

1 Mendelian genetics

Although it has been recognized for thousands of years that traits are passed from organisms to their offspring, the science of genetics began with the experiments of Gregor Mendel. By crossing various strains of peas, Mendel was able to deduce several principles of genetics. The most basic is *unit inheritance*, which states that inheritance is determined by discrete quantities known as *genes* and genes don't mix. For example, when Mendel crossed round peas (i.e. peas with a smooth surface) and wrinkled peas, all of the offspring were round, the offspring were not "kind of wrinkled". In this example, we think there is some gene which controls if the pea is wrinkled or not, and the gene has 2 alleles; round and wrinkled. We call the different values taken by the gene *alleles*. Peas (and humans) have 2 copies of all genes, one from each parent. The set of 2 alleles of a gene for an individual is known as the *genotype* of the individual. In contrast, an observable feature of an organism is referred to as the *phenotype* of the individual. For example, if a pea plant has the allele for round (denoted R) and wrinkled (denoted r) at the pea shape gene, the genotype is Rr (or rR, the order of the symbols is irrelevant). If a pea plant has this genotype, it will appear round, hence we say it has the round phenotype. Another way to think about the relationship between phenotype and genotype is that genotypes help determine the phenotype (in conjunction with environmental effects). The term genotype is also used to refer to the alleles an individual has at more than one gene. If we also know which pairs of alleles were inherited from which parent, then we know the individual's *haplotype*. For example, suppose there are 2 genes, each denoted by a distinct letter, and each with 2 alleles, which we will denote Aa and Bb. If someone has the genotype A, A and B, b then we know one parent must have had the pair of alleles A, B and the other parent must have had the alleles A, b, hence in this case we know haplotype of this individual. In contrast, if the genotype had been A, a and B, b then the parents could have 2 possible configurations for their alleles: A, B and a, b or A, b and a, B. In the latter case we can not deduce the haplotype from the genotype.

If the 2 alleles for a gene are distinct, we say that individual is *heterozygous*, while if the 2 alleles are the same, we say the individual is *homozygous*. For many traits, homozygous genotypes translate into definite phenotypes. For the pea example, if the genotype is rr, the pea will be wrinkled (i.e. the wrinkled phenotype) and if the genotype is RR the phenotype is round. The phenotypes of heterozygous individuals are more complicated. In classical Mendelian genetics, there are 2 ways in which a heterozygous genotype translates into a phenotype and this mapping of genotype to phenotype depends on the trait. For the pea example, the Rr genotype corresponds to a round phenotype, hence we say round is the *dominant* trait and wrinkled is the *recessive* trait. For some traits, heterozygous genotypes imply the individual displays both phenotypes. We call such traits *codominant*. From experimentation, Mendel also deduced that alleles *segregate randomly*, that is, if an individual has 2 distinct alleles for some gene then an offspring of this individual is equally likely to obtain each allele. Mendel's final deduction was that different genes combine independently of one another (this is the principle of *independent assortment*). We will see that Mendel's principles, while useful, are not entirely correct. In particular, the principle of independent assortment is only correct for certain traits. Traits that obey Mendel's first 2 principles are referred to as *Mendelian traits*, and several hundred medical disorders are well modeled as Mendelian traits (i.e. attributable to a single gene which is either dominant or recessive). Cystic fibrosis is a noteworthy example as are some common forms of hearing loss. Many other traits (e.g. height) are far too complex to fit neatly into the Mendelian framework, hence we refer to such traits as *complex traits*.

2 Cell biology

While Mendel's insights regarding the regulation of phenotypes via some unobserved mechanism that is transmitted from parents to offspring was fundamental to our understanding of how some traits are regulated, it left much unexplained. Another fundamental insight into the creation and maintenance life was the observation of cells in living organisms. While cells had been described as early as the mid-1600s, developments in microscopy during the nineteenth century allowed biologists to learn that living organisms are composed of minute, distinct units that have a membrane and structures inside this membrane immersed in a fluid. The structures are called *organelles*, the fluid is called *cytosol* and the collection of organelles and fluid is referred to as the *cytoplasm*. While different types of cells have different organelles, some organelles are common to many cells among complex organisms. For example, mitochondria are organelles found in most human cells that are used by the cell to generate energy so the cell can carry out its functions. Complex organisms, such as humans, are estimated to be composed of trillions of cells. Cells are frequently organized into types, such as the familiar red blood cell, and many types (over 200 currently in humans) have been distinguished. Cells are generally distinguished by the function they carry out in the organism: for example, red blood cells allow for the transportation of oxygen and carbon dioxide throughout the body since these gases will bind to the cell. Complex organisms have special cells, *germ cells*, devoted to the task of reproduction. While there are many single celled organisms, we will focus on multicellular organisms.

