Chapter 1

Heredity, Genes, and DNA

Chapter 1 is a bare-bones synopsis of the fundamental facts, concepts, and terminology from genetics and molecular biology needed to understand the rest of this book. It should be accessible to anyone who knows that hereditary information is carried in genes encoded in the DNA of each cell and is passed from generation to generation in the process of reproduction. The treatment is necessarily schematic and abstract, omitting the many subtleties and empirical facts that make genetics so fascinating. For these, the interested reader should consult an appropriate biology text; some excellent and readable references are cited at the end of the chapter.

1.1 Mendelian genetics

Prior to the amazing advances of molecular biology in the second half of the twentieth century, heredity was understood in terms of Mendelian genetics, the theory founded on Gregor Mendel's seminal experiments of the 1850's and 1860's. Using pollination by hand, Mendel crossed pea plants and observed the relationship of their traits to those of their offspring. Whether as a result of chance, of scientific genius, or a combination of the two, Mendel picked an ideal system to study, for it led him to a theory whose essential insights have withstood the test of time.

Mendel's first smart choice was to work with a limited and discrete set of inherited traits, among which were the texture and the color of the peas. In his varieties, the peas were either wrinkled yellow, wrinkled green, smooth yellow, or smooth green. Mendel first established pure breeding lines that

always produced peas of the same type and then cross-pollinated plants from different lines and studied their progeny through several generations. He made three fundamental observations. First, neither the pea texture nor color traits blended. The progeny of two parent plants, one of which had yellow peas and the other of which had green peas, did not produce peas of intermediate color; in fact, they were invariably yellow. Second, a trait could sometimes disappear in one generation, only to reappear in the next. Thus, when he crossed the yellow pea progeny produced by mating a pure breeding green and pure breeding yellow, he obtained some plants with green as well as yellow peas. Finally, he noticed that offspring could take the trait for texture from one parent, but the trait for color from the other. Mendel did another very important thing. He *counted* the number of plants of each type produced in each generation. His data revealed that the relative proportions of traits in the parent population governed their proportions in the offspring and that these proportions fluctuated around simple integer ratios.

To explain his results, Mendel postulated first that traits are passed down from generation to generation in discrete, indivisible units, which we now call **genes**. Thus, for Mendel's peas, one imagines that there is a "gene" for pea color, with two variants, green and yellow, and a second gene for skin texture, again with two variants, smooth and wrinkled. The different variants of a gene are called **alleles**; thus the gene for pea color is said to have yellow and green alleles. Although in retrospect the gene hypothesis is natural for such a simple system as Mendel's, it was not at all apparent as a general rule, neither at Mendel's time (1865), nor for quite a few years afterward. For instance, Charles Darwin subscribed to a blending theory of inherited traits. In hindsight, blending is an untenable theory because it does not allow the genetic variation observed in natural populations, as we shall see in Chapter 3. But blending is a natural first idea. Many traits—think of physical size—vary over a continuous range of possibilities, or may be affected by the expression of several, interacting genes and by environmental factors such as nutrition. These complex, competing influences blur the discrete and modular nature of the gene. That Mendel's experiments revealed the idea of a discrete unit of heredity is why they are so seminal and beautiful.

Mendel also postulated that organisms carry two versions of each gene, one inherited from each parent. The versions may be identical copies of one another, that is, they may be the same allele, or they may be different. However, when two different alleles are present, Mendel proposed that only one, the **dominant**, as opposed to the **recessive**, allele is expressed physically in

the organism. These postulates explained how, as Mendel observed, a trait expressed in a parent might disappear in its children only to reappear in its grandchildren. Consider, for example, mating two pea plants, one that carries two yellow alleles and one that carries two green alleles. Thus one plant has yellow peas and the other has green peas. The seeds produced from cross pollinating these two plants all carry one allele for yellow and one for green, as they get an allele from each parent, and they will grow into plants with yellow peas, since yellow is dominant. But the seed produced by a cross pollination of this new generation of plants could get a green allele from each parent and hence produce green peas. Thus, the green pea trait disappears in the first generation, only to reappear in the next, due to the masking effect of allele dominance.

Extrapolating from the observation that a child can inherit different traits from different parents, Mendel postulated also that genes for different traits are passed down independently from one another. This is called the independent assortment postulate. What it means in the pea example is that the allele for pea color that a parent passes to its child does not influence the allele it transmits for pea texture.

Now let us add two simple rules for the probabilities with which parental genes are transmitted to offspring. Firstly, assume that each of the two copies of a gene for a trait has an equal probability, namely one-half, to be passed to a child from its parent. In effect, though it is strange to phrase it like this, one can think that the child acquires its two genes for each trait by randomly choosing from its mother one of her two copies and, independently, the other one from its father's two copies. Secondly, interpret the independent assortment postulate as asserting independence in the sense of probability; that is, the random choices of genes are probabilistically independent from trait to trait. Then Mendel's postulates lead to a simple *quantitative* model, which fits empirical data, for the probabilities of all the different possible combinations of genes in the offspring of a mating. The theoretical consequences of these postulates are explained fully in Chapter 3. For now we only illustrate with an example. Suppose that we perform crosses between two parent pea plants. Assume parent 1 carries the alleles Y(ellow) and G(reen) for the color gene and two copies of the S(mooth) allele for the texture gene, while parent 2 carries two Y alleles for color and an S and a W(rinkled) allele for the texture. We denote this situation by labeling parent 1 with the letters $YG\ SS$ and parent 2 with $YY\ WS$; see Figure 1. Following Mendel's postulates, an offspring of the two parents has two copies of the color allele, one

from each parent. Since the second parent has only allele Y, it passes this allele to a child with probability 1. But the first parents passes allele Y and allele G each with probability 1/2. The offspring of a mating of parents 1 and 2 will thus receive YY or GY as its two copies of the color gene, each with probability 1/2. Similarly it will get either SS or WS for its two copies of the texture gene, again each with probability 1/2. Now consider the traits in combination. Independent assortment means that whatever combination YY or YG an offspring receives for the color gene has no effect on the probabilities that it receives SS or WS. Thus all four possible combinations of the allele pairs YY or YG with SS or WS are possible in the offspring, as shown in Figure 1. By independence, the probability of YY WS is the probability of YY times the probability of WS, namely $(1/2) \times (1/2) = 1/4$. Similarly, the probability of each of the other 3 combinations of alleles is 1/4.

