1.7 Genetic variation

Genetic variation is simply a catch-phrase to describe all the observed variability in genetic and genomic characteristics between individuals and populations. In genetic variation lie the keys to understanding differences in traits, such as height or eye color, as well as differences in dispositions to genetic or complex diseases, and genetic markers of heredity and ancestry. Current estimates put the genetic similarity between any two individual humans at approximately 99.9% on average. A 0.1% difference might not seem like a big deal, but remember that the human genome comprises just more than 3 billion base pairs, so a 0.1% difference entails around 3,000,000 base-pair differences!

Many of these differences are likely to be benign, occurring in inconsequential regions of non-functional DNA, or contributing to common differences in human traits, such as height or hair color. Yet some proportion of this variability will cause individuals to harbor recessive genetic diseases, be more or less susceptible to complex disease, or cause them to metabolize medications differently. Therefore, relatively small changes across the vast human genome are the focal point of many modern genomic studies. We learn more about genetic variation in the human population in later chapters on ancestry and trait associations.

1.7.1 Polymorphism

Polymorphism is a word you will hear frequently when you read scientific articles pertaining to genomics. If you consider its Greek roots, the term polymorphism simply means "many forms". In the context of genetics and genomics, it is used to describe the genetic variability that can be found along the human genome (Figure 1.11). A specific location on the genome is called a locus, and therefore, we often say that a gene is found at a particular locus. If we were to sequence a gene in a number if unrelated individuals, we might find small differences among individuals in the nucleotide sequence used to encode the gene. Therefore, we can say that several forms, or alleles, exist for this gene in the population, and that the gene is polymorphic. Let us consider just a single position in one of the coding region (exon) of the gene's nucleotide sequence. If we sample the DNA of the population we might find that 80% of the population has an adenine (A) at this position, while the remaining 20% has a guanine (G) at this position. We would call this a single nucleotide polymorphism (SNP), pronounced "snip", where A is the major allele and G is the minor allele, based on the population frequency. SNPs are the most abundant polymorphisms found in the human genome (Figure 1.11).

It is important to understand that all of the alleles at a position can be "healthy", in that the major alleles are not necessarily biologically preferred or necessary for maintaining proper function. Individuals with the minor allele can be just as biologically robust as those carrying the major allele. The SNP may simply be not functionally relevant, or for coding regions, recall the earlier discussion about the human codon table and how redundancy is built into the genetic code. Since SNPs are known to be associated with disease, it is useful to characterize SNPs across the human genome. The government funded HapMap project (<http://hapmap.ncbi.nlm.nih.gov/>) set out to catalog all of the "normal" variation found among human populations by identifying a large number of SNPs across globally and ethnically distributed human populations. As we will find out first-hand in later chapters, the HapMap project is a highly valuable tool for exploring personal genomics.

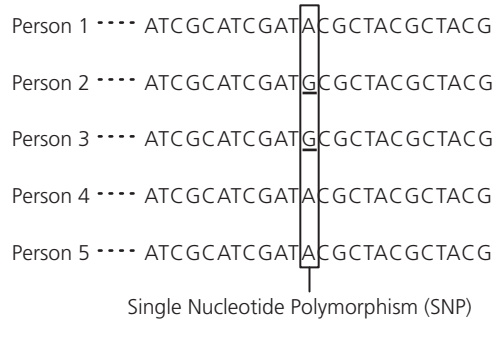

Figure 1.11 Single Nucleotide Polymorphisms: A single base pair mutation (point mutation) that is observed in more than one percent of the population is known as a single nucleotide polymorphism (SNP). These mutations, which most often only have two alleles, arose at some point in human history and have spread throughout human populations. In this example, we observe a SNP with two alleles (A and G).