While the contents of a cell are inside a membrane, this membrane is semipermeable so as to allow the cell to interact with other cells and its environment. Inside a cell many chemicals are present that interact with each other in a set of extremely complex biochemical reactions. These reactions might cause the cell to move or to secrete some chemical into its cellular environment. In addition, due to external sources, the cell might alter the set of reactions taking place in the cell or in some smaller part of the cell. In this sense, a living organism can be thought of as a collection of intertwined biochemical reactions. Which chemicals are present in a cell and their relative quantity are determined by the action of genes. Almost all human cells have a structure inside the cell called the *nucleus* that holds information on the set of alleles for each gene that an individual has received from its parents. All cells are created by the separation of some parent cell, which is called *cell division*. Cells undergo multiple rounds of cell division, and the process from creation of a new cell to when that new cell divides is called the *cell cycle*. As a cell proceeds through the cell cycle there are alterations that the cell undergoes that are visible with a microscope. Sometimes mistakes occur during cell division which cause the offspring cells to have slightly different alleles from the parent cell. Such mistakes are called *mutations* and are a critical feature of natural selection.

3 Genes and chromosomes

Complex organisms have tens of thousands of genes. These genes are mostly located on *chromosomes*, and the spots on the chromosomes where genes are located are referred to as *loci* (locus is the singular). One can observe chromosomes under a microscope in a cell that is dividing. Chromosomes are segments of deoxyribonucleic acid (DNA), and are contained in the nucleus of the cells of eukaryotic organisms. A eukaryotic organism is one whose cells have nuclei, for example, algae, yeast, plants and animals. The other major division of life are referred to as prokaryotes, with the most notable example being bacteria. While most tend to think of bacteria as organisms that are harmful to animals and plants, animal tissue (including humans) is full of bacteria that play an important role in human physiology. The nucleus is surrounded by a semipermeable membrane called the nuclear envelope that regulates what compounds enter and exit the nucleus. In addition to the nuclear DNA organized into chromosomes, in humans, some DNA is found in

the mitochondria. Most genetic investigations ignore this mitochondrial DNA since there are only 37 genes in mitochondrial DNA whereas there are 30,000-35,000 genes on the chromosomes (the exact number is not yet known). The 37 mitochondrial genes are involved in energy synthesis inside the mitochondria, and other genes located on the chromosomes that reside in the nucleus are involved with this process too.

Humans have 46 chromosomes, arranged in 23 pairs of chromosomes. Of these 23 pairs, 1 pair is known as the sex chromosomes and the other 22 are called *autosomal* chromosomes. The centromere is a location near the center of a chromosome that plays an important role during cell division and lacks genes. The centromere can be observed during certain stages of a cell's development, and the shorter end of the chromosome is referred to as the p arm while the longer end is called the q arm. The ends of the arms of the chromosomes are referred to as *telomeres*.

Organisms receive 1 chromosome from each parent, hence, for autosomal chromosomes, the pairs all have the same gene loci, but possibly different alleles. An individual also receives one of the sex chromosomes from each parent, but these are not necessarily pairs. There are 2 sorts of sex chromosomes, denoted X and Y, so there are a total of 24 distinct human chromosomes. An individual with 2 X chromosomes is a female, while an individual with an X and a Y chromosome is a male. Since females have 2 X chromosomes, they always transmit an X chromosome to their offspring, while males contribute either an X or a Y chromosome (which is actually transmitted depends on chance), hence it is the sperm cell which determines the sex of the offspring. There are several serious medical disorders attributable to a gene which resides on the X chromosome: such disorders are referred to as *X-linked* disorders (or X-linked traits more generally). Such disorders are recessive as only one disease allele results in the disease trait. These traits can seem to skip a generation since an affected mother may carry a disease allele but not have the disease phenotype because the other copy of the X chromosome is sufficient to maintain normal function (color blindness is an example of this phenomenon). A woman with a disease allele on an X chromosome can nonetheless transmit the disease allele to her offspring, and if the offspring is a male he will only have the disease allele at this locus and thereby be affected. Humans receive all of their mitochondrial DNA from the maternal source, hence traits determined by mitochondrial genes can exhibit transmission patterns similar to X-linked traits.