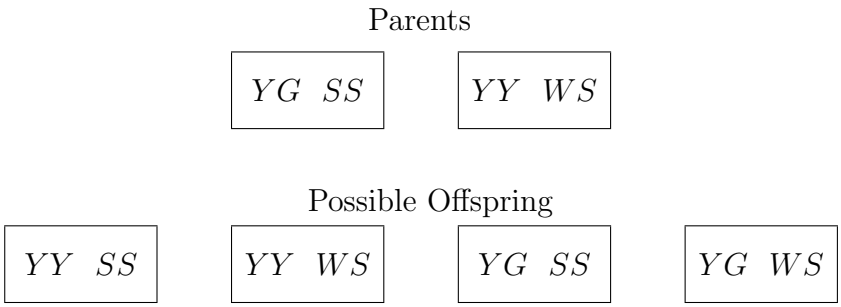

Figure 1. Assortment

It turns out that Mendel's postulates are not always correct. The independent assortment postulate is often true, as for the traits in Mendel's pea experiment, but not universal; oftentimes assortment of alleles is statistically linked, as will be explained shortly. Often, also, one allele is not strictly dominant. However, the idea that there is a fundamental unit of heredity is central in modern biology, and, when modified to account for how reproduction and expression of genes actually works, Mendel's theory continues to provide a correct framework for genetic analysis.

1.2 Genes, chromosomes, and sexual reproduction

In early genetical science, the idea of a gene was an inference from experiments; Mendel and his successors would have had little basis for speculating on the mechanisms by which units of hereditary information were stored or transmitted. But the theory's success suggested that genes exist as real physical entities. This is correct, and the development of genetics and molecular biology has progressively elucidated the physical basis of the gene and of how genes work, down to the molecular level.

The first step in understanding physical mechanisms of inheritance was connecting genes to **chromosomes** and to the behavior of chromosomes in sexual reproduction. In **eukaryotic** cells, that is, cells with nuclei, chromosomes are large complexes of protein and nucleic acid residing in the nucleus. In **prokaryotic** (without nucleus) cells, such as those of bacteria, a chromosome is generally a circular loop of DNA. By the early twentieth century, it was understood that each organism carries a characteristic number of chromosomes and that each of its genes may be physically identified with a precise location in a specific chromosome. Moreover, genes are arranged along chromosomes in a linear fashion. Abstractly, one may think of a chromosome as a line segment and imagine that genes occupy small subintervals of this segment. The position along a chromosome at which a gene occurs is called the **locus** of that gene. In a sense, *gene* and *locus* mean the same thing, with the understanding that the chromosomal material at the locus governs the expression of a particular trait. Alleles are then understood as the alternate forms of a gene that can reside at a locus.

Biologists also observed that the normal body cells of many organisms are **diploid**. In diploid cells the chromosomes come in pairs, each member of the pair being a version of the same chromosome, with the possible exception of the chromosome determining sex. More precisely, each member of a pair carries loci for the same traits, arranged in the same order. A diploid individual therefore carries two versions of the gene for each trait. We shall refer to the members of a chromosome pair as "copies" or "versions" of the same chromosome, although they are not in general exact replicates of one another, because the alleles they carry can differ. For example, this would be the case of one of Mendel's peas carrying one Y and one G allele for pea color; the Y allele sits at the locus for pea color on one copy of a paired chromosome,

while G sits at the corresponding locus on the other copy. Clearly, diploidy is the physical expression of Mendel's hypothesis that organisms carry two, possibly different alleles for each gene.

Diploidy is not a universal rule. Strict diploidy does not even hold for most **dioecious** species, which is the term for a species having individuals of different sexes. For example, the human female contains two copies of the so-called X chromosome, but the male is degenerate (no surprise here!) and carries only one, so that he has only one copy of each gene located on the X chromosome. On the other hand, many plants are **polyploidal**, meaning their cells contain $2n$ copies of each chromosome, where n is greater than 1.

Going in the other direction, there are cells with only one copy of each chromosome and these are called **haploid**. The life cycles of many "simpler" organisms —for example, one-celled eukaryotes (protists), mosses, and fungi—include extensive haploid stages. Prokaryotes are haploid and reproduce asexually, but few eukaryotic species are haploid throughout their entire life cycle, and, indeed, sexual reproduction is not possible without a multiploidal stage. Diploid organisms produce haploid cells for sexual reproduction, as we explain next, and that is the main importance of haploid cells for us.

Sexual reproduction involves the fusing of genetic material from two parents, and biologists also understand this well at the chromosomal and molecular level. The following discussion assumes the ideal model of a diploid organism. For sexual reproduction, organisms make haploid cells called **gametes** —eggs or sperm—from diploid originals, by a type of cell division called **meiosis**. The details of meiosis are intricate, but to understand the flow of hereditary information we need only the following facts. In the first, preliminary stage of meiosis, each copy of each chromosome pair of the parental cell is duplicated, resulting in a cell that contains 4 versions of each chromosome and hence 4 loci for each gene. These copies are then divided among 4 progeny cells, each of which receives one copy of every chromosome, hence contains only one locus for each gene. It is these haploid progeny that are the gametes. Sexual reproduction occurs when two gametes fuse and create a diploid cell, called the **zygote**, which then develops into the mature individual.

Figure 2 presents a caricature of meiosis for an organism with two chromosomes, in the case that each chromosome and its copies remains intact in the passage from generating cell to gametes. It is meant only to show the flow of genetic information and omits the physical processes by which meiosis