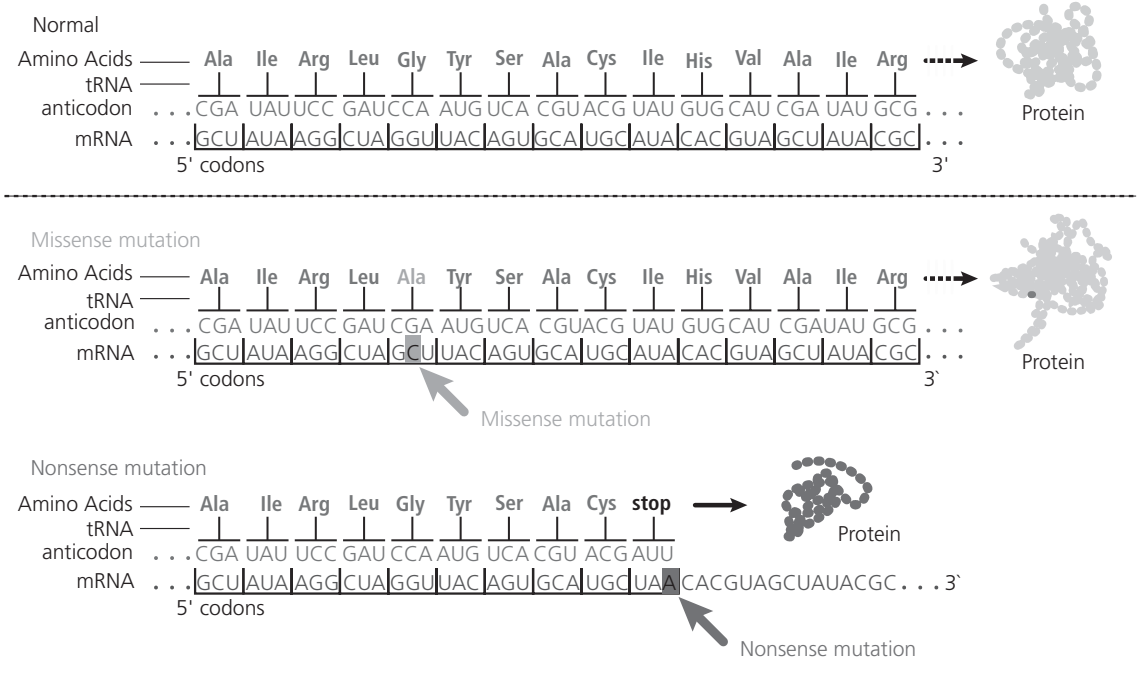

Figure 1.12 Types of coding SNPs: Some SNPs that occur in coding regions can affect the amino acid sequence of the protein. These SNPs are known as non-synonymous SNPs and can result in a change in amino acids (missense) or the introduction or loss of a stop codon (nonsense and nonstop, respectively). Additionally, synonymous SNPs (not pictured in the figure) are SNPs that change the DNA sequence, but do not affect the amino acid sequence (as the same amino acid is used due to redundancy in the genetic code; e.g. a point mutation that changes a codon from GCC to GCA). Image adapted from the National Human Genome Research Institute.

Polymorphisms come into existence by a mutation at a previous moment in evolutionary history that persisted and spread throughout the population (see Box 1.3). These mutations can be caused by any number of factors, such as errors in DNA replication, radiation, mutagenic toxins, structural methylation, or viral activity. Mutations can involve regions of DNA spanning one to many base pairs of DNA. When a mutation involves a single base pair, it is typically referred to as a **point mutation**. Mutations in regions of the genome coding for genes are of particular interest because they have the potential to change the composition, and therefore the function of the protein product encoded by the gene. If a mutation occurs in one of the codons of a gene coding region, it can have one of several possible effects (Figure 1.12). A mutation in this codon might not change the amino acid at all because it might simply mutate one codon into another codon that codes for the same amino acid; this type of mutation is called a **synonymous mutation**. It is generally thought that synonymous mutations are benign, but recent studies have demonstrated that synonymous mutations can impact the structure of the mRNA or affect the translational efficiency of the tRNA. A mutation that causes a region to code for a different amino acid is called a **non-synonymous** or **missense** mutation. The severity of these mutations depends on several factors, such as the physiochemical differences between the original and mutant mutation. Another type of non-synonymous mutation is a mutation that causes an amino-acid coding codon to mutate into a stop codon, which is typically referred to as a **nonsense mutation**; these are often more detrimental as they can result in truncation of the protein encoded by the gene.