The autosomal chromosomes are numbered according to their length, with chromosome 1 being the longest, and the sex chromosomes are chromosome pair 23. A surprising feature of chromosomes in higher organisms (like humans) is that most of the DNA is not part of a gene (over 95% for humans). In addition, most genes are not composed of contiguous segments of DNA, instead there are interruptions in the coding of a gene. Such interruptions are known as *introns*, while the segments of DNA which encode a gene are called *exons*. Some genes can be used by the cell in more than one way since the exons can be assembled in a variety of ways, thereby generating different gene products from the constituent parts of a single gene. This phenomenon is referred to as *alternative splicing* for reasons that will become clear when we discuss RNA. It is currently not clear what is the purpose, if any, of this extragenic DNA, but at least some portions of this DNA regulates the working of the genes. Some of the extragenic DNA was likely deposited by a virus in some ancestor of contemporary humans, and these sequences have been passed down through time since they didn't have a harmful effect on this ancestor. Often the term locus is used to refer to a location on a chromosome even if there is no gene there, and similarly, allele is used to refer to the segment of DNA that an individual has at the locus. Since chromosomes are composed of DNA, further discussion of DNA is necessary.

4 DNA

All life is regulated by DNA. Cells use DNA to make proteins, and the resulting proteins carry out the functions necessary for life. Proteins are built up from a collection of amino acids (there are 20 amino acids), but often incorporate other molecules in their final form (for example, metals and lipids). DNA is composed of a sequence of sugar molecules (2-deoxyribose) joined by phosphate groups. These sugar molecules contain 5 carbon atoms arranged in a ring, and these are referred to by a number between 1' and 5'. To each sugar molecule, at the carbon atom numbered 1', is attached one of four bases; adenine (A), guanine (G), cytosine (C) and thymine (T). The bases A and G are referred to as *purines* while C and T are called *pyrimidines* (due to details regarding the chemical structure of the base). All genetic information is encoded in these 4 bases. A *nucleotide* is a combination of a sugar, a phosphate and a base. Phosphate groups attach to the carbon atoms numbered 3' and 5'. The linear chain of DNA is built by phosphate groups that link the 5' carbon atom of one sugar to the 3' carbon atom of the next sugar. The bases from distinct nucleotides tend to form hydrogen bonds with one another in a specific manner; A binds with T and G binds with C. Due to this pairing, for every sequence of DNA there is a complementary sequence which is defined by substituting a T for an A, an A for a T, a G for a C and a C for a G.

DNA has a tendency to bind to its complementary copy and form a double helix (i.e. a twisted ladder configuration), but the hydrogen bonds which form between the nucleotides from different strands (i.e. the rungs of the twisted ladder) are easily broken (which is necessary for DNA to fulfill its function). While the cell is not dividing (a period of the cell cycle known as interphase), the DNA molecule actually exists as a complex (i.e. a large molecule made up of a number of molecules) with a group of proteins called histones and a group of acidic nonhistone proteins-this complex is referred to as *chromatin*. Basically, the DNA winds around the histone complex, and is further condensed into a structure called a solenoid-these structures are far more compact than free DNA and help protect it from the sometimes harsh environs of the cell. In a DNA molecule, the nucleotides are arranged in a linear fashion along a chemical backbone composed of the alternating phosphates and sugars. If you start reading the bases as you travel along a DNA molecule, then groups of 3 nucleotides are referred to as *codons*. Of course which end you start reading from makes a difference. Recall the carbon atoms are numbered with connections from carbon 5' to carbon 3', so the ends of a DNA segment are referred to as 3' and 5', and you read codons by starting at the 5' end. Moreover, in the double helical formation in which DNA resides, the 2 strands have a reverse orientation in terms of this 5' to 3' directionality. Note that genes are located on both strands and can even overlap in the sense that there is a gene on both of the strands in the same location. We use the term *genome* to refer to the collection of DNA of some organism. The human genome consists of approximately 3 billion basepairs, and about 99.9% of these basepairs are common among humans. Thus, only about 3 million basepairs are responsible for genetic differences among people. Environmental sources (e.g. diet and climate) are also important factors that gives rise to differences observed among humans. Moreover, there can be interactions between genetic factors and environmental factors that impact the phenotype of an organism.

