581-4

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005;310(5746):321-4.

[188] Myers et al (2005), cited above. The originally-reported motif was CCTCCCT, although this is modified in later papers. Myers 2006.

[189] This paradox was first pointed out by Rosie Redfield and colleagues in a 1997 paper, motivated by observations from yeast.

Boulton A, Myers RS, Redfield RJ. The hotspot conversion paradox and the evolution of meiotic recombination. Proceedings of the National Academy of Sciences. 1997;94(15):8058-63

[190] Hotspot selection reference

[191] Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, et al. Fine-scale recombination patterns differ between chimpanzees and humans. Nature Genetics. 2005;37(4):429-34

Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, et al. Comparison of fine-scale recombination rates in humans and chimpanzees. Science. 2005;308(5718):107-11

[192] Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. science. 2008;319(5868):1395-8

Note: to be fair to these earlier papers, several of them invoked the possibility of an unknown trans-acting factor that might be variable within or between species, thereby explaining both varied hotspot use and a solution to the hotspot paradox. For example, Coop et al noted that "A single change in the recombination machinery could create many new hotspots in the genome, counteracting the removal of individual hotspots from the population by biased gene conversion".

[193] Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 2010;327(5967):836-40,

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science. 2010;327(5967):876-9,

Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. Science. 2010;327(5967):835-5,

Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, May CA, et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature Genetics. 2010;42(10):859-63

[194] Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, et al. The landscape of recombination in African Americans. Nature. 2011;476(7359):170-5

[195] Myers et al (2010).

[196] Recent work suggests that PRDM9 has to bind the same hotspots on both homologs for efficient crossover. For this reason, it's particularly bad to lose the *hottest* hotspots, as these are the ones most likely to have double binding. Moreover, these sites are precisely the ones that are lost most rapidly through biased gene conversion. For more on this model see

Baker Z, Przeworski M, Sella G. Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence. bioRxiv. 2022:2022-09.

[197] The ARG is "exact" in the sense that if we make a bunch of assumptions – a version of WF dynamics, a mutation, and recombination model – then it's possible to derive the ARG. But of course, any mathematical model of the world is an approximation of a more-complex reality, so you can think of the ARG as corresponding exactly to our best (but approximate) model of population genetics.

[198] There are infinitely many ARGs that can produce any given data set, and it's very difficult to compute, or even approximate, basic statistical quantities such as the likelihood.

[199] Elsewhere in the literature this model is also referred to as *Li and Stephens* or, following the original paper, the *PAC-likelihood* (for "product of approximate conditionals").

[200] Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213-33

[201] Perspective piece by Yun Song:

Song YS. Na Li and Matthew Stephens on modeling linkage disequilibrium. Genetics. 2016;203(3):1005-6.

[202]The Copying model can be thought of as a **generative** model: i.e., a specific model for the evolutionary process that generates the data. In this way it is analogous to the ARG, which is also a generative model but far more complicated.

[203]The modeling for θ is a bit complicated. The notation is motivated by the tradition definition of θ in population genetics $4N_e\mu$. But here, the expression is intended as a slightly heuristic model of the mismatch probability, and may depend on the nature of the data. For example, if we are looking at ascertained SNPs, we do know that there should be at least 1 mutation per site, somewhere within the observed genealogy, and Li and Stephens suggest scaling θ by the expected genealogy length. Furthermore, θ here is implicitly doing some extra work: it should also be able to incorporate sequencing errors, gene conversions, and other types of deviations from the copying model. You can read more about this in Li and Stephens (2003).

[204]We do this only if $s < S$

[205]HMMs are beyond the scope of this book but some googling will lead you to plenty of tutorials of different flavors, eg [Link].

[206]1000 Genomes Project: [Link];
Haplotype Reference Consortium Consortium" THR. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics. 2016;48(10):1279-83

[207]For already-phased haplotypes, the run-time is proportional to the size of the reference panel K. If we need to perform phasing at the same time, then each individual traces two paths through the reference panel, and the run-time is proportional to K^2. In practice this gets rather slow for large panels. Consequently, there has been a great deal of methods development that uses these (or similar) ideas to develop much faster algorithms.

2.4 Genetic drift in structured populations.

So far, our models have ignored population structure. But of course, individuals do not choose their reproductive partners at random from the entire world's population. This nonrandom mating is referred to as ***population structure****, and over time it leads to differences in allele frequencies.*

Human populations are structured at all levels: between continents and geographic regions, and often between nearby ethnic groups, towns or villages. Here we discuss basic models of structure; we'll revisit these themes in Section 3 with a specific focus on human history.

Humans share a recent African origin. Spoiler Alert! We first need the briefest overview of human genetic history to set the stage.

Humans are descended from populations in sub-Saharan Africa. Around 80,000 years ago part of this population spread out of Africa, and eventually colonized most of the world's land masses. As a result of our shared ancestry, all human populations share much of our genetic variation.

Here's a schematic overview of the relationships among human populations; see Section 3 for more about this topic:

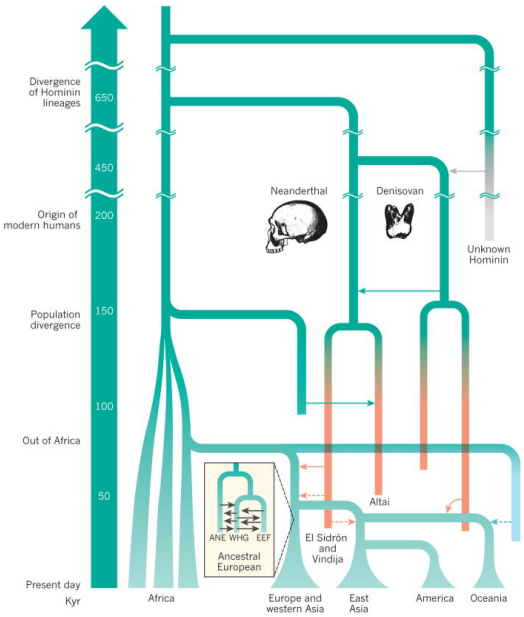


Figure 2.61: **Schematic overview of relationships among human populations.** *Most human populations descend from an ancestral population in sub-Saharan Africa. Non-African populations also briefly contacted archaic humans (Neanderthals and Denisovans) when they reached Eurasia.* *This overview is highly simplified: for example, there has been frequent migration among groups, and many populations have mixed ancestry across branches. Time estimates are approximate. Credit: Figure 2 from Rasmus Nielsen et al (2017) [Link].*

As we shall see in this chapter, the separation time of human populations is actually quite recent in terms of population genetic timescales so that most common genetic variants are shared among all human populations.

Allele frequency variation across populations. As we'll discuss in this chapter, population structure (non-random mating) allows alleles and haplotypes to drift independently in different populations. This leads to differences in allele and haplotype frequencies across populations.

To give you a sense of what this looks like, the next plot shows the allele frequencies in different human populations for a single common SNP. As is typical for intermediate-frequency SNPs, both alleles are present in all sampled populations, but at varying frequencies:

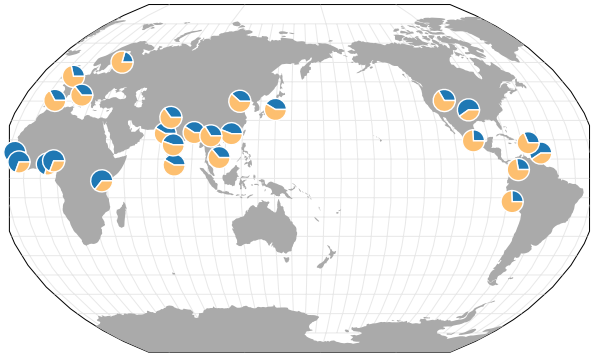


Figure 2.62: **Population allele frequencies at an arbitrary common SNP.** *Each pie chart shows allele frequencies for a 1000 Genomes population sampled at that location. The blue allele at this SNP is ancestral, and yellow derived.* *Credit: Plot made using the Geography of Genetic Variants browser: [Link]. SNP: rs7148516, Blue: A; Yellow: T. A is likely ancestral. Note that the populations plotted in the Americas are not primarily native populations.*

And here's a different visualization, showing allele frequencies for 100 randomly chosen SNPs from the 1000 Genomes Project data [208]. SNPs are sorted by global allele frequency (highest frequencies at the top) and populations from the same continent are show in adjacent columns:

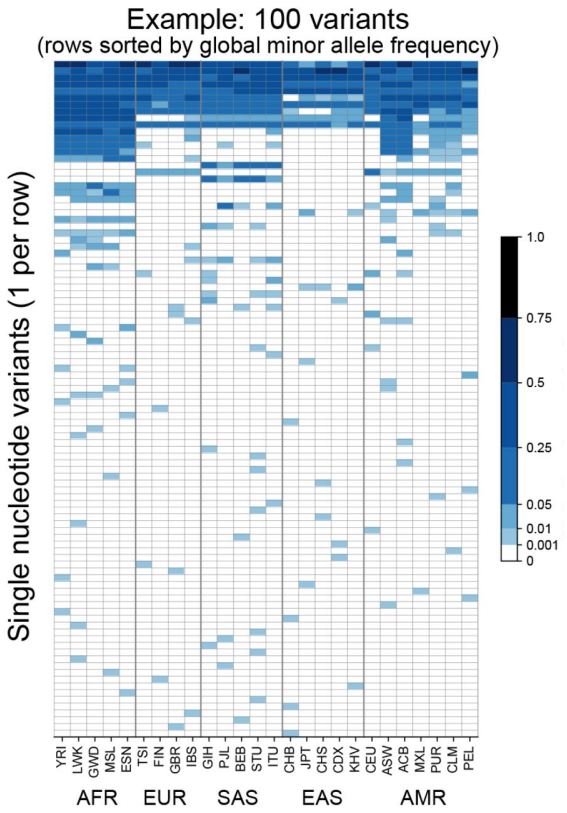


Figure 2.63: **Geographic distribution of 100 random SNPs**. *Rows are Single Nucleotide Variants, columns are populations, grouped into continental groups: Africa; Europe; South Asia; East Asia; Americas. White boxes mean that the derived allele is absent from a particular population sample;* ***the blue color scale indicates allele frequency in each population where the allele is present.*** *Credit: Figure 1 from Arjun Biddanda et al (2020) [Link]. CC-BY 4.0.*

The plots above suggest two key features of the data:

(1) Alleles that are common in one population tend to be common everywhere – notice the solid blue rows at the top of the plot that indicate SNPs that are segregating in most, or all, populations. As we will explain

shortly, this pattern arises because most common alleles arose in sub-Saharan Africa before the human diaspora and were carried everywhere as humans spread around the globe.

(2) Alleles that are rare are usually restricted to a single population or continent – lower down in the plot, the blue bars in each row are usually found in just one or a few populations. This pattern occurs because most rare alleles arose much more recently, after the separation of human populations, and are only found within the populations where they first occurred (or were carried by later migrants).

In the remainder of the chapter we'll discuss models for genetic drift with population structure to try to understand these observations.

Models of population separation and drift. To start thinking about models for allele frequency variation, consider the allele frequencies in two populations. For example, we might compare a pair of closely related populations such as Japanese and Korean; or more distantly related populations such as Japanese and Yoruba (from Nigeria). In each case, we can ask questions such as:

- How do allele frequencies differ between these pairs of populations?
- Are two samples from the same population more closely related (in a coalescent sense) than samples from different populations?

The most basic model for thinking about this is to consider a pair of populations that separated T generations ago from an ancestral population, as you can see in the sketch to the right. We'll start by assuming no migration between the two populations after the split (we'll introduce migration shortly).

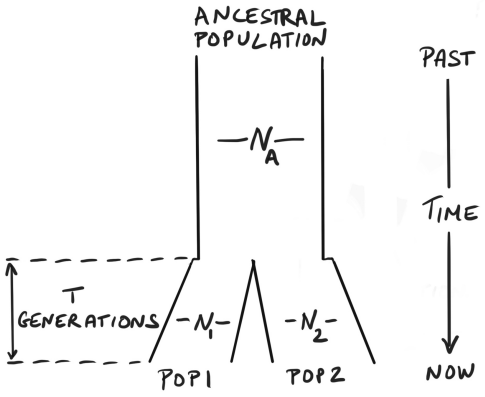


Figure 2.64: **A basic population-split model:** *An ancestral population of size N_A split at time T generations before the present, instantaneously creating two descendant populations, of sizes N_1 and N_2.*

We've discussed two different approaches to understanding drift: the Wright-Fisher forward-in-time approach, and the coalescent approach. Let's use each of these in turn to understand what happens after the population split.

First, in the forward-in-time framework, consider an allele that drifts in the ancestral population to a frequency p_A at the time of the split. Immediately following the split, it is at p_A in each of the descendant populations, *but after that it drifts independently in each population* [a].

This is illustrated in the plot below, which uses the Wright-Fisher model to simulate drift of a single allele with $N = 10,000$: first in an ancestral population (black), and then in two descendant populations (red and blue):

[a] *If Wright-Fisher drift is like a drunk man stumbling aimlessly between 0 and 1, this is now like two drunk men stumbling independently from the same starting position. The allele frequency difference between two populations is analogous to how far apart they get after T steps.*

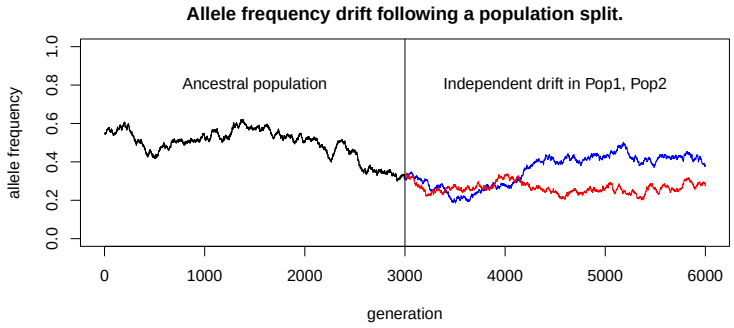


Figure 2.65: **Simulated drift of a single variant.** *Drift in the ancestral population is in black, and drift in the descendant populations is blue and red. Here time is measured forward from left to right; the amount of time plotted after the population split (3000 generations) is similar to the divergence of African and non-African populations.* $N_A = N_1 = N_2 = 10,000$.

That's just one random outcome from this process: what is the overall distribution of allele frequency differences under this model?

There is not a simple, exact mathematical formula for this, but we can get useful insight using the **Nicholson-Donnelly approximation** [209]. This provides a simple model for the present day frequency of an allele in a population (denoted p_T), given that the ancestral frequency was p_A at a time T generations before the present, assuming effective population size N. Nicholson *et al.* suggested that we can approximate this using a normal distribution [210]:

$$p_T \sim \text{Normal}(p_A, \frac{T}{2N} p_A(1-p_A)). \tag{2.53}$$

If p_T falls outside the range $[0,1]$ we think of this as equivalent to loss or fixation of the allele and set the frequency to 0 or 1.

Equation 2.53 may look complicated but is actually pretty intuitive. First, the mean of p_T is simply equal to the starting frequency p_A, since we're assuming no selection.

Meanwhile, the variance of the distribution is $T \cdot p_A(1-p_A)/2N$; this uses an approximation that the variance across T generations is simply T times the WF sampling variance $p_A(1-p_A)/2N$ per generation.

Here's what the model looks for an ancestral allele frequency of 0.55:

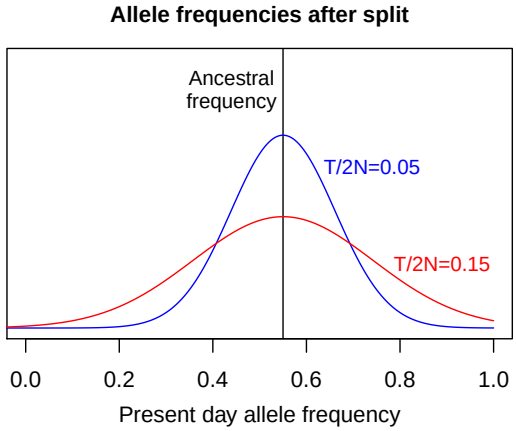


Figure 2.66: **Genetic drift after a population split.** *The plot shows the distributions of possible allele frequencies in two populations of different sizes, both starting from an ancestral frequency of* 0.55. *The population shown in red has larger* $T/2N$ *and shows more drift from the ancestral frequency. The red line approximates the amount of drift in non-African populations since the out-of-Africa migration.*

Example: Tibetans and Han. This basic model predicts the drift of each population from an ancestral population. But in practice we don't know

the ancestral allele frequencies, so we infer drift by comparing frequencies in different modern populations.

One example of this is shown in the plot below, which compares allele frequencies between 50 Tibetans and 40 Han Chinese for about 100,000 SNPs [211]. As you can see, **the allele frequencies are generally close to the diagonal, implying that frequencies are similar in the two populations.** Indeed, about half the scatter around the line actually comes from the limited sample sizes rather than from drift alone (the standard deviations of allele frequency estimates are up to 5% at these sample sizes).

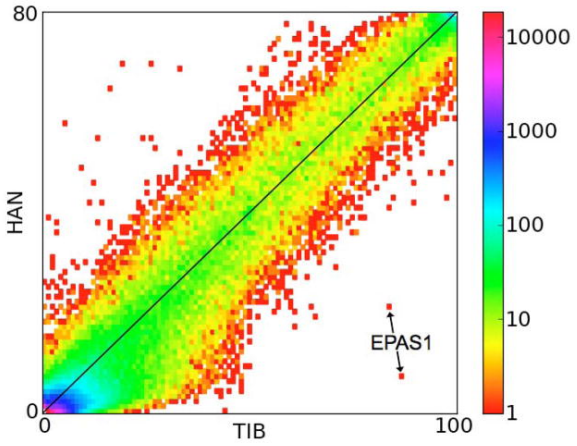


Figure 2.67: **SNP allele frequencies in Tibetans and Han Chinese.** *The plot shows Tibetan and Han allele frequencies for ~100K exonic SNPs. Colors indicate the density of points; notice the high color density along the main diagonal indicating that the vast majority of SNPs have very similar frequencies in the two populations.* *Credit: Figure 1 from Xin Ye et al 2010. [Link] Used with permission.*

While allele frequencies for most SNPs are very similar in the two populations, you'll notice that two SNPs in the **EPAS1 gene** are notable outliers in Tibetans. These outliers were the first indication of a remarkable evolutionary story.

Tibetans, of course, live at high elevations in the Himalayas, and it turns out that the EPAS1 SNPs tag a haplotype involved in local adaptation of Tibetans to altitude. EPAS1 is a transcription factor that plays a central role in regulating red blood cell production, and the haplotype that is common in Tibet increases fitness at high altitude. Natural selection has driven this haplotype to high frequency in Tibet, thus causing it to be an outlier against the genome-wide background of genetic drift [212].

A coalescent interpretation of population splits. So far we have been thinking about drift of allele frequencies forward in time, but it's also helpful to think about how population structure affects coalescence of samples. As before, we'll assume the basic split model shown in Figure 2.64, and to keep things simple, we'll assume that the effective population size is simply N at all times ($N_A = N_1 = N_2 = N$).

Consider two samples: either both from the same population, or both from different populations. When they both come from the same population, they are eligible to start coalescing immediately and, as before, the average coalescence time is simply $2N$ generations (Panels A and B, below).

But if the samples come from different populations, they cannot coalesce during the first T generations (looking backwards in time) until the lineages merge back into the ancestral population (Panel C). At that point, the usual coalescent process starts. Hence, for 2 samples from different populations, the average coalescence time is $2N + T$:

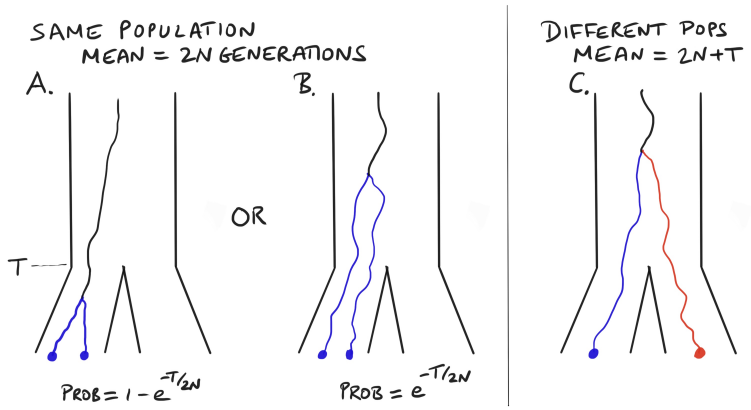


Figure 2.68: **Coalescent times for pairs of samples within and between populations.** *When both samples are from the same population, they either coalesce within their own population* **(A)**, *or back in the ancestral population* **(B)**. *If two samples are drawn from different populations they are not eligible to coalesce until both move into the ancestral population, starting T generations ago* **(C)**.

Under this model, what is the probability that two samples from the same population coalesce in the ancestral population? Using properties of the exponential distribution [213] we can show that this probability is $e^{-T/2N}$. So for example, if we take a person with recent European ancestry [214], then *at a typical locus in their genome there is a very high chance (∼85%) that their two alleles go back as independent lineages into the ancestral African population (Panel B, above).*

How does this look if we consider larger samples? Remember that in a large sample, the first coalescent events occur very quickly, while a few lineages take a long time to coalesce. This means that in a large sample, many lineages coalesce within the population, but the deeper lineages go back into the ancestral population (Panel A):

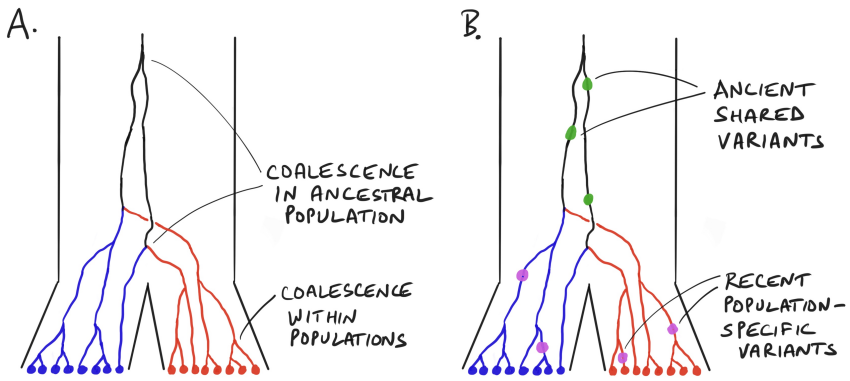


Figure 2.69: **Coalescence of larger samples within and between populations. A.** *Recent coalescences occur within populations, while deeper coalescences are in the ancestral population.* **B.** *Common variants (green) are generally older, and occur in the ancestral population; rare variants (purple) are generally younger, and usually population specific. Note: blue and red lineages are ancestral to samples in one population only; black lineages are ancestral to both.*

This has clear implications for genetic variation (Panel B, above): mutations that occur in the upper parts of the coalescent tree (i.e., older mutations) are usually common, and shared among populations. In contrast, mutations in the lower parts of the tree (younger mutations) are usually rare, and much more likely to be population-specific.

To give you some very rough numbers on this: suppose we sequence m

samples from the same population, then the expected time until the number of lineages goes down to K distinct lineages ($1 \leq K < m$) is

$$2N \sum_{k=m}^{K+1} \frac{2}{k(k-1)}. \tag{2.54}$$

If we start with $m = 1000$ samples in the present day, most of these coalesce very quickly – i.e., within populations. For example, at a time $0.15 \cdot 2N$ generations before the present (i.e., roughly the time of the out-of-Africa dispersal), only ~13 distinct lineages would survive back into the human ancestral population [215] [216]. In other words, each lineage that goes back into the ancestral population is ancestral to a bit less than 10% of the modern sample (on average), so mutations that occur since population splitting would usually be below ~10% frequency.

Meanwhile, the dozen or so deepest lineages would then take a very long time to finally reach an MRCA: almost another $4N$ generations, or ~2 million years. This is why most common genetic variation is old, predates the human diaspora, and is found in all modern populations.

Migration and other complications. I've been describing a highly simplified model in which two populations split at a fixed time T in the past. This simple model is helpful for understanding the main forces at work.

But in truth, real populations are far more complicated. Human structure is somewhat hierarchical, with many populations splitting at different times within and between continents, as in Figure 2.61. Furthermore, as we shall see in Section 3 of the book, populations don't always stay separated: populations exchange migrants or very often undergo major mixing events with other populations.

One important process is **migration** which refers to the movement of individuals (and their alleles) between populations. We can incorporate this into the Wright-Fisher model by defining a migration rate m, per generation.

Then to simulate a new generation, each new allele copy is sampled from its own population with probability $1 - m$, and from another population with probability m:

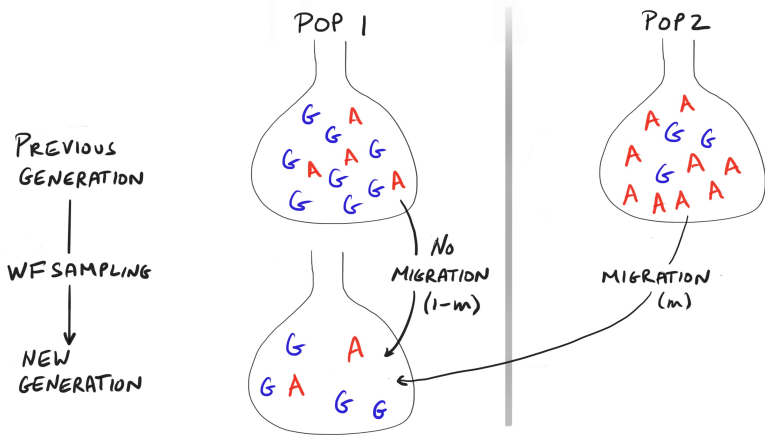


Figure 2.70: **Migration in the Wright Fisher model.** *To simulate a new generation (at bottom), alleles are drawn randomly from one of the parent populations: from the same population with probability from $1 - m$, and from a different population with probability m.*

Equivalently, in the coalescent, lineages switch between populations at rate m per generation. Coalescence can only take place between lineages that are currently in the same population [217] [218]:

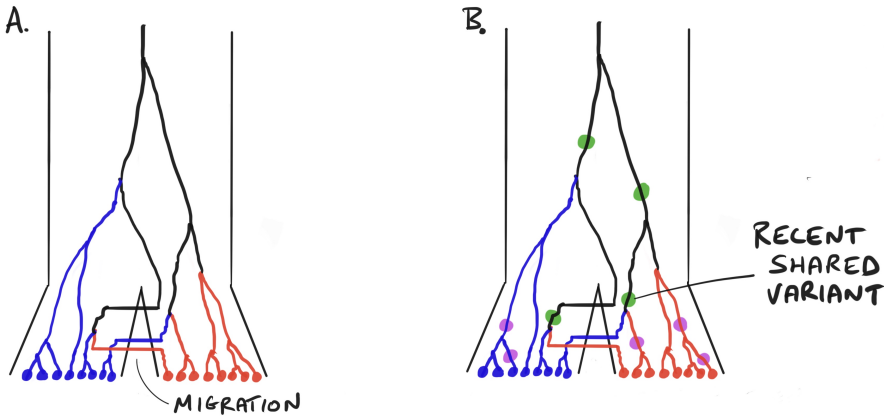


Figure 2.71: **Coalescence in a split model with migration. A.** *In the presence of recent migration, lineages can move between populations at a rate m per generation. They are then eligible to coalesce with lineages in their new population.* **B.** *In the presence of migration, it is possible for recent mutations to be shared among populations, as indicated.*

By moving alleles among populations, migration tends to reduce allele frequency differences among populations, and enables young mutations to move between populations in a way that is not possible in the pure-split model.

Together, these and other related conceptual models help us to understand the effects of a wide range of demographic processes on genetic variation. Using modern software it is now possible to simulate extremely complex models of population histories, including spatial structure, migration, population movements and splits [219].

Measuring population structure: F_{ST}. So far we have been talking about models but, for data analysis, how should we measure the extent of allele frequency differences between populations?

The most widely-used measure of differences between populations is known as F_{ST} (pronounced "F-S-T") [220]. The concept of F_{ST} was developed in the 1930s by Sewall Wright to measure the degree to which random alleles from the same subpopulation are more similar to one another than are random alleles drawn from the total population.

F_{ST} is defined to range from 0 to 1, where **F_{ST}=0 implies no population structure and a value of 1 implies perfect structure**, i.e., that subpopulations are completely fixed for different variants.

Optional: Estimation of F_{ST}

Wright's original formulation referred to F_{ST} as "the correlation between random gametes, drawn from the same subpopulation, relative to the total" [221]. This may sound precise, but there is no unique way to apply Wright's definition to data analysis, and so this idea has spawned many estimators, and many review articles. It's such a mess that I'm tempted to skip the concept entirely, but you can hardly shake a stick around in the population genetics literature without banging into F_{ST}.

One ambiguity is whether the "total" population should refer to the ancestral population, or to an average of modern populations (and if so, which populations to include). Secondly, it's unclear whether our goal should be to estimate an evolutionary parameter that depends on demographic history, or to estimate a simple arithmetic function of the allele frequencies, that can be computed even for individual SNPs. We won't go too far down the F_{ST} rabbit hole here, but I'll sketch out some main ideas [222].

First, consider a situation where multiple populations diverged from a common ancestral population. Let p_k be the present-day allele frequency of a SNP in the kth population; then we can define F_{ST} in terms of *the extent of drift relative to the ancestral population* as follows:

$$F_{ST} = \frac{\text{Var}(p_k)}{p_A(1-p_A)}. \tag{2.55}$$

This expression focuses on the variance in allele frequencies across subpopulations; the denominator $p_A(1-p_A)$ is the maximum possible variance if all subpopulations are fixed for one allele or the other[223]. **This version of F_{ST} measures the variance in allele frequency across subpopulations as a fraction of the maximum possible give the ancestral allele frequencies.**

We can interpret this further using the Nicholson-Donnelly expression $\text{Var}(p_k) \approx (T/2N)p_A(1-p_A)$. Plugging this into Equation 2.55 gives us

$$F_{ST} \approx \frac{T}{2N} \tag{2.56}$$

for small $T/2N$. In contrast, at very large divergence times (for example between species), when all the ancestral variation is either fixed or lost within populations, F_{ST} converges to 1 [224]. Equation 2.55 is not immediately useful for data analysis as it depends on the ancestral frequency p_A, which we cannot observe directly; but it's not hard to estimate this with Bayesian methods [225].

An alternative formulation (and closer to Wright's original framing) is to write F_{ST} in terms of the probability of identity of pairs of alleles between and within subpopulations [226]:

$$F'_{ST} = \frac{H_b - H_w}{H_b} \tag{2.57}$$

where H_w and H_b are the probabilities that two random samples from within a subpopulation, or between subpopulations, are different. The notation H is used here because this is analogous to *heterozygosity*. Although it's not evident at a first glance, this version of F_{ST} is actually a rearrangement of Equation 2.55, but using the total frequency p_t in modern populations instead of the ancestral frequency p_A [227].

Importantly, the expected number of differences between random samples is proportional to their coalescent times, so this expression can be related to average coalescent times within and between populations [228]. **This interpretation of F_{ST} measures the fractional reduction in coalescent times for a pair of samples from the same population compared to a random pair from the total population** [229].

Computing F_{ST} from Equation 2.57 has the advantage that it doesn't depend on the unknown ancestral allele frequency, but it arguably makes estimation *more* difficult because in real applications there is sampling error in both the numerator and the denominator which makes estimation a bit painful. For a helpful summary of moment estimators of F_{ST}, with recommendations, see Bhatia et al (2013).

Turning to data, Bhatia et al (2013) [230] estimated values of F_{ST} between human continental groups. The F_{ST} values are roughly centered around the value of 0.15 that we used above to illustrate our models:

Populations	F_{ST}
Yoruba and Han	0.161
Yoruba and European	0.139
Han and European	0.106

Table 2.6: F_{ST} between human populations. *The data include samples from three populations: Yoruba (from Nigeria), CEU (a sample of individuals from Utah of northwest European descent), and Han Chinese.* Modified from Bhatia et al (2013). [Link]

As you can see, F_{ST} in humans is fairly small (up to about 0.15), even between the most distantly related populations. This reflects the fact that most common variation is shared among all populations.

A second interesting point is that F_{ST} is a bit higher between the African population Yoruba and Han Chinese (0.161), than between Yoruba and Europeans (0.139), even though Europeans and Chinese are descended from the same out-of-Africa migration event. This is because east Asians underwent a stronger bottleneck than Europeans after the out-of-Africa event, resulting in a smaller effective population size and higher F_{ST}.

Third, F_{ST} between populations from the same continent is usually much lower, reflecting more-recent split times and subsequent migration. For example, in the Tibet-Han data set discussed above, the authors estimated that F_{ST} between the two populations is just 0.026.

It would be tempting to interpret F_{ST} values to estimate population split times, using the models described above. But in practice, F_{ST} values depend on a complex mixture of population split times, bottlenecks and migrations. F_{ST} provides a useful summary of the combined impact of all these processes but it's very difficult to untangle the contributions of all these distinct forces in real data [b].

[b] *Ancient DNA has been a game-changer for reconstructing complex population histories, far beyond what is possible using only modern genomes (Chapter 3.3).*

Example: Coalescence between species. To close this chapter, I'll show an example where we can use the coalescent to understand evolutionary splits in a very different context: *between species*.

Before DNA sequencing, the main way that we knew about the evolutionary histories of species was from fossil evidence. But interpretation of the fossil record is often based on just a few fragmentary specimens. It may be unclear how the fossils relate to one another, and to modern populations or species. Fossil evidence continues to be important, but genetic data gives us a powerful complementary type of information for studying our evolutionary history. The accumulation of sequence differences over time, due to mutation, is often called a **molecular clock**.

Figure 2.72: **Our closest relatives:** *female chimpanzee with infant.* Credit: Alain Houle CC BY 4.0 [Link]

Here we'll use genome sequence data to understand the evolutionary relationships among the **great apes**: humans and our closest living relatives: chimpanzees, gorillas, and orangutans.

Until the late 1990s, it was still debated whether humans are more closely related to chimpanzees or to gorillas (orangutans are more distantly related to all three) [231]. DNA sequence data now show that in fact we are most closely related to chimpanzees, with the human and chimpanzee genomes differing at 1.37% of aligned nucleotides compared to 1.75% for

human versus gorilla [232].

This tree shows the evolutionary relationships among the great apes, including that humans and chimpanzees are most-closely related:

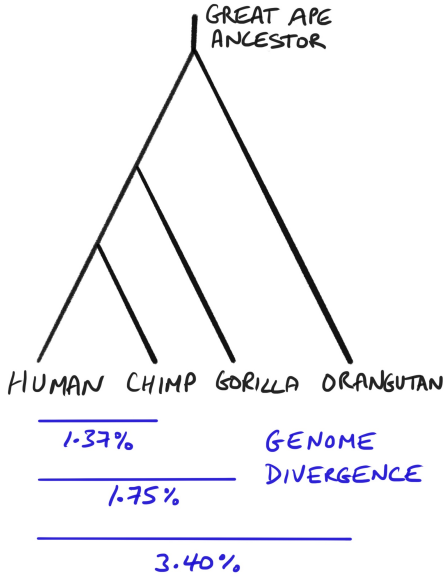


Figure 2.73: **Species tree for the great apes.** *This figure simplifies additional complexity within the nonhuman clades, as there are two recognized chimpanzee species, two gorilla species, and three orangutan species, and additional subspecies of each.* *After Figure 1a from Aylwyn Scally et al (2012) [Link].*

The picture above shows the relationships among the ancestral populations that gave rise to humans and the other great apes (this depiction is known as a **species tree**). But if you look at individual regions of the genome, a very interesting pattern emerges. The branching order for the human, chimpanzee and gorilla sequences vary from region to region across the genome [233]:

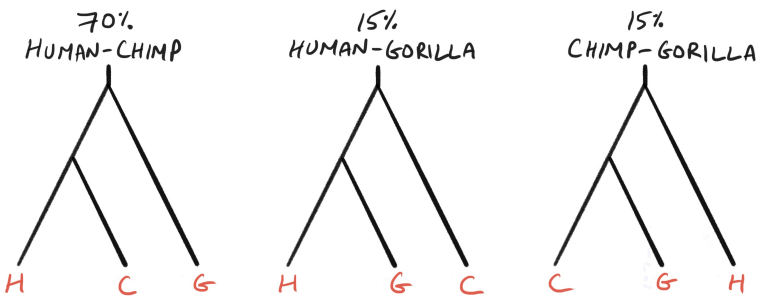


Figure 2.74: **Different parts of the genome support different trees.** *About 70% of the genome supports human and chimpanzee as closest to each other, while the rest supports grouping either human with gorilla or chimpanzee with gorilla.*

About 30% of the genome shows gorilla closer to either human, or chimpanzee. How should we interpret this?

The key to understanding this is to think about the relationships among the different genomes as a coalescent process. First, think about the ancestral lineages for a segment of the human and chimpanzee genomes.

As in the split model we described above, human and chimpanzee cannot coalesce immediately because they come from different species. But unlike our human examples, it is around 6 million years until the human and chimpanzee lineages flow back into an ancestral population. At that

point, the coalescent process you're already familiar with starts: the human and chimpanzee lineages have the opportunity to coalesce, and the average waiting time is an additional $2N_{HC}$ generations, where N_{HC} is the effective population size in the human-chimp ancestral population.