Although SNPs are one of the most prominent types of genetic polymorphism in the human genome, they are far from the only types found (Figure 1.13). SNPs have stolen the stage because the first phase of the personal and medical genomics revolution was enabled by a specific type of genotyping

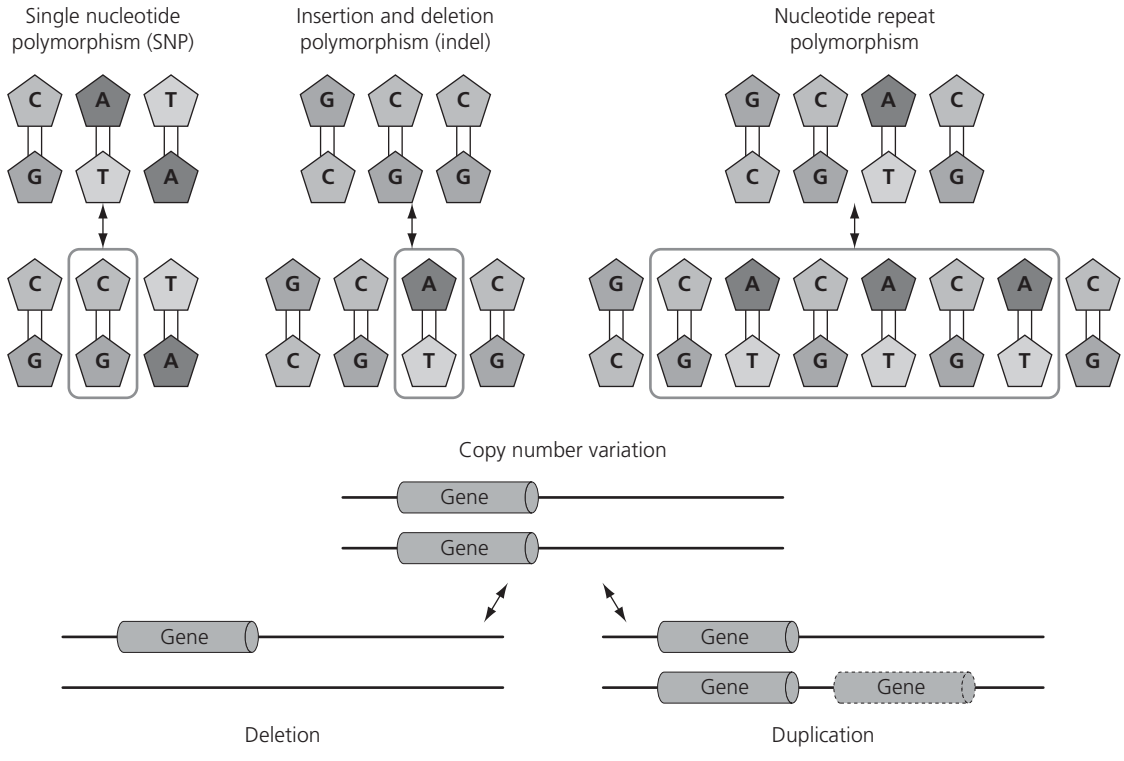

Figure 1.13 Other types of mutations and polymorphisms: In addition to single nucleotide polymorphisms, other types of variations can also alter the DNA and cause a functional change. For instance, indels can form when bases can be inserted or deleted, while the tandem duplication of short segments can produce nucleotide repeat polymorphisms. When these mutations occur in coding regions, they can cause a "frameshift" in codons. Additionally, larger variants can delete or duplicate whole genes or large segments of DNA, which are known as copy number variants and structural variants (see Chapter 10). Reproduced from Hingorani, A. D., Shah, T., Kumari, M., Sofat, R. & Smeeth, L. Translating genomics into improved healthcare. *BMJ* 341, c5945 (2010) with permission from BMJ Publishing Group Ltd.

technology called SNP microarrays, or "SNP chips". With the advent of affordable full-genome sequencing technology, other forms of polymorphism will be easier to measure and evaluate. Important polymorphisms found in the human population include:

- **Single nucleotide polymorphism (SNP):** SNPs entail a class of polymorphisms where a single nucleotide in the human genome is found to be different among members of a population.
- **Insertion/Deletion (indel):** Indel polymorphisms represent a type of genetic variation in which sections of nucleotide sequence are found to be inserted into or deleted from a region in the human genome. Although indel polymorphisms can be associated with disease, especially in coding regions, there are many cases of benign indel polymorphisms in the population.
- **Repeat polymorphisms:** Repeat polymorphisms are repeated patterns of short DNA sequences (e.g. GATAGATA) found in abundance in the human genome. Since the repeat sequences are typically short, ranging from two to five base pairs in length, these repeats are often called short tandem repeats (STR). STR polymorphisms are important for forensics, because STR polymorphisms between individuals serve as the basis for the DNA fingerprinting used by law enforcement agencies. Another larger type of repeat polymorphism, called *Alu* repeats, is also used frequently to study ancestry and population structure.
- **Copy number variations (CNV) and structural variants (SVs):** CNVs describe segments of DNA which have a different number of copies between individual genomes. Although this

Box 1.3 "Molecular Evolution"

Evolution is critical to understanding the origins of species, genome architecture, and even human disease mutations. Evolutionary studies comparing genomes across species have contributed significantly to our understanding of the human genome. Perhaps one of the most critical things to understand is that evolutionary forces are the primary drivers of genetic variability both across species and within populations. The evolutionary forces acting on the human genome have contributed to substantial changes in our species, even within the past several thousand years. Although evolution is typically discussed in terms of millions of years, it is an important note that evolution is ongoing, and that evolutionary forces come to bear on each human generation born into existence today. Additionally, the environment is very important in shaping evolution and our interactions with other species, such as the bacteria in our digestive tracts, can have a profound affect on our evolution. For example, it appears that up to 8% of the human genome is of viral origin, meaning that transposable DNA from ancient viruses was integrated into our ancestral genome some time in the past and has now become part of the blueprint for who we are as a species! A number of evolutionary forces have shaped the human genome and are important for understanding how genomes operate in modern humans.

- **Mutation:** A mutation is simply a change to the DNA sequence of a genome. With regards to evolution, we are primarily concerned with mutations that occur in the germ line of an organism as these mutations can be passed on to offspring.
- **Natural selection:** Natural selection is a key principal in molecular evolution, which pertains to the increase or decrease in particular genetic traits as a function of fitness and reproductive success. Although natural selection is a complex process, it can be thought of as a kind of filter that removes suboptimal alleles from a population so that the population is better adapted to its environment.
- **Genetic drift:** In each generation of a population, there will be individuals who will leave behind more offspring than others simply by chance. After several generations the effects of this random sampling will cause the allele frequencies to "drift" randomly towards higher or lower frequencies, which is called **genetic drift** (Figure 1.14). An important thing to understand about genetic drift is that if we observe changes in allele frequencies within a population, the observed changes are not necessarily the result of natural selection, but rather could simply be explained by random genetic drift. In a sufficiently large population, allele frequencies will tend to drift around some stable population equilibrium (see Box 1.4). The effects of genetic drift are more profound in smaller populations, such as populations that experience a **population bottleneck** (Figure 1.15). One important consequence of this is called the **founder effect**. Imagine a fictional ancient human population of 100 individuals that decided to separate from their main population and establish a new, isolated human settlement in some faraway land. The reduced genetic diversity and smaller pool of reproducing individuals in this population could cause extreme, random shifts in allele frequencies for this population as it expands. We might observe that formerly rare alleles reach fixation at 100% frequency in the population in just a few generations. Such populations can be more prone to certain recessive genetic disorders due to the reduced genetic diversity in their populations.
- **Gene Duplication:** Gene duplication is one of the primary mechanisms for functional innovation in genome evolution. Many of the important gene families in our genome, such as hormones and cell surface receptors (which facilitate the effects of medicinal drugs) came about through gene duplication events in evolutionary history. Gene duplication occurs when any region of genomic DNA containing a gene is duplicated and becomes fixed in the germ line of a species. Duplications can occur as the result of errors during recombination, the activity of retrotransposons, or in some cases entire chromosomes can be duplicated. From an evolutionary standpoint duplications are very interesting because a duplicated "copy" of a gene often has reduced selective pressures acting against it because it is usually functionally redundant to the original copy it was duplicated from. This means that it is more free to acquire mutations and potentially evolve its function towards the formation of a completely novel gene.