5 RNA

The process whereby DNA is used to make proteins is composed of several steps. Proteins are produced by ribosomes, which are organelles found outside the nucleus of the (eukaryotic) cell (prokaryotes also use ribosomes in this fashion), hence there needs to be a way of moving the information from the nucleus to the ribosome. Ribonucleic acid (RNA) is the molecule which allows for this communication. RNA is much like DNA, except the sugar group is ribose instead of 2-deoxyribose and uracil (U) takes the place of thymine

(T). These chemical differences between DNA and RNA imply that RNA doesn't have a strong tendency to form a double helix, unlike DNA. The first step in generating a protein from DNA is when premessenger RNA makes a copy of a DNA segment after the hydrogen bonds that hold the double helix of the double stranded DNA molecule are separated near a gene (the breaking of hydrogen bonds is regulated by proteins called *transcription factors*). After this copying (known as *transcription*), some editing of the segment takes place. Certain sequences of nucleotides indicate that a gene is about to start or end (start and stop codons), so the editing deletes the portions of DNA which are not part of a gene (this is why the alternative splicing terminology is used for some genes). In addition to this editing, a cap structure is attached on the 5' end of the sequence and a poly-A tail (i.e. a variable length sequence of adenine nucleotides) is added to the 3' end. These structures help stabilize the resulting molecule since they can prevent the degradation of the molecule by allowing certain proteins to bind to its 2 ends. This editing and attachment of structures to the ends of the RNA molecule yields messenger RNA (mRNA) which can then pass through the nuclear envelope and potentially associate with ribosomes outside of the nucleus. When an mRNA molecule interacts with a ribosome, transfer RNA (tRNA) helps coordinate amino acids so that a chain based on the sequence of nucleotides present in the mRNA is built (this process is known as *translation*). For each codon in the mRNA molecule, a particular amino acid is attached to the growing chain of amino acids at the ribosome (the collection of naturally occurring amino acids is presented in Table 1). In this way a chain of amino acids is built up by the ribosome forming a *peptide* (a sequence of amino acids). The process whereby DNA is used to produce a protein or active mRNA molecule is referred to as *gene expression*. The mRNA molecule is released when translation is complete, at which point it may associate with proteins that degrade the mRNA molecule, or it may interact with another ribosome and produce more protein. When there is high demand for a gene product, several ribosomes may associate with each other to produce protein from the same mRNA molecule, thereby generating what is called a polyribosome. While we can think of DNA as a linear polymer, RNA has a tendency to fold into more complicated structures, and these structures impact the functioning of the RNA molecule.

6 Proteins

Proteins are molecules composed mostly of amino acids and are created from RNA via a translation system that uses the *genetic code*. While the genetic code is not entirely universal, for example, it is slightly different for human mitochondrial DNA, it is nearly universal for nuclear DNA. The genetic code describes how codons get mapped to amino acids. Since there are 4 distinct nucleotides, there are $4^3 = 64$ possible codons. As noted above, in the process of creating a protein, each codon gets transcribed into one amino acid. Each of the 20 amino acids has a name (e.g. tyrosine), an abbreviated name (e.g. for tyrosine, it is Tyr), and a single letter name (e.g. for tyrosine this is Y). Since there are 64 possible codons and only 20 amino acids, there is some redundancy in the genetic code. This redundancy is not completely without a pattern; as seen in Table 2 it is least sensitive to the nucleotide in the third base and most sensitive to the nucleotide in the first. The start codon is AUG, so all proteins start with the amino acid methionine (whose symbol is M), but this M is sometimes cleaved once the protein has been assembled. When a stop codon is reached (represented by X in Table 2), translation stops, and the protein is released into the cytoplasm of the cell. At that point, proteins undergo changes in shape (conformational changes) and may incorporate other molecules so that the protein can serve its function. Some proteins change their shape depending on the environment, but all seem to undergo immediate changes once released from the ribosome. In addition, proteins can undergo post-translational modifications. These are chemical alterations wrought by other proteins on an existing protein. Unlike DNA (which we think of as a linear segment of bases), the actual shape of a protein has a