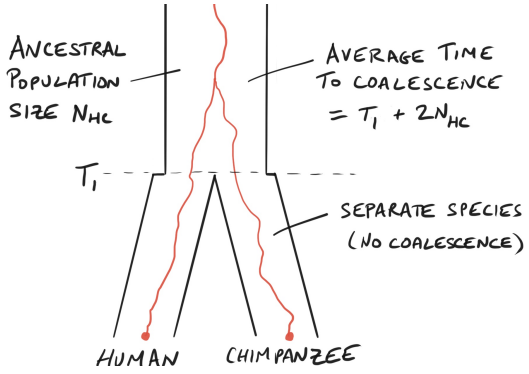


Figure 2.75: **Coalescence of human and chimpanzee lineages**. *Moving backward in time from the present, the lineages are ineligible to coalesce until they flow into the human-chimp ancestral population about 6 million years ago.*

If the human and chimp lineages coalesce quickly, then this always results in the "correct" tree. But if the lineages don't coalesce quickly, they flow back into the human-chimp-gorilla ancestral population. If this happens, all three possible branching patterns are equally likely:

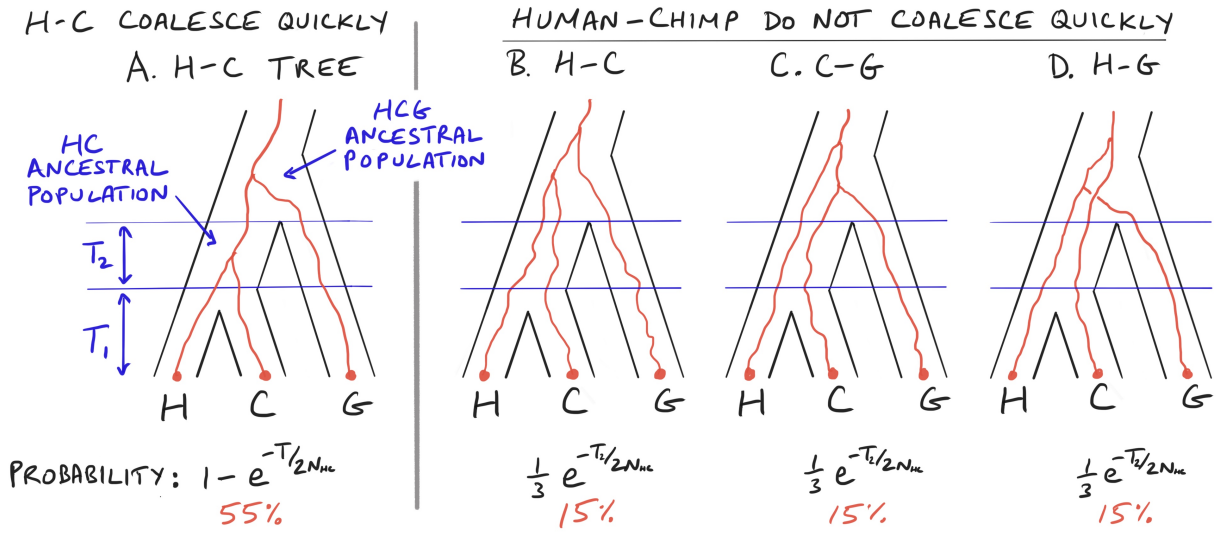


Figure 2.76: Possible coalescent trees relating human, chimpanzee. and gorilla. A. *Human and chimpanzee coalesce in the human-chimp ancestral population, and this ensures that the tree topology matches the overall "correct" relationship among the populations.* **B-D.** *Human and chimpanzee do not coalesce until they flow back into the human-chimp-gorilla ancestral population. When that happens, all three possible trees (with human-chimp, chimp-gorilla, or human-gorilla joining first) are equally likely. The theoretical and actual probabilities for each outcome are shown at the bottom.*

So to summarize, for about 55% of the genome, human and chimp coalesce in the H-C ancestral population. This ensures the "correct" genealogy – meaning that the genealogy matches the species relationships. However, for 45% of the genome the genome, the human and chimp lineages fail to coalesce within the H-C ancestral population and, instead,

flow separately into the H-C-G ancestral population. When that happens, all three possible trees are equally likely, and occur with about 15% probability each [c].

I told you before that 70% of the genome shows a human-chimp pairing: this is the sum of Tree A (55%) and Tree B (15%), while Trees C and D contribute about 15% of the genome each.

[c] *This situation where the local genealogies often differ from the species tree is known as* ***incomplete lineage sorting****.*

Optional math: Probabilities for the four H-C-G tree topologies.

We start by computing the probability of Tree A: i.e., that human and chimp coalesce within the H-C ancestral population. For this we will assume the simplest possible model: constant population size and no population structure. N_{HC} is the effective population size in the human-chimp ancestral population, and we assume that this population existed for T_{HC} generations.

To compute the probability of Tree A we need to compute the probability of a coalescent event for two samples within T_{HC} generations. Using properties of the exponential distribution we can write the probability of a coalescent event at time t as

$$\frac{1}{2N_{HC}} \exp(\frac{-t}{2N_{HC}}) \tag{2.58}$$

where $\exp(x)$ indicates e^x. Then the probability of Tree A equals the probability of $t < T_2$, which we compute by integration:

$$\int_0^{T_2} \frac{1}{2N_{HC}} \exp(\frac{-t}{2N_{HC}}) = 1 - \exp(\frac{-T_2}{2N_{HC}}). \tag{2.59}$$

The remaining probability, $\exp(\frac{-T_{HC}}{2N_{HC}})$, gives us the probability that the human and chimp lineages go back into the H-C-G ancestral population. At that point, there are three lineages (H, C, and G), and any pair of these are equally likely to make the first merger. So the probability of each of these three trees (B, C, D) is simply

$$\frac{1}{3} \exp(\frac{-T_{HC}}{2N_{HC}}). \tag{2.60}$$

It's beyond our scope here, but there has been some fascinating work on the structure of the ancestral great ape populations. While there's still uncertainty in the models, one main result is that the ancestral population sizes were huge: $\sim$120,000 for the human-chimpanzee ancestral population, which is $>$ 6-fold the current human effective size. Consequently, coalescence within that ancestral population was very slow. The human-chimpanzee population split is estimated at $5.5 - 7$ million years ago, and the split from gorilla at $8.5 - 12$ million years ago [234].

Well done! In these last few chapters we have covered the main forces of neutral population genetics! In the remainder of this section of the book we turn our attention to selection. As we shall see, selected alleles are still subject to all the processes we've covered already, but also subject to the guiding hand of natural selection.

Notes and References.

[208] Biddanda A, Rice DP, Novembre J. A variant-centric perspective on geographic patterns of human allele frequency variation. Elife. 2020;9:e60107

[209] Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2002;64(4):695-715

[210] Motivation for the Nicholson-Donnelly Approximation. The variance due to drift in a single generation of the WF model is $p(1-p)/2N$ (using standard properties of binomial sampling). For a sum of independent random variables, the variance of the sum equals the sum of the variances. This rule doesn't really apply here, because the drift is a function of p_t, which depends on the drift in the previous generations. However, if we make the approximation that the drift variance in each generation is constant, and determined by the ancestral frequency, p_A, then the variance over T generations is simply T times the variance in the first generation. This approximation works best for small values of $T/2N$ (for which the allele frequencies don't drift very far from p_A.

[211] Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010;329(5987):75-8

[212] There's also a second fascinating aspect to this story: the selected EPAS1 haplotype is highly divergent from other human haplotypes at this locus, and is believed to have entered the human population by gene flow from a species of archaic hominid known as the Denisovans, which were related to Neanderthals:

Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512(7513):194-7), in a process known as *adaptive introgression*. We'll come back to this when we cover human history.

[213] Recall that coalescent times are exponentially distributed with parameter $1/2N$. The cumulative distribution of the exponential at time T is therefore given by $1 - e^{-T/2N}$; see e.g., [Link].

[214] Here I'm assuming that $T/2N$ since the out-of-Africa migration is around 0.15 time units.

[215] This is calculated using the formula above to compute the expected time to go from $m = 1000$ lineages down to $K = 13$ lineages. You can compute this formula in R using

f <- function(n) { 2/(n*(n-1))}

sum(f(14:1000)).

For simplicity I'm ignoring recent population growth and the out-of-Africa bottleneck. Both events would change the distribution of times but not the overall intuition.

[216] My treatment of this problem is a bit simplistic, for ease of exposition. However there is an extensive literature on the number of lineages at time t, for example:

Jewett EM, Rosenberg NA. Theory and applications of a deterministic approximation to the coalescent model. Theoretical population biology. 2014;93:14-29

Slatkin M. Allele age and a test for selection on rare alleles. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2000;355(1403):1663-8

and references therein.

[217] When there is migration, we can keep track of the number of lineages in each population at any given time (let's call this k_1 and k_2, respectively). Then, going backward in time, migration events from population 1 to population 2 are exponentially distributed at rate mk_1, and mk_2 for the reverse direction. A migration event from 1 to 2 decreases k_1 by one, and increases k_2 by one. Meanwhile, coalescent events occur within populations: e.g., within population 1 at rate $k_1(k_1 - 1)/2$, as usual. We can simulate the next event (coalescence in population 1 or 2, or migration from 1 or from 2) as a process of competing exponentials. Lastly, we can generalize this model to include more populations with an arbitrary matrix of migration rates between populations i and j in each generation.

[218] I'm illustrating the split-plus-migration model here because this is relevant to many human populations. But there's a simpler, classic, model in population genetics called *island migration* in which the populations never merge together, and are subject to migration going back infinitely far in time. In this model, provided that the migration rate is >0 it's guaranteed that eventually the ancestral lineages will happen to collect in one population so that they can merge together. You could motivate the island model by considering populations (for example birds on islands, or butterflies on disconnected systems of serpentine grasslands) that have occupied the same geographic space for a very long time – since long before the joint MRCA of all the populations.

[219] Such as SLiM [Link].

[220] F_{ST} was one of three measures of genetic structure known as Wright's F-statistics. Wright's other F statistics, F_{IS} and

F_{IT}, measure inbreeding of individuals relative to the sub- and total populations, and are less widely used nowadays.

[221] Wright S. The genetical structure of populations. Annals of eugenics. 1949;15(1):323-54

[222] There are various reviews of F_{ST}. I suggest Nicholson et al (2002, cited above) and Bhatia et al (2013), which I relied on for this section

Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Research. 2013;23(9):1514-21;

as well as:

Barton N. Identity and coalescence in structured populations: a commentary on 'Inbreeding coefficients and coalescence times' by Montgomery Slatkin. Genetics Research. 2007;89(5-6):475-7

Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics. 2009;10(9):639-50

[223] To be more precise, this is the variance if there are many subpopulations, each fixed for allele 0 or 1 with probability $1 - p_A$ and p_A respectively or, equivalently, the expected squared difference for each population between its actual allele frequency and the expected value p_A.

[224] We can see that F_{ST} converges to 1 as follows. Eventually every subpopulation either loses the allele (with probability $1 - p_A$) or fixes (with probability p_A). So eventually $\mathrm{Var}(p_k)$ is given by $(1 - p_A)p_A^2 + p_A(1 - p_A)^2 = p_A(1 - p_A)(p_A + 1 - p_A) = p_A(1 - p_A)$. This cancels with the denominator implying that F_{ST} ultimately converges to 1.

[225] Nicholson et al 2002

[226] One advantage of this framing is that it doesn't assume a particular evolutionary model (i.e., population splitting), and is equally applicable for any scenario with structure, such as migration-only models.

[227] To keep this simple we'll consider the frequency in a particular subpopulation p_s as a random variable, and the ancestral or total frequency p_A and p_T, respectively, as fixed parameters. The numerator of Equation 2.55 is $\mathrm{E}[(p_s - p_A)^2]$ by the definition of a variance. Then, noting that $\mathrm{E}[p_s] = p_A$ we have:

$$F_{ST} = \frac{\mathrm{E}[(p_s - p_A)^2]}{p_A(1 - p_A)} = \frac{\mathrm{E}[p_s^2] - 2\mathrm{E}[p_s p_A] + \mathrm{E}[p_A^2]}{p_A(1 - p_A)} = \frac{E[p_s^2] - \mathrm{E}[p_A^2]}{p_A(1 - p_A)}.$$

For Equation 2.57 we note that $H_b = 2p_t(1 - p_t)$ and $H_s = 2p_s(1 - p_s)$, similar to the logic for Hardy-Weinberg. Then

$$F'_{ST} = \frac{2p_t(1 - p_t) - 2\mathrm{E}[2p_s(1 - p_s)]}{2p_t(1 - p_t)} = \frac{\mathrm{E}[p_s^2] - \mathrm{E}[p_t^2] - \mathrm{E}[p_s - p_t])]}{p_t(1 - p_t)} = \frac{\mathrm{E}[p_s^2] - \mathrm{E}[p_t^2]}{p_t(1 - p_t)}$$

[228] See Equations 6 and 8 in Slatkin, M. (1991):

Slatkin M. Inbreeding coefficients and coalescence times. Genetics Research. 1991;58(2):167-75

[229] From Slatkin (1991):

$$F_{ST} = \frac{\bar{t} - \overline{t_w}}{\bar{t}}$$

where $\bar{t}$ is the mean coalescent time for two random samples from the total population and $\overline{t_w}$ is the mean coalescent time for two random samples from the same subpopulation.

[230] Bhatia et al (2013)

[231] A classic paper by Maryellen Ruvolo (1997) discussed incomplete lineage sorting in the human-chimpanzee-gorilla divergence, reporting that 11 out of 14 genomic data sets support the (human, chimpanzee) grouping (see her Table 1):

Ruvolo M. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Molecular biology and evolution. 1997;14(3):248-65

[232] This section draws heavily on work by

Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, et al. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483(7388):169-75

See also

Hobolth A, Christensen OF, Mailund T, Schierup MH. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS genetics. 2007;3(2):e7

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome research. 2011;21(3):349-56

Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499(7459):471-5

[233]The trees at individual genomic regions are known as **gene trees** (although this is a misnomer, since the trees don't correspond to genes *per se*).

[234]There's still quite a bit of uncertainty in these models. One issue is potential changes in mutation rate over time:
Amster G, Sella G. Life history effects on the molecular clock of autosomes and sex chromosomes. Proceedings of the National Academy of Sciences. 2016;113(6):1588-93

2.5 Natural selection: I. Background and models

At its most fundamental level, evolution proceeds through changes in allele frequencies over time. In the next three chapters we will discuss the role of natural selection in shaping genetic variation. This chapter describes basic models of population genetics with selection.

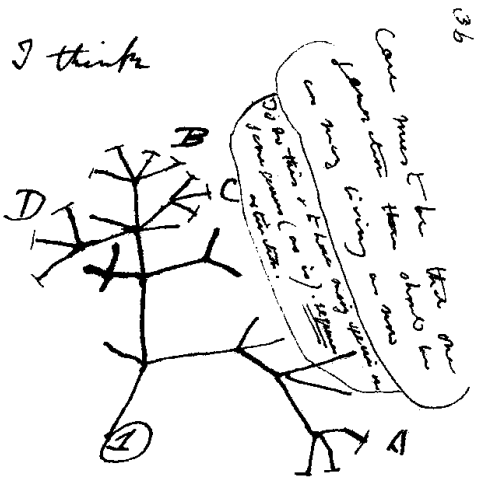


Figure 2.77: **Charles Darwin sketched this evolutionary tree** *in his notebook in 1837, to describe his monumental insight that species evolve from common ancestors.* [Link] *Public Domain.*

Evolution, adaptation, and the modern synthesis. **Charles Darwin**'s 1859 book *On the Origin of Species by Means of Natural Selection* launched a major revolution in the history of science. Darwin articulated two important principles:

(1) that different species evolve from common ancestors, a process that Darwin referred to as "descent with modification"; and

(2) that natural selection and the "struggle for life" provides a driving force for how species change and adapt over time.

These ideas are the fundamental organizing principles of biology: we can understand the similarities and differences among species in terms of the fact that species are descended from common ancestors, while at the same time, their traits evolve over time according to the principles of natural selection [a].

In Darwin's formulation of natural selection (also developed independently by his contemporary Alfred Russell Wallace), populations can adapt over time provided that three conditions are met:

(1) **Variation.** Individuals vary in their phenotypes;

(2) **Inheritance.** The phenotypes are at least partially inherited: i.e., children tend to resemble their parents;

(3) **Competition.** Not all individuals survive or reproduce equally; survival and/or reproductive success depend in part on phenotype;

Under these conditions, the traits that increase the probability of survival or reproduction tend to increase in frequency in the population.

In modern terms, we would say that if there is selection on certain phenotypes, and these phenotypes are (at least partly) controlled by genetic variation, then the genetic variants associated with the preferred phenotypes tend to increase in frequency [b].

Darwin amassed a wealth of evidence for his theory, drawing on natural history, paleontology, biogeography, and other fields. However a crucial gap was that the mechanism of inheritance – i.e., genetics – was not understood at all. At that time, the prevailing model of inheritance was known as "blending inheritance", namely that children represent

[a] *In 1973 the evolutionary biologist Theodosius Dobzhansky famously wrote that "Nothing in Biology Makes Sense Except in the Light of Evolution".*

[b] *Although we usually think of natural selection acting on phenotypes and genotypes, these same principles can act in other domains. In his 1976 book "The Selfish Gene", Richard Dawkins talked about the idea that the principles of natural selection can help to evolve, and spread, ideas in social networks. This idea has become increasingly relevant; you are surely familiar with the term he coined to describe this: "meme".*

some kind of blending, or averaging, of the characteristics of their parents. Blending inheritance would imply a steady loss of phenotypic variation over time, which would be seriously problematic for Darwin's theory since the theory requires the presence of heritable variation. Darwin recognized this as an important gap in his argument, and even endorsed an incorrect alternative model of inheritance called "pangenesis", in which parts of the body emitted particles called gemules that collected in the gonads.

The irony is that, unbeknownst to Charles Darwin, at the same exact time Gregor Mendel was working in Brno (now in the Czech Republic) on the experiments that would lead to Mendelian genetics. His experiments on peas, conducted between 1856 and 1863, showed that genetic information is inherited as discrete packets (i.e., alleles) rather than being blended. In contrast to blending inheritance, Mendelian inheritance means that allelic variation–and hence phenotypic variation–is transmitted from one generation to the next. This insight immediately rescues the Darwinian model. However, Mendel's findings were published in 1866 in an obscure natural history journal published in Brno (*Verhandlungen des naturforschenden Vereines in Brünn*), and were not widely known until the paper was rediscovered in 1900–long after both Darwin and Mendel were dead.

After the rediscovery of Mendel's work, there was a blossoming of genetics in the first half of the 20th Century including, for the first time, a clear understanding of alleles and transmission, a chromosomal theory of inheritance, and some understanding of the connections from genotype to phenotype. Most of the fundamental models of population genetics, including Hardy-Weinberg, the Wright-Fisher model, the basic models of natural selection that we will cover in this chapter, and quantitative genetic models of inheritance that we cover later, all date to this period [c]. This work joining together population genetics with Darwinian evolution in the early-to-mid 20th Century is referred to as the **Modern Synthesis**, and nowadays population genetics and molecular genetics are central pillars of evolutionary biology.

One key insight of the Modern Synthesis is that **evolution results from population genetic processes, played out over long timescales**. In population genetics, we study the forces that change allele frequencies or haplotype frequencies from one generation to the next; accumulated over hundreds, thousands or millions of years this results in adaptive changes, speciation, and everything else in evolutionary biology.

In these next three chapters, we will cover a modern understanding of how natural selection plays out in population genetics, using both theory and examples.

Fitness. In past chapters, our models have assumed that survival and reproduction is independent of genotype. But of course some genotypes do affect the ability of an individual to survive to adulthood, or to reproduce successfully.

Figure 2.78: **Fossil mosquito infected with the malaria plasmodium**, *preserved in amber. Malaria has been a major selective pressure in human history.* *Credit: George Poinar, Jr., [Link] CC BY-SA 2.0.*

[c] *It's striking that most fundamental principles of population genetics can be traced to this period when there was only a rudimentary understanding of genetics, and the molecular details were unknown. In contrast, coalescent theory came rather later (early 1980s), partly stimulated by the emergence of molecular data. The 21st Century has seen huge advances in statistical and computational techniques and the interpretation of modern data.*

To model this, we introduce the concept of **fitness**. Consider an individual at a certain point in the life-cycle (e.g., a newly fertilized egg), with genotype x at a certain variant or set of variants. We define the fitness of genotype x as the expected number of offspring, precisely one generation later, that descend from this individual. In other words, fitness measures the ability of genotype x to survive to reproductive age, to attract mates, and to reproduce successfully through one full turn of the life-cycle.

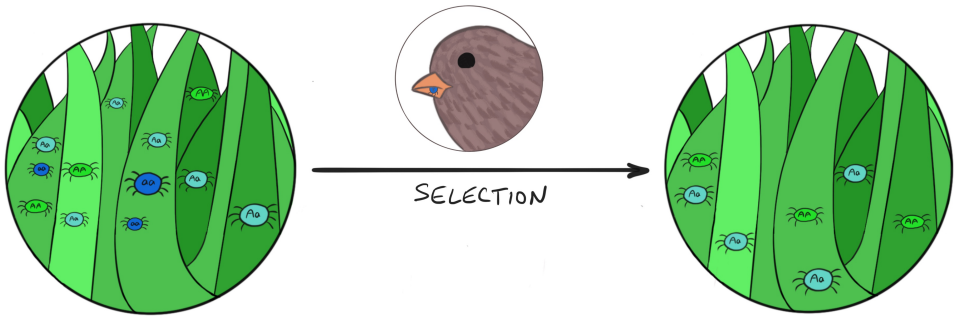


Figure 2.79: **Natural selection and fitness.** *Here, spiders with the aa genotype are blue and stand out from their background; as such they are more likely to be eaten by birds. Hence aa individuals have low fitness, lowering the frequency of the a allele among individuals at reproductive age.* Credit: Lucy Pritchard.

Notice that fitness is defined as an *expected* outcome – importantly, you can *think of fitness as the expected reproductive output for an individual with this genotype, averaging over the possible environments they may experience*, averaging over possible genotypes elsewhere in the genome, and averaging over the good or bad luck experienced by individuals of this genotype throughout their lives: what Hamlet called the "slings and arrows of outrageous fortune".

A basic fitness model. We're now ready to introduce a basic model of selection. We consider a single nucleotide position, with an ancestral allele, A, and a derived allele a.

We model the **relative fitness** of each genotype as follows. AA acts a reference group, defined to have fitness 1, and we measure the fitness of the other genotypes *relative* to that reference [235] [236]:

$$\begin{aligned} \text{Fitness of AA} &= 1 \\ \text{Fitness of Aa} &= 1 + hs \\ \text{Fitness of aa} &= 1 + s \end{aligned} \tag{2.61}$$

Here, s is referred to as the **selection coefficient**, and h is the **dominance coefficient**:
- If s is positive $(s > 0)$ then the derived allele is **advantageous**
- If s is zero then the derived allele is **neutral**
- If s is negative $(s < 0)$ then the derived allele is **deleterious**

Reflecting the sign of s, selection in favor of an advantageous allele is also referred to as **positive selection**; selection against a deleterious allele is **negative selection.**

To give you a sense of scale, the most strongly advantageous derived alleles in humans may have s of up to $\sim$3%. But there are many more ways to break genomes than to improve them: the effects of deleterious

variants can range from just very slightly negative, all the way down to $s = -1$ (which would imply that the derived allele is incompatible with life or reproduction).

Meanwhile, h measures the relative fitness of heterozygotes, and is known as the **dominance coefficient**. If the derived allele a is fully recessive then $h = 0$; and if a is fully dominant then $h = 1$. In rare cases h can be outside the range $[0, 1]$ leading to a special form of selection called balancing selection. Except where stated the figures below assume what is known as an **additive model (h = 0.5)**.

Frequency changes over time. How does selection change allele frequencies and genotype frequencies over time? We'll set p to be the current derived allele frequency, and $q = 1 - p$ as the ancestral frequency.

Genotype frequencies. Before selection the genotype frequencies are given by Hardy Weinberg proportions. The effect of selection is to change the genotype frequencies in proportion to their fitnesses.

For example, the frequency of the *aa* homozygote is p^2 before selection, and proportional to $p^2(1 + s)$ after selection:

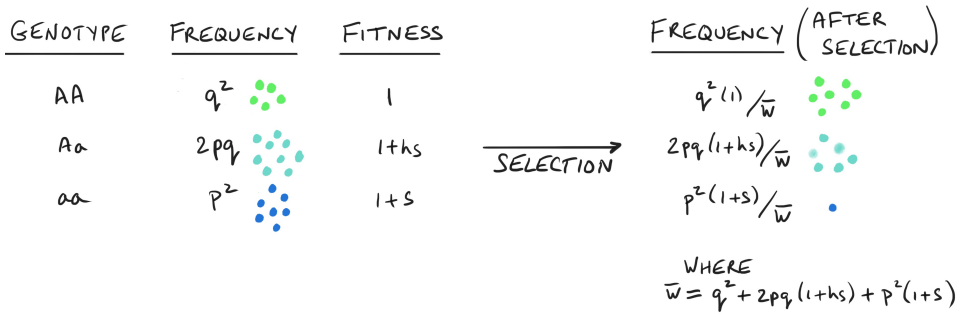


Figure 2.80: **Changes in genotype frequencies due to selection.** *Before selection the genotype frequencies are given by Hardy Weinberg proportions. After selection, the frequencies are multiplied by the genotype fitnesses – illustrated here for $s < 0$. The factor of $\overline{w}$ is used so that the genotype frequencies add to 1.*

By definition, frequencies have to add up to 1, so each of the terms above is divided by the total, a quantity known as the mean fitness:

$$\overline{w} = q^2 \cdot 1 + 2pq \cdot (1 + hs) + p^2 \cdot (1 + s). \tag{2.62}$$

Dividing by $\overline{w}$ simply rescales the frequencies to sum to 1.

Allele frequencies. And what is the expected frequency of a in the next generation? (We'll call this p'.) To get this we add together half the frequency of heterozygotes plus the frequency of *aa* homozygotes:

$$\mathrm{E}[p'] = \frac{pq(1 + hs) + p^2(1 + s)}{\overline{w}} \tag{2.63}$$

This expression isn't particularly illuminating, but we get something more useful if we look at the *change* in allele frequency, Δ_p from one generation to the next:

$$\Delta_p = \mathrm{E}[p'] - p. \tag{2.64}$$

Δ_p tells us whether p is increasing or decreasing over time (depending on whether Δ_p is positive or negative). After a small flurry of algebra [237], we find that

$$\Delta_p = \frac{pqs[p(1-h)+qh]}{\overline{w}}. \tag{2.65}$$

This expression is easier to interpret:

- When $p = 0$ or $q = 0$ there is no allele frequency change. That makes sense, because there's no variation for selection to act on.
- If $s = 0$ there's no selection, and no expected change in allele frequency.
- Third, and most important, if p lies between 0 and 1 we have the intuitive result that *if s is positive, then Δ_p is positive, meaning that the derived allele is favored, and tends to increase in frequency; if s is negative, the derived allele is disfavored and tends to decrease* [238].

What happens over multiple generations? We can iterate Equation 2.65 over multiple generations to predict the trajectory of a selected allele over time. This is known as a **deterministic** model, meaning that it assumes the trajectory of an allele is completely determined by the expectation. As you can see, selection drives favored alleles up towards fixation, and deleterious alleles to loss. The process in which favored alleles are pushed up to fixation is called a **selective sweep**.

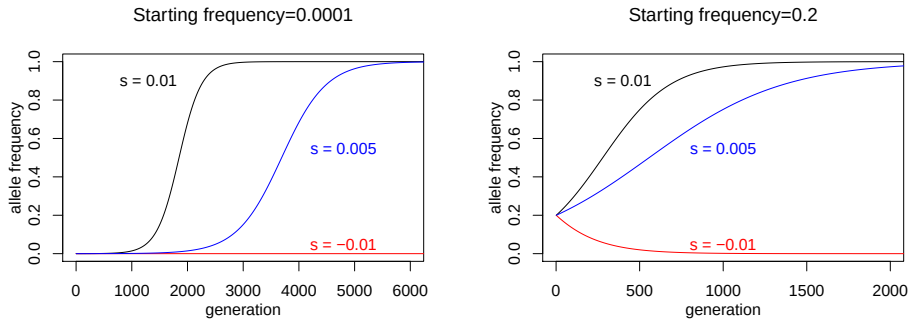


Figure 2.81: **Allele frequency trajectories of selected alleles, over time**. *The blue and black lines show frequency increases of advantageous alleles. The right-hand plot assumes an unreasonably high starting frequency of 0.2 to illustrate that selection drives deleterious alleles (in red) to low frequencies.*

The deterministic model is helpful for understanding the overall process, but it's also important to consider the random effects introduced by drift.

Frequency changes with selection and drift. In Chapter 2.1 I suggested that genetic drift of a new mutation is like a player's winnings over time in a casino. Even for advantageous alleles, the effects of random sampling are extremely important.

Suppose you walk into a casino to play Blackjack, and you play until you either go bust, or beat the house.

For Blackjack, assuming optimal play, players usually have an inherent disadvantage of 0.5–1.0% relative to the casino (the precise value depends

on the house rules). However, there are card-counting strategies that can potentially tilt the odds by 1–2% back towards the player, turning a small player disadvantage into a small player advantage (although these are frowned upon by the casinos [239]). You can think of the default Blackjack game as like selection on a mildly deleterious mutation ($s < 0$), and the game with card counting as like a mildly advantageous mutation ($s > 0$).

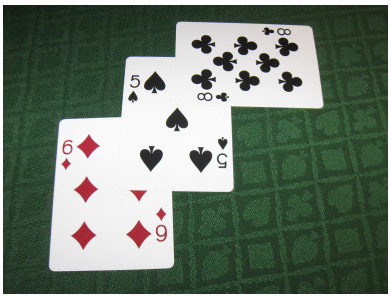


Figure 2.82: **Blackjack cards**. *Credit: Scott5114 [Link] Public Domain*

There are two key points here: First, with a small starting purse, you're likely to go bust quickly, regardless of whether you count cards or not. This is simply because, with small numbers, you're likely to have at least some bad luck that bankrupts you. This reflects the great importance of chance when you're working with small numbers.

But if you get lucky early on, then the power of large numbers starts to take over, and you can start to use a deterministic model to predict how your purse will grow – at least until the casino tosses you out!

Selection in the WF model. The same fundamental processes affect new mutations. So far we have considered the Wright Fisher model for neutral alleles, but it's easy to extend it to allow biased sampling due to selection.

Under the neutral model, if the current allele frequency is p in a population of size $2N$, then the allele frequency in the next generation would be

$$p' \sim \text{Binomial}(p, 2N) \quad [\text{neutral model}] \tag{2.66}$$

With selection, the allele frequency in the next generation is similar but centered on the *expected allele frequency with selection*, $\text{E}(p')$, as given by Equation 2.63:

$$p' \sim \text{Binomial}(\text{E}(p'), 2N) \quad [\text{with selection}] \tag{2.67}$$

Here's what this looks like, for one generation of sampling with relatively strong selection: $s = 0.01$. (A 1% selective advantage may not seem like much, but as we'll discuss shortly there are very few individual changes to the genome that can improve fitness by this much.)

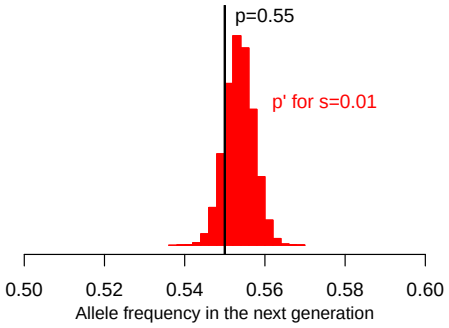


Figure 2.83: **Histogram of binomial sampling outcomes** *(p') after one generation of selection and drift with $s = 0.01$ and $2N = 20,000$. The starting allele frequency $p = 0.55$. (Compare to the neutral case, Figure 2.5.)*

As you can see, the overall distribution of outcomes (in red) is shifted toward higher frequencies than the initial frequency p. However, due to the random sampling process, there is variation in the resulting allele frequency, and even a chance that the frequency actually decreases, despite the upward selection pressure.

Now, let's look at **selection and random drift** together, over multiple generations. The next figure shows simulated trajectories from a starting frequency of 0.5, for a range of different selection coefficients. In the left panel, selection is strong enough that drift has little impact on the selected alleles (blue and red curves). In contrast, in the right panel, selection is just 1/10th as strong, and while the blue curves tend to be higher than the red curves on average, the randomness of drift means that some favored alleles (blue) fare worse than some deleterious alleles (red):

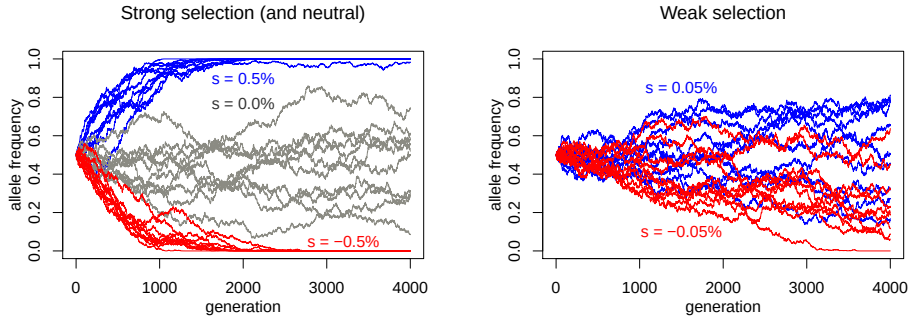


Figure 2.84: **Selection and drift of alleles from starting frequencies of 0.5** *in a population of* $N = 10^4$. *The left panel shows simulated trajectories for relatively strong selection* ($2Ns = 100$ *in blue, and* $2Ns = -100$ *in red); and neutral in gray for comparison. The right panel shows weaker selection* ($2Ns = 10$ *in blue, and* $2Ns = -10$ *in red).*

In the plots above, the directional effect of selection has to fight against the randomness imposed by genetic drift in a finite population. *When selection is strong enough, it overwhelms drift, and allele frequency curves are close to the deterministic trajectory. But when selection is weak, or the population is small, drift can effectively overwhelm selection.*

To quantify this, *a widely-used rule of thumb is that when* $2Ns$ *is in the range of about* -1 *to* 1, *selection is so weak that it is nearly overwhelmed by drift*. Such alleles are referred to as **nearly-neutral**. Alleles with $|2Ns|$ in the range of about 1 to 10 do feel the effects of selection, but are also heavily influenced by drift as you see above.

What sets this scaling for the nearly-neutral range? One way to think about this is that if $2Ns = 1$ then selection effectively adds (or removes) one copy of the alternate allele per generation somewhere in the population [240]. Below this threshold selection is almost entirely ineffective [241].

Most new mutations are lost, even if they are favorable. The last important point is that even strongly favored alleles are vulnerable to the vagaries of random sampling when they are rare [d]. To illustrate this, the simulations shown below started with 1000 new mutations with a 1% selective advantage. Despite the selective advantage, only about 11/1000 of the simulated alleles spread to fixation; the rest were rapidly lost from the population. As you can see, most trajectories stayed below 1% and were lost by drift; in contrast, nearly all of the trajectories that got above 1% went into deterministic growth and reached fixation:

[d] *This is analogous to the card-counter who walks into a casino with a small initial purse. Even if she has a long-term advantage, she is likely to go bust early on.*

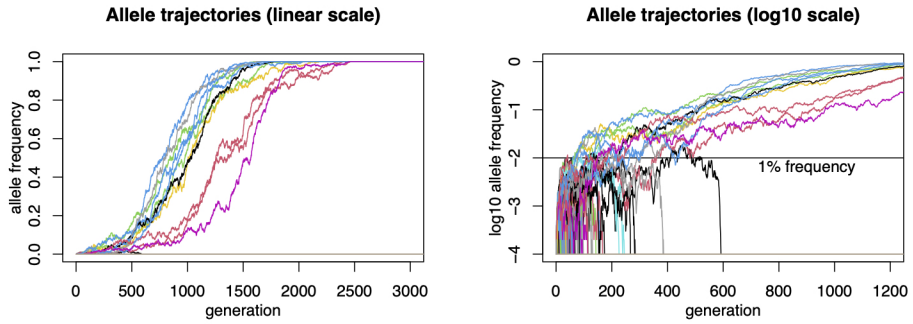


Figure 2.85: **Selection and drift of new favored mutations**. 1000 *simulated allele frequency trajectories with* $s = 0.01$*, starting from a single copy in a population of* $2N = 10^4$*. Around 99% of alleles were lost quickly, and are hard to see as they are effectively on top of each other along the* $y = 0$ *line. The right-hand panel shows the same data but with different axes: note the log*10 *scale on the y-axis to show rare variants more clearly.*

Fixation probabilities with selection. For strongly favored mutations, the probability that a new favored mutation fixes is only about $2hs$: this was 1% in the simulation above, close to the observed rate of $11/1000$. This fixation rate is much higher than the rate for neutral variants (i.e., $1/2N$) but still means that nearly all advantageous mutations are lost. For this reason, adaptation by new mutations can be highly inefficient [242].