Box 1.4 Hardy-Weinberg equilibrium

Developed independently by British mathematician G.H. Hardy and German physician Wilhelm Weinberg, the Hardy-Weinberg Equilibrium (HWE) model is a theoretical mathematical model concerning the probability and distribution genotype allele frequencies in a population. Somewhat analogous to rules of Mendelian inheritance that describe the transmission of alleles at the family level, the HWE establishes a framework that can be used to model and predict genotype frequencies in large, stable populations. To illustrate, let's consider a single SNP locus that has three possible states: homozygous for the minor allele (aa), heterozygous (Aa), or homozygous for the major allele (AA). The parameters of the HWE equations are the frequency of the major allele, denoted p, and the frequency of the minor allele, denoted q, with the relationships between q and p expressed in the equilibrium model expressed as

$$p^2 + 2pq + q^2 = 1$$

Note that because we are dealing in allele frequencies, we can easily infer the frequency of one allele given the frequency of the other. For example, if we measure a SNP in a population and find that 90% of the individuals carry the major allele (A), then we can subtract this proportion from one to determine that 10% of individuals in the population carry the minor allele (a). We can then use the HWE equation to estimate the frequency of heterozygotes ($2pq$) in the population.

$$0.9^2 + 2pq + 0.1^2 = 1$$

$$0.81 + 2pq + 0.01 = 1$$

$$2pq = 1 - 0.82 = 0.18$$

The HWE would assert that the frequencies and relative proportions of these genotypes would remain stable (i.e. in equilibrium) over time if all the assumptions of the HWE are satisfied. The assumptions of the HWE, *all* of which must be satisfied to assert HWE, are:

- The population is (infinitely) large
- The mating patterns between population members are random
- All members of the population have equal reproductive success
- Males and females have similar allele frequencies (more likely on an autosomal locus)
- Mutation is not occurring
- There is no significant migration in or out of the population
- All genotypes have equal fitness (i.e. there is no selection)

At this point in the chapter, even a genomics neophyte should carry enough understanding to suspect that the assumptions of HWE are likely to be violated in the majority of cases. In fact, this is why HWE is useful, because deviation from HWE is often suggestive that the locus has been affected by non-equilibrium forces such as mutation or evolutionary selection. One of the most straightforward means to statistically evaluate deviations from HWE is to perform a Pearson's chi-squared test to look for significant deviation between the observed genotype frequencies, compared to the genotype frequencies that would expected under HWE. It is also reasonable to use HWE as a basis for inferring population genotype frequencies where only the frequency of a single allele for a locus is known or reported by a genetic association study, for example.

can pertain to any segment of DNA ranging from a few kilobases to even megabases in size, it is easiest to conceptualize CNVs in terms of gene regions. Since humans are diploid, they typically inherit two copies of any one gene; one from your father and the other from your mother. However, an individual may only have one copy of a gene because the other copy was deleted in one of the parental genomes, or have three copies of a gene because the gene was duplicated. CNVs can be inherited or can happen spontaneously, the latter of which is known as *de novo* CNVs, due to errors in the DNA replication machinery. CNVs are implicated in a number of diseases, such as schizophrenia, and are suspected to underlie many more genetic disorders. Structural variants are rearrangements of DNA that can affect large regions of DNA, similar to CNVs, including insertions, deletions, and inversions that can affect entire genes. We will discuss CNVs and SVs further in Chapter 11.