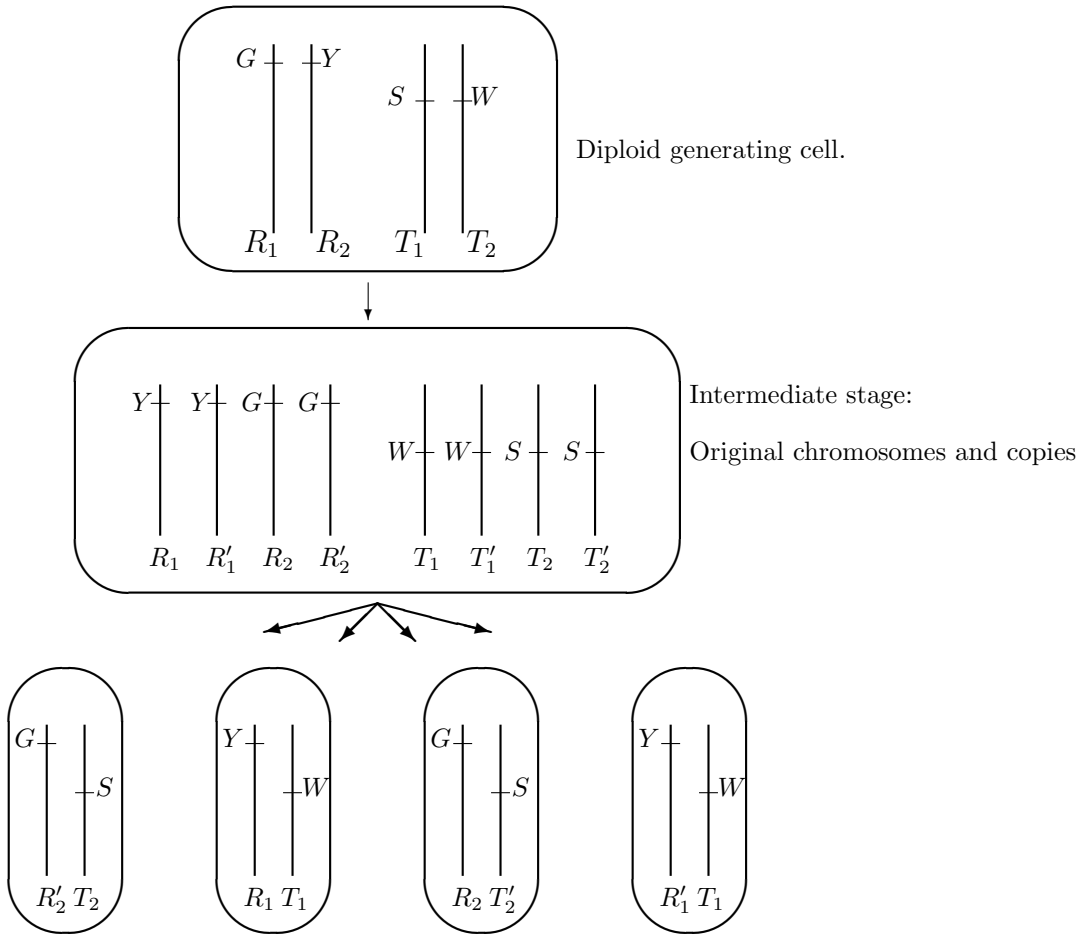

Figure 2. Caricature of meisosis without recombination.

actually takes place. The cell at the top is the diploid cell before meiosis. It contains copies $R1$ and $R2$ of chromosome R and copies $T1$ and $T2$ of chromosome T. These would be in the cell nucleus, but even this is not shown.

As meiosis begins each chromosome is copied, resulting in a cell with 4 separated versions of each chromosome. The result is shown in the middle diagram; R_1' is the copy of R_1, R_2' is the copy of R_2, etcetera. In the final

stages, the chromosomes and their copies are assorted into four gametes, each receiving one chromosome of each type. In figure 2, R_2' and T_2 find their way into one gamete, R_2 and T_2 into another, and so on. Since in this picture it is assumed that each primed chromosome is an identical copy of the corresponding primed component, two of the final gametes in Figure 2 are of type R_2T_2 and two of type R_1T_1. However, it could have happened instead that the four gametes contained the pairs (R_1, T_2'), (R_1', T_1'), (R_2, T_1) and (R_2, T_2), in which case we would have obtained one gamete of each of the four possible types, R_1T_1, R_1T_2, R_2T_1, and R_2T_2.

This picture of meiotic division neatly explains Mendel's independent assortment postulate, at least for two traits whose loci are on different chromosome pairs, if it is assumed that the R and T types of chromosomes are inherited independently by the gametes, in other words, that different chromosomes are independently assorted. This assumption is very reasonable if we assume there is no physical linkage of one member of the R-pair with another of the T-pair, because then there is no reason for any one of the possible gametes (R_1, T_1), (R_1, T_2), (R_2, T_1) and (R_2, T_2) to be more likely than any other. If chromosomes are independently assorted, then so are the alleles they carry and Mendel's independent assortment postulate follows as a consequence. To illustrate, suppose that the cell at the top of Figure 2 represents a diploid cell in one of Mendel's peas and that the locus governing pea color is on the R. In Figure 2, the two copies of this locus are shown on R_1 and R_2, with respective alleles Y and G. Likewise, suppose the T chromosome carries the locus for pea texture, with the specific alleles W and S, as shown. As half of this individual's gametes have a copy of R_1 and half a copy of R_2, half will carry allele Y and the other half allele G. And similarly, half of the gametes will carry allele W, and half S. Since the chromosomes carrying these alleles assort themselves independently, the same will automatically be true of the alleles for color and texture carried on these chromosomes.

Meiosis also explains when independent assortment fails and to what degree. Consider two loci ℓ_1 and ℓ_2 along the same chromosome. Imagine these loci are on the paired chromosomes R_1 and R_2 of Figure 2, and suppose that copy R_1 carries allele A at locus ℓ_1 and allele C at locus ℓ_2, while R_2 carries allele a at its copy of ℓ_1 and allele c at ℓ_2, as in the two chromosomes at the top of Figure 3. If the caricature of Figure 2 were always true and R_1 and R_2 were always copied intact in meiosis, there could be no assortment of the alleles A and a with C and c. Each gamete would contain either A and C or a and c, depending on whether it receives a copy of R_1 or R_2. Geneticists

describe this situation by saying that the alleles at loci ℓ_1 and ℓ_2 are stricly linked.

In actual fact, strict linkage is not observed even for loci on the same chromosome, and to explain this we need to complete our picture of meisosis with a discussion of **recombination**. Recall that in meiosis each copy of the pair is duplicated, at which point the cell will contain two replicates of R_1 and two of R_2. Before being separated into the four gamete progeny, these four chromosomes form a complex in which it is possible for a piece of one of the replicates of R_1 to exchange places with the analogous piece of one of the replicates of R_2. If this happens, we still have four versions of the same chromosome; one is a straight copy of R_1, one a straight copy of R_2, but the other two are complementary amalgams of R_1 and R_2. These are the four versions that are then divided amongst the four gametes. Figure 3 is a schematic explanation of a recombination event. On the top are the two original members, R_1 and R_2, of a chromosome pair. On the bottom are two new chromosomes resulting from an interchange of material between them. The figure shows how recombination can separate loci on the same chromosome. Whereas alleles A and C both reside on R_1, and alleles a and c on R_2, one of the recombined chromosomes carries A and c, while the other carries a and C.