A general formula for fixation probabilities with selection and drift was developed by the Japanese population geneticist Motoo Kimura in the 1950s (we'll hear from Kimura again soon, when we get to the Neutral Theory) [243]. For a new mutation with $h = 0.5$ the Kimura formula simplifies to

$$\text{Probability of fixation} = \frac{1 - e^{-s}}{1 - e^{-2Ns}}. \tag{2.68}$$

You can see this plotted here:

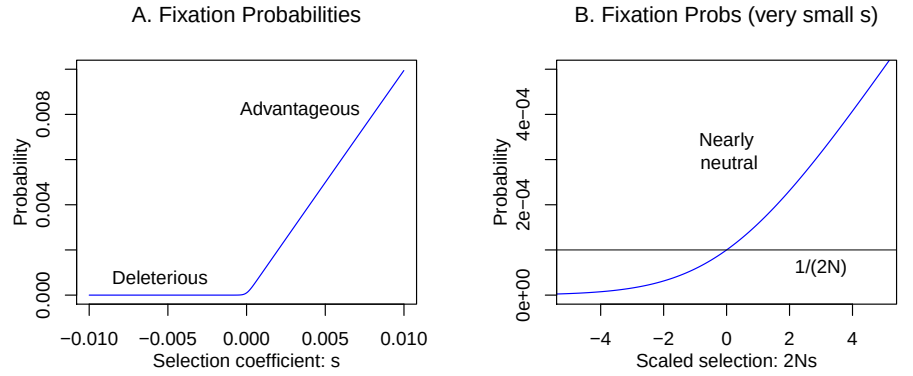


Figure 2.86: **Fixation probabilities of new mutations** *(here* $2N = 10^4$ *and* $h = 0.5$*).* **A.** *Fixation probabilities across a wide range of s.* **B.** *Nearly-neutral range: Same plot highly magnified near* $s = 0$*, with x-axis in units of* $2Ns$ *instead of s. The horizontal line at* $1/2N$ *shows the fixation probability for neutral mutations.*

The left-hand plot above illustrates that when selection is strong, the model does not depend on population size: strongly deleterious alleles have essentially no chance of fixing, and strongly advantageous mutations fix with probability $\sim 2hs$. Chance *does* matter for favored mutations, but only because it determines whether they start to spread when they are extremely rare [244].

But we see something quite different in the right-hand plot. This shows what happens in the **nearly-neutral range**, where selection is weak compared to drift (roughly $|2Ns| < 1$). These alleles drift very much like neutral alleles, and selection only modestly increases or decreases their chances of fixation [245].

We close this chapter with a deeper discussion of negative selection; we'll return to positive selection and balancing selection in the next chapter.

Purifying selection: protecting the genome against mutation. As we discussed in Chapter 1.5, our genomes suffer a barrage of mutations in every generation – around 70 per child. The vast majority of these are close to neutral, but among those with functional effects, the vast majority have deleterious effects. I think this is intuitive: if you introduce random typos into a written document, you're far more likely to reduce the quality of the writing than to improve it!

For this reason, the most common form of selection is against **deleterious** variants. The term "deleterious" refers to variants with fitness $s < 0$, and includes everything from severe disease-causing mutations to millions of variants across the genome with tiny effects on phenotypes and mildly negative effects on fitness. Selection against deleterious variants is referred to as **negative selection**; or sometimes **purifying selection** because it cleanses deleterious mutations from the genome.

Under a strictly deterministic model, a new deleterious mutation would not increase in frequency at all. But in practice, natural selection is competing against the randomness of genetic drift, and some deleterious alleles do manage to drift up to higher frequencies [e]. For this reason, at any given time, some variants segregating in a population are actually deleterious, but they tend to be at lower frequencies than neutral variants.

[e] *Going back to the gambling metaphor, even a completely rubbish player might win some money by luck early in a game, but they are extremely unlikely to keep winning indefinitely.*

Here you can see simulations comparing genetic drift of 1000 neutral variants (panels A and B) and 1000 deleterious variants with a fitness disadvantage of 5% (panels C and D). In each plot, all the trajectories were started from a single copy at time 0.

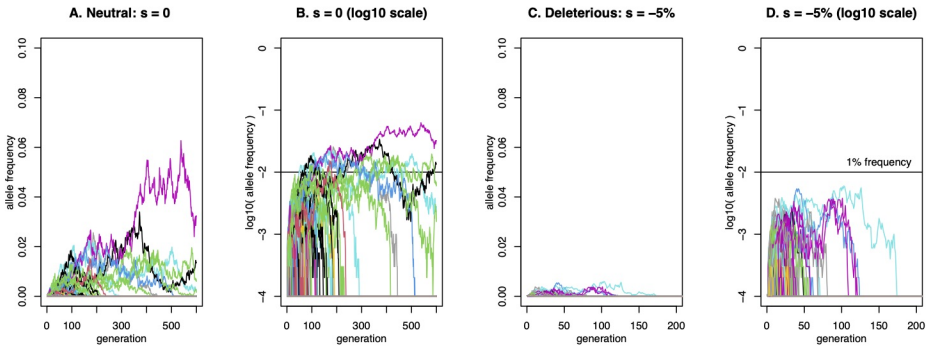


Figure 2.87: **Selection and drift of new mutations: neutral (panels A and B) and deleterious (panels C and D).** *Here panels A and B show the same data, but with the y-axis of B plotted on a log-scale to show more detail about rare variants. The same is true for C and D.*
Parameters: 1000 *simulated allele frequency trajectories for each panel, starting from a single copy at time* 0 *in a population of* $2N = 10^4$.

As you can see, the deleterious variants (C and D) are held at lower frequencies and are removed from the population much faster than the neutral variants.

Here, a useful approximation is that deleterious variants can drift up to a maximum frequency on the order of $\sim 1/(2N \cdot hs)$*, corresponding to selective removal of about one copy of the derived allele per generation.* This corresponds to 0.4% in Panels C and D above, which you can verify is close to the highest frequencies across the 1000 replicates.

The SFS for deleterious alleles. The plots above show trajectories of mutations over time, but in practice it's much easier to measure the distribution of allele frequencies across many different variants at a single point in time.

This is shown in the next plot, with theoretical distributions for neutral sites (red), nearly-neutral (blue), mildly deleterious (black) [246]. You can see that at low frequencies all three curves are similar, but at higher frequencies selection greatly reduces the numbers of deleterious alleles:

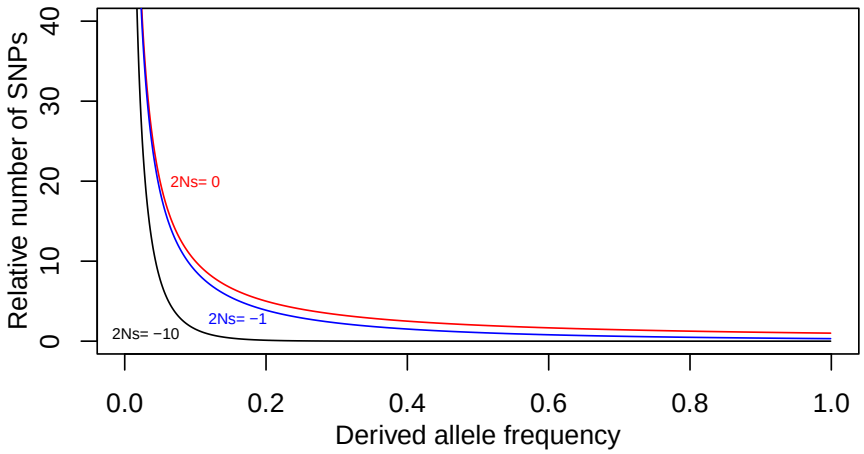


Figure 2.88: **Theoretical distributions for numbers of variants as a function of allele frequency, with weak purifying selection**. *The expected number of variants between frequencies p_1 and p_2 in a region of L basepairs is $4N\mu L$ times the integral from p_1 and p_2. Curves computed from theory in Sawyer and Hartl (1992) [Link].*

Hence, for a given number of base pairs, we see fewer total SNPs at deleterious sites, and the SNPs we do see tend to be at low frequencies.

We can also see similar effects in real data. The plot below, from 2005 [247], was one of the first to show the **site frequency spectrum (SFS)** [f] for sites with different levels of constraint [248].

This analysis tests the hypothesis that missense (nonsynonymous) variants are under purifying selection, and uses synonymous and noncoding variants as controls that are less often constrained [249]. Under this hypothesis, we would expect more of the missense variants to be at low frequencies.

[f] *Recall from Chapter 2.2 that the SFS shows the fraction of SNPs at each allele frequency.*

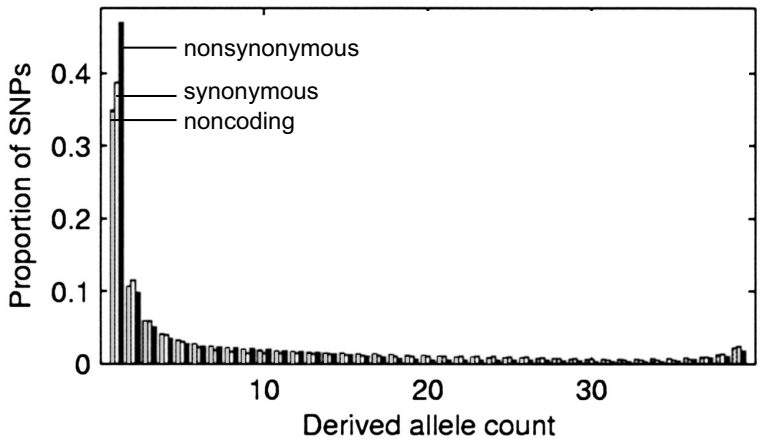


Figure 2.89: **SFS for different types of SNPs** *in a sample of size 40. Note that the plot is drawn differently than the theoretical plot, as here the histograms add up to 1 within each category. This plotting style emphasizes the relative shift toward rare variants for nonsynonymous SNPs. Credit: Modified Figure 1 from Scott Williamson et al (2005) [Link].*

Indeed, as you can see, around 48% of missense sites are singletons in

this data set, compared to around 35–38% of noncoding and synonymous sites [250]. *This reflects the fact that a large fraction of missense variants are under purifying selection.*

Sequence conservation between species is an important indicator of function. The SFS analysis is useful for showing that a *type* of variant (such as missense mutations) is under purifying selection, but it's not very useful for testing at individual sites [251].

However, recall that selection is extremely effective at preventing deleterious variants from fixing. So an alternative is to use **sequence conservation** between distant species to identify regions or sites that are functionally important. If we compare distantly related species, then a large fraction of neutral sites will show differences, but sites that are **functionally constrained** are much more likely to be shared.

This concept has been used to identify functional regions of the genome: for example important regulatory enhancers, or protein domains that are particularly crucial for protein function [252].

For example, the plot below shows sequence conservation between mouse and four distantly related vertebrates in the region around the TBX2 and TBX4 genes (highly conserved master regulators of limb development). Regions marked in blue are exons, are regions in red are putative noncoding elements. The boxed regions were shown to have regulatory activity in transgenic mouse experiments [253].

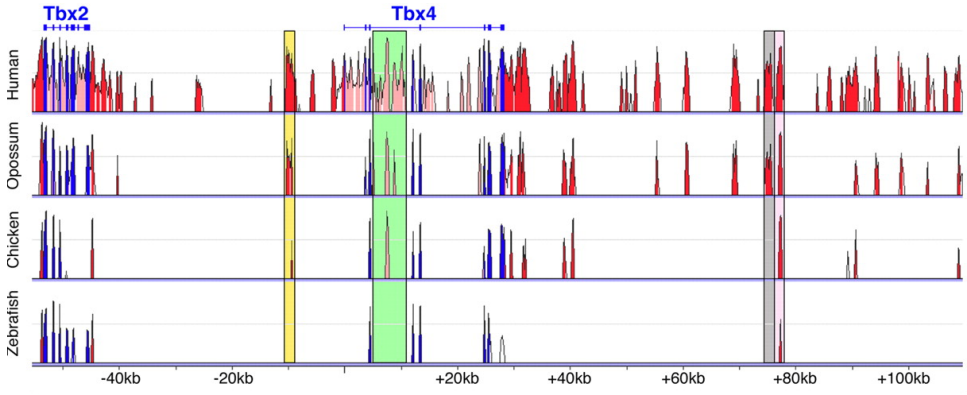


Figure 2.90: **Genome sequence conservation identifies functional elements.** *Each track shows regions with high sequence identity between mouse and the indicated species (the y-axis of each track ranges between 70 – 100% sequence identity). Coding exons are shown in blue and noncoding conserved regions in red.*
Credit: Figure 2 from Douglas Menke et al (2008) [Link]; CC BY.

Nearly-neutral mutations and the limits of natural selection. We've been talking about how natural selection tends to purge deleterious mutations. But as I discussed above, for variants with very weak selection, the vagaries of drift become more important than selection; we refer to these weakly selected variants as **nearly-neutral**.

There's no hard cutoff for a variant to be "nearly-neutral", but as I noted above, a common definition is $|2Ns| \leq 1$.

Here it's worth pausing to reflect on the fact that selection is an extraordinarily efficient process. To put this into numbers, if the human effective population size N_e is $\sim 15,000$, this implies that selection is efficient

down to around 3×10^{-5}, or three extra individuals surviving or reproducing per 100,000. Variants with s smaller than this are nearly-neutral.

And yet, while a single variant with a fitness cost of 10^{-5} is almost inconsequential, the combined impact of many nearly-neutral mutations can have meaningful effects. In particular, **the existence of the nearly-neutral zone places important limits on the extent to which natural selection can optimize genomes.**

This is especially true for species with small effective population size, including humans, since the size of the nearly-neutral zone depends on N. These species are much worse at safe-guarding their genomes from weakly deleterious mutations compared to species with larger populations including fruit flies, yeast, or E. coli.

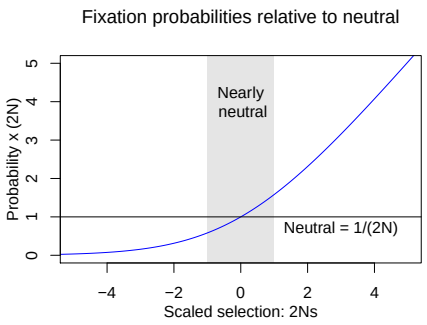


Figure 2.91: **The nearly-neutral zone**: *fixation probabilities of new mutations. A slightly deleterious mutation with* $2Ns = -1$ *is nearly as likely* ($0.58\times$) *as a new neutral variant to fix. A slightly advantageous mutation* ($2Ns = 1$) *is about* $1.6\times$ *more likely to fix than neutral.*

One setting where nearly-neutral mutations are relevant is for something called **codon bias**. As you know, different DNA triplets can code for the same amino acid (e.g., GGA, GGC, GGG, and GGT all encode glycine). Mutations that switch between alternative triplets encoding the same amino acid are referred to as synonymous. However, it turns out that some synonymous codons are slightly preferred over others, likely because they enable greater translation accuracy or speed. Preferences are species-specific and correlate with the abundances of the corresponding tRNAs [254].

These codon preferences can result in a *very slight selective benefit* to using one synonymous codon instead of another. But you can imagine that the fitness consequence of switching, for example, a single GGA to GGG in a single gene, is very very small. In consequence, the ability for a species to maintain codon usage bias depends on its effective population size – for species with sufficiently large N_e, codon switches can lie outside the nearly-neutral zone. As a result, many species with large N_e, such as in Drosophila, can maintain strong codon bias across the genome, while species with small N_e including humans cannot [255].

A second example comes from the difficulty that genomes have in controlling the spread of **transposable elements** (TEs). TEs are DNA elements that can copy themselves and reinsert the copies elsewhere in the genome, usually via an RNA intermediate [256]. While TE insertions do occasionally have salubrious effects [257], on the whole they are considered **selfish DNA**: they replicate because they can, but they do not benefit the host genome. Quite remarkably, it's estimated that more that 2/3 of the human genome was originally derived from transposable element insertions [258]. Moreover, 10% of your genome is made up by copies of just a single 300 bp element called **Alu**, which is present about in about 1 million copies [259]! Although a few Alu copies play functional roles in gene regulation, Alus are primarily parasitic elements.

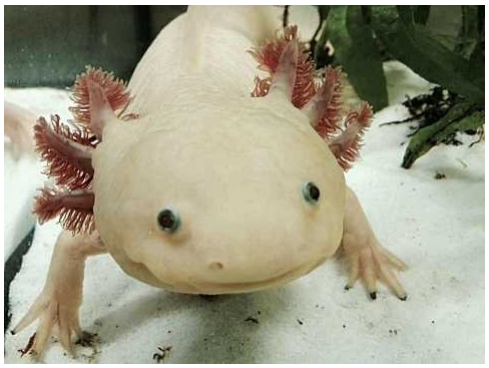


Figure 2.92: **Axolotl genomes are 10-fold larger than ours**. *The axolotl, a model organism for limb regeneration, has a huge 32GB genome chock-full of millions of transposable elements. This fact also provides the opportunity for a gratuitous axolotl photo.* Credit: th1098 [Link], CC BY-SA 3

The key problem is that the selective costs of most new TE insertions are very small. When an Alu is copied into a new location, there is a slight chance that it inserts into a functional region such as an exon, in which case it will probably be deleterious [260], and be removed by selection. But

if the new copy inserts into nonfunctional sequence, the added cost of the new copy is almost negligible – mainly the tiny cost of replicating a few hundred basepairs of additional DNA at every cell division [261] .

In fact, the marginal cost of each new Alu is so small that selection cannot effectively prevent individual Alus from fixing. At the same time, however, there are substantial genome-wide costs to carrying and replicating millions of TEs. In consequence, genomes have evolved trans-acting mechanisms for epigenetic silencing of TEs to try to reduce their rates of spreading [262].

A third example is **evolution of the mutation rate** [263]. The mutation rate depends on a number of factors: the rate of spontaneous damage and copying errors, as well as the ability of cells to fix these errors. These factors – in particular the complex machinery that cells use to prevent and repair errors – are of course evolved properties of organisms. What factors determine the evolution of the mutation rate?

Given that mutation is an essential component of evolution, you might think that some amount of mutation is helpful. That may be true for the long-term survival of a species, but from the viewpoint of an individual – which is what matters for natural selection – the overall effect of mutation is negative. The mutations your kids inherit may have no impact on their fitness, but if they do impact fitness, then it's *much* more likely that they have a negative effect than a positive effect.

Consider a new variant that makes the DNA repair machinery very slightly worse – such a variant is known as a **mutator** allele. Let's suppose this mutator increases the average genome-wide number of mutations by a single mutation. We can estimate that this mutator variant would decrease fitness by around 10^{-5}, which puts it in the nearly-neutral range for humans, and selection wouldn't be very good at removing it [264]. In contrast, a mutation that adds 10 new mutations per generation would have a ten-fold higher fitness cost and would be much more visible to selection [265].

This process creates what Michael Lynch has termed a **drift barrier**: natural selection cannot reduce the mutation rate indefinitely because below a certain point, any improvement to the mutation rate is nearly-neutral, and hence mainly governed by drift. **The mutation rate at which the drift barrier kicks in depends on population size.** Indeed, data on mutation rates of different organisms suggest that mutation rates are determined by the drift barrier model, as species with larger population sizes tend to have lower mutation rates:

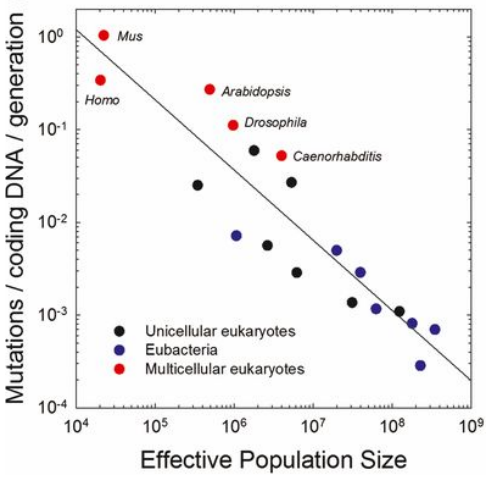


Figure 2.93: **Relationship between mutation rate and effective population size.** *The mutation rate here refers to the total number of mutations at protein-coding positions, per generation.* *Credit: Figure 1c from Way Sung et al 2012. [Link]*

Genetic load. Given these limits of natural selection, each of our genomes contains many deleterious variants. These come in two main categories:

- Each of us has *a unique personal collection of deleterious variants that are currently drifting at low to moderate frequencies* (and will eventually be removed from the population by natural selection).
- Like any other species, theory argues that *humans must also carry many fixed variants that are weakly deleterious*, but within the nearly-neutral range where selection is ineffective.

Together, these deleterious variants are referred to as genetic load.

It's been argued that the second of these categories – fixed nearly-neutral variants – leads to an evolutionary paradox. If we make the plausible assumption that many more mutations are slightly bad than slightly good, then we should predict an inexorable increase in genetic load over evolutionary time. One famous paper had the colorful title *"Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over?"* [266]. But clearly we're still around after 4 billion years of evolution so this argument cannot be fully correct. While the details are still not entirely clear, this argument likely under-states the ability of weakly advantageous mutations to counteract the accumulation of load [267]

Meanwhile, our burden of segregating deleterious variants is responsible for the genetic contributions to phenotypic variation and disease – and is something we'll come back to in much greater detail in Section 4 of the book.

In this chapter we have covered basic models of selection, with and without drift, and an overview of negative selection. In the next chapter we turn to a deeper consideration of positive selection.

Notes and References.

[235]In these models, the alleles compete against each other, but we assume that the population size is fixed by exogenous factors–perhaps food or other resources–and that selection at the variant in question does not directly drive population growth. This is referred to as "soft selection", and the genotype fitnesses are measured relative to one another. In contrast, in *hard selection* models, the genotypes have absolute fitness values, and this means that the population can grow, or grow faster, as fitter alleles increase in frequency. Soft selection models are theoretically more tractable, and usually a good approximation in humans where fitness gains from any single variant tend to be very small. Hard selection may be relevant in other situations–for example in modeling growth of E. coli on antibiotics, where an antibiotic resistance allele can allow a dramatic increase in growth rate.

[236]You'll often see this model parameterized slightly differently, denoting the fitness of each genotype by w with a subscript: i.e., w_{AA}, w_{Aa}, w_{aa}. But in the soft selection case what matters is the fitness of each genotype relative to the others, so we set the ancestral homozygote to be a *reference group*, and divide all three fitnesses by w_{AA}. Now the fitnesses are 1, w_{Aa}/w_{AA}, w_{aa}/w_{AA}, which we rewrite as $1, 1+hs, 1+s$. (We can do this provided that we don't have the special case of symmetric balancing selection $w_{AA} = w_{aa} \neq w_{Aa}$).

[237]First, recall that we want to compute $\Delta_p = \mathrm{E}[p'] - p$ where

$$\mathrm{E}[p'] = \frac{pq(1+sh) + p^2(1+s)}{q^2 + 2pq(1+sh) + p^2(1+s)} \tag{2.69}$$

We simplify the notation by using $\overline{w}$ in place of the denominator (pronounced w-bar, and referred to as "mean fitness"), and simplifying:

$$\begin{aligned} \overline{w} &= q^2 + 2pq(1+sh) + p^2(1+s) && (2.70)\\ &= q^2 + 2pq + 2pqsh + p^2 + p^2 s && (2.71) \end{aligned}$$

Noting that $p + q = 1$ and $q^2 + 2pq + p^2 = 1$ we simplify this to

$$\overline{w} = 1 + 2pqsh + p^2 s \tag{2.72}$$

Now we're ready to start calculating Δ_p as follows:

$$\begin{aligned} \Delta_p &= \frac{pq(1+sh) + p^2(1+s)}{\overline{w}} - p \times \frac{\overline{w}}{\overline{w}} && (2.73)\\ &= [pq(1+sh) + p^2(1+s) - p[1 + 2pqsh + p^2 s]/\overline{w} && (2.74)\\ &= p[q(1+sh) + p(1+s) - 1 - 2pqsh - p^2 s]/\overline{w} && (2.75)\\ &= p[q + qsh + p + ps - 1 - 2pqsh - p^2 s]/\overline{w} && (2.76)\\ &= p[qsh + ps - 2pqsh - p^2 s]/\overline{w} && (2.77)\\ &= ps[qh + p - 2pqh - p^2]/\overline{w} && (2.78)\\ &= ps[qh + pq - 2pqh]/\overline{w} && (2.79)\\ &= pqs[h + p - 2ph]/\overline{w} && (2.80)\\ &= pqs[h(1-2p) + p]/\overline{w} && (2.81)\\ &= pqs[h(q-p) + p]/\overline{w} && (2.82)\\ &= pqs[p(1-h) + qh]/\overline{w} && (2.83) \end{aligned}$$

which gives us the desired result.

[238]We assume that h is in the range of $[0, 1]$; in the next chapter we'll discuss balancing selection, which can happen when h is outside the range $[0, 1]$. Also note that $\overline{w}$ is positive under reasonable conditions.

[239]Overview of card counting: [Link], and an example of a card-counting technique: [Link]. And a classic movie scene about counting cards from Rain Man: [Link].

[240]To be more precise, if the allele is at frequency p, selection would add or remove $2Nsp$ copies in expectation. So for a common allele this is of order 1.

[241]A second intuition for why $2Ns = 1$ represents the lower bound for selection is that the expected change in allele frequency ($\mathrm{E}(\Delta_p)$ due to selection is on the order of $sp(1-p)$, while the variance in allele frequency due to drift ($\mathrm{Var}(\Delta_p)$ is $p(1-p)/2N$. So the expected change due to selection trumps the change in variance when $2Ns >> 1$.

[242]A nice description of the math for the haploid case is given by Otto and Whitlock (1997). Otto and Whitlock also point out that the fixation rate of new mutations is much higher in growing populations, and this is probably important in some ecological settings. See also Pritchard et al (2010) for further discussion of these issues:

Otto SP, Whitlock MC. The probability of fixation in populations of changing size. Genetics. 1997;146(2):723-33

Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology. 2010;20(4):R208-15

[243]Kimura M. Some problems of stochastic processes in genetics. The Annals of Mathematical Statistics. 1957:882-901

Kimura M. On the probability of fixation of mutant genes in a population. Genetics. 1962;47(6):713

[244]For strong positive selection, if the alleles are lucky enough to reach more than a handful of copies then the deterministic dynamics take over, and this randomness at very low numbers is independent of N. In fact the dynamics at very low sample numbers are often modeled as branching processes, ignoring the total population size. When $s > 0$, the branching process either goes extinct quickly or goes to infinity (i.e., fixation).

[245]You may be wondering what happened to the distinction between census population size N and effective population size N_e. I've been focusing on the ideal Wright-Fisher model where they are the same. For more general models both can matter: the initial frequency of a mutation depends on N (i.e., it is $1/2N$), but the rate of the drift depends on N_e. It's worth noting that N_e is a useful hack that gives us insight into complicated models, while not always being a perfect approximation. For example, fixation probabilities of advantageous alleles can be dramatically different with population size changes in a way that is not modeled by the neutral N_e. You can see this by noting that exponential growth (which is not well-modeled by a single N_e) gives new mutations a big boost; the same will be true to a smaller extent even with fluctuating population sizes (where N_e is traditionally computed as the harmonic mean of N); see Otto and Whitlock (1997). Meanwhile, Simons et al explored the interactions between selection, drift and population size changes, and found complicated effects on genetic load:

Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nature Genetics. 2014;46(3):220-4.

[246]The theoretical prediction for the number of sites at frequency p given mutational input $4N\mu$ is

$$4N\mu \frac{1 - e^{-2\gamma(1-x)}}{(1 - e^{-2\gamma})p(1-p)} \tag{2.84}$$

where $\gamma = 2Ns$. You can find derivations for this leading up to Equation 11 of Sawyer and Hartl (1992), and Equations 33 and 35 in the review by Senupathy and Hannenhalli (2008):

Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics. 1992;132(4):1161-76

Sethupathy P, Hannenhalli S. A tutorial of the poisson random field model in population genetics. Advances in bioinformatics. 2008;2008

[247]Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proceedings of the National Academy of Sciences. 2005;102(22):7882-7

[248]Recall from Chapter 2.2 that the SFS can be used to estimate population histories. Since the SFS is also influenced by selection, the demographic analysis would usually be restricted to putatively neutral sites, such as synonymous or noncoding sites.

[249]For real data we don't (yet) know the actual selection coefficients for most types of sites, but it's common to use synonymous and noncoding sites as proxies for a more-neutral baseline. While these sites may occasionally have functional effects such as altering splicing or transcription factor binding, they usually have little selection compared to coding sites.

[250]Note: It's not entirely clear why the noncoding sites have fewer singletons than synonymous in this analysis. I suspect it may reflect differences in sequence composition and mutation rates between exons and noncoding regions rather than major differences in functional constraint

Harpak A, Bhaskar A, Pritchard JK. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genetics. 2016;12(12):e1006489).

[251]If we see a common variant at a site then we can be confident this site is not under selective constraint. But even neutral sites generally don't have common variants so this test lacks sensitivity. However, there are new approaches that can detect strong selection in very large samples:

Agarwal I, Przeworski M. Mutation saturation for fitness effects at human CpG sites. Elife. 2021;10:e71513

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv. 2022:2022-03

[252]These methods are no longer as widely used for predicting gene regulation as recent improvements in functional genomics are far more interpretable, including providing cell-type specific information. Nonetheless the general principles are still important.

[253]Menke DB, Guenther C, Kingsley DM. Dual hindlimb control elements in the Tbx4 gene and region-specific control of bone size in vertebrate limbs. Development. 2008

[254]e.g., Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008;134(2):341-52.

[255]Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics. 2006;7(2):98-108

Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Molecular biology and evolution. 2008;25(3):568-79

Hershberg R, Petrov DA. Selection on codon bias. Annual review of genetics. 2008;42:287-99

Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, et al. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Molecular biology and evolution. 2018;35(5):1092-103

[256]Platt RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. Chromosome Research. 2018;26:25-43

[257]Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philosophical Transactions of the Royal Society B. 2020;375(1795):20190347

[258]de Koning AJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics. 2011;7(12):e1002384

[259]Deininger P. Alu elements: know the SINEs. Genome biology. 2011;12(12):1-12

[260]Deininger PL, Batzer MA. Alu repeats and human disease. Molecular genetics and metabolism. 1999;67(3):183-93

[261]There is some tiny cost from the fact that it has to be copied every time the cell divides: the nucleotides, the energetic cost, and the copying time. If the Alu inserts inside an intron, it is must also be transcribed every time the gene is transcribed. Pairs of nearby Alu elements also occasionally trigger incorrect chromosome pairing and recombination

Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, et al. Human genomic deletions mediated by recombination between Alu elements. The American Journal of Human Genetics. 2006;79(1):41-53

Kim S, Cho CS, Han K, Lee J. Structural variation of Alu element and human disease. Genomics & informatics. 2016;14(3):70. Another potential issue arises from inverted Alu repeats in mRNA can form double stranded RNA (dsRNA). Since dsRNA is a hallmark of some viruses (and not ordinarily present in human mRNA), this can trigger an inappropriate (auto)immune response. There is an entire machinery evolved to edit dsRNA to reduce double-strand pairing

Chung H, Calis JJ, Wu X, Sun T, Yu Y, Sarbanes SL, et al. Human ADAR1 prevents endogenous RNA from triggering translational shutdown. Cell. 2018;172(4):811-24

[262]e.g., Yang F, Wang PJ. Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline. In: Seminars in cell & developmental biology. vol. 59. Elsevier; 2016. p. 118-25.

[263]Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proceedings of the National Academy of Sciences. 2012;109(45):18488-92

Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nature Reviews Genetics. 2016;17(11):704-14

[264]To get a ballpark estimate, let's suppose that mutations in 1% of the genome would have an average deleterious effect on fitness of 10^{-3}. Assuming these numbers, each new mutation in the genome produces an average fitness cost of 10^{-5}, per generation (usually zero, and occasionally much higher, depending on where the mutation lands). There's an additional complication which is that the precise selective effect that a mutator allele experiences as the result of the mutations it produces is slightly more complicated because it can experience those effects over multiple generations. However in a recombining organism, it recombines away from the damage it produces at a rate of $1/2$ per generation. Lynch et al (2016) give the fitness effect of a mutator allele as being $\approx 2s\Delta(U_D)$, where s is average fitness effect of a new mutation, $\Delta(U_D)$ is the change in genome-wide mutation number caused by the mutator, and the factor of 2 reflects the average number of generations that the mutator is in the same genome as the mutations it causes. (Lynch 2016)

[265]For examples of mutator evolution in action see e.g.,

Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. Elife. 2017;6:e24284

Sasani TA, Ashbrook DG, Beichman AC, Lu L, Palmer AA, Williams RW, et al. A natural mutator allele shapes mutation spectrum variation in mice. Nature. 2022;605(7910):497-502

[266]Kondrashov AS. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? Journal of theoretical biology. 1995;175(4):583-94

[267]One hypothesis is that protein evolution involves a lot of weakly deleterious substitutions that are repaired by very slightly advantageous compensatory mutations that maintain overall function.

2.6 Natural selection: II. Positive selection and adaptation

The previous chapter gave an introduction to the basic model of fitness and selection, and the role of purifying selection. Here we explore positive selection in greater detail, illustrated with key examples in humans.

Figure 2.94: **Inuit seal hunter in Noatak, Alaska, 1929.** *During the past* 70,000 *years, humans successfully colonized a wide range of different ecosystems, including arctic regions, deserts, and tropical rainforests.* *Credit: Edward Curtis 1929. [Link], Public Domain.*

Positive selection and adaptation. We now come to the type of selection that is arguably the most interesting and – as we discuss in the next chapter – the most argued-about form of selection: positive selection in favor of advantageous phenotypes and alleles.

Positive selection is the central organizing force of evolutionary change. It drives populations to adapt to their environments, and over longer evolutionary timescales it drives the evolution of new forms and functions at all levels: for example, the emergence of multicellular eukaryotes and of animals; the transition from fish to amphibians that enabled vertebrates to move onto land; the evolution of primates, of apes, and of humans.

Key evolutionary changes in the human lineage include the transition to bipedalism; bigger brains; changes in body size and shape, musculature, body hair and so on; enhanced capacity for language and highly complex social structures. All of these changes are genetically encoded and were presumably driven, at least in part, by positive selection. Moreover, we now have many examples where aspects of these genetic transitions have been elucidated, although there is still much more to be learned.

In more recent human evolution, during the last ∼70,000 years, humans have spread around the globe to inhabit nearly all the world's land masses and ecosystems. Prehistoric humans successfully colonized a huge range of **environments**: environments with extreme cold and ice, or extreme heat and humidity; high altitude in Tibet, the Andes, east Africa, and elsewhere; tropical rainforests; deserts. Humans subsist on a wide range of **foods**, and encounter a diversity of infectious **pathogens**. All of these factors must have exerted strong selective pressures on human populations, driving both genetic adaptations–as well as cultural adaptations such as innovations in clothing, hunting and agriculture [268] [a].

[a] *William Clark, of the Lewis and Clark expedition, marveled at the ability of the native Mandan people (in present day North Dakota) to survive extreme cold: "This morning a boy of 13 years of age Came to the fort with his feet frozed, haveing Stayed out all night without fire, with no other Covering than a Small Robe goat skin leagens & a pr. Buffalow Skin mockersons– The Murcery Stood at 72° below the freesing point– Several others Stayed out all night not in the least hurt, This boy lost his Toes only...those people has ancered to bare more Cold than I thought it possible for man to indure."–Jan. 10, 1805.* *Credit: Journals of the Lewis and Clark Expedition: [Link].*

Genetic adaptation proceeds mainly through two general types of processes: **selective sweeps** and **polygenic adaptation**. In a selective sweep, selection drives a strongly advantageous allele from low to high frequency in a population. In contrast, polygenic adaptation is driven by small shifts in allele frequencies spread across many loci, and is most relevant for complex traits. In practice, these two models are extremes along a spectrum, and adaptation may often proceed through a mixture of both types of processes.

Here we describe the features of both types of adaptation, as well as a third model, **balancing selection** which can drive both short term direc-

tional selection, and long-term stable polymorphism.

Signatures of sweeps in genome data. We've already covered basic models of positive selection in the last chapter. But, in practice, how can we find signals of positive selection in data?