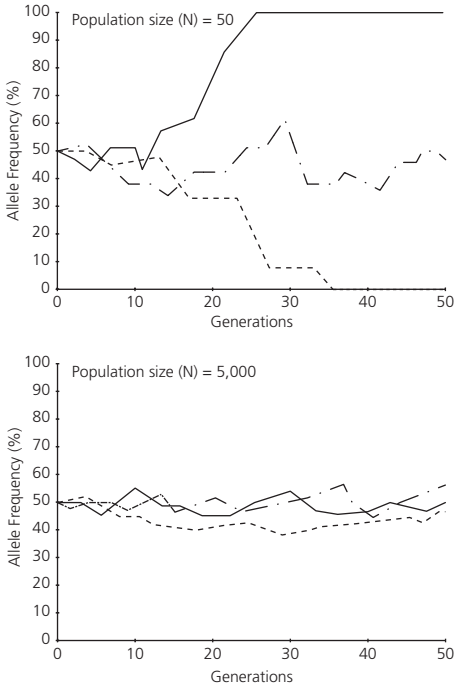

Figure 1.14 Effects of genetic drift on allele frequencies: These graphs show the effect of genetic drift on allele frequencies in a small (top; N = 50) and large (bottom; N = 5000) human population for three distinct SNP variants exhibiting Hardy-Weinberg equilibrium. In this example, each allele begins with a population frequency of 50%, and the frequency begins to vary over successive populations due to the effects of genetic drift. The effects of random sampling error are much more drastic in small population, causing one of the alleles to drift towards fixation (solid line), and another towards elimination (i.e. loss of the allele from the population) (dashed line). In the larger population, the effects of genetic drift due to sampling error are less drastic, and the allele frequencies do not vary far from their original population frequency over successive generations.

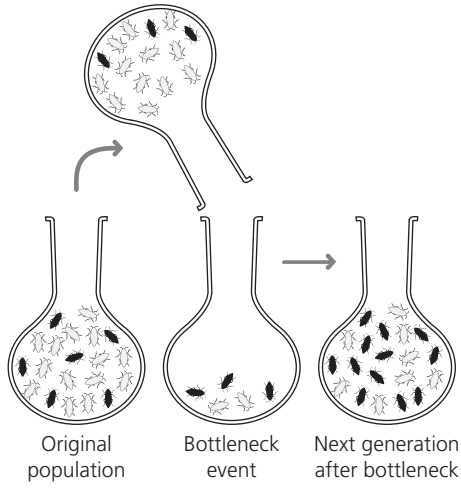

Figure 1.15 Population bottleneck: This phrase describes a scenario in which a population's size is substantially reduced for at least one generation, perhaps due to a pandemic disease, war, natural disaster, or some other event. The reduced population will reduce the amount of total genetic variation in the population, and, as illustrated in Figure 1.14, the effects of genetic drift on allele frequencies can be much more drastic in smaller populations. In this illustration, insects of the same species are contained within a bottle, and are distinguished by their external color, which is controlled by a single allele. Even if a population quickly recovers back to a large population size in subsequent generations, the effects of the population bottleneck, in terms of reduced genetic variation, can be apparent in the genomic makeup of a population. The genomic patterns of modern populations of European ancestry seem to bear hallmarks of a historical population bottleneck, marked by reduced overall patterns of heterozygosity compared to other worldwide populations.

1.7.2 Linkage disequilibrium

Thus far, we've discussed the different types of mutations independently, where a single mutation occurs and may spread throughout the population until it reaches high enough frequency to become a polymorphism. However, the mechanics of genomic inheritance implies that the inheritance of any two polymorphisms may not be independent. To illustrate how this may come about, consider two germline point mutations that occur on the same copy of a chromosome 100 bases apart. When a chromosome is transmitted to a sperm or an egg cell during meiosis (see Section 1.6—Replication and reproduction), these two neighboring SNPs will then be transmitted together, unless there is a recombination event in between. Furthermore, the probability of recombination is relatively low for small sections (such as the two SNPs 100 base pairs apart): there are, on average, 35 recombination events per meiosis, which corresponds to roughly one recombination every 100 Mb. Over time, this process continues and leads to the correlated inheritance of SNPs, which are known as linked SNPs (Figure 1.16). More broadly, this phenomenon is also termed linkage disequilibrium (LD); when two SNPs are inherited randomly (unlinked), they are said to be in equilibrium. Thus, a high level of LD implies that two SNPs are "linked" and the measure of linkage is known as R^2, which is the level of correlation between the

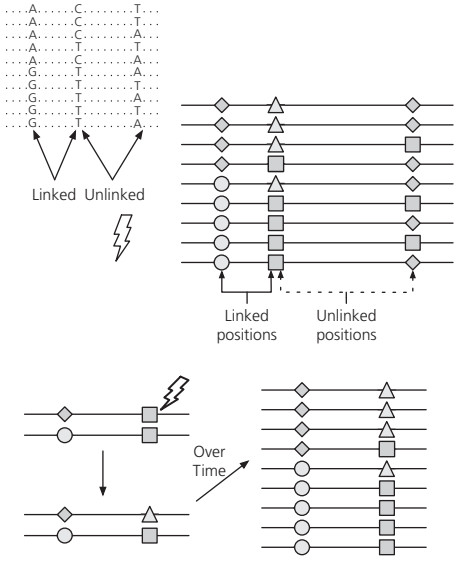

Figure 1.16 Polymorphisms are often inherited together due to linkage disequilibrium: As mutations arise, they are inherited together with nearby alleles, unless a recombination event occurs in between. Over time, as these alleles spread throughout the population, the presence of one of the alleles is correlated or "linked" with the other. These SNPs are said to be in linkage disequilibrium (LD).