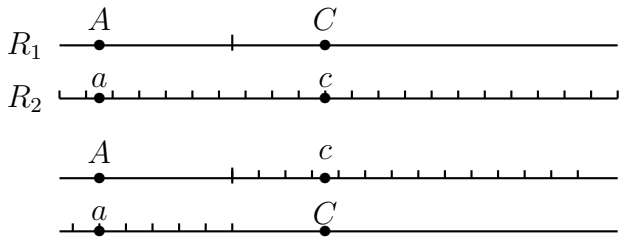

Figure 3. Interchange of chromosomal material in recombination

Recombinations and where they are located may be treated as random events. They can occur throughout the chromosome and may or may not separate two given loci. For any two given loci along a chromosome, there will be some probability, called the *recombination frequency*, that they end up on different copies of the chromosome in a gamete. The less this probability, the greater the degree with which the alleles are linked. Thus, genes on the same chromosome can assort themselves among different gametes in reproduction, but will not assort independently.

Recombination has been very important to the development of genetics. Geneticists studying a particular organism want to know on which chromosome and where on that chromosome each gene locus sits; they want, in other words, to construct *maps* from known genes to chromosomal locations. Nowadays, DNA sequencing technologies help researchers order genes along the chromosome. However, traditional—that is, non-molecular—genetics techniques, combining pedigree analysis with examination of chromosomes, have led to detailed and accurate gene maps, and they exploit recombination in an essential way. The simple idea is that alleles at loci which are close together will not get assorted as frequently by recombination as those farther apart. Thus recombination frequencies are a measure the distance of loci from one another and help to establish the relative positions of genes along a chromosome. We do not address these traditional techniques in this text. Different mathematical issues in DNA and protein sequence analysis are discussed in Chapters 5, 6, and 7.

We defined a gene above abstractly, as a unit of heredity. Using our understanding of how chromosomes function we can make a more physically concrete definition. Here is the formulation from the glossary of W.J. Ewens' monograph, *Population Genetics*, Methuen, London (1969): a gene is "a minute zone of a chromosome which is the fundamental unit of heredity. A gene partially or wholly governs the expression of a certain character or characters in an individual." This definition still does not take us down to the molecular level, nor explain how inheritance is carried on the chromosome, but it is adequate to understand much of population genetics theory.

(Note on terminology: In Ewens' definition, "gene" is practically synonymous with "locus"; the term "allele" then refers to the different trait-determining units that can occupy a locus. This is the usage common in molecular genetics. However, as Richard Dawkins points out in, *The Ancestor's Tale*, Houghton-Mifflin, New York, 2004, p. 47, many biologists use "gene" and "allele" are almost interchangeably. One may speak, say, of the gene for green pea color, as well as the allele for pea color. Our usage is more in line with that of the molecular geneticists, but we often find it convenient to refer to alleles as genes, or versions of genes. No confusion should result.)

1.3 Genotypes, phenotypes, polymorphisms

The concepts and terms defined in this section are basic to genetics and to the applications discussed in this text. Learn them well. Confident use of them can even give the impression that you know something about biology!

The **genotype** of an individual is a list of the alleles it carries at the loci of its chromosomes. A genotype thus summarizes the genetic endowment unique to an individual. In practice, one almost always discusses genotypes *with respect to* a fixed, usually small, number of loci or traits. For example, the genotype of one of Mendel's peas, with respect to the gene for pea color only, is completely specified by the two alleles it carries for pea color. The possible genotypes in this case are denoted by YY, YG, and GG, refering respectively to individuals with two yellow alleles, one yellow and one green allele, and two green alleles. A genotype for both pea color and texture together would also list the alleles a pea plant carries for texture. For example, a plant with genotype $YYSW$ carries two green alleles for color and one smooth and one wrinled allele for texture. Genotypes are usually denoted in this fashion, as a list of letters standing for the alleles. One may also discuss the genotypes of haploid gametes similarly; thus the genotype of a pea plant gamete carrying one allele for green color and one allele for smooth texture would be GS. In discussing the genotypes of diploid organisms, the following terminology is fundamental. An diploid individual bearing a genotype, AA, with identical alleles at a locus, is said to be **homozygous** at that locus; if the genotype is AB, where A and B are different the individual is **heterozygous** at that locus.

It is important to distinguish an organism's genotype from its **phenotype**. The phenotype is the set of the organism's physical, biochemical, and behavioral traits—how it looks and functions in the world. As with the word *genotype*, the term *phenotype* can be applied either in a broad sense to the whole organism or, in a restricted sense, to a specified set of traits being studied. Despite a popular misconception that it's all in the genes, that we are our genes, the relationship between genotype and phenotype, between genes and traits, is complicated and varied. First, even with respect to traits which are determined with fair precision by genes, the map from genotype to phenotype is not one-to-one. Individuals with different genotypes can have the same phenotype. This happens because of the phenomenon of allelic dominance. For example, in Mendel's pea plants, the allele for yellow color dominates that for green. This means that if a pea plant is a heterozygote for

color, that is, possesses one allele for green and one for yellow, it will have yellow peas, just the same as a homozygous plant with two yellow alleles. Thus heterozygotes and homozygotes for yellow will be phenotypically indistinguishable. It also happens that smooth skin dominates wrinkled skin in Mendel's peas. Because of this, if you examine every genotype that appears in Figure 1, you will see that every organism with these genotypes will have the same phenotype with respect to pea color and texture; they will all have smooth, yellow peas. Only by crossing the rightmost offspring with itself, is it possible to obtain in the next generation a $GG\ WW$ individual whose phenotype will display both recessive traits. Another complicating factor in the relationship between genotype and phenotype is environment. The expression of an individual's genetic endowment, as it develops into a mature individual and lives its life, is mediated by the environment. Environmental influences will cause even genetically identical individuals—in other words, clones—to have different phenotypes; for example, supplied with different amounts of nutrition, clones might grow to different sizes. A third complication is the complex way in which genes affect traits. There are of course traits controlled by a single gene; the pea traits studied by Mendel appear to be examples—a very fortunate fact for Mendel as it led him to the idea of discrete hereditary units. But other traits may be governed by the interaction of many genes, so that various combinations of alleles can cause subtle difference between individuals. These are called **quantitative traits**, and it is in general complicated to sort out their genetical basis. Finally, there are molecular processes, such as *DNA methylation* or *histone modification*, which affect gene expression and hence the phenotype, and which also can be passed from one generation to the next, although the underyling genotype remains unchanged. The study of how non-genetic and environmental factors influence gene expression is called *epigenetics*.

Let us now step back from a focus on individuals and consider populations. Fix a particular locus or gene to consider in a population of individuals of the same species. The situation in which two or more alleles at this locus are present in the population is called a **polymorphism**. Actually, a polymorphism is defined to occur only if the frequency of the most common allele is less than or equal to 0.95. This restriction is imposed to limit the term to those situations in which the diversity of alleles across a population is significant, since for almost any locus in a large population there will be a least a few individuals with a variant gene, perhaps one recurrently introduced by rare mutations. In practice, it seems that geneticists loosen the

95% rule when discussing rare genetic diseases.