The key insight here is that *strong positive selective drives very rapid allele frequency changes that would be extraordinarily unlikely for a neutral allele meandering under the random effects of genetic drift*. The next plot illustrates this for simulated data:

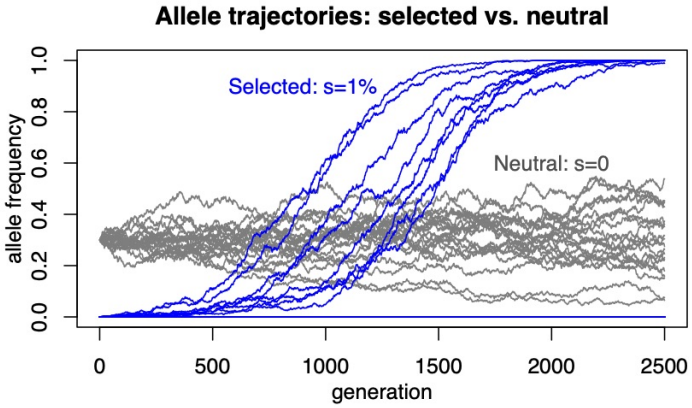


Figure 2.95: **Rapid increase in allele frequency for favored alleles versus neutral.** *This simulation compares trajectories for favored alleles with $s = 1\%$, starting from new mutations, compared to common neutral alleles. Simulations: 1000 favored alleles with $s = 1\%$ each starting at a frequency of $1/2N$ at time 0, versus 20 neutral alleles, starting at frequency 0.3. Only about 1% of the favored alleles spread to fixation; the remainder are on the $y = 0$ line. Population size $N = 10,000$.*

As you see above, the neutral alleles drift along aimlessly, while a favored allele rushes toward fixation [269].

Crucially, *this very rapid change in allele frequency distorts patterns of genetic variation in a large region in predictable ways.* As the selected variant spreads rapidly through the population, it drags a haplotype up to high frequency along with it. This means that nearby neutral variants on the same haplotype are also dragged up to high frequency, in a process known as **genetic hitchhiking** [270]:

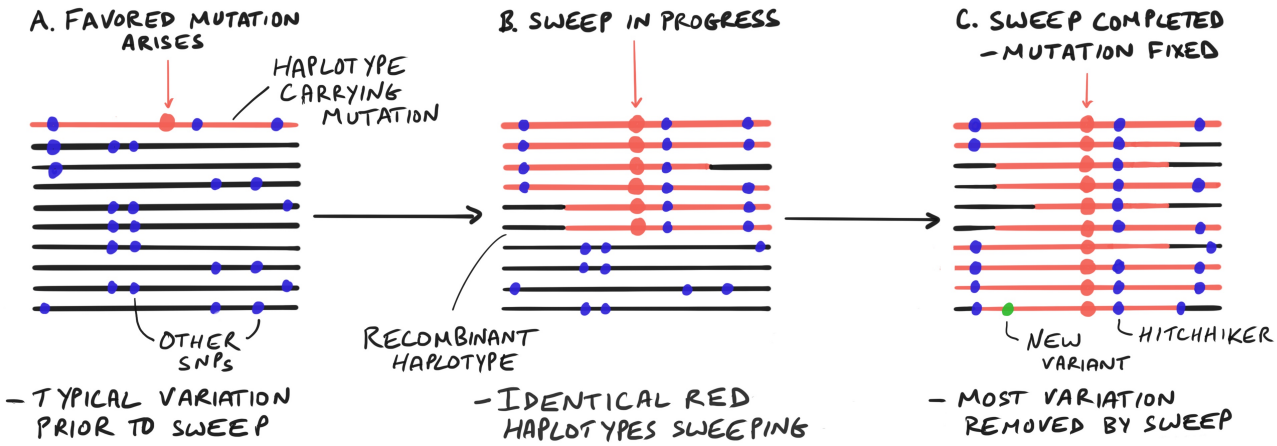


Figure 2.96: Sweeps reduce variation in a linked region. *When a new favored mutation spreads through the population, it increases frequency very rapidly, and drags a long haplotype to high frequency along with it. Red indicates the haplotype on which the favored mutation occurred, as well as its descendants; to some extent this gets whittled down by recombination (black segments). We cannot observe the red versus black coloring directly, but we can infer this from the haplotype structure.*

While the sweep is in progress the favored allele sits on a long, nearly identical haplotype. This contrasts markedly with relatively normal haplotypes carrying the ancestral allele [271]. Next, as the sweep completes, it essentially wipes out variation in a window around it, aside from any rare variants that arose during the sweep.

The size of the affected region depends on the speed of the sweep versus the local rate of recombination. A very fast sweep (large s) carries a large haplotype with it, simply because recombination does not have time to chop it down very far; similarly, the sweep region would also be larger in regions with a low recombination rate. The reduction in heterozygosity as a function of distance x from a selected site can be approximated by $1 - e^{-\tau r x}$ where $\tau = 2log(2N)/s$, and r is the local recombination rate [272]. Notice that the size of the swept region depends on the ratio r/s:

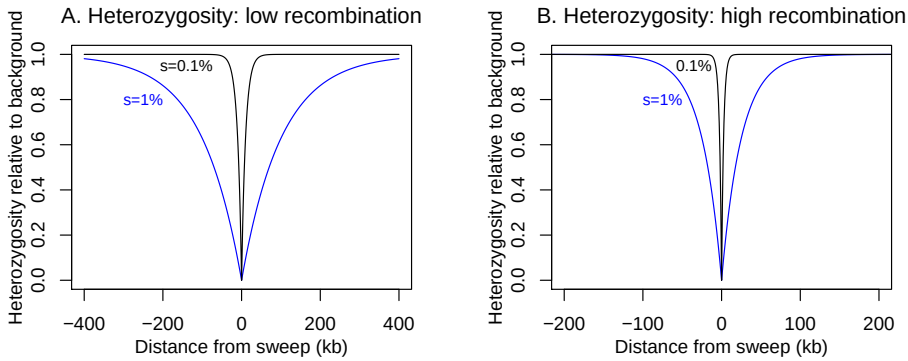


Figure 2.97: **Reduced genetic diversity around the site of a sweep due to hitchhiking.** *The plots show the expected reduction in heterozygosity relative to the average heterozygosity at neutral sites, assuming a recently completed sweep at position zero. The size of the affected region depends on s and on the local recombination rate.* *Parameters:* $r = 0.5 \times 10^{-8}/bp$ *(left plot), and* $2 \times 10^{-8}/bp$ *(right plot);* $N = 10,000$. *Figure inspired by Coop (2020) [Link], Fig. 13.5.*

Starting from this intuition about how sweeps impact patterns of variation, there has been a huge amount of work on methods for detecting selective sweeps using genetic data [273]. In short, these methods use a variety of features in the data to detect positive selection:

- *long haplotypes with low genetic variability around a putative selected allele*; these contrast with typical patterns of variation on haplotypes carrying the ancestral allele at the selected site (ongoing sweep);
- *low genetic diversity in a genomic region around a recently fixed site* (recently completed sweep);
- *most variants in a region are young and at low frequency* (recently completed sweep);
- *large allele frequency differences at a selected site, and potentially nearby sites*, between populations where the sweep is occurring compared to control populations, or in time series data from ancient DNA (ongoing or completed sweep)

We illustrate these principles using several examples of recent selection in humans, highlighting some of the key selective pressures as well as signals in the genetic data.

Selection pressures due to diet. Diet has been a major driver of selection in recent human evolution. As humans spread around the globe to inhabit virtually all possible ecosystems, they were forced to learn to survive on a wide array of different foods. Further enormous shifts in diet were driven by the transition to **agriculture**, starting in the past 5,000-10,000 years, in many parts of the world. Several potential signals of selection have been hypothesized as relating to diet, including at the FADS locus, which is involved in metabolism of fatty acids [274] and at Amylase1, which is involved in starch digestion [275].

The clearest diet-related signal is at the **lactase** locus. Lactase is the enzyme that is responsible for digesting the sugar lactose, which is present in milk. Most mammals stop consuming milk (and lactose) after weaning, and expression of the lactase gene is generally turned off in adults.

The first known evidence for dairy farming is in Anatolia (modern day Turkey) in the early Neolithic, about 9,000 years ago. Dairy farming subsequently became important in many places, including in Europe, in India and the Middle East, and in east Africa. This, in turn, provided a strong selective pressure for early humans to be able to digest milk throughout life. Consequently, several different **regulatory mutations** that cause the lactase gene to be expressed throughout life have spread to intermediate or high frequency in different farming populations. These regulatory mutations are often referred to as **lactase persistence** alleles as they cause lactase to persist throughout life.

The strongest signals of selection on lactase are found in Europe. As it happens, there are now extensive genotype data from early European populations, collected from skeletons [b]; this allows a rare opportunity to track the selective spread of an allele directly using allele frequencies in ancient DNA. Analysis by Iain Mathieson shows the remarkable spread of the lactase allele during the past 5,000 years to a modern frequency above 80% in some parts of Europe:

Figure 2.98: **Maasai herder, with cattle.** *The lactase locus has been a target of selection in the Maasai and other east African farming populations.* *Credit: Nicor: [Link]. CC BY 3.0.*

[b] *We'll cover ancient DNA in Chapter 3.3.*

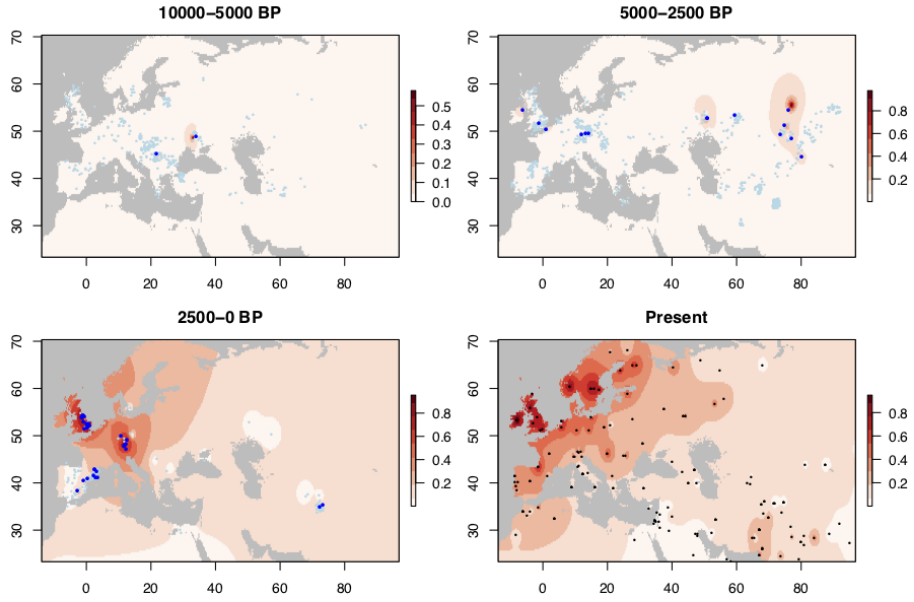


Figure 2.99: **Increase in allele frequency of the lactase variant in Europe (from ancient DNA).** *The maps indicate locations of samples with ancestral alleles (light blue) and derived alleles (dark blue) in each time period (BP = years Before Present); as well as spatially smoothed maps of the derived allele frequency in each period (red color scale beside each panel).* *Credit: Iain Mathieson blog post (2019) [Link]. Used with permission of the author.*

Recall that the rate of spread of a favored allele depends on its selective advantage s. Using the data shown above, Mathieson estimated s for the lactase persistence allele at 2%–3% in central and western Europe, consistent with other estimates using haplotype patterns in modern data [276]. This makes lactase persistence one of most strongly selected traits in recent human evolution [277].

Signals of selection at lactase are also found in other dairy farming populations. For example, dairy farming was practiced in east Africa by around 6,000 years ago and is a major food source for several east African groups [278]. As you can see below, the lactase locus shows strong signals of a sweeping haplotype in Tanzania: the derived allele sits on a long shared haplotype, in sharp contrast with the much higher diversity on ancestral haplotypes [279]:

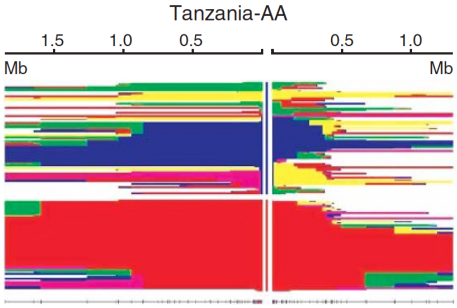


Figure 2.100: **Signal of a partial lactase sweep in a Tanzanian population.** *Visualization of haplotype sharing around a sweep-in-progress. The derived allele at the likely selected site is indicated in red at the center of the plot, and the ancestral allele in blue. The x-axis indicates location within the region in Mb; rows are haplotypes.* ***Shared blocks of color indicate shared haplotypes extending outward from the selected site.*** *Credit: From Figure 5, Sarah Tishkoff et al (2007) [Link] Used with permission.*

The selected variants in African populations are distinct from the European variant, indicating that they arose from independent mutations, rather than being carried in by migration.

Selection on pigmentation: SLC24A5. Another important target of natural selection in human evolution is on **skin pigmentation**, as well as hair and eye coloring. Globally, populations that live close to the equator, including in our ancestral range in Africa, tend to have darker pigmentation. Populations at higher (and lower) latitudes tend to have lighter pigmentation [280].

Variation in pigmentation seems to have been driven by strong selective pressures [281]. In regions with intense sunlight it is advantageous to have darker skin as this protects against ultraviolet (UV) damage from the sun. In addition to skin damage, excess UV radiation also degrades folic acid, deficiencies of which cause neural tube defects during pregnancy. However, too little UV is also bad, as UV catalyzes Vitamin D production; Vitamin D plays an important role in bone development, reproductive health, and other traits.

Remarkably, around ten different genes involved in pigmentation of skin, hair, or eyes show either clear or suggestive signals of selection in some part of the world [282]. In the case of related traits such as blue eyes and blond hair with signals of selection in Europe, it's unclear if UV, or some other factor such as sexual selection, was the main driver of selection.

Among the most striking pigmentation signals is a *missense variant at the*

gene SLC24A5. The derived allele (in red), which causes lighter pigmentation, has swept to high frequencies throughout western Eurasia.

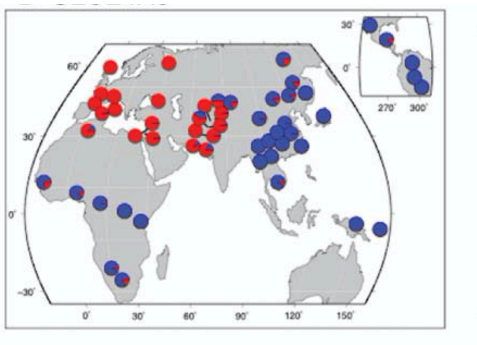


Figure 2.101: **Selective sweep in western Eurasia at SLC24A5.** *The derived missense allele is shown in red.* *The populations, for example in the Americas, are representatives of indigenous populations. From Graham Coop et al 2009, Figure 2B. CC BY 4.*

The allele frequency differences between populations at SLC24A5 are among the most significant frequency differences anywhere in the genome, reflecting very strong selection for the derived allele [283].

As expected for a recently completed sweep, this event has swept away genetic variation in a large region around SLC24A5 in Europeans (red line in panel A); this contrasts sharply with more typical levels of variation in other populations. The role of SLC24A5 in pigmentation is supported by human association data, and a zebrafish knockout:

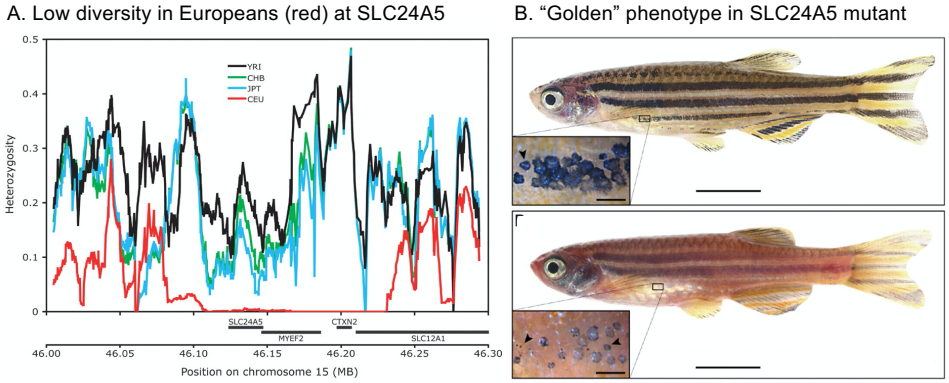


Figure 2.102: **Selective sweep at the pigmentation gene SLC24A5. A.** *Near-zero genetic diversity in a European population (red, CEU) near SLC24A5. An African (YRI) and two east Asian (CHB, JPT) populations have more normal patterns of genetic diversity across this region.* **B.** *Mutations in SLC24A5 cause changes in melanophore coloring, as seen in a zebrafish mutant ("golden") at bottom, versus the wild type at top.* *Heterozygosity in Panel A is measured at pre-ascertained SNPs. From Figures 5A, 1A, B. Rebecca Lamason et al (2005) [Link] Used with permission.*

Together, the lactase and SLC24A5 examples illustrate classic features of selective sweeps: rapid changes in allele frequencies at selected sites; large sweeping haplotypes for sweeps in progress (lactase); removal of variation in regions around completed sweeps (SLC24A5).

But as we shall see next, not all sweeps show these characteristics.

Soft sweeps. What would happen if a mutation is not immediately favored: instead it drifts along for a while, and only becomes favored some time later, after an environmental change? In this case, during the drift phase, the favored allele potentially has time to recombine onto multiple haplotype backgrounds. Then, when it *does* sweep, it carries multiple haplotypes with it, and the overall footprint of selection is greatly

reduced. This scenario may seem contrived, but there are many phenotypes that are favored only in specific environments, for example in the presence of a pathogen, a specific food source, or at high altitude – and otherwise neutral [284] [c].

A second scenario that could reduce the sweep signal is if multiple functionally equivalent mutations arise at about the same time and sweep together. These would likely occur on different haplotypes and the sweep signal would be greatly reduced.

In terminology developed by Pleuni Pennings and Joachim Hermisson in 2005, both of these scenarios are described as **soft sweeps** [285]. This term evokes the image of a mutation, or mutations, that sweep to fixation, without greatly disturbing the variation at nearby sites. This contrasts with the classical sweep model described above, which we now refer to as a **hard sweep**.

[c] *For more about high altitude adaptation see EPAS1 in Chapters 2.4 and 3.3.*

Figure 2.103: Soft sweeps have minimal impact on variation in the linked region. *In one type of soft sweep, the mutation is initially neutral, and drifts to low or intermediate frequency in the population. During this time, recombination shuffles it onto multiple haplotypes. Selection then turns on, driving the allele to fixation, but without a strong hitchhiking effect.*

Selection for malaria resistance: Duffy. One likely soft sweep occurred at the **DARC gene**, which encodes a cell-surface protein named **Duffy**, found on the surface of red blood cells. Duffy serves as a cell surface receptor for a class of chemokines, a type of signaling molecule [286].

Duffy also plays a critical role in **malaria** infection and resistance. One species of malaria, **Plasmodium vivax**, binds the Duffy protein on the surface of red blood cells and uses it to enter cells. This property leads to a truly remarkable story of selection at the Duffy locus.

Most sub-Saharan Africans carry a derived variant near the Duffy locus that disrupts a DNA binding site for the transcription factor GATA1. GATA1 plays an important role in erythroid (red blood cell) development, and loss of this particular binding site eliminates Duffy expression specifically in erythrocytes while maintaining Duffy expression in other cell types [287]. This variant is known as the *Duffy null* allele (abbreviated FY*0). The lack of Duffy expression in the null allele has the crucial effect

of blocking entry of vivax malaria into red blood cells. Thus Duffy null individuals are resistant to vivax malaria [288] [289].

The next striking fact about Duffy null is that it shows extreme frequency divergence between populations: it is essentially fixed in most sub-Saharan African populations, and essentially absent outside Africa. The null variant has the largest population differentiation of any high-frequency African allele, anywhere in the genome [290] (green allele in Panel A, below).

This might make you think that Duffy has been the target of a completed sweep in most of Africa: a sweep that presumably started after the major out-of-Africa migrations around 70,000 years ago. However studies of genetic variation show something surprising: namely that the Duffy null allele is carried on two distinct major haplotypes, with additional low frequency variants (maroon circles in Panel B):

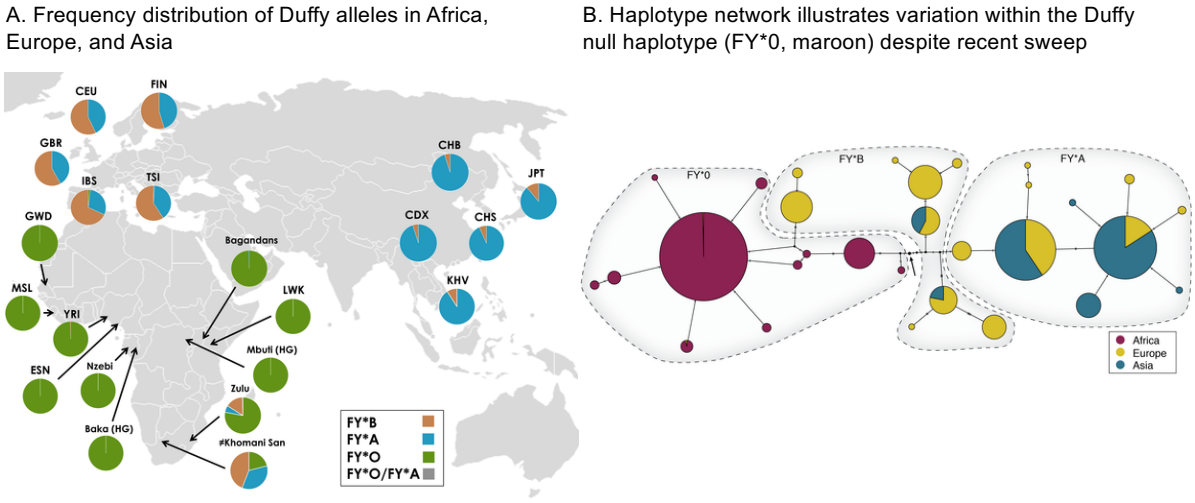


Figure 2.104: Soft sweep of the Duffy null allele in sub-Saharan Africa. A. *The Duffy null allele, shown here in green, swept to near fixation in many African populations subsequent to the out-of-Africa migrations. Non-African populations have a mix of A and B alleles (brown and red), corresponding to alternate missense variants. The B allele is ancestral.* **B.** *Visualization of SNP variation seen on each of the three allelic backgrounds in a 5kb region around the FY*0 site. Each circle represents a different haplotype, and its area is proportional to frequency. Line segments between the haplotype circles indicate similarity. Notice that haplotype variation on the FY*0 haplotype is nearly as high as on the A and B backgrounds, arguing against a recent hard sweep at this locus.*

Credit: Figs 1, 2a from Kimberly McManus et al (2017) [Link]. CC BY 4.0

Instead, the high level of variation on the Duffy null background suggests that this is a soft sweep. One study estimated that two major Duffy null haplotypes were drifting at low frequency (0.1%) prior to the onset of strong selection around 45,000 years ago, at which point the Duffy null allele swept to fixation [291].

There's one more surprising twist in this story, namely that Plasmodium vivax malaria is mainly found in Asia and Latin America – not in Africa! (The main malaria parasite in Africa is a different species, P. falciparum.) So why was there such strong selection at the Duffy locus, specifically in Africa?

Recent work shows that P. vivax is currently found in African chimpanzees

and gorillas [292]. So one plausible model is that vivax malaria jumped into humans around 45,000 years ago, and drove strong selection on pre-existing variation at the Duffy locus. Subsequent fixation of the Duffy null allele provided such powerful disease resistance that the human adaptation ultimately eliminated vivax malaria from human populations in Africa!

Infectious diseases as major drivers of selection. Malaria has long been a major cause of global disease and mortality [293]. As such, malaria has exerted strong selective effects on several other genes in addition to Duffy, including α-globin, β-globin, and G6PD, coming up next.

Beyond malaria, infectious diseases in general are potent agents of natural selection. Like all species, we are continually barraged by a range of pathogens – viruses, bacteria, fungi, protists – that evolve rapidly to outwit our inbuilt defenses. During active infections, our bodies combat pathogens using a mixture of so-called innate and adaptive immune systems. While outside the scope of this book, it's interesting to note that during an infection our **adaptive immune systems** harness the principles of evolution, including genome modification and proliferation, to rapidly evolve B and T cells that recognize the infectious agents. This allows our own immune systems to adapt on the same timescales as rapidly evolving pathogens.

Moreover, the biological systems that combat infections are, themselves, finely tuned by evolution. As a result, several examples of selection in humans relate to pathogens and immunity: in addition to the malaria examples these include the MHC system which we discuss in the next chapter, the Toll-like receptor complex involved in innate immunity, and others.

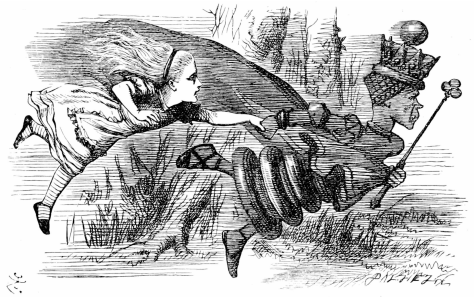


Figure 2.105: **The Red Queen Hypothesis** *proposes that pathogens and their hosts are constantly evolving molecular systems to counter each in infection and defense. The name alludes to Lewis Carroll's Red Queen in Through The Looking Glass who tells Alice that "**here, you see, it takes all the running you can do, to keep in the same place.**"* *Credit: Illustration by Sir John Tenniel, 1871. Public Domain.*

Balancing selection. So far we have been focusing on positive selection that always favors one allele over the other. But what happens if the heterozygote fitness is higher than both homozygotes, or lower than both homozygotes?

To be more precise, recall our fitness model from the previous chapter, where the three genotypes have fitness 1, $1+hs$, and $1+s$, respectively. We have been looking at models where h is in the range $[0,1]$. But if h is outside this range, then the heterozygotes are either better, or worse, than both homozygotes, leading to some very interesting models.

To understand this, recall from the previous chapter that we computed the expected change in allele frequency, Δ_p, from one generation to the next:

$$\Delta_p = \frac{pqs[p(1-h)+qh]}{\overline{w}}. \tag{2.85}$$

Δ_p tells us how allele frequency p changes over time. If h is in the range of $[0,1]$, then we get simple directional selection: Δ_p is always increasing (if s is positive) or always decreasing (if s is negative). But if h is out-

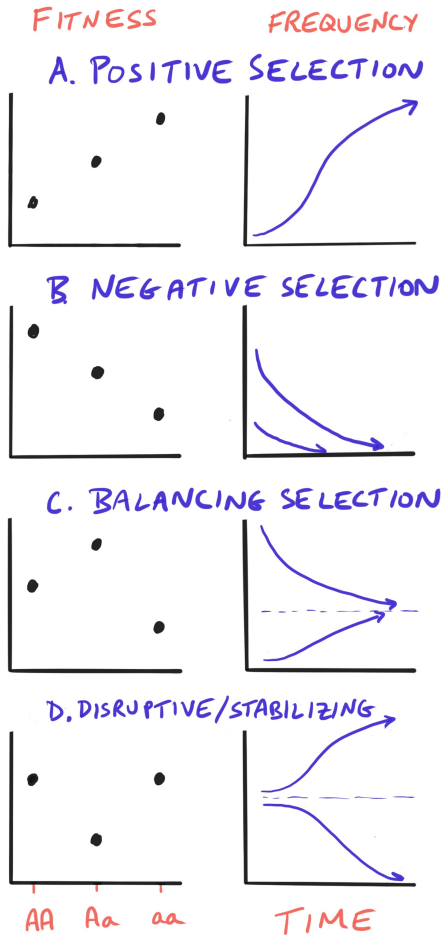


Figure 2.106: **Overview of models.** ***Left:*** *Relative genotype fitnesses (y-axis), for AA, Aa, and aa genotypes.* ***Right:*** *Simplified frequency trajectories (y-axis) of the a allele over time.*

side this range, then the direction of change depends on allele frequency. There are two main scenarios, shown here:

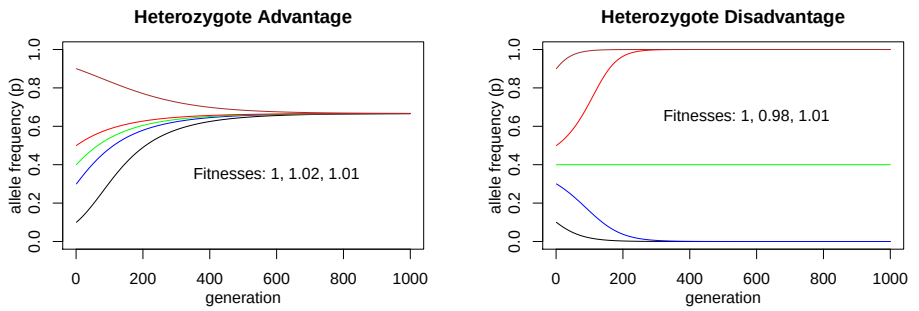


Figure 2.107: **Allele frequency trajectories from different starting points. Left:** *With heterozygote advantage, the allele trajectories converge to a stable equilibrium.* **Right:** *With heterozygote disadvantage, the trajectories diverge away to fixation or loss depending on the starting frequency. In the absence of drift, the green trajectory stays precisely at the unstable equilibrium.* *Parameters: $s = 0.01$, $h = 2$, $\hat{p} = 2/3$ (left) and $s = 0.01$, $h = -2$, $\hat{p} = 0.4$ (right).*

In the left plot, the heterozygote is fitter than both homozygotes, and the allele frequencies converge toward a fixed stable point. This is known as **heterozygote advantage** and leads to **balancing selection**.

When the heterozygote is less fit than both homozygotes, the population evolves toward fixation of one allele or the other, depending on the starting point. This is called **heterozygote disadvantage** or **disruptive selection**.

In both cases we can solve for an **equilibrium** frequency – i.e., a value of p for which the allele frequencies don't change under this model [294]:

$$\hat{p} = \frac{h}{2h - 1}. \tag{2.86}$$

(You can see here that when h is inside the range $[0, 1]$ – i.e., directional selection – this equation does not produce meaningful allele frequencies between 0 and 1.)

We can also look at this with drift. With balancing selection (left), the population converges to the equilibrium and stays there. But when the heterozygote is less fit (right), we see that $\hat{p}$ is an **unstable equilibrium**. As soon as the green trajectory drifts slightly away from $\hat{p}$, selection pushes it rapidly toward fixation or loss.

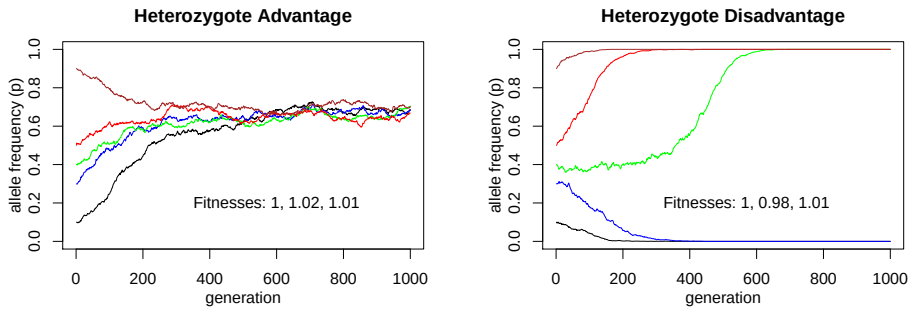


Figure 2.108: **Simulated frequency trajectories with drift.** ***Left****: even with drift, heterozygote advantage allows stable balanced polymorphism that can persist for long timescales.* ***Right****: Drift causes trajectories to move away from the unstable equilibrium; in this case both fixation or loss are possible for the green trajectory.* *Parameters as above, and $2N = 20,000$.*

To summarize: if the heterozygote has higher fitness than both homozygotes, then this leads to a stable polymorphic equilibrium. In practice this can last for many millions of years.

But if the heterozygote is worse than both homozygotes, then there is an unstable polymorphic equilibrium, which can result in fixation of either allele depending on the starting allele frequency. This type of selection is usually difficult to detect in practice [295]. However, a similar model will become important when we study **stabilizing selection** later in the book [d]. In that case the unstable equilibrium is at 0.5 and selection acts against minor alleles [296].

[d] *Stabilizing selection is likely the main form of selection acting against variation in gene expression and many complex traits (Chapter 4.8).*

Balancing selection examples. Perhaps the best-known example of balancing selection is for **sickle cell disease**. Sickle cell disease is caused by a missense mutation in the β-globin gene (also known as *HBB*), which encodes a major subunit of hemoglobin, the molecule that red blood cells use to transport oxygen [297] [298]. In heterozygotes, the sickle cell mutation provides strong defense against both vivax and falciparum malaria without major side effects. However, individuals who are homozygotes for the missense mutation suffer devastating symptoms including hemolytic crisis, severe pain, kidney disease, and stroke.

The protective nature of the sickle allele was first documented in the 1950s [299]. A recent global study confirms that sickle heterozygotes benefit from extremely strong protection against malaria (odds ratio of 0.14, p-value=10^{-225}), left panel below [300]:

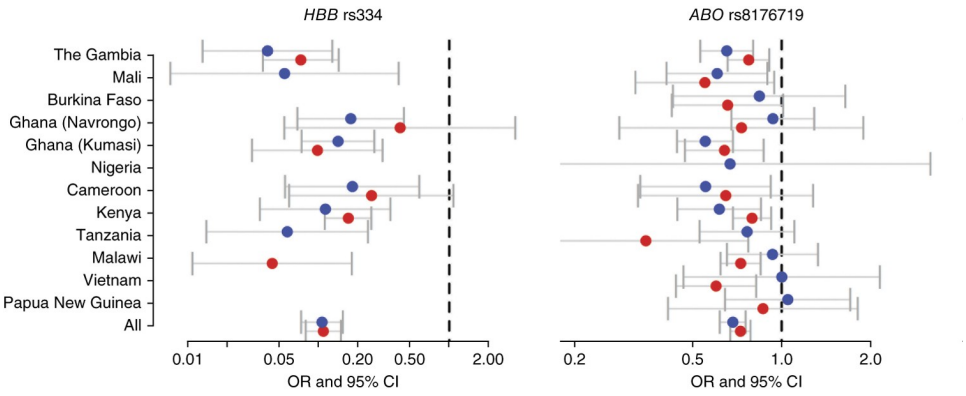


Figure 2.109: **Protective effects of HBB sickle and ABO type-O against malaria.** *The x-axes show estimated odds ratios for the risk of cerebral malaria (red) and severe malarial anemia (blue), for samples from different countries. Estimates to the left of the vertical dashed lines indicate protective effects.* *The odds-ratio is (approximately) a measure of the risk for individuals with this genotype compared to controls; values < 1 indicate protection relative to controls. Grey bars show confidence intervals. Credit: From Figure 1, Malaria Genomic Epidemiology Network (2014) [Link] Used with permission.*

The sickle cell allele is common in most of central/western Africa, peaking at around 15% frequency in Angola [301]. Based on this we can estimate the frequency of sickle cell disease (i.e., sickle homozygotes) using the Hardy-Weinberg rule, as $\sim$2.25% (i.e., $0.15^2 \times 100$).

We can use Equation 2.86 to estimate h. Rearranging that expression gives us $h = p/(2p-1)$: hence h=$-$0.21. Assuming $s \approx -1$, this implies a heterozygote fitness of 1.21, which is an extraordinarily large fitness effect for humans [302].

Several other alleles are protective against malaria in heterozygotes but cause disease in homozygotes. In addition to the "classic" sickle cell missense mutation, a variety of other rarer mutations affect either the **ff**-globin or **fi**-globin genes, and produce varying levels of sickle cell-like disease. These diseases are referred to as **α- or β-thalassemia**. Like the sickle cell mutation, they are mainly found in individuals with ancestry

from malaria-endemic regions, and are also likely spread by heterozygote advantage. Lastly, mutations in the enzyme **G6PD**, which plays a role in glucose metabolism, are also protective against malaria in heterozygotes but cause pathologies in homozygotes. Balancing selection likely maintains variation in all these genes [303].

Ancient trans-species balancing selection. There's one more unique feature of balancing selection: if the selective pressures are stable, they can maintain polymorphism for extremely long times.

One of the best examples of this is the **ABO gene**, which is responsible for the ABO blood groups. ABO is an enzyme that modifies sugar attachments on cell surface proteins called glycoproteins. Two functional alleles, A and B, differ by a pair of missense variants that lead to different glycoprotein modifications. The third allele, O, carries a frameshift mutation that obliterates enzyme activity. You are likely familiar with the ABO system in the context of blood donations, as some combinations of blood types are incompatible donors and recipients: this is because unfamiliar glycoproteins can trigger immune reactions against donor blood cells. As shown in the figure above, the O allele at ABO is protective against malaria; more broadly the ABO alleles are associated with many different traits.