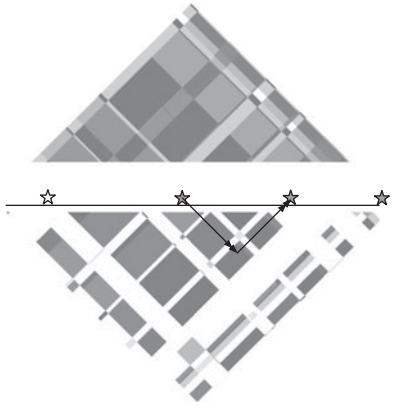

Figure 1.17 Haplotype blocks: The phenomenon of linkage disequilibrium (LD) often results in the correlated inheritance of nearby variants. In this schematic, a stretch of DNA with SNPs (represented by stars) is shown. Below is a commonly used display of linkage disequilibrium patterns: darker shaded regions correspond to more tightly linked SNPs. In this example, the white star SNP is unlinked to the gray star SNPs, which are all linked with each other. To read the diagram, follow down the diagonal from a SNP until the opposite diagonal of the other SNP of interest. Here, because the triangle is dark between the two marked SNPs, the two SNPs are in high LD.

SNPs. A perfect correction (fully linked SNPs, where genotype A of SNP 1 is always observed with genotype B of SNP 2) occurs when $R^2 = 1.0$, and two random SNPs will have $R^2 = 0$.

As we progress through our discussion of personal genomics, the concept of LD will become crucial for clinical and phenotype risk analysis, as well as ancestry analysis. One major reason for this is the development of **haplotype blocks** in a population (Figure 1.17). These blocks are segments of SNPs that are often inherited as a group; thus, we would only need to measure one SNP (a **tag SNP**) to predict the genotype of another (in a process known as **imputation**). In reality, the situation is not always as clear-cut: unless we can find a SNP in perfect LD ($R^2 = 1$), we may not be able to fully and accurately impute our SNP of interest. Additionally, different populations will have different patterns of linkage disequilibrium: populations that have undergone bottlenecks typically have longer haplotype blocks than those that have not. As we will see, these haplotype blocks will be used in the design of genotype-phenotype association experiments (Chapter 2) and the application of these associations to our personal genomes (Chapter 6), as well as ancestry analysis (Chapter 5).

Further reading

Unfortunately, there is more fascinating breadth and depth to genomics that could not be covered by this primer. We suggest the following materials and resources for those inclined to expand their study and understanding of molecular biology and genomics.

Alberts, B. (2008) *Molecular biology of the cell*. New York, NY: Garland Pub.

Berger, S. L. (2000) Gene regulation. Local or global? Nature 408, 412–13, 415.

Brown, T. A. (2007) *Genomes Three*. London: Taylor Francis.

Chen, K. & Rajewsky, N. (2007) The evolution of gene regulation by transcription factors and microRNAs. *Nature Reviews Genetics* 8, 93–103.

Collins, F. S. (2011) *The Language of Life*. New York, NY: Harper Perennial.

Hingorani, A. D., Shah, T., Kumari, M., Sofat, R. & Smeeth, L. (2010) Translating genomics into improved healthcare. *BMJ* 341, c5945.

Lynch, M. (2007) *The origins of genome architecture*. Sunderland, MA: Sinauer Associates Inc).

Nei, M. & Kumar, S. (2000) *Molecular evolution and phylogenetics*. New York: Oxford University Press.

Ridley, M. (2006) *Genome.* New York, NY: Harper Perennial.

Watson, J. D., Caudy, A. A., Myers, R. & Witkowski, J. A. (2007) *Recombinant DNA*. W. H. Freeman.