1.4 Heredity at the molecular level

In the last half of the twentieth century, an understanding of the gene at the molecular level emerged, following upon the discovery in the 1940's that deoxyribonucleic acid, DNA, is the heredity-carrying material of chromosomes. Although there is still much to learn, we know a lot today about how DNA stores hereditary information in chromosomes and how the cell 'reads' that information to acquire its inherited traits. The following two statements summarize the essentials of what the reader should understand in order to appreciate the biological sequence models studied in this text.

1. DNA (deoxyribonucleic acid) is a linear, unbranched polymer consisting of two complementary chains of the nucleotides. For the purposes of genetics, DNA may be visualized as complementary strings of letters from the DNA alphabet $\{A, G, C, T\}$, each letter corresponding to one of four types of nucleotides. (Nucleotides will be explained below.)

2. A gene is a segment of chromosomal DNA that holds the instructions for the production of a **polypeptide**. Polypeptides are the linear chains of amino acids making up proteins, so, in essence, genes code for proteins. The code is contained in the order in which the letters A, G, C, and T appear along the gene.

   In this interpretation, the locus of a gene is that specific stretch of DNA where the code for its associated polypeptide is located. The alleles of a gene are variant DNA sequences at its locus. Different alleles will cause the cell to produce different variants of the associated polypeptides. *Expression of a gene* means production of the protein it codes for.

Statement 2 expresses the modern, molecular-level definition of a gene called the **one gene, one polypeptide theory**; to repeat, a gene is a segment of DNA that governs the production of a polypeptide. We will see later that this theory needs to be reconsidered and expanded in light of the recent discoveries of genomics.

In the remainder of this section, we explain statements 1 and 2 in more detail and discuss proteins as well.

1.4.1 DNA

To the reader who has not already studied molecular biology, the definition of DNA in statement 1 above is not likely to be too informative. To elucidate, it is necessary to explain what a nucleotide is, how nucleotides link in chains, and what complementary chains are. We will explain in enough detail to make the second point of statement 1 clear. To understand in broad outline how DNA functions as hereditary material and to formulate statistical models for DNA, it suffices to picture a DNA molecule using a directed strings of letters from the alphabet $\{A, G, C, T\}$. The letters A and G stand respectively for the two purine molecules adenine and guanine, C and T for the pyrimidine molecules cytosine and thymine. These are relatively small, nitrogenous, cyclic molecules and, in the context of discussing DNA, are collectively referred to as **bases**. (Here, "base" is used in the sense of a fundamental building block, not in the sense of a base as opposed to an acid.)

DNA molecules are formed from linking together small subunits called **nucleotides**. A nucleotide is a molecule consisting of a deoxyribose sugar molecule to which is attached one of the bases, A, G, C, or T and also a phosphate group. It is not necessary to understand all the chemical terminology here, only the following picture.

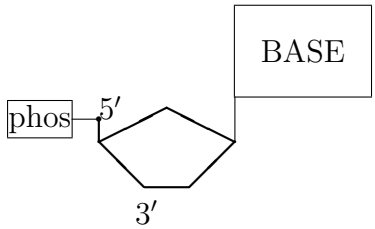

Figure 4: A nucleotide

In Figure 4, the pentagon together with the little post on the left labeled $5'$ represents the deoxyribose sugar. Atoms sit at the vertices of the pentagon and at $5'$ and the edges represent bonds. A base (A, G, C, or T) bonds to the sugar at the vertex shown on the right, and the phosphate group bonds to the carbon atom sitting at the vertex labelled $5'$. We have labeled one other vertex on the deoxyribose pentagon as $3'$. The $3'$ and $5'$ labels are conventions of biochemistry.

The manner in which nucleotides link together in a chain is now simple to describe. The $3'$ carbon atom of one nucleotide bonds to the phosphate group

of the next one; (the phosphate group is modified in the process, but this is not important). Any number of nucleotides can be linked in this manner.

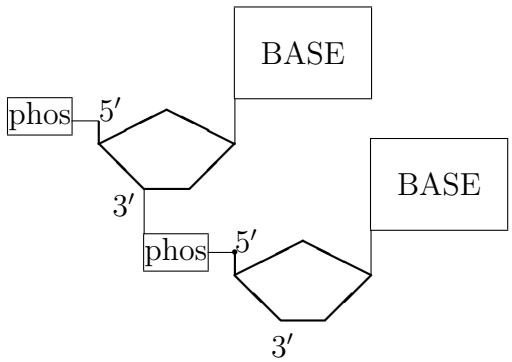

Figure 5: Linked nucleotides

As a consequence, a single-stranded chain of DNA is essentially a sequence of bases supported on a backbone composed of linked sugar and phosphate molecules. Abandoning all pretense of correctly drawing molecules, we may represent a single strand as in the following example:

$$5' - TAGGTTAGGCTATTAGGCTGA - 3'$$

Here, we have suppressed all representation of the backbone and of chemical bonds. We show only the sequence of bases along the chain, and we retain a $5'$ and a $3'$ at the ends to indicate the direction in which the nucleotides are linked. Thus, the $5'$ carbon of the nucleotide supporting the first base A on the left is unlinked, or at least the nucleotide to which it links is not shown. The $3'$ carbon of this first nucleotide then links via a phosphate group to the $5'$ carbon of the next nucleotide supporting G; the $3'$ carbon of the nucleotide for G then bonds via a phosphate group to the $5'$ carbon of the next nucleotide, which supports the base C, and so on down the line.

The asymmetry inherent in the 5' to 3' linkage gives an orientation to a single strand of DNA. We will adopt the convention that if single-stranded DNA is written without indication of its orientation, the $5'$ to $3'$ direction is assumed. The previous example would this be represented simply as $TAGGTTAGGCTATTAGGCTGA$. This convention brings us to the main point: single-stranded DNA can be thought of as a word read from left to right using letters from the alphabet $\{A, G, C, T\}$.

In its normal state, DNA consists of two complementary nucleotide chains. What happens is that two chains link to one another by means of hydrogen bonds between the bases. But the two chains are complementary in that the adenines of one chain bond only with the thymines of the other and the guanines of one chain only with the cytosines of the other. To illustrate, here is an example of the representation of a double-stranded DNA segment.

$$5' - TAGGTTAGGCTATTAGGCTGA - 3'$$
$$3' - ATCCAATCCGATAATCCGACT - 5'$$

The top strand is just the one used in the single stranded example above. We imagine that hydrogen bonds connect each base in the top strand with the complementary base below it, but the bonds are not represented. The complementarity of the strands is evident; only A is paired with T and only C is paired with G. Note also that the bottom strand reads in the opposite direction from the top. The sugar/phosphate backbones not shown lie to the outside of both chains. In its regular state the whole assembly twists, forming the famous double helix.

