Preface

This book is about the genetic differences between human genomes.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

motivated important new avenues of theory in many areas.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

deeper understanding.

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

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

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

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

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

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