Curiously, it turns out that the ABO alleles are actually shared among different species of apes and old world monkeys. Analysis shows that the alleles from the different species are actually shared in a single ancient coalescent tree, reaching back at least 20 million years, and probably longer [304]! This type of deep, ancestral sharing of alleles is known as a **trans-species polymorphism** and is extremely rare in the human genome overall:

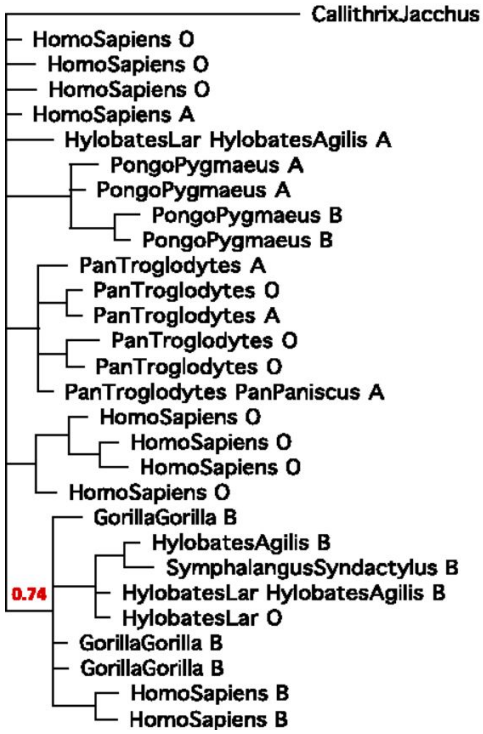


Figure 2.110: **Trans species polymorphism in the ABO blood group system.** *This shows a phylogenetic tree of sequences from different ape species for exon 7 of the ABO gene which determines A/B/O blood type.* ***Notice that most of the B alleles from different species cluster together in the lower part of the tree, indicating that they descend from a shared ancestral mutation.*** *A and O mutations are shuffled together suggesting that O alleles may have arisen repeatedly (although the tree structure cannot be confidently determined in this clade. Species names: Callithrix (marmoset monkey); Homo (human); Hylobates (gibbon); Pongo (orangutan); Pan (chimpanzee); Gorilla (gorilla); Symphalangus (siamang). Credit: Figure 3A from Laure Ségurel et al 2012. [Link]*

The extreme age of this polymorphic system indicates that it cannot be neutral, and must instead be preserved by some form of balancing selection. However, the precise explanation remains mysterious, as there is no obvious advantage to heterozygotes. We do know that cell surface molecules such as the glycoproteins modified by ABO, are frequently used as cellular entry points for pathogens (as for Duffy), and this may explain why blood group O provides some degree of malaria protection. It's possible that the different alleles provide protection against distinct pathogens, and that this creates selection pressure for maintaining diverse alleles. The details remain to be discovered [305] [306].

So far, we have been describing examples where individual alleles are strongly selected. We close this chapter by considering a different mode of adaptation that depends on the joint action of many variants across the genome.

Polygenic adaptation. These examples of strong positive selection are biologically important, and illustrate essential concepts. But they represent a rather special class of selective events: all of these are variants that – on their own – exert major effects on specific traits.

The most dramatic examples of positive selection are usually associated with genes that play some critical, unique role in a selected process. For example, lactase is *the* critical enzyme involved in digestive breakdown on the main sugar in milk. Duffy is *the* critical receptor involved in vivax entry into erythrocytes. SLC24A5 is one of a handful of genes with strong impact on pigmentation and minimal pleiotropic effects.

However, this situation where a single gene plays a central role in a specific trait without major unintended consequences is the exception rather than the rule. Aside from rare genetic diseases, **most phenotypes are highly polygenic** [e]. The inheritance of most traits is due to thousands of variants across the genome, each with only tiny effects on the trait.

This includes most traits that vary in populations, including for example: morphometrics such as height, weight, and body shape; molecular and cellular traits such as hormone levels, lipid levels, or blood cell counts; risk for most diseases, including cardiovascular disease, diabetes, psychiatric conditions; and even behavioral traits.

For these traits, a person's expected phenotype can be modeled as a sum of thousands of pluses and minuses, depending on their alleles at every contributing variant: this is known as a **polygenic score**.

When selection acts on a polygenic trait, the effect of selection is to increase polygenic scores in the population. This occurs mainly through small shifts in allele frequencies, spread across thousands of variants [307]:

[e] *We'll cover the genetics of polygenic traits in much more detail starting in Chapter 4.4.*

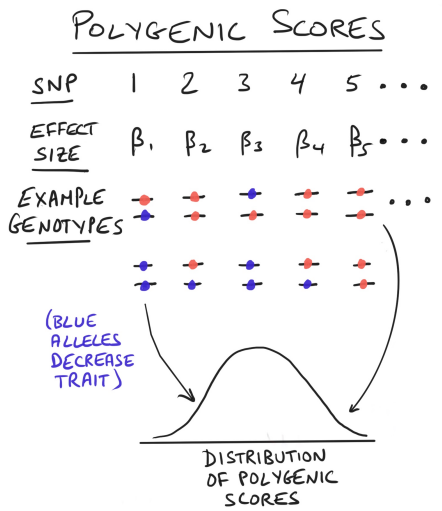


Figure 2.111: **The genetics for many traits can be modeled using polygenic scores.** *Here, blue alleles decrease, and red alleles increase the expected phenotype by some amount β per allele. An individual's polygenic score is a sum across the relevant allelic effects, and predicts the genetic component of their phenotype.*

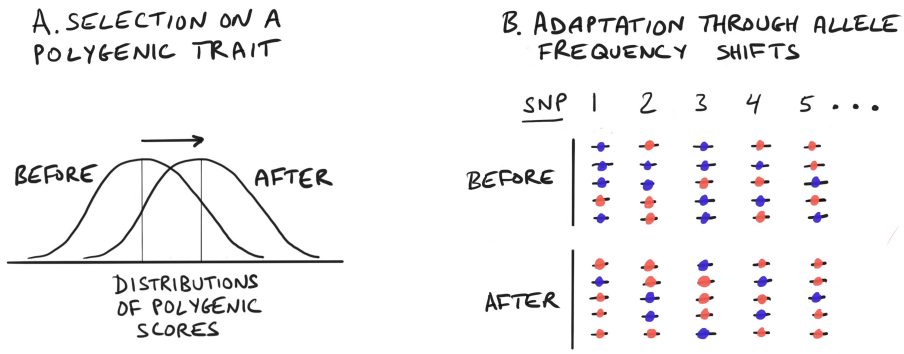


Figure 2.112: **Polygenic adaptation. A.** *Selection on a polygenic trait drives the mean polygenic score up or down, depending on the direction of selection.* **B.** *The shift in polygenic scores mainly occurs through small changes in allele frequencies at thousands of sites that contribute to the trait (**here, increasing the number of red alleles at each site**).*

One important feature of polygenic adaptation is that it proceeds extremely rapidly compared to conventional sweeps. This is because for conventional sweeps it can take hundreds of generations for a suitable, favored mutation to reach intermediate frequencies; in contrast, for a polygenic trait there is often a great deal of genetic variation present at the onset of selection (Chapter 4.4).

For this reason, polygenic adaptation is the main mechanism underlying the enormous responses to artificial selection that are commonly seen in plant and animal breeding. Farm animals including meat and dairy cattle, pigs, and chickens; as well as crop plants such as maize and soy, have undergone enormous improvements in yield due to artificial selection on polygenic traits. One example is shown at right, based on a remarkable study of maize, conducted at the University of Illinois continuously since 1896. As you can see, this study observed huge phenotypic changes within just 100 generations of artificial selection [308].

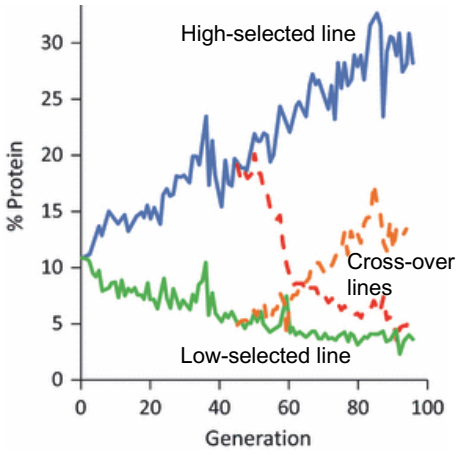


Figure 2.113: **Polygenic adaptation: the Illinois Maize experiment.** *Starting in 1896, lines were selected for either high, or low, protein content. In the cross-over lines, the direction of selection was reversed partway through the experiment.* *Credit: Modified from Figure 4 of Andrew Hendry et al 2011. [Link]*

It seems certain that complex phenotypes must be under a constant assault of selective pressures in one direction or another, though not necessarily in consistent directions in time and space. But despite the likely importance of polygenic adaptation as a mechanism, it has been difficult to detect clear signals in human data: the frequency shifts at most individual variants are very small and cannot be detected by traditional methods for detecting sweeps. There has been progress with alternative approaches, but these are still a work in progress [309].

Well done! In this chapter we have covered some of the main mechanisms for positive selection and adaptation, with examples. Next we examine the overall extent of different forms of selection.

Notes and References.

[268] Fan S, Hansen ME, Lo Y, Tishkoff SA. Going global by adapting local: A review of recent human adaptation. Science. 2016;354(6308):54-9

[269] The average fixation time for a strongly selected allele is $4\ln(2N)/s$, compared to $4N$ for a neutral allele: see Equation 10.30 in Coop (2020); also see simulations in Teshima and Przeworski (2006)

Coop G. Population and Quantitative Genetics; 2020

Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps? Genome research. 2006;16(6):702-12

[270] This term was coined in a classic 1974 paper

Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genetics Research. 1974;23(1):23-35

[271] Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832-7

Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS biology. 2006;4(3):e72

[272] A detailed derivation is beyond our scope, but the key idea is that τ gives the fixation time in the deterministic model, so τrx measures the ability for recombination to chop up the region at distance x within the course of the sweep. For more on this see Coop (2020), Chapter 13. For a very nice application to detecting sweeps, and further helpful citations see

Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome research. 2005;15(11):1566-75.

[273] For example see Voight et al (2006), Fan et al 2016, and

Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010;365(1537):185-205

Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. science. 2006;312(5780):1614-20

Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. BMC biology. 2017;15:1-10

[274] Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jørgensen ME, et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science. 2015;349(6254):1343-7

Mathieson S, Mathieson I. FADS1 and the timing of human adaptation to agriculture. Molecular biology and evolution. 2018;35(12):2957-70

Mathieson I, Day FR, Barban N, Tropf FC, Brazel DM, eQTLGen Consortium, et al. Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the FADS locus. Nature human behaviour. 2023;7(5):790-801

[275] Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nature Genetics. 2007;39(10):1256-60

Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nature Genetics. 2015;47(8):921-5

Pajic P, Pavlidis P, Dean K, Neznanova L, Romano RA, Garneau D, et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. Elife. 2019;8:e44628

[276] Mathieson I. Estimating time-varying selection coefficients from time series data of allele frequencies. bioRxiv. 2020

Segurel L, Guarino-Vignon P, Marchi N, Lafosse S, Laurent R, Bon C, et al. Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. PLoS biology. 2020;18(6):e3000742

[277] Until recently it has been difficult to do similar analyses for other selected variants, or in other parts of the world, as we have less dense sampling of ancient DNA outside Europe. However, this is now changing: for an application in east Asia see

Cong PK, Bai WY, Li JC, Yang MY, Khederzadeh S, Gai SR, et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. Nature Communications. 2022;13(1):2939. Furthermore, we have little data before $\sim$10,000 years ago, limiting the aDNA approach to sweeps that are recent.

[278] Bleasdale M, Richter KK, Janzen A, Brown S, Scott A, Zech J, et al. Ancient proteins provide evidence of dairy consumption in eastern Africa. Nature communications. 2021;12(1):632

[279] Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nature genetics. 2007;39(1):31-40

Schlebusch CM, Sjödin P, Skoglund P, Jakobsson M. Stronger signal of recent selection for lactase persistence in Maasai than in Europeans. European Journal of Human Genetics. 2013;21(5):550-3

[280]Crawford NG, Kelly DE, Hansen ME, Beltrame MH, Fan S, Bowman SL, et al. Loci associated with skin pigmentation identified in African populations. Science. 2017;358(6365):eaan8433

[281]Jones P, Lucock M, Veysey M, Beckett E. The vitamin D–folate hypothesis as an evolutionary model for skin pigmentation: an update and integration of current ideas. Nutrients. 2018;10(5):554

Jablonski NG. The evolution of human skin pigmentation involved the interactions of genetic, environmental, and cultural variables. Pigment Cell & Melanoma Research. 2021;34(4):707-29

[282]Nielsen et al (2005),

Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome research. 2009;19(5):826-37

Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354(6313):760-4

Ju D, Mathieson I. The evolution of skin pigmentation-associated variation in West Eurasia. Proceedings of the National Academy of Sciences. 2021;118(1):e2009227118

[283]One question is why the SLC24A5 variant is not found in east Asia. It appears that the SLC24A5 variant arose after the separation of west and east Eurasian populations, and that to some extent east Asians adapted to higher latitudes via mutations in different genes.

[284]For reviews see e.g.,

Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology. 2010;20(4):R208-15,

Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends in ecology & evolution. 2013;28(11):659-69

Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods in Ecology and Evolution. 2017;8(6):700-16

and for a classic example in sticklebacks see

Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484(7392):55-61.

[285]Orr HA, Betancourt AJ. Haldane's sieve and adaptation from the standing genetic variation. Genetics. 2001;157(2):875-84

Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005;169(4):2335-52

Przeworski M, Coop G, Wall JD. The signature of positive selection on standing genetic variation. Evolution. 2005;59(11):2312-23

[286]Langhi DM, Orlando Bordin J. Duffy blood group and malaria. Hematology. 2006;11(5-6):389-98

[287]Tournamille C, Colin Y, Cartron JP, Le Van Kim C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. Nature genetics. 1995;10(2):224-8

[288]Spencer HC, Miller LH, Collins WE, Knud-Hansen C, McGinnis MH, Shiroishi T, et al. The Duffy blood group and resistance to Plasmodium vivax in Honduras. The American Journal of Tropical Medicine and Hygiene. 1978;27(4):664-70

[289]A similar mechanism exists for HIV, which uses the CCR5 cell surface protein to enter CD4+ T cells. Individuals who are homozygotes for the CCR5 null allele (about 1% of Europeans) are HIV resistant.

Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science. 1996;273(5283):1856-62

[290]Pioneering work on Duffy by Martha Hamblin and Anna Di Rienzo in 2000 and 2002 showed, surprisingly, that Duffy did not show the expected signals of a hard sweep. Instead they proposed that the two major null haplotypes likely predated the onset of selection. My text relies on updated population genetic analysis, including Fst analysis and model estimates by Kimberly McManus et al (2017); Coop 2009 for genome-wide measures; estimated selection coefficient from Hodgson et al (2014):

Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. The American Journal of Human Genetics. 2000;66(5):1669-79

Hamblin MT, Thompson EE, Di Rienzo A. Complex signatures of natural selection at the Duffy blood group locus. The American Journal of Human Genetics. 2002;70(2):369-83

McManus KF, Taravella AM, Henn BM, Bustamante CD, Sikora M, Cornejo OE. Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. PLoS genetics. 2017;13(3):e1006560

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The role of geography in human adaptation. PLoS genetics. 2009;5(6):e1000500

Hodgson JA, Pickrell JK, Pearson LN, Quillen EE, Prista A, Rocha J, et al. Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. Proceedings of the Royal Society B: Biological Sciences. 2014;281(1789):20140930

[291]McManus et al (2017)

[292]Reservoir populations of P vivax can be found in African great apes:

Prugnolle F, Rougeron V, Becquart P, Berry A, Makanga B, Rahola N, et al. Diversity, host switching and evolution of Plasmodium vivax infecting African great apes. Proceedings of the National Academy of Sciences. 2013;110(20):8123-8

Liu W, Li Y, Shaw KS, Learn GH, Plenderleith LJ, Malenke JA, et al. African origin of the malaria parasite Plasmodium vivax. Nature communications. 2014;5(1):3346

[293]Globally in 2016 there were 216 million reported cases of malaria, and 445,000 deaths: [Link]

[294]To identify values of p for which $\Delta_p = 0$, with h and s fixed, we set

$$\frac{pqs[p(1-h)+qh]}{\overline{w}} = 0 \tag{2.87}$$

Noting that $q = 1 - p$, and assuming that $\overline{w} > 0$ for sensible biological parameters, we see immediately that two trivial solutions are

$$\hat{p} = 0 \tag{2.88}$$

$$\hat{p} = 1. \tag{2.89}$$

Next, let's consider the cases where $p \neq 0$ and $p \neq 1$. We further assume that $s \neq 0$. (The $1, 1+hs, 1+s$ parameterization used in this book has a slight oddity in that it does not allow the heterozygote to have a fitness different than 1 if $s = 0$.) Then we can divide both sides by p, q, and s, and multiply by $\overline{w}$, to yield

$$p(1-h) + qh = 0 \tag{2.90}$$

$$p(1-h) + (1-p)h = 0 \tag{2.91}$$

$$\hat{p} = \frac{h}{2h-1} \tag{2.92}$$

Note that this equilibrium for p is outside $[0, 1]$ and thus not relevant for an allele frequency, unless either $h < 0$ or $h > 1$. This is a stable equilibrium (i.e. balanced polymorphism) if $hs > 0$ and otherwise an unstable equilibrium.

[295]There are some good examples of disruptive selection in nature: for example, in the evolution of bird beak sizes

Hendry AP, Huber SK, De Leon LF, Herrel A, Podos J. Disruptive selection in a bimodal population of Darwin's finches. Proceedings of the Royal Society B: Biological Sciences. 2009;276(1657):753-9.

[296]Simons YB, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS biology. 2018;16(3):e2002985

Simons YB, Mostafavi H, Smith CJ, Pritchard JK, Sella G. Simple scaling laws control the genetic architectures of human complex traits. bioRxiv. 2022:2022-10

[297]Rees DC, Williams TN, Gladwin MT. Sickle-cell disease. The Lancet. 2010;376(9757):2018-31

[298]Under normal conditions, two units of the β-globin protein, along with two units of α-globin, join together to form the hemoglobin molecule, which is responsible for carrying oxygen in red blood cells. In individuals who are homozygous for the β-globin mutation, especially under low oxygen conditions, their hemoglobin molecules can stick together to form polymers. This in turn leads the red blood cells to change shape from a disc-like shape to a sickle-like shape. The sickling reduces oxygen-carrying capacity, and blocks blood vessels, leading a variety of severe symptoms. In individuals who are heterozygotes, only half of the β-globin proteins carry the mutation, and the tendency for red blood cells to sickle is greatly reduced under normal conditions. Importantly however, infection by the malaria parasite causes low oxygenation within the cell and causes sickling specifically of the infected cells. These can then be removed by the spleen, thereby helping to clear infection. Prior to modern medicine these children had very low survival rates. In recent years, treatment options have greatly improved, giving new hope for this devastating disease, although treatment is expensive and equitable access remains highly problematic.)

[299]Allison AC. Protection afforded by sickle-cell trait against subtertian malarial infection. British medical journal. 1954;1(4857):290

[300]The Malaria Genomic Epidemiology Network (2014) reported a huge reduction in severe malaria among sickle heterozygotes compared to non-sickle controls (odds ratio of 0.14, p-value=10^{-225}).

MalariaGen. Reappraisal of known malaria resistance loci in a large multicenter study. Nature genetics. 2014;46(11):1197-204

MalariaGen. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nature communications. 2019;10(1):5732

[301]Piel et al (2013) used spatial smoothing to estimate allele frequencies on global maps, as local sample sizes are often small. Their highest estimate at any location was 18% in northern Angola, but with high uncertainty, while they are more confident in estimates around 15%:

Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Dewi M, et al. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. The Lancet. 2013;381(9861):142-51

[302]A recent study of selection at sickle uses a slightly lower allele frequency and concludes the following: "If we take the 21% HbAS average prevalence in Gabon, it translates to a HbS frequency p = 0.105 and to a selection coefficient s = 0.12, ... a figure comparable to that of 0.11 found by Cavalli-Sforza and Bodmer"

Elguero E, Délicat-Loembet LM, Rougeron V, Arnathau C, Roche B, Becquart P, et al. Malaria continues to select for sickle cell trait in Central Africa. Proceedings of the National Academy of Sciences. 2015;112(22):7051-4

[303]More on G6PD:

Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science. 2001;293(5529):455-62

[304]There's a similar polymorphism in old world monkeys and it's likely that the origin goes back even further, to the ancestor of apes and monkeys.

Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, et al. The ABO blood group is a trans-species polymorphism in primates. Proceedings of the National Academy of Sciences. 2012;109(45):18493-8

[305]It has been suggested that other types of pressures, such as gut pathogen interactions may also be important in maintaining the system. For discussion of selective pressures see

Segurel L, Gao Z, Przeworski M. Ancestry runs deeper than blood: the evolutionary history of ABO points to cryptic variation of functional importance. Bioessays. 2013;35(10):862-7.

[306]For more examples of ancient balancing selection see

Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science. 2013;339(6127):1578-82

Fortier AL, Pritchard JK. Ancient Trans-Species Polymorphism at the Major Histocompatibility Complex in Primates. bioRxiv. 2022:2022-06

[307]Pritchard et al (2010);

Hayward LK, Sella G. Polygenic adaptation after a sudden change in environment. Elife. 2022;11:e66697

[308]Illinois Maize study lab website: [Link];

Moose SP, Dudley JW, Rocheford TR. Maize selection passes the century mark: a unique resource for 21st century genomics. Trends in plant science. 2004;9(7):358-64

Hendry AP, Kinnison MT, Heino M, Day T, Smith TB, Fitt G, et al. Evolutionary principles and their practical application. Evolutionary Applications. 2011;4(2):159-83

[309]Most promising, there has been interesting work on detecting polygenic shifts for specific traits, but these are still challenging to apply in practice:

Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. Elife. 2019;8:e39725

Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife. 2019;8:e39702

Cox SL, Ruff CB, Maier RM, Mathieson I. Genetic contributions to variation in human stature in prehistoric Europe. Proceedings of the National Academy of Sciences. 2019;116(43):21484-92

Chen M, Sidore C, Akiyama M, Ishigaki K, Kamatani Y, Schlessinger D, et al. Evidence of polygenic adaptation in Sardinia at height-associated loci ascertained from the Biobank Japan. The American Journal of Human Genetics. 2020;107(1):60-71

2.7 Natural Selection III. Genome-wide extent of selection

We have now touched on the main types of natural selection, and I have already hinted at a key question: how important are each of these in practice? Here we tackle this key question.

First though, it's helpful to give some historical context [310].

By the 1960s, much of the basic theory of population genetics had already been developed, but molecular techniques for measuring genetic variation were extremely limited. Consequently, population genetics was largely a theoretical field [311]. Little was known about the relative importance of the fundamental processes: mutation, recombination, migration, and drift; negative selection, positive selection, and balancing selection.

This started to change with the invention of **gel electrophoresis**, a technique that made it possible to measure protein variation on gels [312]. The first examples came from humans and flies in 1966 [313]. These first studies were followed by a flurry of electrophoresis studies in a wide range of organisms – so many that this was cheekily referred to as the "find 'em and grind 'em" approach [314].

Before the electrophoresis era it was anticipated that most protein variants would be subject to strong selection. Thus the default state would be a wildtype allele and perhaps additional rare deleterious variants; meanwhile there would be occasional rapid sweeps, and perhaps balancing selection in some genes [315].

Given these expectations, it was a surprise to find that protein variation is widespread in most species. For example, in 1966 Lewontin and Hubby estimated that around $^1/_4$ to $^1/_3$ of genes were polymorphic within populations of the fly Drosophila pseudoobscura [316]. One possibility was that this might indicate huge amounts of balancing selection, but this conclusion was controversial.

A complementary insight came from emerging data on protein differences between species. By the mid-1960s it was becoming apparent that proteins tend to accumulate amino acid substitutions steadily over evolutionary time. This was referred to by Zuckerkandl and Pauling in 1965 as the **molecular clock** [317]. One vivid illustration of the molecular clock was published by Richard Dickerson, below, in 1971 [318]:

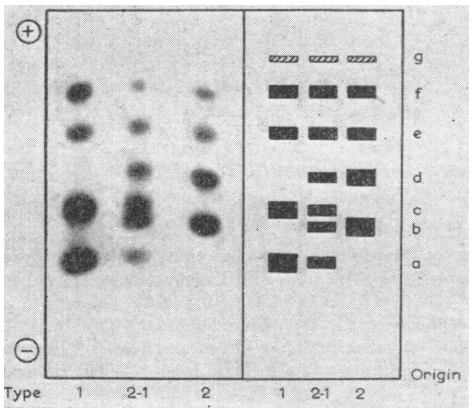


Figure 2.114: **Enzyme Polymorphisms in Man (1966).** *Gel electrophoresis, as shown here, made it possible to survey genetic variation for the first time.* *The vertical lanes show banding patterns for two alleles at the phosphoglucomutase enzyme: homozygotes in lanes 1, and 2; heterozygotes are a mixture of both patterns: 2-1. Experimental data at left, and a schematic of the banding patterns at right. The alleles were reported at frequencies 0.75 and 0.25 respectively, in human populations. Credit: Fig. 68 from Harry Harris (1966). [Link] Used with permission.*

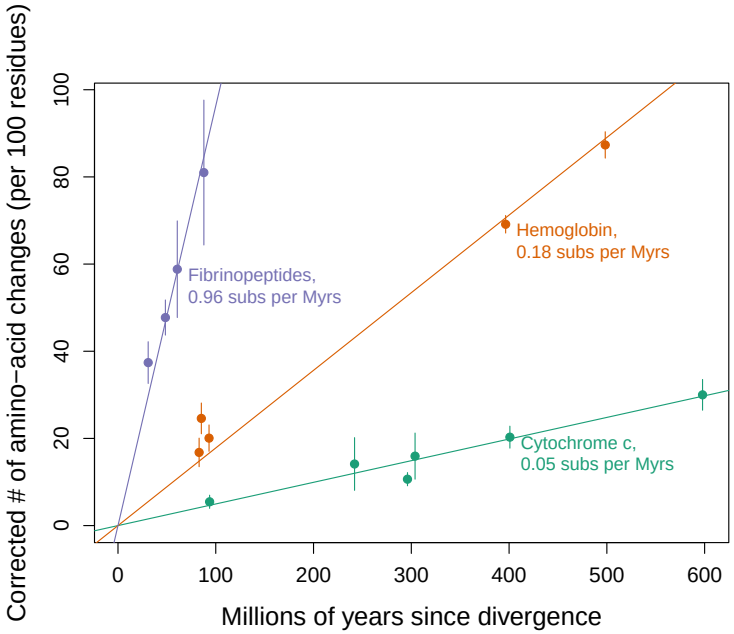


Figure 2.115: **One of the first demonstrations of the molecular clock, from 1971.** *The x-axis shows divergence times of pairs of species, as estimated from the fossil record; the y-axis shows fractions of amino acid differences in three proteins. The analysis was important for showing that protein differences accumulate roughly linearly over evolutionary time, but at different rates for different proteins.* *Credit: Figure 5.3 from Graham Coop in Population and Quantitative Genetics [Link], CC BY 3.0; based on Dickerson (1971).*

Of course it was possible that these protein changes were all adaptive, but even in the 1960s there were reasons to doubt this. The gene Cytochrome C, shown above, is found in a wide range of eukaryotes and fulfills a conserved role in the electron transport system in the mitochondria. King and Jukes (1968) noted that experiments comparing Cytochrome C proteins from different species could detect no functional differences. They hypothesized that the observed substitutions are mainly at positions that *do not* have a functional impact, and have fixed by neutral drift [319]. Their proposal contrasts sharply with an adaptive model of protein evolution, where one might expect most substitutions to be functional.

The Neutral Theory of Molecular Evolution. Together, these observations stimulated a paradigm shift in the late 1970s in how people thought about the main forces acting on genetic variation – and especially the role of genetic drift. These new ideas were articulated in particular by the Japanese scientist Motoo Kimura, who dubbed this the Neutral Theory of Molecular Evolution [320]. **In short, he proposed that most new mutations are either approximately neutral, or deleterious; advantageous mutations are very rare and contribute only a tiny fraction of polymorphism and differences between species.**

As stated by Kimura (1983) [321]: *"The neutral theory asserts that a great majority of evolutionary changes at the molecular level...are caused not by Darwinian selection but by random drift of selective neutral or nearly neutral mutants.... (P)olymorphisms are mainly due to mutations that are nearly enough neutral... that their behavior and fate are mainly determined by mutation and random drift...*

On the topic of selection he clarified that: *"The theory does not ... assume that selection plays no role; however, it does deny that any appreciable fraction of molecular change is due to positive selection or that molecular polymorphisms are determined by balanced selective forces... selective constraints imposed by negative selection are a very important part of the neutralist explanation..."*

It's hard to overstate the impact this model has had on how we think about genetic variation. The Neutral Theory provides an intellectual framework for thinking about modeling, and a null hypothesis for data analysis. It is no longer controversial that most new mutations are neutral and that, of those that are not neutral, most are selected against. As we'll discuss in Part 3 of the book, these properties allow us to use genetic variation as a tool for studying population structure and history while largely ignoring the role of selection.

That said, it's worth noting that early conceptions of the Neutral Theory under-appreciated the importance of some processes in shaping patterns of variation [322]. One important early addition came from Tomoko Ohta's work emphasizing the importance of **nearly neutral mutations**. Starting in 1973, Ohta argued that many mutations may have selection coefficients that are close to, but not precisely, 0. These can have important consequences: for example, recalling that selection is ineffective when $|4Ns|$ is less than about 1, we see that weakly deleterious variants fix at a higher rate in species with small N than in species with large N, which can affect substitution rates in different lineages (Chapter 2.5) [323].

Another under-appreciated area was the role of **linked selection** (which we'll cover in this chapter), and a third is the role of **polygenic stabilizing selection and adaptation** (Chapter 2.7).

The original theory also predated modern understandings of genome architecture, as well as the central importance of **gene regulation** in phenotypic variation and evolution.

And despite the Neutral's Theory's importance as a null hypothesis, significant effort in the last 50 years has been devoted to understanding its limitations. There has been a great deal of work aimed both at measuring overall rates of positive selection, as well as at elucidating the specific genetic changes that underlie adaptations [324]. ***Even if only a small fraction of polymorphisms and substitutions are positively selected, the most interesting biology lies in those exceptions****: for many evolutionary biologists, a sense of awe at the power of Darwinian adaptation is what got us excited about biology in the first place!*

Figure 2.116: **Camouflaged cicada on tree.** *Although the neutral theory provides a powerful framework for modeling molecular evolution,* ***it does not deny the central importance of Darwinian adaptation*** *– in this case driving adaptation of the cicada to be almost perfectly camouflaged in its natural habitat.* Credit: Henk Monster [Link] CC BY 3.0.

Substitution rates and the molecular clock. As shown above, proteins (and DNA sequences) tend to accumulate changes roughly linearly in time, though the rates differ between proteins. This observation would be puzzling if most substitutions are adaptive: why should adaptation occur at a roughly constant rate over hundreds of millions of years, while the organisms themselves, ecosystems, and parameters such as effective population size, vary hugely over time? The Neutral Theory provides a simple model for this.

First, we need to derive the substitution rate for purely neutral sequences. Suppose we sequence a neutral region of the genome in two species that diverged T generations. How many differences do we expect to see be-

tween the two species?

Mutations arise at a rate μ per base pair per generation. Let's look first at fixation events in Species 1. Suppose that the population size of Species 1 is $2N$; then across the entire population of Species 1 we get $2N\mu$ new mutations per base pair each generation. Recall from Chapter 2.1 that new mutations will ultimately fix with a probability equal to their starting frequency: i.e., $1/2N$. Hence, the rate of fixation of mutations is

$$\text{Fixation rate} = [\text{Total rate of new muts}] \times [\text{Fixation prob. of muts}] \quad (2.93)$$

$$= 2N\mu \times \frac{1}{2N} \quad (2.94)$$

$$= \mu \quad (2.95)$$

The population size, N here, cancels out, leading to the crucial result that neutral mutations fix at a rate μ per generation per site, regardless of population size.

Similarly if we compare two species that have diverged for T generations, then at neutral sites the expected frequency of differences is $2\mu T$ [325]. The factor of 2 reflects that fixation events occur in *both* lineages for T generations [326].

Now let's focus specifically on nonsynonymous changes [a]. Think about what happens if a gene contains some positions where mutations would be neutral, and others where mutations would be deleterious: for example mutations in a functional binding pocket of an enzyme might strongly disrupt function, while a change between similarly charged amino acids in an unstructured region might be neutral. Let's suppose that a fraction of λ of all changes are neutral, and $1 - \lambda$ are sufficiently deleterious that they have essentially no chance of fixing [327]. Now we find that mutations fix at a rate

[a] *Recall that nonsynonymous (=missense) substitutions change the amino acid encoded at a position, while synonymous substitutions do not. For example, CCC→GCC changes proline to alanine (nonsynonymous); but CCC→CCG maintains proline (synonymous).*

$$\text{Fixation rate} = \lambda\mu \quad (2.96)$$

per generation, and the expected number of substitutions per site between species is

$$2\lambda\mu T. \quad (2.97)$$

If we convert μ from a per-generation rate to a per-year rate, and assume that this is roughly constant across the phylogeny, and across genes, then *this predicts that substitutions accumulate linearly in time, where the rates are proportional to the fraction of neutral sites. This results in the molecular clock, where the slope is proportional to λ* [328].

$\mathbf{d_n/d_s}$ as an estimator for amino acid constraint. If we want to use Equation 2.97 to estimate λ we need to know both the divergence time T and gene-specific mutation rate μ. Unfortunately we don't always have good estimates of these.

But we can get a better estimator of λ by simply comparing the substitution rates for synonymous and nonsynonymous sites within the same

gene. This is captured in a measure called d_n/d_s (also known as K_a/K_s [329]). Here d_n is the expected number of nonsynonymous substitutions per nonsynonymous site, and d_s is the corresponding number for synonymous substitutions. Then d_n/d_s gives the ratio of the two rates [b].

[b] *Note that d_n and d_s are adjusted for the effective numbers of nonsynonymous and synonymous sites, based on the numbers of possible mutations that would/would not change the encoded protein. Thus d_n/d_s should be 1 in the absence of selection* [330].

To interpret d_n/d_s, let's first make the simplifying assumption that all synonymous mutations are neutral [331]. Then the expected synonymous divergence between two species would $E(d_s) = 2\mu T$, as above.

For nonsynonymous sites in the same gene, the expected nonsynonymous divergence would be $E(d_n) = 2\lambda\mu T$. So the ratio of the expected values tells us that [332]:

$$\frac{d_n}{d_s} = \lambda. \tag{2.98}$$

Notice that since λ represents the *fraction* of neutral sites, d_n/d_s must be between 0 and 1 under this model.

Indeed, this is the case for most genes, as you can see in this plot showing the distribution of d_n/d_s values in mammals [333]:

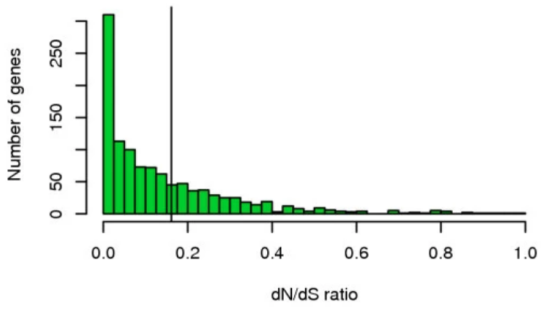


Figure 2.117: **Distribution of $\mathbf{d_n/d_s}$ across human genes.** *The plot shows a histogram of estimated d_n/d_s across genes, measured in the human lineage. The vertical line indicates the mean.* *d_n/d_s is measured on the human lineage since the common ancestor of human, mouse, and pig. Credit: From Fig. 5 of Frank Jørgenson et al (2005) [Link]; CC BY 2.0.*

This study reported a genome-wide average of about $d_n/d_s = 0.14$, which we could interpret to mean that about 14% of amino acid substitutions were effectively neutral.

Positive selection, d_n/d_s, and the MK test. In the absence of positive selection d_n/d_s is always ≤ 1. But what happens if some nonsynonymous mutations are actually favored by selection? Intuitively, you might expect that selection should increase divergence at nonsynonymous sites, and could potentially push $d_n/d_s > 1$. This suggests a test for adaptive evolution of protein sequences: Can we find genes for which d_n/d_s is significantly > 1?

To understand this, let's consider a simple extension of the model to three categories of sites:
- A fraction λ_0 are neutral
- A fraction λ_a are advantageous with selection coefficient s
- A fraction $1 - \lambda_0 - \lambda_a$ are strongly deleterious

Recall that favored mutations fix with probability s [334]. Hence, favored mutations arise at a rate $2N\lambda_a\mu$ per generation, and fix at a rate $2N\lambda_a\mu \cdot s$.

So the expected divergence at nonsynonymous sites in time $2T$ will be $\lambda_0 \cdot 2\mu T + 2N\lambda_a s \cdot 2\mu T$, compared to $2\mu T$ at synonymous sites, and

$$\frac{d_n}{d_s} = \lambda_0 + 2Ns\lambda_a. \tag{2.99}$$

To make this concrete, suppose that $\lambda_0 = 0.2$; and further suppose that 1% of nonsynonymous mutations in a gene have a selective advantage $s = 0.1\%$ in a population of 10^4. Then $2Ns\lambda_a = 0.2$, and d_n/d_s is 0.4. In this example, even though many of the sites are fixed by positive selection, d_n/d_s is still much less than 1.