A pair of complementary nucleotides linked by hydrogen bonds is called a base pair, abbreviated *bp*. The base pair or kilo base pair (*kb*) is commonly used as a unit of length when discussing DNA. Thus, we would say that the example used in the previous paragraph is 21 *bp* long.

1.4.2 DNA, genes, and proteins

The one gene/one polypeptide theory was encapsulated in summary statement 2 above. How does this molecular-level definition connect to that in section 1.3, defining a gene as a unit of heredity governing the expression of a trait? Before answering this question, we look more closely at what proteins are and how genes store the information for making proteins.

Chemically, proteins are polypeptides, which means that they are linear chains of amino acids joined by peptide bonds. An amino acid consists of an amine functional group, NH_2, a carboxyl group, $COOH$, and a side group generically denoted by R, all linked via a carbon atom. The side group determines the type of amino acid and its chemical properties. There are twenty, biologically important amino acids, each of which has a standard, single capital letter abbreviation. The names of the amino acids and their respective abbreviations are shown in Table 1. There are a few other amino

acids which appear in proteins, but the 20 shown are those directly coded for in genetic DNA

The amino acids in a polypeptide are linked by a *peptide bond*, in which the nitrogen (N) of the amine group of one amino acid bonds to the carbon atom (C) of the carboxyl group of the next in the chain. Therefore there is an unbound amine group at one end of the polypeptide, and an unbound carboxyl group at the other end; the terminal unbound amine group is called the *N-terminus*, and the unbound carboxyl group is called the *C-terminus*. The sequence of amino acids along a polypeptide chain is called its **primary structure**, and because of the N-terminus and C-terminus distinction between its ends, it has a direction. In the cell, proteins are synthesized in the $N-$ to $-C$ terminus direction. Thus, the convention is to represent the primary structure as a list of the amino acids from the N-terminus to the C-terminus. If each amino acid is represented by its one letter abbreviation in Table 1, the result is an ordered sequence, called a *word*, in the amino acid alphabet. from the amino acid alphabet.

There are further levels of protein structure The linear chains comprising a protein are coiled or arrayed in sheets—this is the**secondary structure**—which are then folded over each other and held together with cross-linking bonds to create a characteristic 3-dimensional shape (**tertiary structure**) that is the key to how the protein functions in the cell. The biologist studying a protein wants ultimately to understand its 3-dimensional structure. However, it seems the primary structure of a protein determines its native secondary and tertiary structures, and therefore primary stucture is the focus of bioinformatic study of proteins.

DNA carries the hereditary information for making a protein by storing in its sequence of bases the information to recover the protein's primary structure. DNA accomplishes this by means of the famous **genetic code**. Every three-letter "word" or **codon** from the DNA alphabet $\{A, T, C, G\}$ codes either for one of the 20 amino acids or a signal to start or stop coding. By convention, codons are always represented in the $5' - 3'$ direction because DNA synthesis takes place in this direction. Starting at the $5'$-end of a strand of DNA and "reading" it codon by codon thus produces an ordered list of amino acids, which, if actually strung together, would make a protein. Any segment of a DNA molecule that begins with a start codon and ends with a stop codon and is $3n$ bp long for some integer n, is called an **open reading frame**. In the one gene/one polypeptide theory, a gene is an open reading frame in the chromosomal DNA that gets translated into an actual

protein. The transcription mechanism that accomplishes this translation is a complicated affair involving the molecule RNA. We will make a few, brief remarks about it in the next section. Note however, that not every open reading frame in chromosomal DNA is a gene. The protein transcription mechanism somehow knows, presumably by way of regulatory regions in the DNA near the gene, which open reading frames to transcribe.

The genetic code is also included in Table 1. In listing the codons for the amino acids, an asterisk indicates that any base may occupy that position; for example, GCT, GCA, GCG, and GCC all code for alanine. The codons for start and stop are not shown; they are TAA and TAG for start and TGA for stop. Notice that the genetic code is many-to-one. There are 64 possible codons (first math problem: why?) and only 20 amino acids.

Amino acid	Abbr.	Codons
Alanine	A	GC*
Arginine	R	AGA,AGG,CG*
Asparagine	N	AAT, AAC
Aspartic acid	D	GAT, GAC
Cysteine	C	TGT, TGC
Glutamic acid	E	GAA, GAG
Glutamine	Q	CAA, CAG
Glycine	G	GG*
Histidine	H	CAT, CAC
Isoleucine	I	ATT, ATC, ATT
Leucine	L	CT*, TTA, TTG
Lysine	K	AAA, AAG
Methionine	M	ATG
Phenylalanine	F	TTT, TTC
Proline	P	CC*
Serine	S	TC*, AGT, AGC
Threonine	T	AC*
Tryptophan	W	TGG
Tyrosine	Y	TAT, TAC
Valine	V	GT*

Table 1.1: Amino Acids and the Genetic Code

Exercise 1. Verify that the example introduced on page 12 of a single strand of DNA is an open reading frame and write down the sequence of amino acids it codes for. Identify all triplets inside the DNA segment that code for start.

Biologically, proteins are the molecules which carry out all the major functions of a cell. They govern cell structure and facilitate regulatory and immume mechanisms, transport and movement. Enzymes, which catalyze the chemical reactions in a cell, are proteins. In short, the proteins in a cell largely determine its traits—what it does, how well it performs its functions, how it looks—and the traits of the organism to which it belongs. Thus, in the one gene/one polypeptide theory, genes pass on traits by encoding the information for making proteins.

Sickle-cell anemia in humans provides a classic and dramatic illustration of the one gene/one polypeptide theory. A specific gene in the human genome encodes instructions for fabrication of the beta sheet of hemoglobin. In normal hemoglobin, glutamic acid occupies the sixth position in its beta chain (you don't have to know what this is—it's just a part of the hemoglobin molecule), and it is encoded for in the gene by the DNA triple GAG. The allele responsible for sickle-cell anemia differs from the normal allele at a single nucleotide only! Affected individuals carry the codon GTG instead of GAG for amino acid six in the beta chain, causing a substitution of valine for the normal glutamic acid, as one can see from Table 1. This one substitution so affects the three dimensional structure of hemoglobin as to cause sickle-cell anemia.

