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

[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 het-

erogeneity 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

[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 func-

tional 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

[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

[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.

[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

$$\frac{n!}{k!(n-k)!}p^k q^{n-k}, \tag{5.75}$$

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

[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 pub-

lished 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

$$\int_{t=0}^{\infty} t \cdot \lambda \mathrm{e}^{-\lambda t} \, dt = \lambda^{-1}. \tag{5.76}$$

[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].

$$\text{number of mutations} \sim Poisson(\mu L b_i) \tag{5.77}$$

[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:

$$\begin{aligned}
\mathrm{E}[m] &= \int_0^{\infty} \mathrm{E}[m|T] \, \Pr[T] \mathrm{d}t & (5.78)\\
&= \int_0^{\infty} (2\mu LT) \, \Pr[T] \mathrm{d}t & (5.79)\\
&= 2\mu L \int_0^{\infty} T \, \Pr[t] \mathrm{d}t & (5.80)\\
&= 2\mu L \, \mathrm{E}[T] & (5.81)\\
&= 2\mu L 2N = 4N\mu L & (5.82)
\end{aligned}$$

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.

[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.

[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 2.47 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 2.49 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

[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

$$\mathrm{E}[p'] = \frac{pq(1+sh) + p^2(1+s)}{q^2 + 2pq(1+sh) + p^2(1+s)} \tag{5.83}$$

We simplify the notation by using $\overline{w}$ in place of the denominator (pronounced w-bar, and referred to as "mean fitness"), and simplifying:

$$\begin{aligned}
\overline{w} &= q^2 + 2pq(1+sh) + p^2(1+s) && (5.84)\\
&= q^2 + 2pq + 2pqsh + p^2 + p^2 s && (5.85)
\end{aligned}$$

Noting that $p + q = 1$ and $q^2 + 2pq + p^2 = 1$ we simplify this to

$$\overline{w} = 1 + 2pqsh + p^2 s \tag{5.86}$$

Now we're ready to start calculating Δ_p as follows:

$$\begin{aligned}
\Delta_p &= \frac{pq(1+sh) + p^2(1+s)}{\overline{w}} \quad - \quad p \times \frac{\overline{w}}{\overline{w}} && (5.87)\\
&= [pq(1+sh) + p^2(1+s) - p[1 + 2pqsh + p^2 s]/\overline{w} && (5.88)\\
&= p[q(1+sh) + p(1+s) - 1 - 2pqsh - p^2 s]/\overline{w} && (5.89)\\
&= p[q + qsh + p + ps - 1 - 2pqsh - p^2 s]/\overline{w} && (5.90)\\
&= p[qsh + ps - 2pqsh - p^2 s]/\overline{w} && (5.91)\\
&= ps[qh + p - 2pqh - p^2]/\overline{w} && (5.92)\\
&= ps[qh + pq - 2pqh]/\overline{w} && (5.93)\\
&= pqs[h + p - 2ph]/\overline{w} && (5.94)\\
&= pqs[h(1-2p) + p]/\overline{w} && (5.95)\\
&= pqs[h(q-p) + p]/\overline{w} && (5.96)\\
&= pqs[p(1-h) + qh]/\overline{w} && (5.97)
\end{aligned}$$

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 ($E(\Delta_p)$ due to selection is on the order of $sp(1-p)$, while the variance in allele frequency due to drift ($Var(\Delta_p)$ 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

$$4N\mu \frac{1 - e^{-2\gamma(1-x)}}{(1 - e^{-2\gamma})p(1-p)} \tag{5.98}$$

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 ef-

fect 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.

[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 amy-

lase 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

$$\frac{pqs[p(1-h)+qh]}{\overline{w}} = 0 \tag{5.99}$$

Noting that $q = 1 - p$, and assuming that $\overline{w} > 0$ for sensible biological parameters, we see immediately that two trivial solutions are

$$\hat{p} = 0 \tag{5.100}$$

$$\hat{p} = 1. \tag{5.101}$$

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

$$p(1-h) + qh = 0 \tag{5.102}$$

$$p(1-h) + (1-p)h = 0 \tag{5.103}$$

$$\hat{p} = \frac{h}{2h-1} \tag{5.104}$$

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

[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 substitu-

tions. 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