In fact, we need a *lot* of selection to detect it using d_n/d_s. For example, suppose that 1% of nonsynonymous mutations have an advantage of $s = 1\%$. We now predict that $2Ns\lambda_a = 2.0$ and d_n/d_s will be 2.2, and we reject neutrality.

High $\mathbf{d_n/d_s}$ at MHC genes. It's quite unusual in mammals for selection to be strong enough to drive d_n/d_s above 1, but a famous example occurs in genes of the Major Histocompatibility Complex (MHC) [335]. MHC genes play an **essential role in defense against infection** by presenting fragments of proteins known as peptides for surveillance by T cells. T cells are trained to ignore peptides from our own proteomes; but when they detect foreign peptides they initiate an immune response.

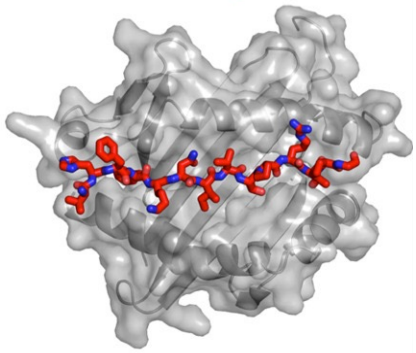


Figure 2.118: **Peptide presentation by MHC.** *MHC proteins (here in gray) play an essential role in the immune system by presenting short peptides (red/blue) for inspection by T cells. MHC proteins must be able to successfully bind a highly diverse and rapidly evolving array of foreign peptides.* *Credit: Figure 3e of Meriem Attaf et al (2015) [Link]. CC BY 4.0*

Crucially, different MHC alleles have different potential binding repertoires. Thus, the universe of peptides that you can present to T cells depends on your genotype across the six MHC genes involved in antigen presentation. There is an overall advantage to having different alleles at each MHC gene, as it expands the potential space of antigens you can present, and particular MHC alleles may especially effective against particular pathogens. All of these factors have led to huge selection pressure for allelic diversity in the MHC, driving ancient balancing selection, similar to the ABO story in the last chapter [336].

Given the strong selection pressure in favor of functional diversity, it should come as no surprise that there is enormous nonsynonymous diversity at functional sites in the MHC genes. A classic 1988 paper by Austin Hughes and Masatoshi Nei examined d_n and d_s between highly diverged human alleles for three MHC genes. They predicted high d_n within the peptide binding region (PBR), but not in the rest of the protein where the function is more conserved [337].

	Peptide Binding (L=57)		Not PBR (L=125)		Exon 4 (L=92)	
	d_n	d_s	d_n	d_s	d_n	d_s
MHC-A	**13.3**	**3.5**	1.6	2.5	1.6	9.5
MHC-B	**18.1**	**7.1**	2.4	6.9	0.5	1.5
MHC-C	**8.8**	**3.8**	4.8	10.5	1.0	2.1

Table 2.7: High $\mathbf{d_n/d_s}$ in MHC genes. *Average d_n and d_s between different human alleles in three MHC genes.* *"Peptide Binding" refers to sites within the PBR; "Not PBR" corresponds to other sites in Exons 2 and 3 that do not contact the peptide; sites in Exon 4 also do not contact the peptide. L indicates numbers of sites. Standard errors for most comparisons were ~2. Modified from Hughes and Nei (1988) [Link].*

Consistent with this logic, you can see above that d_n is larger than d_s within the

peptide binding region, and lower elsewhere. This indicates that there is frequent adaptive evolution within the peptide binding region, and main selective constraint in the structural regions of these genes.

However, more generally, testing for $d_n/d_s > 1$ is not a very powerful test because it's highly unusual to see so many adaptive changes in one small region; secondly, we may not know in advance which sites are likely to be evolving adaptively as we do in the example above (but see [338, 339]).

Tests contrasting polymorphism and divergence. A paper by John McDonald and Martin Kreitman in 1991 suggested a more powerful test for selection by contrasting variation within and between species, now known as the **McDonald-Kreitman or MK test** [340]. The key concept is that selective sweeps occur quickly compared to drift, and so they are more likely to be observed in a data set as differences between species than as polymorphic sites within species.

In one version, the MK test considers gene sequences for multiple individuals from each of two species. Variants can be classified as being either *fixed differences* between the species, or *polymorphic* within one of the species [c]. Similar to the model we used before, the null hypothesis will be that a fraction λ of new nonsynonymous mutations are neutral, and $1 - \lambda$ are strongly deleterious.

[c] *A fixed difference is a variant where all individuals in one species have one variant, while all individuals in the other species have a different variant. A polymorphism would be variable within the sample from one or other species.*

Assuming that selection against deleterious variants is strong enough, we don't expect to see deleterious variants as polymorphisms, and certainly not as fixed differences. In this scenario, we expect the ratio of nonsynonymous to synonymous to be the same (i.e., λ) for both polymorphisms and fixed differences (Panel A):

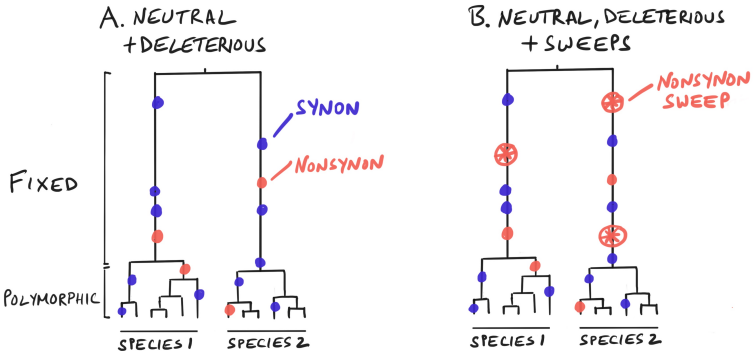


Figure 2.119: **Overview of the MK test. A.** *In the baseline model (Neutral + Strongly deleterious) the expected ratio of nonsynonymous:synonymous variants is the same in the fixed and polymorphic categories.* **B.** *In the model with positive selection, there is a greater fraction of nonsynonymous sites among the fixed differences.*

But if some nonsynonymous sites are positively selected, then these will tend to sweep through populations very quickly (Panel B). Because they sweep quickly, it's rare that one would be just in the process of sweeping right now, and much more likely that they would be fixed differences [d]. For this reason we expect that *positively selected variants will increase the fraction of nonsynonymous variants among the fixed differences.*

[d] *You can look at Figure 2.95 to see that selected variants fix much faster than neutral variants.*

Consistent with the positive selection model, the first application of the MK approach found a much higher fraction of nonsynonymous substitutions *between* Drosophila species (41% of substitutions) compared to nonsynonymous polymorphisms within species (just 5% of variants):

	Fixed	Polymorphic
Nonsynonymous	7	2
Synonymous	17	42
% Nonsynon.	**41.2%**	**4.7%**

Table 2.8: Excess of nonsynonymous substitutions in the Drosophila ADH gene. *The table shows the numbers of nonsynonymous and synonymous variants that are either polymorphic within species, or fixed between species (P-value for a test of independence is* 0.006*).* *Modified from McDonald and Kreitman (1991) [Link].*

The authors interpreted this as evidence that positive selection at ADH drives nonsynonymous fixations that accumulate as an excess of between-species differences [341] [342].

Since then, MK analyses in the genome-wide era have revealed rampant positive selection in Drosophila: likely as many as 50% of nonsynonymous differences between species were fixed by positive selection [343].

For humans, in contrast, it seems that a much smaller fraction of nonsynonymous differences between humans and other primates were fixed by positive selection: likely in the range of 0–10%, although the precise fraction is still a matter of debate [344]. *This work shows that the great majority of nonsynonymous substitutions in primates are effectively neutral.*

Linked selection: background selection and hitchhiking. This brings us to our last major selection topic, **linked selection**, which deals with the effects of selection–both positive and negative–at nearby sites. Here, we ask: *How does selection affect the patterns of genetic diversity at nearby neutral sites* [345]*?*

Our story begins around the same time as development of the McDonald-Kreitman test, when an observation from Drosophila presented an important new challenge to the Neutral Theory. A 1992 paper by David Begun and Chip Aquadro showed that regions of the fruit fly (Drosophila) genome with low recombination rates tend to have low genetic diversity [346].

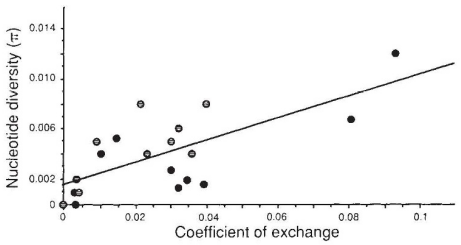


Figure 2.120: **Classic plot of the relationship between recombination rate and genetic diversity in fruit flies (1992).** *The x-axis shows a measure of local recombination rate; the y-axis is average pairwise heterozygosity, π; each data point is a different sequenced locus. The null hypothesis that the slope is* 0 *is rejected with* $p = 0.0007$. *Credit: Fig. 1 of David Begun and Charles Aquadro (1992) [Link] Used with permission.*

Similar patterns are also seen in humans (side panel) [347].

Begun and Aquadro proposed that this observation is evidence for widespread **genetic hitchhiking with selective sweeps**. Recall from Chapter 2.6 that when a favored mutation sweeps rapidly through the population, it carries a surrounding haplotype with it, up to high frequency, in a process known as hitchhiking. As a sweep nears completion it eliminates

genetic variation in a window around the selected site (Figure 2.96).

The size of the window is inversely related to $r \times 2log(2N)/s$, where r is the local recombination rate, N is the population size, and s is the selection coefficient. This is intuitive: **when r is high, recombination breaks up the sweeping haplotype much more efficiently than when r is low.**

Begun and Aquadro hypothesized that sweeps are scattered randomly across the genome. When they sequenced a locus with low recombination rate it was much more likely to fall within the window of a recent sweep (and therefore, have low diversity) than when they sequenced a locus with high recombination rate. They concluded that "Hitch-hiking thus seems to occur over a large fraction of the *Drosophila* genome and may constitute a major constraint on levels of genetic variation".

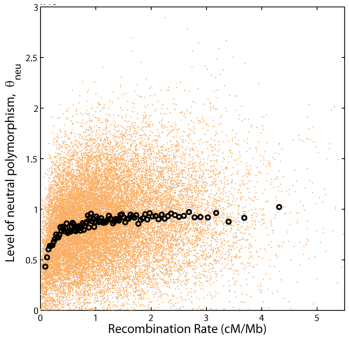


Figure 2.121: **Human genetic diversity is also reduced in regions of low recombination.** *Black dots show binned averages of the genetic diversity ($\theta \times 10^{-3}$) as a function of local recombination rate. Orange dots show raw data in a sliding window across the genome.* Credit: *Fig. 1B of James Cai et al (2009); CC BY 4.*

Background selection. However, the next year an alternative explanation, dubbed *background selection*, was proposed by Brian Charlesworth and colleagues [348]. The essential concept of background selection is that when deleterious mutations arise, they may drift briefly but are unlikely to contribute to the population long-term. As those variants are eventually purged, any linked neutral variants are lost too.

One helpful way to think about this is that, at any given locus, *the copies of this locus present in the population today are primarily descended from past copies of the locus that did not carry deleterious mutations. Thus, deleterious mutations can be thought of as reducing effective population size within a linked region.*

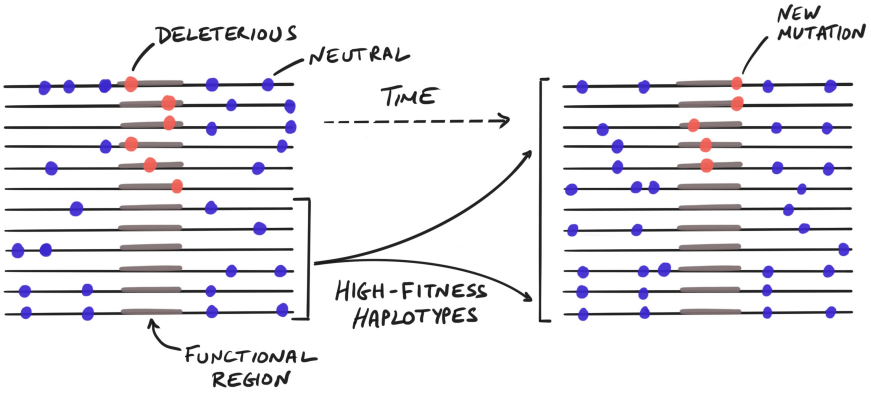


Figure 2.122: **Background selection.** *At any given time a fraction of chromosomes carry deleterious variants (red mutations). These chromosomes have low fitness and don't contribute much to future populations in the long term. Neutral variants linked to deleterious variants will eventually be removed by selection. Meanwhile new deleterious variants continue to arise by mutation.*

To model this, let's first look at background selection in a region without recombination [349]. Define f as the total fraction of chromosomes that carry deleterious mutations. A simple model [350] suggests that at equilibrium

$$f \approx L\mu/hs, \tag{2.100}$$

where L is the number of basepairs that can produce deleterious mutations, μ is the mutation rate per base pair, and hs is the selective disadvantage for a heterozygote [e].

Provided that selection is strong enough that individual deleterious mutations don't persist long in the population ($hs >> 1$), you can think of

[e] *In a minor abuse of notation, in this section we use $hs > 0$ to indicate a selective disadvantage. The derivation requires that selection is considerably stronger than drift, i.e., $hs >> 1/N$.*

this as reducing the effective population size locally by a factor $1 - f$. Then we can approximate the expected pairwise genetic diversity, $\mathrm{E}[\pi]$, in this region as

$$\mathrm{E}[\pi] = \pi_0(1 - \frac{L\mu}{hs}), \tag{2.101}$$

where $\mathrm{E}[\pi_0]$ is what the expected genetic diversity would have been if there were no background selection.

What happens with recombination? Let's say we sequence a region that is at a recombination fraction r from a conserved functional element. Now things are more complicated, because a neutral variant in the sequenced region could be rescued by recombining away from a linked deleterious mutation. After a flurry of math [351], the expected diversity is found to be

$$\mathrm{E}[\pi] = \pi_0\left[1 - \frac{L\mu}{hs(1 + r/hs)^2}\right]. \tag{2.102}$$

The last part of this expression, $\frac{L\mu}{hs(1+r/hs)^2}$, represents the proportional decrease in variation due to background selection. *This has the intuitive form that the impact of background selection increases with the deleterious mutation rate $L\mu$, and decreases with recombination distance r.* The relationship with selection strength is more complicated [352].

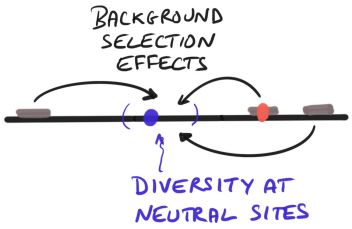


Figure 2.123: **Genetic diversity at neutral sites is reduced by background selection from linked functional elements.** *Each functional element reduces expected diversity by a factor of* $(1 - \frac{L_i\mu}{hs(1+r_i/hs)^2})$.

Next, the total strength of background selection experienced at a site depends on the cumulative contributions from all linked functional loci (for example, all coding exons, conserved gene regulatory elements, etc). The total reduction in π is a product of the contributions from each functional element:

$$\mathrm{E}[\pi] \approx \pi_0 \prod_{i=1}^{M}\left[1 - \frac{L_i\mu}{hs(1 + r_i/hs)^2}\right]. \tag{2.103}$$

where i indexes each of M linked functional elements [353].

Does this model fit real data?

In a 2009 paper, Graham McVicker and colleagues used this approach to predict the background selection effect of constrained regions across the human genome [354]. As you see from Equation 2.103, the strength of background selection at any specific location depends on the local landscape of linked functional elements. This can be used to predict genetic diversity at neutral sites across the genome, depending on the number, size, and genetic distance to nearby functional elements in the genome sequence. This reduction in diversity is commonly written as **B** [355]:

$$B = \frac{\mathrm{E}[\pi]}{\pi_0} \tag{2.104}$$

The plot below shows an updated version of McVicker's analysis [356]. As you can see, the background selection model provides a remarkably good prediction of the landscape of genetic diversity in humans:

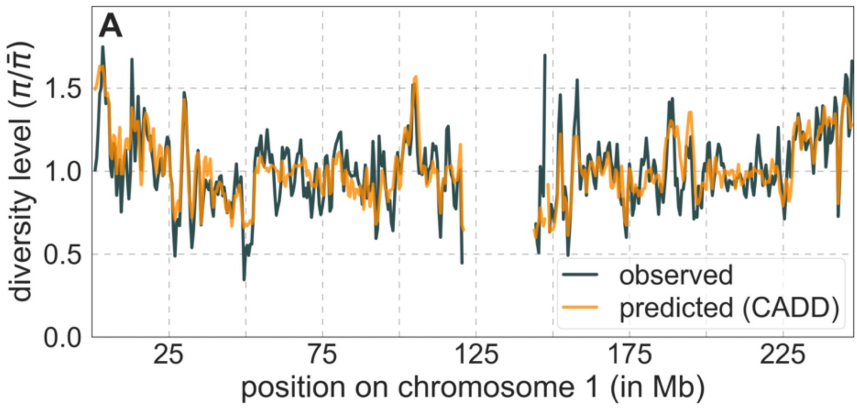


Figure 2.124: **Human genetic diversity predicted from background selection model.** *Genetic diversity along chromosome 1 is plotted in teal; predictions from a background selection model are in orange.* *Data are for Yoruba (from Nigeria). The y-axis is genetic diversity π divided by the genome-wide average. The data are plotted in megabase-sized windows. The gap at the center of the plot is due to repetitive regions near the centromere. Credit: Figure 2 from David Murphy et al (2021) [Link]. CC BY 4.*

This is actually quite a dramatic effect, with the model accounting for most variation in genetic diversity (at megabase scale) across the genome, emphasizing the importance of linked selection in shaping genetic variation.

Background selection or hitchhiking? Thus, in both humans and flies (and other species) we see a strong relationship between neutral sequence diversity, and the local recombination rate and density of nearby functional sequence. This is compelling evidence that linked selection plays a major role in shaping genetic diversity across the genome. But it leaves us wondering how much of the linked selection effect is due to background selection versus hard sweeps [357].

One way of distinguishing these is to note that if hard sweeps are important, then there should be a dip in diversity specifically near the sites of completed sweeps. This is different from the general depletion of variation due to background selection. We don't know which sites have had recently completed sweeps, but we could hypothesize that these would be enriched at recent nonsynonymous substitutions. Under this hypothesis, diversity near nonsynonymous substitutions would reflect a mixture of neutral and selected signals. To summarize: *if an appreciable fraction of nonsynonymous substitutions on the human lineage are recently completed hard sweeps, then we should see lower diversity near those sites compared to a model with background selection only* [358].

But, instead, the genetic diversity around nonsynonymous substitutions can be predicted entirely from the background selection model. The plot below shows the average of genetic diversity around all nonsynonymous substitutions, along with the predictions under background selection. There's a dip at the center of the plot, but this is only because nonsynonymous sites are generally in regions with lots of functional sequences – as you can see the data are extremely similar to the background selection model:

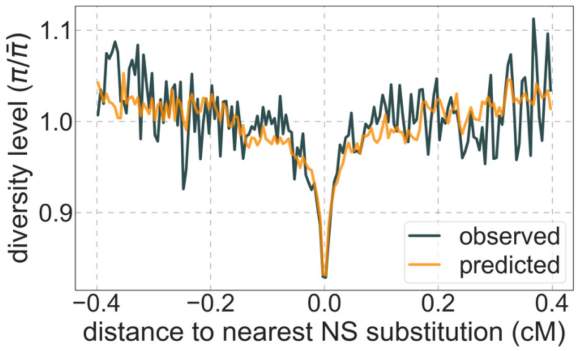


Figure 2.125: **Average levels of genetic diversity at nonsynonymous substitutions and comparison to predictions from background selection.** *Genetic diversity at nonsynonymous substitutions (teal) is accurately predicted from a background selection model (orange). The close fit at nonsynonymous sites argues against frequent hard sweeps driving nonsynonymous substitutions.* *Credit: Figure 3 from David Murphy et al (2021) [Link]. CC BY 4.*

This analysis would have had power to detect a signal if as much as 10% of nonsynonymous variants had swept with strong selection (s=1%) [359], though it has less power to detect soft sweeps. In contrast, this type of analysis *does* show a clear signal in Drosophila, where selective sweeps seem to be much more common [360].

In summary, this analysis and the MK results argue that at most a small fraction (<10%) of nonsynonymous variants in humans were fixed by strong positive selection. Similarly, genomewide selection scans reveal relatively few unequivocal examples of recent sweeps in noncoding regions. This leaves open the possibility of positive selection through soft sweeps, much weaker hard sweeps, and polygenic adaptation as we'll discuss next.

Concluding remarks. In this section of the book we have covered some of the core principles for understanding genetic variation. One remarkable aspect of population genetics is that many of the fundamental concepts extend logically from the basic genetic and population processes: mutation; Mendelian segregation; linkage; random mating and population structure; and different forms of selection.

That said, while many key concepts were already understood 50 years ago, it has taken much longer to determine the relative importance of the different processes–in particular the impact of genetic drift, linkage, and the different types of selection in shaping patterns of variation and evolution–and many aspects of this are important areas of research now that we have much richer genome data, and modern tools from functional genomics.

I think it's fair to say that versions of the Neutral Theory now provide the central structure for most models of genetic variation: at least 90% of new single nucleotide mutations are essentially neutral, and most of what is not neutral is deleterious. However, we also now know that linked selection in the genome, mainly from background selection, is pervasive, so that diversity in most of the genome is reduced relative to the maximum possible under a fully neutral model.

What then, is the role of positive selection? Even if only a tiny fraction of variants are positively selected, we do know that the natural world, including humans, show an astonishing diversity of forms. Organisms are

amazing molecular machines, and exquisitely adapted to their environments. This must happen through forms of adaptation. As I discussed at length, we do now have compelling examples of the various forms of positive selection acting in humans: including hard and soft sweeps, and ancient balancing selection.

However, my personal reading of the data is that strong hard selection on individual loci has been rare in the human genome during the past $\sim$ 200,000 years when we can best detect it. Many of the exceptions where we do see sweep signals are at genes where a single protein plays an exceptional role in some process–for example Duffy, which serves as a specific receptor for vivax malaria; or lactase which plays an essential role in digesting lactose. The relative importance of different modes of selection seems to vary greatly across species, and hard sweeps may be less important in humans than in some other species that have been studied, including flies and stickleback fish.

It's possible that environmental pressures acting on human populations are often variable and inconsistent across space and time, and thus it is rare for selection be both strong and consistent enough over the many thousands of years that are required for hard sweeps in a species with our long generation time. This hypothesis may be consistent with recent work on ancient DNA identifying many short-term selective frequency shifts [361]. This work suggests that perhaps much of the recent selection has taken the form of partial soft sweeps – which would not show up clearly in most analyses [362].

Lastly, it is likely that most human adaptation comes through polygenic shifts of complex traits. We do know that the genetic variation in most phenotypes, aside from monogenic genetic diseases, is highly polygenic. It must surely be the case that environmental pressures are continually pushing optimal phenotypes around in some high-dimensional phenotype space as conditions change. However, polygenic adaptation leaves little trace in the data and, at the time of writing, detection remains difficult [363]. We will consider the population genetics of polygenic traits in detail later in the book.

Well done! You have now completed the population genetics section of the book! These main principles are useful for understanding all aspects of human genetic variation. In the next section we'll focus on application of these principles for understanding the genetic structure and history of human populations.

Figure 2.126: **Exquisite adaptation of the spicebush swallowtail caterpillar.** *This caterpillar discourages would-be predators using pigmented spots that mimic snake eyes.* *Credit: Michael Hodge [Link] CC BY 2.*

Notes and References.

[310]A short history of population genetics:
Charlesworth B, Charlesworth D. Population genetics from 1966 to 2016. Heredity. 2017;118(1):2-9

[311]In 1963 Dick Lewontin who, a few years later, helped introduce electrophoresis into population genetics, lamented the plight of population genetics in the absence of data: *"In many ways the lot of the theoretical population geneticist of 1963 is a most unhappy one. For he is employed, and has been employed for the last thirty years, in polishing with finer and finer grades of jeweler's rouge these three colossal monuments of mathematical biology...By the end of 1932 Haldane, Fisher, and Wright had said everything of truly fundamental importance about the theory of genetic change in populations and it is due mainly to man's infinite capacity to make more and more out of less and less, that the rest of us are not currently among the unemployed."* As quoted in Singh and Krimbas, Evolutionary Genetics: From molecules to morphology, Chapter 11; the original does not seem to be online.

[312]A short history of electrophoresis:
Charlesworth B, Charlesworth D, Coyne JA, Langley CH. Hubby and Lewontin on protein variation in natural populations: when molecular genetics came to the rescue of population genetics. Genetics. 2016;203(4):1497-503

[313]Harris H. C. Genetics of Man Enzyme polymorphisms in man. Proceedings of the Royal Society of London Series B Biological Sciences. 1966;164(995):298-310
Hubby JL, Lewontin RC. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics. 1966;54(2):577
Lewontin RC, Hubby JL. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics. 1966;54(2):595

[314]Charlesworth et al (2016).

[315]One viewpoint, motivated by observations of balanced inversion polymorphisms in Drosophila pseudoobscura, by Dobzhansky, emphasized the importance of balancing selection.

[316]Lewontin and Hubby (1966).

[317]Zuckerkandl and Pauling called this the "molecular evolutionary clock", though this is usually shortened to "molecular clock" in modern usage [REF]. See also the Kumar NRG review 2005:
Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. Journal of theoretical biology. 1965;8(2):357-66
Kumar S. Molecular clocks: four decades of evolution. Nature Reviews Genetics. 2005;6(8):654-62

[318]Dickerson RE. The structure of cytochrome c and the rates of molecular evolution. Journal of Molecular Evolution. 1971;1:26-45

[319]King JL, Jukes TH. Non-Darwinian Evolution: Most evolutionary change in proteins may be due to neutral mutations and genetic drift. Science. 1969;164(3881):788-98.

[320]Two key papers in 1968 helped to outline this: Kimura (1968); King and Jukes (1968). In the longer run, Kimura became most influential due to his continued work on this, including his 1983 book.
Kimura M, et al. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624-6

[321]The quotes are from the Introduction to Kimura (1983):
Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983

[322]One recent review is strongly critical of the Neutral Theory, in part for under-appreciating the role of linked selection:
Kern AD, Hahn MW. The neutral theory in light of natural selection. Molecular biology and evolution. 2018;35(6):1366-71
however, to the extent that the linked selection signal is due to background selection it can actually be viewed as a natural extension of the Neutral Theory:
Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution. 2019;73(1):111-4

[323]So far we have been following the Neutral Theory in treating mutations as either neutral, or strongly deleterious. However, starting in 1973, another Japanese scientist Tomoko Ohta emphasized the role of nearly-neutral mutations in protein evolution (Ohta 1973 paper, and later Annals review). In contrast to this simplest model, she argued that many amino acid substitutions may be weakly selected – i.e., with $|2Ns|$ around 1 or less. Notice that the "drift barrier" model discussed earlier is closely related to this model. The Nearly Neutral model allows for much more complexity in protein evolution: for example we can expect higher substitution rates in populations with smaller effective population sizes. Hence in the Nearly Neutral model, λ, the fraction of approximately neutral sites, is no longer a fixed property of a gene,

but instead increases or decreases depending on changes in N_e. Secondly, the fixation of nearly neutral mutations can lead to clumping of substitutions over time, because the substitution of one weakly deleterious mutation may be followed by substitution of weakly advantageous compensatory mutations nearby.

[324]Sackton TB. Studying natural selection in the era of ubiquitous genomes. Trends in Genetics. 2020;36(10):792-803

[325]Technically, here, T is the average coalescent time for lineages from each of the two species, rather than the species split time.

[326]Note that in data analysis, the number of sequence differences between two species is actually a lower bound on the number of substitutions, as there may be "multiple hits": i.e., positions that have had multiple substitutions; there are many statistical methods to account for this.

[327]Variants are sufficiently deleterious that they have essentially no chance of fixing if $s << -1/N$.

[328]It's long been observed that the molecular clock is not *precisely* clocklike. The strongest version of the molecular clock model would suggest that substitutions occur at a constant rate, uniformly in time (technically, as a Poisson Process with a fixed rate). In practice, substitutions tend to be more clumped within a phylogeny than expected under the ideal clock model; this is referred to as the **overdispersed molecular clock**. Early work documenting this argued that the overdispersed clock was evidence against the Neutral Model, and in favor of bursts of adaptive evolution Gillespie (1989) but later work has argued that much of this can be explained by a combination of effects, including gene- and lineage-specific changes in mutation rates, as well as substitutions of nearly neutral mutations, as in Ohta's Nearly Neutral Theory. For recent work in this area see work from Bedford and colleagues. Note that Bedford et al found stronger overdispersion at nonsynonymous sites than synonymous, indicating that these are not purely mutational effects. Secondly they found stronger overdispersion in mammals than in flies, than in yeast; this pattern suggests that overdispersion may be stronger in small populations than in large populations, which is perhaps the opposite of what we might expect if the overdispersion were mainly due to bursts of adaptation.

Gillespie JH. Lineage effects and the index of dispersion of molecular evolution. Molecular biology and evolution. 1989;6(6):636-47

Bedford T, Wapinski I, Hartl DL. Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. Genetics. 2008;179(2):977-84

Bedford T, Hartl DL. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in Drosophila. Molecular biology and evolution. 2008;25(8):1631-8

[329]The traditional notation dN/dS or d_N/d_S notation introduces multiple notational clashes: d is a distance and not a derivative; N and S refer to nonsynonymous and synonymous sites and not population size or selection. For this reason I use lower case, subscript n and s. In general the usage should hopefully be clear from context.

[330]Here I'm skating over many complexities in estimating d_n/d_s. First, it varies across papers whether these distances are treated as expected outcomes of an evolutionary process, or the realized numbers of substitutions. Even if it's the latter, these are still difficult to estimate due to the possibilities of multiple substitutions occurring at the same sites, and variation in the rates of transitions, transversions, and other mutation types. Lastly, one should be cautious when estimating ratios of random variables – for example the simple estimator can blow up for short genes if we don't observe any synonymous substitutions.

[331]You might reasonably worry about non-neutral effects on synonymous sites, including codon bias, or exonic splicing enhancers that overlap synonymous sites; but in aggregate these are generally weak compared to selective constraint on amino acid sequences so using synonymous sites as a baseline is generally a useful approximation.

[332]In practice d_n/d_s is usually estimated as a ratio of estimates, namely $\hat{d_n}/\hat{d_s}$. Interpreting this is a bit more tricky because obviously the estimate comes with sampling variation, and as a ratio of random variables the estimator is a biased estimator of λ.

[333]Jørgensen FG, Hobolth A, Hornshøj H, Bendixen C, Fredholm M, Schierup MH. Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. BMC biology. 2005;3:1-15

[334]Chapter 2.5. As before we take $h = 0.5$

[335]In humans the MHC is also known as the HLA or Human Leukocyte Antigen complex. The MHC/HLA is the main focus for transplant matching in organ donations because it is essential for distinguishing self from non-self antigens. The MHC is also the major driver of autoimmune disease – the immune system treads a delicate balance between sensitive immune surveillance for pathogens versus the risk of autoimmunity.

[336]Like at ABO, distinct allelic lineages have likely persisted for $>$ 20 million years, and there is enormous genetic diversity in the MHC region, with nucleotide diversity reaching well above 1% – more than 10-fold the genome-wide average.background;

Jensen JM, Villesen P, Friborg RM, Mailund T, Besenbacher S, Schierup MH, et al. Assembly and analysis of 100 full MHC haplotypes from the Danish population. Genome research. 2017;27(9):1597-607

Norman PJ, Norberg SJ, Guethlein LA, Nemat-Gorgani N, Royce T, Wroblewski EE, et al. Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. Genome research. 2017;27(5):813-23

[337]Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335(6186):167-70.

[338]One other fascinating example of high d_n/d_s is found in the PRDM9 zinc fingers, which you will recall from Chapter 2.3 play the critical role of directing recombination events:

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS genetics. 2009;5(12):e1000753.

[339]For this reason there has been a great deal of work on improving power to identify particular sites that are subject to positive selection, even within genes that are constrained at most positions eg:

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007;24(8):1586-91

Sackton TB. Studying natural selection in the era of ubiquitous genomes. Trends in Genetics. 2020;36(10):792-803.

[340]McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351(6328):652-4.

The MK test built on other contemporaneous work, including notably the HKA test

Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116(1):153-9

[341]It's beyond our scope here, but there has been a great deal of work on more complicated models that extend this basic idea. One weakness of the original MK test is that it ignores the fact that deleterious variants are much more likely to be polymorphic than to be substitutions: this in turn reduces power to detect an excess of nonsynonymous substitutions. However, it's possible to improve the test by considering only common variants, or to use the polymorphism data to estimate a distribution of selection coefficients to make more-powerful MK tests, eg:

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153-7

Messer PW, Petrov DA. Frequent adaptation and the McDonald–Kreitman test. Proceedings of the National Academy of Sciences. 2013;110(21):8615-20

[342]However it's worth noting that as the tests become more powerful, they also become more sensitive to model assumptions. One key vulnerability is variation in ancestral population sizes: for example, a small ancestral population size could allow more weakly deleterious variants to fix, and conversely for a large ancestral population size:

Eyre-Walker A. Changing effective population size and the McDonald-Kreitman test. Genetics. 2002;162(4):2017-24

[343]Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the Drosophila genome? PLoS genetics. 2009;5(6):e1000495

Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular biology and evolution. 2009;26(9):2097-108

[344]Eyre-Walker and Keightley (2009) write that analysis of the human data "...reveals little evidence for adaptive substitutions. However, the true frequency of adaptive substitutions in human-coding DNA could be as high as 40%, because estimates based on current polymorphism may be strongly downwardly biased by a decrease in the effective population size along the human lineage." Boyko et al (2008) estimated 9% in their baseline model. Uricchio et al (2019) estimated 13%. Again, it's important to take all of these estimates with caution as the MK test is easily misled by changes in N_e, which affect the rates of fixation of nearly neutral variants.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS genetics. 2008;4(5):e1000083

Uricchio LH, Petrov DA, Enard D. Exploiting selection at linked sites to infer the rate and strength of adaptation. Nature ecology & evolution. 2019;3(6):977-84

[345]Interactions between selected sites, or between selected sites and nearby neutral sites are sometimes referred to as **Hill-Robertson interference**, based on early work showing that selection at linked sites tends to reduce the efficacy of selection at both sites.

Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetics Research. 1966;8(3):269-94

Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974;78(2):737-56

[346]Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992;356(6369):519-20

[347]Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS genetics. 2009;5(1):e1000336

[348]Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289-303

Charlesworth B. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. Journal of Heredity. 2013;104(2):161-71

[349]Theory on background selection: Charlesworth et al (1993);

Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141(4):1605-17

Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genetics Research. 1996;67(2):159-74

Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, et al. A genomic map of the effects of linked selection in Drosophila. PLoS genetics. 2016;12(8):e1006130

Buffalo V, Kern AD. A Quantitative Genetic Model of Background Selection in Humans. bioRxiv. 2023:2023-09

[350]Note that for consistency with the background selection literature, and to simplify the notation, we use $s > 0$ in this section to indicate a deleterious allele, i.e., that fitnesses $1, 1-hs, 1-s$, with $h \in (0,1]$ and $s > 0$ indicate a deleterious derived allele.

We can solve for f by noting that the input of new deleterious mutations per generation is $2NL\mu$, and the number of deleterious mutations removed by selection is $N \cdot 2f(1-f) \cdot hs$ (the latter uses Hardy Weinberg, assuming that f is low enough that most deleterious mutations are heterozygous). At equilibrium, input equals output, and solving for f we get $f \approx 2NL\mu/hs$.

[351]This is Equation 11 from Nordborg et al (1996); see also Hudson and Kaplan (1994)

[352]This expression implies the interesting result that for a fixed distance r, the background selection effect is strongest when $hs = r$. In other words, at nearby functional elements (small r), small values of hs remove the most variation because the deleterious variants can drift up to become relatively common before ultimately being removed. But at large distances, only strong selection really matters: if selection is weak the linked variants have time to recombine to other chromosomes. Thus, assuming a recombination rate of 1cM/MB, at 100kb from a function region, weakly deleterious variants with $hs = 0.1\%$ would have the most impact but at 1MB distance variants with stronger effects, $hs = 1\%$, would have the most impact.

[353]Coop 2020 Eq 13.13; Nordborg, Elyashiv et al (2016) Eq 2. Note that for computational purposes it is common to use the further approximation that $1 - x \approx e^{-x}$ and then to rewrite this in the form $\exp\sum x_i$.

[354]McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics. 2009;5(5):e1000471

[355]It's sometimes known as McVicker's B, which is an example of Stigler's Law of Eponymy [Link].

[356]Murphy DA, Elyashiv E, Amster G, Sella G. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. Elife. 2022;12:e76065

[357]For examples of contrasting views see Lohmueller et al (2011), Enard et al (2014)

Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS genetics. 2011;7(10):e1002326

Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome research. 2014;24(6):885-95

and additional references as follow.