1.4.3 RNA and genes

RNA, ribonucleic acid, is another biologically important nucleic acid. Its structure very much resembles that of DNA. In its single stranded form, it is again a linear chain of nucleotides, but the sugar in the nucleotide is ribose rather than dioxyribose and the base uracil (U) replaces the base thymine (T) appearing in DNA. Thus, a single strand of RNA may be represented as a string of letters, read from left to right, from the alphabet $\{A, G, C, U\}$. Base pairing of nucleotides, A with U and G with C, occurs also in RNA, but the typical RNA molecule is not the two-stranded, double helix of typical DNA. RNA structures are more diverse, and although they contain regions where two strands align by base pairing and twist into a double helix, these regions are typically on the order of tens of base pairs.

RNA is the hereditary material of many viruses, but in all higher organisms it stars in a number of auxiliary roles. For example, RNA is the crucial molecule in the mechanisms, now fairly well understood, for transcribing genes into proteins. In this process, the double stranded DNA of a gene unzips and is transcribed into a so-called **messenger RNA** that is formed on the DNA template by complementary base pairing; a nucleotide bearing A in the DNA gives rise to a nucleotide bearing U in the messenger RNA, T in the DNA gives rise to A in the RNA, etc. Messenger RNA now has the code for making the protein, which it carries to the ribosomes for actual protein production. For this reason, messenger RNA is called a coding RNA. There are also many different types of non-coding RNA: ribosomal RNA, which is a major constituent of the ribosomes where proteins are produced; transfer RNA, which transports amino acids to the ribosomes; small nucleolar and small cytoplasmic RNA, and other forms. RNA molecular biology is developing dramatically, as new types of RNA are being discovered and synthesized, and as their structures and functions are being resolved.

1.4.4 The genome, genomics, and the gene

Molecular genetics has progressed dramatically in recent years, in part because of ever improving technology for sequencing, that is, for determining the order of bases in DNA. The first draft of the approximately 3 billion base pairs of modern human DNA was completed in 2001. It is now possible, in a reasonable amount of time and with high accuracy, to obtain model sequences for the entirety of the chromosomal DNA of a species. It is even possible to process the highly degraded DNA from the bones of prehistoric humans and Neanderthals, and to reconstruct sequences that are sufficiently reliable for studying how and to what extent human and Neanderthals were related. Such proficiency in sequencing makes it possible to study the entire genetic complement of an organism as an object in its own right. This is the science of **genomics**, the study of the **genome**. Precise definitions of the term, *genome*, vary from author to author. For example, Robert F. Weaver, in *Molecular Biology*, second edition, McGraw-Hill, 2002, defines it abstractly as the "complete set of genetic information from a genetic system." A more concrete variant identifies the genome of an organism with the sum total of its chromosomal DNA, considered as a long sequence of bases. If every part of this sequence conveyed information that affected inherited traits these two definitions would be synonomous, but, as we shall see, it is not clear this is

the case. Since, in practice, genomics seems concerned with understanding the structure and function of all parts of inherited DNA sequences, we shall 'genome' in the sense of this second definition.

The current discoveries of genomics about the structure of the genome and about the functions of its different parts are challenging our understanding of what a gene is and of how genes are translated into the phenotype of an individual. Before genomics got started one might have imagined the genome as an organized array of closely spaced genes—in the sense of the one gene/one polypeptide definition—somewhat like the books in a neatly arranged library with minimal empty shelving. One surprising discovery is how sparse protein coding regions actually are in the genomes of higher organisms; they occupy but small islands in a sea of intergenic DNA. For example, estimates of protein coding genes as a fraction of the human genome come in at 2 percent or under. Another, related discovery is how weakly the number of genes correlates with the complexity of an organism and with genome size. Using sequence data and experiments on gene expression, it is possible to estimate the number of genes in a genome without having to understand what each gene is and what the protein it determines does. The exact number is a subject of debate, but the human genome is thought to contain between $20,000$ and $25,000$ genes, with most estimates at the lower end. The genome of a species of nematode worms—abundant, generally microscopic, ecologically important worms with simple body plans and nervous systems—also seems to have about $20,000$ genes. (However, when noncoding DNA is included, the human genome is about 30 times as large as that of the nematode.) Moreover, even genes themselves do not have a simple structure. They are composed of interspersed regions called **introns** and **exons**. The RNA transcribed directly from the gene—called the *primary transcript*—is edited before being translated into a protein and the introns correspond to those portions of the DNA whose RNA transcripts are edited out and never expressed; the **exons** are the regions of the gene whose DNA code is actually (ex)pressed in a protein.

At first, the abundance of intergenic DNA was confusing. Some of the functions of this 'non-coding' DNA were understood. For example, there are promoter regions that function in the expression of genes, and there are regions transcribed into the different types of non-coding RNA described above. Other short sequences define segments to which various proteins bind for regulating gene expression and for determining how the DNA is physically packaged in the chromosome. There are identifiable types of subsequences:

pseudogenes, which look very much like known genes but are not expressed; sequences called transposons which have the ability to move about in the DNA as a unit; and so-called repeat sequences, which feature repetitions of small DNA words and occur at many places in the genome. Still there is much DNA that can be cannot be assigned to one of these classes, and whose function, if any, is not known. We would like to think it matters, because we have more of it than nematodes, and we would like to think of ourselves as the superior species! But an early hypothesis, the *junk DNA* theory, postulated that the bulk of this DNA in fact has no function; it somehow arose through extraneous duplication of functional material, insertion of base pairs, whatever, and, having no deleterious effect, is not weeded out by selective pressures as the genome evolves. According to this view, the appropriate metaphor for the genome is not an orderly library, but the large study of a rich and eccentric collector. There are plenty of books (genes) on the shelves, to be sure, but they appear haphazardly among an overwhelming larger collection of old machine parts, half-finished manuscripts, worm-eaten extra copies, and miscellaneous bric-a-brac.

The genome as a whole is being intensely studied. The ENCODE project (*Enc*yclopedia *o*f *D*NA *E*lements) is a multi-team effort sponsored by the National Human Genome Research Institute, to catalog the functional elements of the genome. Its results largely contradict the junk DNA hypothesis and point to a more complex picture of genome function than a focus on genes as protein encoders might suggest. In the first phase, ENCODE teams analyzed about 1% of the the human genome closely, trying to identify and map all functional elements. The summary results of this project are reported in Volume 17, issue 6, of *Genome Research*, which appeared in June, 2007. This issue includes informative commentary and review articles; I cite in particular *ENCODE: more genetic empowerment*, by G.M. Weinstock, and *Origin of phenotypes: genes and transcripts*, by T.R. Gingeras. One fact that emerges is that a great deal of the genome, in intergenic regions as well as genic regions, is transcribed into RNA. These transcripts include of course the messenger RNA that helps turn the genetic code into propteins. Other transcripts are the RNA molecules discussed above whose functions we already know. But the functions of most transcripts are not known, and biologists have begun referring to them collectively as TUFs—Transcripts of Unknown Function. How these transcripts fit into the genome is also suprisingly complex. Many overlap with gene transcripts from both introns and exons; that is different transcripts are produced from overlapping parts of

the genome sequence. Others are taken from intergenic regions. Researchers are starting to propose that it is this network of transcripts and regulatory regions, not just the array of coded proteins, that regulates the phenotype, and that the sophistication of this network structure, rather than number of proteins, is the source of the increased informational content expressed in more complex organisms. Some also propose that the gene should be defined, not narrowly as a protein coding region, but more broadly as a transcript, a piece of RNA decoded from genetic DNA.

ENCODE continued after 2007 with a second phase to see if the results it obtained extended to the whole human genome. Thirty research papers appeared in September 2012 reporting on the result of many teams working on this project. The news article, Pennisi, E, "Genomics: ENCODE project writes eulogy for junk DNA," *Science* (**337**), Issue 6099, 1159-1161, summarized the results briefly. ENCODE finds that about 80% of the human genome serves some function and that about 76% of the full genome is transcribed.

The ENCODE estimates of functionality are hotly debated. The ENCODE project probes for evidence of functionality in non-coding regions of the genome mostly with biochemical tools. However, analysis from the evolutionary perspective leads to much lower estimates of the functional fraction of the human genome. For example, in "An Upper Limit on the Functional Fraction of the Human Genome," *Genome Biol. Evol.*,(2017) 9(7):1880-1885, D. Graur takes data on mutation rates and uses them in a simple model of the relationship between fertility necessary to maintain a population and reduction in fitness due to deleterious mutations to conclude that at most a quarter of the human genome is functional, and probably much less. Whatever the case, there are many exciting and interesting questions opened up by genomic sequencing, and they will undoubtedly require further refinements of the concept of gene and the understanding of heredity.

1.4.5 Polymorphisms at the molecular level

In the synopsis of pre-molecular genetics in section 1.3, we introduced the term *polymorphism* to describe variation across a population of the alleles at a locus. Similarly, any variation across a population in the base pairs at a site or in a region of the genome is called a polymorphism. Polymorphisms occuring inside genes in the genome underlie allele polymorphisms, since they cause different variants of a protein to be produced. But polymorphisms also

occur at non-gene sites, and such sites are often called **DNA markers**. DNA markers are used extensively in genetic research to help locate the position of genes that control specific traits, especially those implicated in diseases.

Polymorphisms in chromosomal DNA occur in a number of forms. A **single nucleotide polymorphism**, abbreviated SNP, is a polymorphism in the nucleic acids present at a specific base pair. Sickle-cell anemia is an example of a SNP, because, as explained, it is caused by a substitution at a specific base pair in the gene coding for the beta sheet of hemoglobin. Polymorphisms occur also at satellite DNA sites. DNA satellites are locations in which a short DNA word is repeated over and over, as in $CACACACACACACACACA$, a tandem repeat. (The word *satellite* comes from the term *satellite bands* used to describe the bands by which repeats are revealed in centrifugal fractioning of DNA.) Satellites come as *microsatellites*, up to 20kb long with repeat units up to 25bp, and *minisatellites*, less than 150bp with small repeat units, typically tandem repeats. They can be highly polymorphic in the number of repeat units, and so they serve as useful DNA markers.

A third type of molecular polymorphism is the **restriction frament length polymorphism**, or RFLP. There is a class of enzymes, called restriction enzymes, which cut DNA at specific sequences of bases called recognition sequences. For example, the recognition sequence of the *Alu*1 enzyme is

$$5' - AGCT - 3'$$
$$3' - TCGA - 5'$$

and when it encounters this sequence it cuts through the DNA between the second and third base pairs. If a DNA segment is mixed with the enzyme and allowed sufficient time to react, it will be cut at every recognition sequence and hence digested into a number of small fragments. An RFLP is a polymorphism in the locations in the genome of of the recognition sequence for a restriction enzyme; that is, the locations of the recognition sequence will differ among members of the same population. One way such a polymorphism can arise is through point mutations. Imagine an individual whose genome contains recognition sequences for *Alu*1 starting at sites labelled $\ell_1, \ell_2, \ldots, \ell_n$. Now think of the population of its progeny after many generations, when mutations have accumulated in the genomes. For example, at one point an individual might be born with a mutation in a base pair at site ℓ_3, so that instead of $AGCT$ appearing there, $GGCT$ appears instead. Thus ℓ_3 will no

longer be a recognition sequence for this individual, nor, barring a mutation that restores the recognition sequence, for any of her progeny bearing a copy of the mutated chromosome. Conversely, a mutation might give rise to a recognition sequence of *Alu*1 at a new location. RFLP's are widely used in genetical studies, in part because there is nice technology for picking them out. The distribution of fragments lengths of DNA digested by a restriction enzyme can be read read off by gel electrophoresis. Clearly the distribution of fragment lengths depends on the locations of the recognition sequences, so individuals with differing locations can be easily identified.

1.5 Notes, References, and Further Reading

The information in this chapter is standard. Among sources I consulted are: T.A. Brown, *Genomes 3*, Wiley-Liss, New York, 2007; Robert F. Weaver, *Molecular Biology*, second edition, McGraw-Hill, New York, 2002; Purves, Orians, & Heller, *Life: The Science of Biology*, third edition, W.H. Freeman and Co., 1992; the web site of J. Kimball's biology text (for sickle cell anemia), `http://users.rcn.com/jkimball.ma.ultranet/BiologyPages`. The book by Brown makes excellent backgroung reading for the student interested in the problems about analysis of DNA sequences addressed in this text. The other books are clear and very approachable texts. I also took information from W.J. Ewens, *Population Genetics*, Methuen, London (1969). *The Ancestor's Tale*, Mariner Books, Houghton Mifflin Company, New York, 2004, by Richard Dawkins, is a well-written popular, but serious account of what the comparative study of DNA sequences tells us about the history of life. *Neanderthal Man: In Search of Lost Genomes*, by Svante Pääbo, is a fascinating personal account of the science of reconstructing genomes from fossil bones and its application to the study of prehistoric human populations.

The understanding of genes and gene regulation, of the function of genomic material and of RNA is developing rapidly. Outside of the technical literature, regular consultation of *Scientific American*, *Science*, and *Nature* will alert the reader to the latest major developments.