[358]This method was pioneered by

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. science. 2011;331(6019):920-4

Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans. PLoS genetics. 2011;7(2):e1001302

Here I present results from the updated analysis by Murphy et al (2021).

[359]Or 25% with moderate selection (s=0.1%). The power analyses are from Hernandez (2011)

[360]Elyashiv et al (2016) estimated that 4% of missense substitutions were fixed by strong selection, and 35% by weak selection.

[361] Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528(7583):499-503

[362] There is a large literature on selection scans in humans and primates, using a variety of analysis techniques and data, and reaching different conclusions on the frequency, strength, and types of selection. Some of these discrepancies may reflect poor calibration of some studies, but my suspicion is that much of this probably reflects a lot of weak, soft, selection that forces variants up or down in frequency but rarely to fixation. This would lead to low power and poor replication across study types. It's plausible that a lot of this selection is actually the tail-end of the distribution of polygenic effects.

[363] Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. Elife. 2019;8:e39725

Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife. 2019;8:e39702

Cox SL, Ruff CB, Maier RM, Mathieson I. Genetic contributions to variation in human stature in prehistoric Europe. Proceedings of the National Academy of Sciences. 2019;116(43):21484-92

Chen M, Sidore C, Akiyama M, Ishigaki K, Kamatani Y, Schlessinger D, et al. Evidence of polygenic adaptation in Sardinia at height-associated loci ascertained from the Biobank Japan. The American Journal of Human Genetics. 2020;107(1):60-71

Hayward LK, Sella G. Polygenic adaptation after a sudden change in environment. Elife. 2022;11:e66697

Part 3.

Human population history and structure

In Part 3 we turn our attention to ***human population structure and history***[f].

We discuss how concepts from population genetics, combined with modern genomic technologies, have rewritten our understanding of human history and prehistory. Genetics provides insights that are completely distinct from the classical approaches in paleontology, archaeology, and history.

In this section of the book we will emphasize ***inference****: how can we apply population genetics to modern genetic data to learn about structure and history?*

We'll use these principles to discuss key ***examples****: archaic hominids and their relations with modern humans; deep population structure in Africa and the (relatively) recent origin of non-Africans; to the migrations of Pacific Islanders in the last millennium.*

Specifically, we will cover the following:

Chapter 3.1: ***Population structure:*** *the genetic structuring of modern human populations, resulting from ancestry, drift, and mixture.*

Chapter 3.2: ***Inference of population histories:*** *a tour of how we can use genetic variation in modern humans to reveal the history of our species.*

Chapter 3.3: ***Ancient DNA:*** *how new technologies for retrieving DNA from bones have reshaped our understanding of human prehistory.*

[f] I expect to release this section in late 2023/ early 2024 –JKP.

3.1 Population structure and ancestry estimation.

3.2 Inferring human prehistory from genetic data.

3.3 Digging deeper into human history: Ancient DNA.

Part 4.

Genetics of phenotypic variation and disease

*In Part 4 we turn our attention to the **genetics of phenotypic variation**. We'll cover three main categories of traits: monogenic diseases, cancer, and complex traits, with particular emphasis on complex traits.*

As you read you should pay attention to the themes that repeat, but with key differences, across the different categories:

- *The number, allele frequencies and molecular mechanisms of variants;*
- *The types of selection that are most relevant;*
- *The study designs used to identify causal genes and variants;*
- *The main conceptual approaches to data analysis, and major insights.*

Specifically, we cover the following:

*Chapter 4.1: A **Starter Pack** of trait genetics: an introduction to the topics in this section.*

*Chapter 4.2: The **genetics of monogenic diseases**: mapping approaches, the major mechanisms, and selection.*

*Chapter 4.3: The **genetics of cancer**, emphasizing aspects of this huge field that intersect our main themes including somatic mutation and selection.*

*Chapter 4.4: **Quantitative Genetics**: statistical models for the inheritance of polygenic traits, including heritability and artificial selection.*

*Chapter 4.5–4.7: An overview of the main approaches for studying **complex traits**, and major emerging themes: GWAS; SNP heritability; regulatory genomics and the mechanisms of variant effects; and polygenic scores.*

*Chapter 4.8: We close with more on the **population genetics of complex traits** including stabilizing selection and polygenic adaptation.*

I expect to release these chapters in 2024 –JKP.

Notes and References

[310]A short history of population genetics:
Charlesworth B, Charlesworth D. Population genetics from 1966 to 2016. Heredity. 2017;118(1):2-9

[311]In 1963 Dick Lewontin who, a few years later, helped introduce electrophoresis into population genetics, lamented the plight of population genetics in the absence of data: *"In many ways the lot of the theoretical population geneticist of 1963 is a most unhappy one. For he is employed, and has been employed for the last thirty years, in polishing with finer and finer grades of jeweler's rouge these three colossal monuments of mathematical biology...By the end of 1932 Haldane, Fisher, and Wright had said everything of truly fundamental importance about the theory of genetic change in populations and it is due mainly to man's infinite capacity to make more and more out of less and less, that the rest of us are not currently among the unemployed."* As quoted in Singh and Krimbas, Evolutionary Genetics: From molecules to morphology, Chapter 11; the original does not seem to be online.

[312]A short history of electrophoresis:
Charlesworth B, Charlesworth D, Coyne JA, Langley CH. Hubby and Lewontin on protein variation in natural populations: when molecular genetics came to the rescue of population genetics. Genetics. 2016;203(4):1497-503

[313]Harris H. C. Genetics of Man Enzyme polymorphisms in man. Proceedings of the Royal Society of London Series B Biological Sciences. 1966;164(995):298-310
Hubby JL, Lewontin RC. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics. 1966;54(2):577
Lewontin RC, Hubby JL. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics. 1966;54(2):595

[314]Charlesworth et al (2016).

[315]One viewpoint, motivated by observations of balanced inversion polymorphisms in Drosophila pseudoobscura, by Dobzhansky, emphasized the importance of balancing selection.

[316]Lewontin and Hubby (1966).

[317]Zuckerkandl and Pauling called this the "molecular evolutionary clock", though this is usually shortened to "molecular clock" in modern usage [REF]. See also the Kumar NRG review 2005:
Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. Journal of theoretical biology. 1965;8(2):357-66
Kumar S. Molecular clocks: four decades of evolution. Nature Reviews Genetics. 2005;6(8):654-62

[318]Dickerson RE. The structure of cytochrome c and the rates of molecular evolution. Journal of Molecular Evolution. 1971;1:26-45

[319]King JL, Jukes TH. Non-Darwinian Evolution: Most evolutionary change in proteins may be due to neutral mutations and genetic drift. Science. 1969;164(3881):788-98.

[320]Two key papers in 1968 helped to outline this: Kimura (1968); King and Jukes (1968). In the longer run, Kimura became most influential due to his continued work on this, including his 1983 book.
Kimura M, et al. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624-6

[321]The quotes are from the Introduction to Kimura (1983):
Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983

[322]One recent review is strongly critical of the Neutral Theory, in part for under-appreciating the role of linked selection:

Kern AD, Hahn MW. The neutral theory in light of natural selection. Molecular biology and evolution. 2018;35(6):1366-71

however, to the extent that the linked selection signal is due to background selection it can actually be viewed as a natural extension of the Neutral Theory:

Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution. 2019;73(1):111-4

[323]So far we have been following the Neutral Theory in treating mutations as either neutral, or strongly deleterious. However, starting in 1973, another Japanese scientist Tomoko Ohta emphasized the role of nearly-neutral mutations in protein evolution (Ohta 1973 paper, and later Annals review). In contrast to this simplest model, she argued that many amino acid substitutions may be weakly selected – i.e., with $|2Ns|$ around 1 or less. Notice that the "drift barrier" model discussed earlier is closely related to this model. The Nearly Neutral model allows for much more complexity in protein evolution: for example we can expect higher substitution rates in populations with smaller effective population sizes. Hence in the Nearly Neutral model, λ, the fraction of approximately neutral sites, is no longer a fixed property of a gene, but instead increases or decreases depending on changes in N_e. Secondly, the fixation of nearly neutral mutations can lead to clumping of substitutions over time, because the substitution of one weakly deleterious mutation may be followed by substitution of weakly advantageous compensatory mutations nearby.

[324]Sackton TB. Studying natural selection in the era of ubiquitous genomes. Trends in Genetics. 2020;36(10):792-803

[325]Technically, here, T is the average coalescent time for lineages from each of the two species, rather than the species split time.

[326]Note that in data analysis, the number of sequence differences between two species is actually a lower bound on the number of substitutions, as there may be "multiple hits": i.e., positions that have had multiple substitutions; there are many statistical methods to account for this.

[327]Variants are sufficiently deleterious that they have essentially no chance of fixing if $s << -1/N$.

[328]It's long been observed that the molecular clock is not *precisely* clocklike. The strongest version of the molecular clock model would suggest that substitutions occur at a constant rate, uniformly in time (technically, as a Poisson Process with a fixed rate). In practice, substitutions tend to be more clumped within a phylogeny than expected under the ideal clock model; this is referred to as the **overdispersed molecular clock**. Early work documenting this argued that the overdispersed clock was evidence against the Neutral Model, and in favor of bursts of adaptive evolution Gillespie (1989) but later work has argued that much of this can be explained by a combination of effects, including gene- and lineage-specific changes in mutation rates, as well as substitutions of nearly neutral mutations, as in Ohta's Nearly Neutral Theory. For recent work in this area see work from Bedford and colleagues. Note that Bedford et al found stronger overdispersion at nonsynonymous sites than synonymous, indicating that these are not purely mutational effects. Secondly they found stronger overdispersion in mammals than in flies, than in yeast; this pattern suggests that overdispersion may be stronger in small populations than in large populations, which is perhaps the opposite of what we might expect if the overdispersion were mainly due to bursts of adaptation.

Gillespie JH. Lineage effects and the index of dispersion of molecular evolution. Molecular biology and evolution. 1989;6(6):636-47

Bedford T, Wapinski I, Hartl DL. Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. Genetics. 2008;179(2):977-84

Bedford T, Hartl DL. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in Drosophila. Molecular biology and evolution. 2008;25(8):1631-8

[329]The traditional notation dN/dS or d_N/d_S notation introduces multiple notational clashes: d is a distance and not a derivative; N and S refer to nonsynonymous and synonymous sites and not population size or selection. For this reason I use lower case, subscript n and s. In general the usage should hopefully be clear from context.

[330]Here I'm skating over many complexities in estimating d_n/d_s. First, it varies across papers whether these distances are treated as expected outcomes of an evolutionary process, or the realized numbers of substitutions. Even if it's the latter, these are still difficult to estimate due to the possibilities of multiple substitutions occurring at the same sites, and variation in the rates of transitions, transversions, and other mutation types. Lastly, one should be cautious when estimating ratios of random variables – for example the simple estimator can blow up for short genes if we don't observe any synonymous substitutions.

[331]You might reasonably worry about non-neutral effects on synonymous sites, including codon bias, or exonic splicing enhancers that overlap synonymous sites; but in aggregate these are generally weak compared to selective constraint

on amino acid sequences so using synonymous sites as a baseline is generally a useful approximation.

[332]In practice d_n/d_s is usually estimated as a ratio of estimates, namely $\hat{d_n}/\hat{d_s}$. Interpreting this is a bit more tricky because obviously the estimate comes with sampling variation, and as a ratio of random variables the estimator is a biased estimator of λ.

[333]Jørgensen FG, Hobolth A, Hornshøj H, Bendixen C, Fredholm M, Schierup MH. Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. BMC biology. 2005;3:1-15

[334]Chapter 2.5. As before we take $h = 0.5$

[335]In humans the MHC is also known as the HLA or Human Leukocyte Antigen complex. The MHC/HLA is the main focus for transplant matching in organ donations because it is essential for distinguishing self from non-self antigens. The MHC is also the major driver of autoimmune disease – the immune system treads a delicate balance between sensitive immune surveillance for pathogens versus the risk of autoimmunity.

[336]Like at ABO, distinct allelic lineages have likely persisted for $>$ 20 million years, and there is enormous genetic diversity in the MHC region, with nucleotide diversity reaching well above 1% – more than 10-fold the genome-wide average.background;

Jensen JM, Villesen P, Friborg RM, Mailund T, Besenbacher S, Schierup MH, et al. Assembly and analysis of 100 full MHC haplotypes from the Danish population. Genome research. 2017;27(9):1597-607

Norman PJ, Norberg SJ, Guethlein LA, Nemat-Gorgani N, Royce T, Wroblewski EE, et al. Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. Genome research. 2017;27(5):813-23

[337]Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335(6186):167-70.

[338]One other fascinating example of high d_n/d_s is found in the PRDM9 zinc fingers, which you will recall from Chapter 2.3 play the critical role of directing recombination events:

Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS genetics. 2009;5(12):e1000753.

[339]For this reason there has been a great deal of work on improving power to identify particular sites that are subject to positive selection, even within genes that are constrained at most positions eg:

Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007;24(8):1586-91

Sackton TB. Studying natural selection in the era of ubiquitous genomes. Trends in Genetics. 2020;36(10):792-803.

[340]McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351(6328):652-4.

The MK test built on other contemporaneous work, including notably the HKA test

Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116(1):153-9

[341]It's beyond our scope here, but there has been a great deal of work on more complicated models that extend this basic idea. One weakness of the original MK test is that it ignores the fact that deleterious variants are much more likely to be polymorphic than to be substitutions: this in turn reduces power to detect an excess of nonsynonymous substitutions. However, it's possible to improve the test by considering only common variants, or to use the polymorphism data to estimate a distribution of selection coefficients to make more-powerful MK tests, eg:

Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153-7

Messer PW, Petrov DA. Frequent adaptation and the McDonald–Kreitman test. Proceedings of the National Academy of Sciences. 2013;110(21):8615-20

[342]However it's worth noting that as the tests become more powerful, they also become more sensitive to model assumptions. One key vulnerability is variation in ancestral population sizes: for example, a small ancestral population size could allow more weakly deleterious variants to fix, and conversely for a large ancestral population size:

Eyre-Walker A. Changing effective population size and the McDonald-Kreitman test. Genetics. 2002;162(4):2017-24

[343]Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the Drosophila genome? PLoS ge-

netics. 2009;5(6):e1000495

Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular biology and evolution. 2009;26(9):2097-108

[344]Eyre-Walker and Keightley (2009) write that analysis of the human data "...reveals little evidence for adaptive substitutions. However, the true frequency of adaptive substitutions in human-coding DNA could be as high as 40%, because estimates based on current polymorphism may be strongly downwardly biased by a decrease in the effective population size along the human lineage." Boyko et al (2008) estimated 9% in their baseline model. Uricchio et al (2019) estimated 13%. Again, it's important to take all of these estimates with caution as the MK test is easily misled by changes in N_e, which affect the rates of fixation of nearly neutral variants.

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS genetics. 2008;4(5):e1000083

Uricchio LH, Petrov DA, Enard D. Exploiting selection at linked sites to infer the rate and strength of adaptation. Nature ecology & evolution. 2019;3(6):977-84

[345]Interactions between selected sites, or between selected sites and nearby neutral sites are sometimes referred to as **Hill-Robertson interference**, based on early work showing that selection at linked sites tends to reduce the efficacy of selection at both sites.

Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetics Research. 1966;8(3):269-94

Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974;78(2):737-56

[346]Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992;356(6369):519-20

[347]Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS genetics. 2009;5(1):e1000336

[348]Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289-303

Charlesworth B. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. Journal of Heredity. 2013;104(2):161-71

[349]Theory on background selection: Charlesworth et al (1993);

Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141(4):1605-17

Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genetics Research. 1996;67(2):159-74

Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, et al. A genomic map of the effects of linked selection in Drosophila. PLoS genetics. 2016;12(8):e1006130

Buffalo V, Kern AD. A Quantitative Genetic Model of Background Selection in Humans. bioRxiv. 2023:2023-09

[350]Note that for consistency with the background selection literature, and to simplify the notation, we use $s > 0$ in this section to indicate a deleterious allele, i.e., that fitnesses $1, 1-hs, 1-s$, with $h \in (0,1]$ and $s > 0$ indicate a deleterious derived allele.

We can solve for f by noting that the input of new deleterious mutations per generation is $2NL\mu$, and the number of deleterious mutations removed by selection is $N \cdot 2f(1-f) \cdot hs$ (the latter uses Hardy Weinberg, assuming that f is low enough that most deleterious mutations are heterozygous). At equilibrium, input equals output, and solving for f we get $f \approx 2NL\mu/hs$.

[351]This is Equation 11 from Nordborg et al (1996); see also Hudson and Kaplan (1994)

[352]This expression implies the interesting result that for a fixed distance r, the background selection effect is strongest when $hs = r$. In other words, at nearby functional elements (small r), small values of hs remove the most variation because the deleterious variants can drift up to become relatively common before ultimately being removed. But at large distances, only strong selection really matters: if selection is weak the linked variants have time to recombine to other chromosomes. Thus, assuming a recombination rate of 1cM/MB, at 100kb from a function region, weakly deleterious variants with $hs = 0.1\%$ would have the most impact but at 1MB distance variants with stronger effects, $hs = 1\%$, would have the most impact.

[353]Coop 2020 Eq 13.13; Nordborg, Elyashiv et al (2016) Eq 2. Note that for computational purposes it is common to use the further approximation that $1 - x \approx e^{-x}$ and then to rewrite this in the form $\exp\sum x_i$.

[354]McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolu-

tion. PLoS genetics. 2009;5(5):e1000471

[355]It's sometimes known as McVicker's B, which is an example of Stigler's Law of Eponymy [Link].

[356]Murphy DA, Elyashiv E, Amster G, Sella G. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. Elife. 2022;12:e76065

[357]For examples of contrasting views see Lohmueller et al (2011), Enard et al (2014)
Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS genetics. 2011;7(10):e1002326
Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome research. 2014;24(6):885-95
and additional references as follow.

[358]This method was pioneered by
Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. science. 2011;331(6019):920-4
Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans. PLoS genetics. 2011;7(2):e1001302
Here I present results from the updated analysis by Murphy et al (2021).

[359]Or 25% with moderate selection (s=0.1%). The power analyses are from Hernandez (2011)

[360]Elyashiv et al (2016) estimated that 4% of missense substitutions were fixed by strong selection, and 35% by weak selection.

[361]Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528(7583):499-503

[362]There is a large literature on selection scans in humans and primates, using a variety of analysis techniques and data, and reaching different conclusions on the frequency, strength, and types of selection. Some of these discrepancies may reflect poor calibration of some studies, but my suspicion is that much of this probably reflects a lot of weak, soft, selection that forces variants up or down in frequency but rarely to fixation. This would lead to low power and poor replication across study types. It's plausible that a lot of this selection is actually the tail-end of the distribution of polygenic effects.

[363]Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. Elife. 2019;8:e39725
Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife. 2019;8:e39702
Cox SL, Ruff CB, Maier RM, Mathieson I. Genetic contributions to variation in human stature in prehistoric Europe. Proceedings of the National Academy of Sciences. 2019;116(43):21484-92
Chen M, Sidore C, Akiyama M, Ishigaki K, Kamatani Y, Schlessinger D, et al. Evidence of polygenic adaptation in Sardinia at height-associated loci ascertained from the Biobank Japan. The American Journal of Human Genetics. 2020;107(1):60-71
Hayward LK, Sella G. Polygenic adaptation after a sudden change in environment. Elife. 2022;11:e66697

Bibliography

[1] Milo R, Phillips R. Cell biology by the numbers. Garland Science; 2015.

[2] Nurk S, Koren S, Rhie A, Rautiainen M, Bzikadze AV, Mikheenko A, et al. The complete sequence of a human genome. Science. 2022;376(6588):44-53.

[3] Erlich Y, Zielinski D. DNA Fountain enables a robust and efficient storage architecture. Science. 2017;355(6328):950-4.

[4] Kim J, Bae JH, Baym M, Zhang DY. Metastable hybridization-based DNA information storage to allow rapid and permanent erasure. Nature Communications. 2020;11(1):1-8.

[5] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596(7873):583-9.

[6] Barbosa-Morais NL, Irimia M, Pan Q, Xiong HY, Gueroussov S, Lee LJ, et al. The evolutionary landscape of alternative splicing in vertebrate species. Science. 2012;338(6114):1587-93.

[7] Merkin J, Russell C, Chen P, Burge CB. Evolutionary dynamics of gene and isoform regulation in Mammalian tissues. Science. 2012;338(6114):1593-9.

[8] Li YI, Knowles DA, Humphrey J, Barbeira AN, Dickinson SP, Im HK, et al. Annotation-free quantification of RNA splicing using LeafCutter. Nature Genetics. 2018;50(1):151-8.

[9] Rosenberg AB, Patwardhan RP, Shendure J, Seelig G. Learning the sequence determinants of alternative splicing from millions of random sequences. Cell. 2015;163(3):698-711.

[10] Jaganathan K, Panagiotopoulou SK, McRae JF, Darbandi SF, Knowles D, Li YI, et al. Predicting splicing from primary sequence with deep learning. Cell. 2019;176(3):535-48.

[11] Zeng T, Li YI. Predicting RNA splicing from DNA sequence using Pangolin. Genome Biology. 2022;23(1):1-18.

[12] Calof AL, Santos R, Groves L, Oliver C, Lander AD. Cornelia de Lange syndrome: Insights into neural development from clinical studies and animal models. In: Neurodevelopmental Disorders. Elsevier; 2020. p. 129-57.

[13] Muniz L, Nicolas E, Trouche D. RNA polymerase II speed: a key player in controlling and adapting transcriptome composition. The EMBO journal. 2021;40(15):e105740.

[14] Zhou J, Troyanskaya OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature Methods. 2015;12(10):931-4.

[15] Kelley DR, Snoek J, Rinn JL. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. Genome Research. 2016;26(7):990-9.

[16] Avsec Ž, Weilert M, Shrikumar A, Krueger S, Alexandari A, Dalal K, et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. Nature Genetics. 2021;53(3):354-66.

[17] Sobreira DR, Joslin AC, Zhang Q, Williamson I, Hansen GT, Farris KM, et al. Extensive pleiotropism and allelic heterogeneity mediate metabolic effects of IRX3 and IRX5. Science. 2021;372(6546):1085-91.

[18] Fulco CP, Nasser J, Jones TR, Munson G, Bergman DT, Subramanian V, et al. Activity-by-contact model of enhancer–promoter regulation from thousands of CRISPR perturbations. Nature genetics. 2019;51(12):1664-9.

[19] Ward LD, Kellis M. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science. 2012;337(6102):1675-8.

[20] Rands CM, Meader S, Ponting CP, Lunter G. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. PLoS Genetics. 2014;10(7):e1004525.

[21] Piovesan A, Caracausi M, Antonaros F, Pelleri MC, Vitale L. GeneBase 1.1: a tool to summarize data from NCBI gene datasets and its application to an update of human gene statistics. Database (Oxford). 2016;2016.

[22] Ransohoff JD, Wei Y, Khavari PA. The functions and unique features of long intergenic non-coding RNA. Nature Reviews Molecular cell biology. 2018;19(3):143-57.

[23] Ponting CP, Haerty W. Genome-wide analysis of human long noncoding RNAs: a provocative review. Annual Review of Genomics and Human Genetics. 2022;23.

[24] Liu N, Lee CH, Swigut T, Grow E, Gu B, Bassik MC, et al. Selective silencing of euchromatic L1s revealed by genome-wide screens for L1 regulators. Nature. 2018;553(7687):228-32.

[25] Chuong EB, Elde NC, Feschotte C. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science. 2016;351(6277):1083-7.

[26] Bartonicek N, Rouet R, Warren J, Loetsch C, Rodriguez GS, Walters S, et al. The retroelement Lx9 puts a brake on the immune response to virus infection. Nature. 2022:1-9.

[27] Li R, Bitoun E, Altemose N, Davies RW, Davies B, Myers SR. A high-resolution map of non-crossover events reveals impacts of genetic diversity on mammalian meiotic recombination. Nature Communications. 2019;10(1):3900.

[28] Lewontin R. The dream of the human genome: doubts about the Human Genome Project. The New York review of books. 1992;39(10):31-40.

[29] Roberts L. The Human Genome. Controversial from the start. Science. 2001;291:1182-8.

[30] Waterston RH, Lander ES, Sulston JE. On the sequencing of the human genome. Proceedings of the National Academy of Sciences. 2002;99(6):3712-6.

[31] Myers EW, Sutton GG, Smith HO, Adams MD, Venter JC. On the sequencing and assembly of the human genome. Proceedings of the National Academy of Sciences. 2002;99(7):4145-6.

[32] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. Nature. 2001;409(6822):860-921.

[33] Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, et al. The sequence of the human genome. Science. 2001;291(5507):1304-51.

[34] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. Nature. 2004;431(7011):931-45.

[35] Ballouz S, Dobin A, Gillis JA. Is it time to change the reference genome? Genome Biology. 2019;20(1):1-9.

[36] Green RE, Krause J, Briggs AW, Maricic T, Stenzel U, Kircher M, et al. A draft sequence of the Neandertal genome. Science. 2010;328(5979):710-22.

[37] Rosenberg NA. Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Annals of human genetics. 2006;70(6):841-7.

[38] Hellenthal G, Busby GB, Band G, Wilson JF, Capelli C, Falush D, et al. A genetic atlas of human admixture history. science. 2014;343(6172):747-51.

[39] Mallick S, Li H, Lipson M, Mathieson I, Gymrek M, Racimo F, et al. The Simons genome diversity project: 300 genomes from 142 diverse populations. Nature. 2016;538(7624):201-6.

[40] Lek M, Karczewski KJ, Minikel EV, Samocha KE, Banks E, Fennell T, et al. Analysis of protein-coding genetic variation in 60,706 humans. Nature. 2016;536(7616):285-91.

[41] Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;581(7809):434-43.

[42] Popejoy AB, Fullerton SM. Genomics is failing on diversity. Nature. 2016;538(7624):161-4.

[43] Edwards A. Anecdotal, Historical and Critical Commentaries on Genetics: GH Hardy (1908) and Hardy–Weinberg Equilibrium. Genetics. 2008;179(3):1143.

[44] Jonsson H, Magnusdottir E, Eggertsson HP, Stefansson OA, Arnadottir GA, Eiriksson O, et al. Differences between germline genomes of monozygotic twins. Nature Genetics. 2021;53(1):27-34.

[45] 1000 Genomes Project Consortium. A global reference for human genetic variation. Nature. 2015;526(7571):68.

[46] Scally A, Dutheil JY, Hillier LW, Jordan GE, Goodhead I, Herrero J, et al. Insights into hominid evolution from the gorilla genome sequence. Nature. 2012;483(7388):169-75.

[47] Frisse L, Hudson R, Bartoszewicz A, Wall J, Donfack J, Di Rienzo A. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. The American Journal of Human Genetics. 2001;69(4):831-43.

[48] Carlson CS, Eberle MA, Rieder MJ, Smith JD, Kruglyak L, Nickerson DA. Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. nature Genetics. 2003;33(4):518-21.

[49] Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. Comprehensive identification and characterization of diallelic insertion–deletion polymorphisms in 330 human candidate genes. Human Molecular Genetics. 2005;14(1):59-69.

[50] Montgomery SB, Goode DL, Kvikstad E, Albers CA, Zhang ZD, Mu XJ, et al. The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes. Genome Research. 2013;23(5):749-61.

[51] Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, et al. Global diversity, population stratification, and selection of human copy-number variation. Science. 2015;349(6253):aab3761.

[52] Stefansson H, Helgason A, Thorleifsson G, Steinthorsdottir V, Masson G, Barnard J, et al. A common inversion under selection in Europeans. Nature Genetics. 2005;37(2):129-37.

[53] Salm MP, Horswell SD, Hutchison CE, Speedy HE, Yang X, Liang L, et al. The origin, global distribution, and functional impact of the human 8p23 inversion polymorphism. Genome Research. 2012;22(6):1144-53.

[54] Wellenreuther M, Bernatchez L. Eco-evolutionary genomics of chromosomal inversions. Trends in Ecology & Evolution. 2018;33(6):427-40.

[55] Coble MD, Loreille OM, Wadhams MJ, Edson SM, Maynard K, Meyer CE, et al. Mystery solved: the identification of the two missing Romanov children using DNA analysis. PloS One. 2009;4(3):e4838.

[56] Rogaev EI, Grigorenko AP, Faskhutdinova G, Kittler EL, Moliaka YK. Genotype analysis identifies the cause of the "royal disease". Science. 2009;326(5954):817-7.

[57] Mukamel RE, Handsaker RE, Sherman MA, Barton AR, Zheng Y, McCarroll SA, et al. Protein-coding repeat polymorphisms strongly shape diverse human phenotypes. Science. 2021;373(6562):1499-505.

[58] Agarwal I, Fuller ZL, Myers S, Przeworski M. Relating pathogenic loss-of function mutations in humans to their evolutionary fitness costs. bioRxiv. 2022.

[59] Perry GH, Dominy NJ, Claw KG, Lee AS, Fiegler H, Redon R, et al. Diet and the evolution of human amylase gene copy number variation. Nature Genetics. 2007;39(10):1256-60.

[60] Usher CL, Handsaker RE, Esko T, Tuke MA, Weedon MN, Hastie AR, et al. Structural forms of the human amylase locus and their relationships to SNPs, haplotypes and obesity. Nature Genetics. 2015;47(8):921-5.

[61] Pajic P, Pavlidis P, Dean K, Neznanova L, Romano RA, Garneau D, et al. Independent amylase gene copy number bursts correlate with dietary preferences in mammals. Elife. 2019;8:e44628.

[62] Torres EM, Williams BR, Amon A. Aneuploidy: cells losing their balance. Genetics. 2008;179(2):737-46.

[63] Posynick BJ, Brown CJ. Escape from X-chromosome inactivation: an evolutionary perspective. Frontiers in Cell and Developmental Biology. 2019;7:241.

[64] Bellott DW, Hughes JF, Skaletsky H, Brown LG, Pyntikova T, Cho TJ, et al. Mammalian Y chromosomes retain widely expressed dosage-sensitive regulators. Nature. 2014;508(7497):494-9.

[65] Ferguson-Smith MA, Trifonov V. Mammalian karyotype evolution. Nature Reviews Genetics. 2007;8(12):950-62.

[66] Yunis JJ, Prakash O. The origin of man: a chromosomal pictorial legacy. Science. 1982;215(4539):1525-30.

[67] Ventura M, Catacchio CR, Sajjadian S, Vives L, Sudmant PH, Marques-Bonet T, et al. The evolution of African great ape subtelomeric heterochromatin and the fusion of human chromosome 2. Genome Research. 2012;22(6):1036-49.

[68] Carbone L, Harris RA, Vessere GM, Mootnick AR, Humphray S, Rogers J, et al. Evolutionary breakpoints in the gibbon suggest association between cytosine methylation and karyotype evolution. PLoS Genetics. 2009;5(6):e1000538.

[69] Carbone L, Alan Harris R, Gnerre S, Veeramah KR, Lorente-Galdos B, Huddleston J, et al. Gibbon genome and the fast karyotype evolution of small apes. Nature. 2014;513(7517):195-201.

[70] Chmátal L, Gabriel SI, Mitsainas GP, Martínez-Vargas J, Ventura J, Searle JB, et al. Centromere strength provides the cell biological basis for meiotic drive and karyotype evolution in mice. Current Biology. 2014;24(19):2295-300.

[71] Shendure J, Balasubramanian S, Church GM, Gilbert W, Rogers J, Schloss JA, et al. DNA sequencing at 40: past, present and future. Nature. 2017;550(7676):345-53.

[72] Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nature Reviews Genetics. 2016;17(6):333-51.

[73] Dohm JC, Peters P, Stralis-Pavese N, Himmelbauer H. Benchmarking of long-read correction methods. NAR Genomics and Bioinformatics. 2020;2(2):lqaa037.

[74] Gorzynski JE, Goenka SD, Shafin K, Jensen TD, Fisk DG, Grove ME, et al. Ultrarapid nanopore genome sequencing in a critical care setting. New England Journal of Medicine. 2022;386(7):700-2.

[75] Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, et al. Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Scientific Reports. 2018;8(1):1-14.

[76] Teissandier A, Servant N, Barillot E, Bourc'his D. Tools and best practices for retrotransposon analysis using high-throughput sequencing data. Mobile DNA. 2019;10(1):1-12.

[77] Lee H, Schatz MC. Genomic dark matter: the reliability of short read mapping illustrated by the genome mappability score. Bioinformatics. 2012;28(16):2097-105.

[78] Tubbs A, Nussenzweig A. Endogenous DNA damage as a source of genomic instability in cancer. Cell. 2017;168(4):644-56.

[79] Gates KS. An overview of chemical processes that damage cellular DNA: spontaneous hydrolysis, alkylation, and reactions with radicals. Chemical Research in Toxicology. 2009;22(11):1747-60.

[80] Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. Drift-barrier hypothesis and mutation-rate evolution. Proceedings of the National Academy of Sciences. 2012;109(45):18488-92.

[81] Roach JC, Glusman G, Smit AF, Huff CD, Hubley R, Shannon PT, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. Science. 2010;328(5978):636-9.

[82] Ségurel L, Wyman MJ, Przeworski M. Determinants of mutation rate variation in the human germline. Annu Rev Genomics Hum Genet. 2014;15(1):47-70.

[83] Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, et al. Rate of de novo mutations and the importance of father's age to disease risk. Nature. 2012;488(7412):471-5.

[84] Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, et al. Parental influence on human germline de novo mutations in 1,548 trios from Iceland. Nature. 2017;549(7673):519-22.

[85] Weinberg W. Zur vererbung des zwergwuchses. Arch Rassen-u Gesel Biolog. 1912;9:710-8.

[86] Crow JF, Denniston C. Mutation in human populations. Advances in Human Genetics 14. 1985:59-123.

[87] Risch N, Reich E, Wishnick M, McCarthy J. Spontaneous mutation and parental age in humans. American Journal of Human Genetics. 1987;41(2):218.

[88] Shimmin LC, Chang BHJ, Li WH. Male-driven evolution of DNA sequences. Nature. 1993;362(6422):745-7.

[89] Gao Z, Moorjani P, Sasani TA, Pedersen BS, Quinlan AR, Jorde LB, et al. Overlooked roles of DNA damage and maternal age in generating human germline mutations. Proceedings of the National Academy of Sciences. 2019;116(19):9491-500.

[90] Kaplanis J, Ide B, Sanghvi R, Neville M, Danecek P, Coorens T, et al. Genetic and chemotherapeutic influences on germline hypermutation. Nature. 2022;605(7910):503-8.

[91] Belyeu JR, Brand H, Wang H, Zhao X, Pedersen BS, Feusier J, et al. De novo structural mutation rates and gamete-of-origin biases revealed through genome sequencing of 2,396 families. The American Journal of Human Genetics. 2021;108(4):597-607.

[92] Mitra I, Huang B, Mousavi N, Ma N, Lamkin M, Yanicky R, et al. Patterns of de novo tandem repeat mutations and their role in autism. Nature. 2021;589(7841):246-50.

[93] Jaiswal S, Natarajan P, Silver AJ, Gibson CJ, Bick AG, Shvartz E, et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. New England Journal of Medicine. 2017;377(2):111-21.

[94] Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nature Protocols. 2014;9(11):2586-606.

[95] Abascal F, Harvey LM, Mitchell E, Lawson AR, Lensing SV, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021;593(7859):405-10.

[96] Sun JX, Helgason A, Masson G, Ebenesersdóttir SS, Li H, Mallick S, et al. A direct characterization of human mutation based on microsatellites. Nature Genetics. 2012;44(10):1161-5.

[97] Steely CJ, Watkins S, Baird L, Jorde L. The Mutational Dynamics of Short Tandem Repeats in Large, Multigenerational Families. bioRxiv. 2021.

[98] Kristmundsdottir S, Jonsson H, Hardarson MT, Palsson G, Beyter D, Eggertsson HP, et al. Sequence variants affecting the genome-wide rate of germline microsatellite mutations. Nature Communications. 2023;14(1):3855.

[99] Fu Q, Mittnik A, Johnson PL, Bos K, Lari M, Bollongino R, et al. A revised timescale for human evolution based on ancient mitochondrial genomes. Current Biology. 2013;23(7):553-9.

[100] Fontana GA, Gahlon HL. Mechanisms of replication and repair in mitochondrial DNA deletion formation. Nucleic Acids Research. 2020;48(20):11244-58.

[101] Gymrek M, Willems T, Guilmatre A, Zeng H, Markus B, Georgiev S, et al. Abundant contribution of short tandem repeats to gene expression variation in humans. Nature Genetics. 2016;48(1):22-9.

[102] Carvalho C, Lupski JR. Mechanisms underlying structural variant formation in genomic disorders. Nature Reviews Genetics. 2016;17(4):224-38.

[103] Ottaviani D, LeCain M, Sheer D. The role of microhomology in genomic structural variation. Trends in Genetics. 2014;30(3):85-94.

[104] Kidd JM, Graves T, Newman TL, Fulton R, Hayden HS, Malig M, et al. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. Cell. 2010;143(5):837-47.

[105] Roa BB, Garcia CA, Pentao L, Killian JM, Trask BJ, Suter U, et al. Evidence for a recessive PMP22 point mutation in Charcot–Marie–Tooth disease type 1A. Nature Genetics. 1993;5(2):189-94.

[106] Lupski JR, Reid JG, Gonzaga-Jauregui C, Rio Deiros D, Chen DC, Nazareth L, et al. Whole-genome sequencing in a patient with Charcot–Marie–Tooth neuropathy. New England Journal of Medicine. 2010;362(13):1181-91.

[107] Porubsky D, Höps W, Ashraf H, Hsieh P, Rodriguez-Martin B, Yilmaz F, et al. Recurrent inversion polymorphisms in humans associate with genetic instability and genomic disorders. Cell. 2022;185(11):1986-2005.

[108] Gao Z, Wyman MJ, Sella G, Przeworski M. Interpreting the dependence of mutation rates on age and time. PLoS biology. 2016;14(1):e1002355.

[109] Wu FL, Strand AI, Cox LA, Ober C, Wall JD, Moorjani P, et al. A comparison of humans and baboons suggests germline mutation rates do not track cell divisions. PLoS Biology. 2020;18(8):e3000838.

[110] de Manuel M, Wu FL, Przeworski M. A paternal bias in germline mutation is widespread across amniotes and can arise independently of cell divisions. bioRxiv. 2022.

[111] Vraneković J, Božović IB, Grubić Z, Wagner J, Pavlinić D, Dahoun S, et al. Down syndrome: parental origin, recombination, and maternal age. Genetic Testing and Molecular Biomarkers. 2012;16(1):70-3.

[112] Kuliev A, Zlatopolsky Z, Kirillova I, Spivakova J, Janzen JC. Meiosis errors in over 20,000 oocytes studied in the practice of preimplantation aneuploidy testing. Reproductive biomedicine online. 2011;22(1):2-8.

[113] Gruhn JR, Zielinska AP, Shukla V, Blanshard R, Capalbo A, Cimadomo D, et al. Chromosome errors in human eggs shape natural fertility over reproductive life span. Science. 2019;365(6460):1466-9.

[114] Greaney J, Wei Z, Homer H. Regulation of chromosome segregation in oocytes and the cellular basis for female meiotic errors. Human Reproduction Update. 2018;24(2):135-61.

[115] Nagaoka SI, Hassold TJ, Hunt PA. Human aneuploidy: mechanisms and new insights into an age-old problem. Nature Reviews Genetics. 2012;13(7):493-504.

[116] Webster A, Schuh M. Mechanisms of aneuploidy in human eggs. Trends in cell biology. 2017;27(1):55-68.

[117] Zielinska AP, Holubcova Z, Blayney M, Elder K, Schuh M. Sister kinetochore splitting and precocious disintegration of bivalents could explain the maternal age effect. Elife. 2015;4:e11389.

[118] Patel J, Tan SL, Hartshorne GM, McAinsh AD. Unique geometry of sister kinetochores in human oocytes during meiosis I may explain maternal age-associated increases in chromosomal abnormalities. Biology Open. 2016;5(2):178-84.

[119] Wang S, Hassold T, Hunt P, White MA, Zickler D, Kleckner N, et al. Inefficient crossover maturation underlies elevated aneuploidy in human female meiosis. Cell. 2017;168(6):977-89.

[120] So C, Menelaou K, Uraji J, Harasimov K, Steyer AM, Seres KB, et al. Mechanism of spindle pole organization and instability in human oocytes. Science. 2022;375(6581):eabj3944.

[121] Bennabi I, Terret ME, Verlhac MH. Meiotic spindle assembly and chromosome segregation in oocytes. Journal of Cell Biology. 2016;215(5):611-9.

[122] Zwick ME, Salstrom JL, Langley CH. Genetic variation in rates of nondisjunction: association of two naturally occurring polymorphisms in the chromokinesin nod with increased rates of nondisjunction in Drosophila melanogaster. Genetics. 1999;152(4):1605-14.

[123] Malik HS. The centromere-drive hypothesis: a simple basis for centromere complexity. Centromere. 2009:33-52.

[124] Kursel LE, Malik HS. The cellular mechanisms and consequences of centromere drive. Current opinion in cell biology. 2018;52:58-65.

[125] Lampson MA, Black BE. Cellular and molecular mechanisms of centromere drive. In: Cold Spring Harbor symposia on quantitative biology. vol. 82. Cold Spring Harbor Laboratory Press; 2017. p. 249-57.

[126] Hurst LD. Selfish centromeres and the wastefulness of human reproduction. PLoS Biology. 2022;20(7):e3001671.

[127] Soodyall H, Nebel A, Morar B, Jenkins T. Genealogy and genes: tracing the founding fathers of Tristan da Cunha. European Journal of Human Genetics. 2003;11(9):705-9.

[128] Chen N, Juric I, Cosgrove EJ, Bowman R, Fitzpatrick JW, Schoech SJ, et al. Allele frequency dynamics in a pedigreed natural population. Proceedings of the National Academy of Sciences. 2019;116(6):2158-64.

[129] Harpak A, Bhaskar A, Pritchard JK. Mutation rate variation is a primary determinant of the distribution of allele frequencies in humans. PLoS Genetics. 2016;12(12):e1006489.

[130] Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, et al. The genetic legacy of the Mongols. The American Journal of Human Genetics. 2003;72(3):717-21.

[131] Balaresque P, Poulet N, Cussat-Blanc S, Gerard P, Quintana-Murci L, Heyer E, et al. Y-chromosome descent clusters and male differential reproductive success: young lineage expansions dominate Asian pastoral nomadic populations. European Journal of Human Genetics. 2015;23(10):1413-22.

[132] Leffler EM, Bullaughey K, Matute DR, Meyer WK, Segurel L, Venkat A, et al. Revisiting an old riddle: what determines genetic diversity levels within species? PLOS Biology. 2012;10(9):e1001388.

[133] Messer PW. SLiM: simulating evolution with selection and linkage. Genetics. 2013;194(4):1037-9.

[134] Haller BC, Messer PW. SLiM 3: forward genetic simulations beyond the Wright–Fisher model. Molecular Biology and Evolution. 2019;36(3):632-7.

[135] Kingman JFC. The coalescent. Stochastic processes and their applications. 1982;13(3):235-48.

[136] Kingman JF. Origins of the coalescent: 1974-1982. Genetics. 2000;156(4):1461-3.

[137] Hudson RR. Testing the constant-rate neutral allele model with protein sequence data. Evolution. 1983:203-17.

[138] Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 1983;23(2):183-201.

[139] Tajima F. Evolutionary relationship of DNA sequences in finite populations. Genetics. 1983;105(2):437-60.

[140] Hudson RR. Gene genealogies and the coalescent process. Oxford surveys in evolutionary biology. 1990;7(1):44.

[141] Hudson R. The how and why of generating gene genealogies. Mechanisms of molecular evolution. 1993:23-36.

[142] Nordborg M. Coalescent theory. Handbook of Statistical Genomics: Two Volume Set. 2019:145-30.

[143] Hudson RR. A new proof of the expected frequency spectrum under the standard neutral model. Plos One. 2015;10(7):e0118087.

[144] Ramachandran S, Deshpande O, Roseman CC, Rosenberg NA, Feldman MW, Cavalli-Sforza LL. Support from the relationship of genetic and geographic distance in human populations for a serial founder effect originating in Africa. Proceedings of the National Academy of Sciences. 2005;102(44):15942-7.

[145] Waldman S, Backenroth D, Harney É, Flohr S, Neff NC, Buckley GM, et al. Genome-wide data from medieval German Jews show that the Ashkenazi founder event pre-dated the 14th century. Cell. 2022;185(25):4703-16.

[146] Slatkin M, Hudson RR. Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. Genetics. 1991;129(2):555-62.

[147] Tennessen JA, Bigham AW, O'connor TD, Fu W, Kenny EE, Gravel S, et al. Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science. 2012;337(6090):64-9.

[148] Kreitman M. Nucleotide polymorphism at the alcohol dehydrogenase locus of Drosophila melanogaster. Nature. 1983;304(5925):412-7.

[149] Halldorsson BV, Palsson G, Stefansson OA, Jonsson H, Hardarson MT, Eggertsson HP, et al. Characterizing mutagenic effects of recombination through a sequence-level genetic map. Science. 2019;363(6425):eaau1043.

[150] Pritchard JK, Przeworski M. Linkage disequilibrium in humans: models and data. The American Journal of Human Genetics. 2001;69(1):1-14.

[151] Bulik-Sullivan BK, Loh PR, Finucane HK, Ripke S, Yang J, of the Psychiatric Genomics Consortium SWG, et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nature Genetics. 2015;47(3):291-5.

[152] Hudson RR. Properties of a neutral allele model with intragenic recombination. Theoretical Population Biology. 1983;23(2):183-201.

[153] McVean GA. A genealogical interpretation of linkage disequilibrium. Genetics. 2002;162(2):987-91.

[154] Chakravarti A, Buetow K, Antonarakis S, Waber P, Boehm C, Kazazian H. Nonuniform recombination within the human beta-globin gene cluster. American Journal of Human Genetics. 1984;36(6):1239.

[155] Jeffreys AJ, Kauppi L, Neumann R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. Nature Genetics. 2001;29(2):217-22.

[156] McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, Donnelly P. The fine-scale structure of recombination rate variation in the human genome. Science. 2004;304(5670):581-4.

[157] Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. A fine-scale map of recombination rates and hotspots across the human genome. Science. 2005;310(5746):321-4.

[158] Boulton A, Myers RS, Redfield RJ. The hotspot conversion paradox and the evolution of meiotic recombination. Proceedings of the National Academy of Sciences. 1997;94(15):8058-63.

[159] Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, et al. Fine-scale recombination patterns differ between chimpanzees and humans. Nature Genetics. 2005;37(4):429-34.

[160] Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, et al. Comparison of fine-scale recombination rates in humans and chimpanzees. Science. 2005;308(5718):107-11.

[161] Coop G, Wen X, Ober C, Pritchard JK, Przeworski M. High-resolution mapping of crossovers reveals extensive variation in fine-scale recombination patterns among humans. science. 2008;319(5868):1395-8.

[162] Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, et al. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science. 2010;327(5967):836-40.

[163] Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, et al. Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. Science. 2010;327(5967):876-9.

[164] Parvanov ED, Petkov PM, Paigen K. Prdm9 controls activation of mammalian recombination hotspots. Science. 2010;327(5967):835-5.

[165] Berg IL, Neumann R, Lam KWG, Sarbajna S, Odenthal-Hesse L, May CA, et al. PRDM9 variation strongly influences recombination hot-spot activity and meiotic instability in humans. Nature Genetics. 2010;42(10):859-63.

[166] Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, et al. The landscape of recombination in African Americans. Nature. 2011;476(7359):170-5.

[167] Baker Z, Przeworski M, Sella G. Down the Penrose stairs: How selection for fewer recombination hotspots maintains their existence. bioRxiv. 2022:2022-09.

[168] Li N, Stephens M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. Genetics. 2003;165(4):2213-33.

[169] Song YS. Na Li and Matthew Stephens on modeling linkage disequilibrium. Genetics. 2016;203(3):1005-6.

[170] Consortium" THR. A reference panel of 64,976 haplotypes for genotype imputation. Nature Genetics. 2016;48(10):1279-83.

[171] Biddanda A, Rice DP, Novembre J. A variant-centric perspective on geographic patterns of human allele frequency variation. Elife. 2020;9:e60107.

[172] Nicholson G, Smith AV, Jónsson F, Gústafsson Ó, Stefánsson K, Donnelly P. Assessing population differentiation and isolation from single-nucleotide polymorphism data. Journal of the Royal Statistical Society Series B: Statistical Methodology. 2002;64(4):695-715.

[173] Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZXP, Pool JE, et al. Sequencing of 50 human exomes reveals adaptation to high altitude. Science. 2010;329(5987):75-8.

[174] Huerta-Sánchez E, Jin X, Asan, Bianba Z, Peter BM, Vinckenbosch N, et al. Altitude adaptation in Tibetans caused by introgression of Denisovan-like DNA. Nature. 2014;512(7513):194-7.

[175] Jewett EM, Rosenberg NA. Theory and applications of a deterministic approximation to the coalescent model. Theoretical population biology. 2014;93:14-29.

[176] Slatkin M. Allele age and a test for selection on rare alleles. Philosophical Transactions of the Royal Society of London Series B: Biological Sciences. 2000;355(1403):1663-8.

[177] Wright S. The genetical structure of populations. Annals of eugenics. 1949;15(1):323-54.

[178] Bhatia G, Patterson N, Sankararaman S, Price AL. Estimating and interpreting FST: the impact of rare variants. Genome Research. 2013;23(9):1514-21.

[179] Barton N. Identity and coalescence in structured populations: a commentary on 'Inbreeding coefficients and coalescence times' by Montgomery Slatkin. Genetics Research. 2007;89(5-6):475-7.

[180] Holsinger KE, Weir BS. Genetics in geographically structured populations: defining, estimating and interpreting FST. Nature Reviews Genetics. 2009;10(9):639-50.

[181] Slatkin M. Inbreeding coefficients and coalescence times. Genetics Research. 1991;58(2):167-75.

[182] Ruvolo M. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. Molecular biology and evolution. 1997;14(3):248-65.

[183] Hobolth A, Christensen OF, Mailund T, Schierup MH. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden Markov model. PLoS genetics. 2007;3(2):e7.

[184] Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome research. 2011;21(3):349-56.

[185] Prado-Martinez J, Sudmant PH, Kidd JM, Li H, Kelley JL, Lorente-Galdos B, et al. Great ape genetic diversity and population history. Nature. 2013;499(7459):471-5.

[186] Amster G, Sella G. Life history effects on the molecular clock of autosomes and sex chromosomes. Proceedings of the National Academy of Sciences. 2016;113(6):1588-93.

[187] Otto SP, Whitlock MC. The probability of fixation in populations of changing size. Genetics. 1997;146(2):723-33.

[188] Pritchard JK, Pickrell JK, Coop G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology. 2010;20(4):R208-15.

[189] Kimura M. Some problems of stochastic processes in genetics. The Annals of Mathematical Statistics. 1957:882-901.

[190] Kimura M. On the probability of fixation of mutant genes in a population. Genetics. 1962;47(6):713.

[191] Simons YB, Turchin MC, Pritchard JK, Sella G. The deleterious mutation load is insensitive to recent population history. Nature Genetics. 2014;46(3):220-4.

[192] Sawyer SA, Hartl DL. Population genetics of polymorphism and divergence. Genetics. 1992;132(4):1161-76.

[193] Sethupathy P, Hannenhalli S. A tutorial of the poisson random field model in population genetics. Advances in bioinformatics. 2008;2008.

[194] Williamson SH, Hernandez R, Fledel-Alon A, Zhu L, Nielsen R, Bustamante CD. Simultaneous inference of selection and population growth from patterns of variation in the human genome. Proceedings of the National Academy of Sciences. 2005;102(22):7882-7.

[195] Agarwal I, Przeworski M. Mutation saturation for fitness effects at human CpG sites. Elife. 2021;10:e71513.

[196] Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, et al. A genome-wide mutational constraint map quantified from variation in 76,156 human genomes. bioRxiv. 2022:2022-03.

[197] Menke DB, Guenther C, Kingsley DM. Dual hindlimb control elements in the Tbx4 gene and region-specific control of bone size in vertebrate limbs. Development. 2008.

[198] Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell. 2008;134(2):341-52.

[199] Chamary JV, Parmley JL, Hurst LD. Hearing silence: non-neutral evolution at synonymous sites in mammals. Nature Reviews Genetics. 2006;7(2):98-108.

[200] Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Molecular biology and evolution. 2008;25(3):568-79.

[201] Hershberg R, Petrov DA. Selection on codon bias. Annual review of genetics. 2008;42:287-99.

[202] Galtier N, Roux C, Rousselle M, Romiguier J, Figuet E, Glémin S, et al. Codon usage bias in animals: disentangling the effects of natural selection, effective population size, and GC-biased gene conversion. Molecular biology and evolution. 2018;35(5):1092-103.

[203] Platt RN, Vandewege MW, Ray DA. Mammalian transposable elements and their impacts on genome evolution. Chromosome Research. 2018;26:25-43.

[204] Sundaram V, Wysocka J. Transposable elements as a potent source of diverse cis-regulatory sequences in mammalian genomes. Philosophical Transactions of the Royal Society B. 2020;375(1795):20190347.

[205] de Koning AJ, Gu W, Castoe TA, Batzer MA, Pollock DD. Repetitive elements may comprise over two-thirds of the human genome. PLoS genetics. 2011;7(12):e1002384.

[206] Deininger P. Alu elements: know the SINEs. Genome biology. 2011;12(12):1-12.

[207] Deininger PL, Batzer MA. Alu repeats and human disease. Molecular genetics and metabolism. 1999;67(3):183-93.

[208] Sen SK, Han K, Wang J, Lee J, Wang H, Callinan PA, et al. Human genomic deletions mediated by recombination between Alu elements. The American Journal of Human Genetics. 2006;79(1):41-53.

[209] Kim S, Cho CS, Han K, Lee J. Structural variation of Alu element and human disease. Genomics & informatics. 2016;14(3):70.

[210] Chung H, Calis JJ, Wu X, Sun T, Yu Y, Sarbanes SL, et al. Human ADAR1 prevents endogenous RNA from triggering translational shutdown. Cell. 2018;172(4):811-24.

[211] Yang F, Wang PJ. Multiple LINEs of retrotransposon silencing mechanisms in the mammalian germline. In: Seminars in cell & developmental biology. vol. 59. Elsevier; 2016. p. 118-25.

[212] Lynch M, Ackerman MS, Gout JF, Long H, Sung W, Thomas WK, et al. Genetic drift, selection and the evolution of the mutation rate. Nature Reviews Genetics. 2016;17(11):704-14.

[213] Harris K, Pritchard JK. Rapid evolution of the human mutation spectrum. Elife. 2017;6:e24284.

[214] Sasani TA, Ashbrook DG, Beichman AC, Lu L, Palmer AA, Williams RW, et al. A natural mutator allele shapes mutation spectrum variation in mice. Nature. 2022;605(7910):497-502.

[215] Kondrashov AS. Contamination of the genome by very slightly deleterious mutations: why have we not died 100 times over? Journal of theoretical biology. 1995;175(4):583-94.

[216] Fan S, Hansen ME, Lo Y, Tishkoff SA. Going global by adapting local: A review of recent human adaptation. Science. 2016;354(6308):54-9.

[217] Coop G. Population and Quantitative Genetics; 2020.

[218] Teshima KM, Coop G, Przeworski M. How reliable are empirical genomic scans for selective sweeps? Genome research. 2006;16(6):702-12.

[219] Maynard Smith J, Haigh J. The hitch-hiking effect of a favourable gene. Genetics Research. 1974;23(1):23-35.

[220] Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, Schaffner SF, et al. Detecting recent positive selection in the human genome from haplotype structure. Nature. 2002;419(6909):832-7.

[221] Voight BF, Kudaravalli S, Wen X, Pritchard JK. A map of recent positive selection in the human genome. PLoS biology. 2006;4(3):e72.

[222] Nielsen R, Williamson S, Kim Y, Hubisz MJ, Clark AG, Bustamante C. Genomic scans for selective sweeps using SNP data. Genome research. 2005;15(11):1566-75.

[223] Oleksyk TK, Smith MW, O'Brien SJ. Genome-wide scans for footprints of natural selection. Philosophical Transactions of the Royal Society B: Biological Sciences. 2010;365(1537):185-205.

[224] Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, Shamovsky O, et al. Positive natural selection in the human lineage. science. 2006;312(5780):1614-20.

[225] Booker TR, Jackson BC, Keightley PD. Detecting positive selection in the genome. BMC biology. 2017;15:1-10.

[226] Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, Jørgensen ME, et al. Greenlandic Inuit show genetic signatures of diet and climate adaptation. Science. 2015;349(6254):1343-7.

[227] Mathieson S, Mathieson I. FADS1 and the timing of human adaptation to agriculture. Molecular biology and evolution. 2018;35(12):2957-70.

[228] Mathieson I, Day FR, Barban N, Tropf FC, Brazel DM, eQTLGen Consortium, et al. Genome-wide analysis identifies genetic effects on reproductive success and ongoing natural selection at the FADS locus. Nature human behaviour. 2023;7(5):790-801.

[229] Mathieson I. Estimating time-varying selection coefficients from time series data of allele frequencies. bioRxiv. 2020.

[230] Segurel L, Guarino-Vignon P, Marchi N, Lafosse S, Laurent R, Bon C, et al. Why and when was lactase persistence selected for? Insights from Central Asian herders and ancient DNA. PLoS biology. 2020;18(6):e3000742.

[231] Cong PK, Bai WY, Li JC, Yang MY, Khederzadeh S, Gai SR, et al. Genomic analyses of 10,376 individuals in the Westlake BioBank for Chinese (WBBC) pilot project. Nature Communications. 2022;13(1):2939.

[232] Bleasdale M, Richter KK, Janzen A, Brown S, Scott A, Zech J, et al. Ancient proteins provide evidence of dairy consumption in eastern Africa. Nature communications. 2021;12(1):632.

[233] Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, Silverman JS, et al. Convergent adaptation of human lactase persistence in Africa and Europe. Nature genetics. 2007;39(1):31-40.

[234] Schlebusch CM, Sjödin P, Skoglund P, Jakobsson M. Stronger signal of recent selection for lactase persistence in Maasai than in Europeans. European Journal of Human Genetics. 2013;21(5):550-3.

[235] Crawford NG, Kelly DE, Hansen ME, Beltrame MH, Fan S, Bowman SL, et al. Loci associated with skin pigmentation identified in African populations. Science. 2017;358(6365):eaan8433.

[236] Jones P, Lucock M, Veysey M, Beckett E. The vitamin D–folate hypothesis as an evolutionary model for skin pigmentation: an update and integration of current ideas. Nutrients. 2018;10(5):554.

[237] Jablonski NG. The evolution of human skin pigmentation involved the interactions of genetic, environmental, and cultural variables. Pigment Cell & Melanoma Research. 2021;34(4):707-29.

[238] Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, Absher D, et al. Signals of recent positive selection in a worldwide sample of human populations. Genome research. 2009;19(5):826-37.

[239] Field Y, Boyle EA, Telis N, Gao Z, Gaulton KJ, Golan D, et al. Detection of human adaptation during the past 2000 years. Science. 2016;354(6313):760-4.

[240] Ju D, Mathieson I. The evolution of skin pigmentation-associated variation in West Eurasia. Proceedings of the National Academy of Sciences. 2021;118(1):e2009227118.

[241] Messer PW, Petrov DA. Population genomics of rapid adaptation by soft selective sweeps. Trends in ecology & evolution. 2013;28(11):659-69.

[242] Hermisson J, Pennings PS. Soft sweeps and beyond: understanding the patterns and probabilities of selection footprints under rapid adaptation. Methods in Ecology and Evolution. 2017;8(6):700-16.

[243] Jones FC, Grabherr MG, Chan YF, Russell P, Mauceli E, Johnson J, et al. The genomic basis of adaptive evolution in threespine sticklebacks. Nature. 2012;484(7392):55-61.

[244] Orr HA, Betancourt AJ. Haldane's sieve and adaptation from the standing genetic variation. Genetics. 2001;157(2):875-84.

[245] Hermisson J, Pennings PS. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. Genetics. 2005;169(4):2335-52.

[246] Przeworski M, Coop G, Wall JD. The signature of positive selection on standing genetic variation. Evolution. 2005;59(11):2312-23.

[247] Langhi DM, Orlando Bordin J. Duffy blood group and malaria. Hematology. 2006;11(5-6):389-98.

[248] Tournamille C, Colin Y, Cartron JP, Le Van Kim C. Disruption of a GATA motif in the Duffy gene promoter abolishes erythroid gene expression in Duffy–negative individuals. Nature genetics. 1995;10(2):224-8.

[249] Spencer HC, Miller LH, Collins WE, Knud-Hansen C, McGinnis MH, Shiroishi T, et al. The Duffy blood group and resistance to Plasmodium vivax in Honduras. The American Journal of Tropical Medicine and Hygiene. 1978;27(4):664-70.

[250] Dean M, Carrington M, Winkler C, Huttley GA, Smith MW, Allikmets R, et al. Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the CKR5 structural gene. Science. 1996;273(5283):1856-62.

[251] Hamblin MT, Di Rienzo A. Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. The American Journal of Human Genetics. 2000;66(5):1669-79.

[252] Hamblin MT, Thompson EE, Di Rienzo A. Complex signatures of natural selection at the Duffy blood group locus. The American Journal of Human Genetics. 2002;70(2):369-83.

[253] McManus KF, Taravella AM, Henn BM, Bustamante CD, Sikora M, Cornejo OE. Population genetic analysis of the DARC locus (Duffy) reveals adaptation from standing variation associated with malaria resistance in humans. PLoS genetics. 2017;13(3):e1006560.

[254] Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, et al. The role of geography in human adaptation. PLoS genetics. 2009;5(6):e1000500.

[255] Hodgson JA, Pickrell JK, Pearson LN, Quillen EE, Prista A, Rocha J, et al. Natural selection for the Duffy-null allele in the recently admixed people of Madagascar. Proceedings of the Royal Society B: Biological Sciences. 2014;281(1789):20140930.

[256] Prugnolle F, Rougeron V, Becquart P, Berry A, Makanga B, Rahola N, et al. Diversity, host switching and evolution of Plasmodium vivax infecting African great apes. Proceedings of the National Academy of Sciences. 2013;110(20):8123-8.

[257] Liu W, Li Y, Shaw KS, Learn GH, Plenderleith LJ, Malenke JA, et al. African origin of the malaria parasite Plasmodium vivax. Nature communications. 2014;5(1):3346.

[258] Hendry AP, Huber SK, De Leon LF, Herrel A, Podos J. Disruptive selection in a bimodal population of Darwin's finches. Proceedings of the Royal Society B: Biological Sciences. 2009;276(1657):753-9.

[259] Simons YB, Bullaughey K, Hudson RR, Sella G. A population genetic interpretation of GWAS findings for human quantitative traits. PLoS biology. 2018;16(3):e2002985.

[260] Simons YB, Mostafavi H, Smith CJ, Pritchard JK, Sella G. Simple scaling laws control the genetic architectures of human complex traits. bioRxiv. 2022:2022-10.

[261] Rees DC, Williams TN, Gladwin MT. Sickle-cell disease. The Lancet. 2010;376(9757):2018-31.

[262] Allison AC. Protection afforded by sickle-cell trait against subtertian malarial infection. British medical journal. 1954;1(4857):290.

[263] MalariaGen. Reappraisal of known malaria resistance loci in a large multicenter study. Nature genetics. 2014;46(11):1197-204.

[264] MalariaGen. Insights into malaria susceptibility using genome-wide data on 17,000 individuals from Africa, Asia and Oceania. Nature communications. 2019;10(1):5732.

[265] Piel FB, Patil AP, Howes RE, Nyangiri OA, Gething PW, Dewi M, et al. Global epidemiology of sickle haemoglobin in neonates: a contemporary geostatistical model-based map and population estimates. The Lancet. 2013;381(9861):142-51.

[266] Elguero E, Délicat-Loembet LM, Rougeron V, Arnathau C, Roche B, Becquart P, et al. Malaria continues to select for sickle cell trait in Central Africa. Proceedings of the National Academy of Sciences. 2015;112(22):7051-4.

[267] Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, et al. Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science. 2001;293(5529):455-62.

[268] Ségurel L, Thompson EE, Flutre T, Lovstad J, Venkat A, Margulis SW, et al. The ABO blood group is a trans-species polymorphism in primates. Proceedings of the National Academy of Sciences. 2012;109(45):18493-8.

[269] Segurel L, Gao Z, Przeworski M. Ancestry runs deeper than blood: the evolutionary history of ABO points to cryptic variation of functional importance. Bioessays. 2013;35(10):862-7.

[270] Leffler EM, Gao Z, Pfeifer S, Ségurel L, Auton A, Venn O, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. Science. 2013;339(6127):1578-82.

[271] Fortier AL, Pritchard JK. Ancient Trans-Species Polymorphism at the Major Histocompatibility Complex in Primates. bioRxiv. 2022:2022-06.

[272] Hayward LK, Sella G. Polygenic adaptation after a sudden change in environment. Elife. 2022;11:e66697.

[273] Moose SP, Dudley JW, Rocheford TR. Maize selection passes the century mark: a unique resource for 21st century genomics. Trends in plant science. 2004;9(7):358-64.

[274] Hendry AP, Kinnison MT, Heino M, Day T, Smith TB, Fitt G, et al. Evolutionary principles and their practical application. Evolutionary Applications. 2011;4(2):159-83.

[275] Berg JJ, Harpak A, Sinnott-Armstrong N, Joergensen AM, Mostafavi H, Field Y, et al. Reduced signal for polygenic adaptation of height in UK Biobank. Elife. 2019;8:e39725.

[276] Sohail M, Maier RM, Ganna A, Bloemendal A, Martin AR, Turchin MC, et al. Polygenic adaptation on height is overestimated due to uncorrected stratification in genome-wide association studies. Elife. 2019;8:e39702.

[277] Cox SL, Ruff CB, Maier RM, Mathieson I. Genetic contributions to variation in human stature in prehistoric Europe. Proceedings of the National Academy of Sciences. 2019;116(43):21484-92.

[278] Chen M, Sidore C, Akiyama M, Ishigaki K, Kamatani Y, Schlessinger D, et al. Evidence of polygenic adaptation in Sardinia at height-associated loci ascertained from the Biobank Japan. The American Journal of Human Genetics. 2020;107(1):60-71.

[279] Charlesworth B, Charlesworth D. Population genetics from 1966 to 2016. Heredity. 2017;118(1):2-9.

[280] Charlesworth B, Charlesworth D, Coyne JA, Langley CH. Hubby and Lewontin on protein variation in natural populations: when molecular genetics came to the rescue of population genetics. Genetics. 2016;203(4):1497-503.

[281] Harris H. C. Genetics of Man Enzyme polymorphisms in man. Proceedings of the Royal Society of London Series B Biological Sciences. 1966;164(995):298-310.

[282] Hubby JL, Lewontin RC. A molecular approach to the study of genic heterozygosity in natural populations. I. The number of alleles at different loci in Drosophila pseudoobscura. Genetics. 1966;54(2):577.

[283] Lewontin RC, Hubby JL. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. Genetics. 1966;54(2):595.

[284] Zuckerkandl E, Pauling L. Molecules as documents of evolutionary history. Journal of theoretical biology. 1965;8(2):357-66.

[285] Kumar S. Molecular clocks: four decades of evolution. Nature Reviews Genetics. 2005;6(8):654-62.

[286] Dickerson RE. The structure of cytochrome c and the rates of molecular evolution. Journal of Molecular Evolution. 1971;1:26-45.

[287] King JL, Jukes TH. Non-Darwinian Evolution: Most evolutionary change in proteins may be due to neutral mutations and genetic drift. Science. 1969;164(3881):788-98.

[288] Kimura M, et al. Evolutionary rate at the molecular level. Nature. 1968;217(5129):624-6.

[289] Kimura M. The neutral theory of molecular evolution. Cambridge University Press; 1983.

[290] Kern AD, Hahn MW. The neutral theory in light of natural selection. Molecular biology and evolution. 2018;35(6):1366-71.

[291] Jensen JD, Payseur BA, Stephan W, Aquadro CF, Lynch M, Charlesworth D, et al. The importance of the neutral theory in 1968 and 50 years on: a response to Kern and Hahn 2018. Evolution. 2019;73(1):111-4.

[292] Sackton TB. Studying natural selection in the era of ubiquitous genomes. Trends in Genetics. 2020;36(10):792-803.

[293] Gillespie JH. Lineage effects and the index of dispersion of molecular evolution. Molecular biology and evolution. 1989;6(6):636-47.

[294] Bedford T, Wapinski I, Hartl DL. Overdispersion of the molecular clock varies between yeast, Drosophila and mammals. Genetics. 2008;179(2):977-84.

[295] Bedford T, Hartl DL. Overdispersion of the molecular clock: temporal variation of gene-specific substitution rates in Drosophila. Molecular biology and evolution. 2008;25(8):1631-8.

[296] Jørgensen FG, Hobolth A, Hornshøj H, Bendixen C, Fredholm M, Schierup MH. Comparative analysis of protein coding sequences from human, mouse and the domesticated pig. BMC biology. 2005;3:1-15.

[297] Jensen JM, Villesen P, Friborg RM, Mailund T, Besenbacher S, Schierup MH, et al. Assembly and analysis of 100 full MHC haplotypes from the Danish population. Genome research. 2017;27(9):1597-607.

[298] Norman PJ, Norberg SJ, Guethlein LA, Nemat-Gorgani N, Royce T, Wroblewski EE, et al. Sequences of 95 human MHC haplotypes reveal extreme coding variation in genes other than highly polymorphic HLA class I and II. Genome research. 2017;27(5):813-23.

[299] Hughes AL, Nei M. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature. 1988;335(6186):167-70.

[300] Oliver PL, Goodstadt L, Bayes JJ, Birtle Z, Roach KC, Phadnis N, et al. Accelerated evolution of the Prdm9 speciation gene across diverse metazoan taxa. PLoS genetics. 2009;5(12):e1000753.

[301] Yang Z. PAML 4: phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution. 2007;24(8):1586-91.

[302] McDonald JH, Kreitman M. Adaptive protein evolution at the Adh locus in Drosophila. Nature. 1991;351(6328):652-4.

[303] Hudson RR, Kreitman M, Aguadé M. A test of neutral molecular evolution based on nucleotide data. Genetics. 1987;116(1):153-9.

[304] Bustamante CD, Fledel-Alon A, Williamson S, Nielsen R, Todd Hubisz M, Glanowski S, et al. Natural selection on protein-coding genes in the human genome. Nature. 2005;437(7062):1153-7.

[305] Messer PW, Petrov DA. Frequent adaptation and the McDonald–Kreitman test. Proceedings of the National Academy of Sciences. 2013;110(21):8615-20.

[306] Eyre-Walker A. Changing effective population size and the McDonald-Kreitman test. Genetics. 2002;162(4):2017-24.

[307] Sella G, Petrov DA, Przeworski M, Andolfatto P. Pervasive natural selection in the Drosophila genome? PLoS genetics. 2009;5(6):e1000495.

[308] Eyre-Walker A, Keightley PD. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. Molecular biology and evolution. 2009;26(9):2097-108.

[309] Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, et al. Assessing the evolutionary impact of amino acid mutations in the human genome. PLoS genetics. 2008;4(5):e1000083.

[310] Uricchio LH, Petrov DA, Enard D. Exploiting selection at linked sites to infer the rate and strength of adaptation. Nature ecology & evolution. 2019;3(6):977-84.

[311] Hill WG, Robertson A. The effect of linkage on limits to artificial selection. Genetics Research. 1966;8(3):269-94.

[312] Felsenstein J. The evolutionary advantage of recombination. Genetics. 1974;78(2):737-56.

[313] Begun DJ, Aquadro CF. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature. 1992;356(6369):519-20.

[314] Cai JJ, Macpherson JM, Sella G, Petrov DA. Pervasive hitchhiking at coding and regulatory sites in humans. PLoS genetics. 2009;5(1):e1000336.

[315] Charlesworth B, Morgan M, Charlesworth D. The effect of deleterious mutations on neutral molecular variation. Genetics. 1993;134(4):1289-303.

[316] Charlesworth B. Background selection 20 years on: the Wilhelmine E. Key 2012 invitational lecture. Journal of Heredity. 2013;104(2):161-71.

[317] Hudson RR, Kaplan NL. Deleterious background selection with recombination. Genetics. 1995;141(4):1605-17.

[318] Nordborg M, Charlesworth B, Charlesworth D. The effect of recombination on background selection. Genetics Research. 1996;67(2):159-74.

[319] Elyashiv E, Sattath S, Hu TT, Strutsovsky A, McVicker G, Andolfatto P, et al. A genomic map of the effects of linked selection in Drosophila. PLoS genetics. 2016;12(8):e1006130.

[320] Buffalo V, Kern AD. A Quantitative Genetic Model of Background Selection in Humans. bioRxiv. 2023:2023-09.

[321] McVicker G, Gordon D, Davis C, Green P. Widespread genomic signatures of natural selection in hominid evolution. PLoS genetics. 2009;5(5):e1000471.

[322] Murphy DA, Elyashiv E, Amster G, Sella G. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. Elife. 2022;12:e76065.

[323] Lohmueller KE, Albrechtsen A, Li Y, Kim SY, Korneliussen T, Vinckenbosch N, et al. Natural selection affects multiple aspects of genetic variation at putatively neutral sites across the human genome. PLoS genetics. 2011;7(10):e1002326.

[324] Enard D, Messer PW, Petrov DA. Genome-wide signals of positive selection in human evolution. Genome research. 2014;24(6):885-95.

[325] Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, et al. Classic selective sweeps were rare in recent human evolution. science. 2011;331(6019):920-4.

[326] Sattath S, Elyashiv E, Kolodny O, Rinott Y, Sella G. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in Drosophila simulans. PLoS genetics. 2011;7(2):e1001302.

[327] Mathieson I, Lazaridis I, Rohland N, Mallick S, Patterson N, Roodenberg SA, et al. Genome-wide patterns of selection in 230 ancient Eurasians. Nature. 2015;528(7583):499-503.