Table 1: The naturally occurring amino acids, with abbreviations and single letter designator.

Name	abbreviation	designator
Aspartic Acid	Asp	D
Glutamic Acid	Glu	E
Lysine	Lys	K
Arginine	Arg	R
Histidine	His	H
Asparagine	Asn	N
Glutamine	Gln	Q
Serine	Ser	S
Threonine	Thr	T
Tyrosine	Tyr	Y
Cysteine	Cys	C
Glycine	Gly	G
Alanine	Ala	A
Valine	Val	V
Leucine	Leu	L
Isoleucine	Ile	I
Proline	Pro	P
Methionine	Met	M
Phenylalanine	Phe	F
Tryptophan	Trp	W

Table 2: The genetic code. The protein associated with each codon is indicated by its single letter name and is presented to the right of the codon. An X indicates a stop codon and the start codon is AUG.

codon	protein	codon	protein	codon	protein	codon	protein
UUU	F	CUU	L	AUU	I	GUU	V
UUC	F	CUC	L	AUC	I	GUC	V
UUA	L	CUA	L	AUA	I	GUA	V
UUG	L	CUG	L	AUG	M	GUG	V
UCU	S	CCU	P	ACU	T	GCU	A
UCC	S	CCC	P	ACC	T	GCC	A
UCA	S	CCA	P	ACA	T	GCA	A
UCG	S	CCG	P	ACG	T	GCG	A
UAU	Y	CAU	H	AAU	N	GAU	D
UAC	Y	CAC	H	AAC	N	GAC	D
UAA	X	CAA	Q	AAA	K	GAA	E
UAG	X	CAG	Q	AAG	K	GAG	E
UGU	C	CGU	R	AGU	S	GGU	G
UGC	C	CGC	R	AGC	S	GGC	G
UGA	X	CGA	R	AGA	R	GGA	G
UGG	W	CGG	R	AGG	R	GUG	G

tremendous impact on the function of the protein, and the final shape in a particular environment is not easily deduced from the sequence of amino acids, although certain rules do seem to hold.

An amino acid is composed of an amino group (NH_2), a carboxyl group (COOH), a hydrogen atom, and a residue, and these bind at a common carbon atom (denoted C_α). The particular residue determines the particular amino acid. There are 4 groups of amino acids that are commonly distinguished: basic (K, R and H), acidic (D and E), uncharged polar (N, Q, S, T, Y and C) and nonpolar neutral (the remainder). Amino acids within a group tend to behave in the same fashion, for example, nonpolar neutral amino acids are hydrophobic, meaning they tend to avoid association with water molecules. Each amino acid has a unique set of chemical properties. For example, strong chemical bonds (called disulphide bonds, or disulphide bridges) tend to form between cysteine residues common to a single protein (or across proteins), and these bonds can impact the final shaped of the protein. A peptide chain grows by forming bonds (known as *peptide bonds*) between the amino group of one amino acid and the carboxyl group of the adjacent amino acid (giving off water). These peptide bonds form a backbone along which the residues reside. All of the atoms in the peptide bond (NH-CO) lie in the same plane, so the 3 dimensional structure of a peptide depends on only 2 angles per C_α . There are only certain combinations of these 2 angles which are allowed due to the residues interfering with one another in space (which are referred to as *steric constraints*), and these constraints give rise to the commonly encountered 3 dimensional structures in proteins. Glycine is largely free of steric constraints because its residue is so small (it is a single hydrogen atom). Two major structures built from the backbone of peptide chains (referred to as *secondary structures*) are commonly encountered; α -helixes (which are helixes of residues) and β -sheets (in which residues line up roughly co-linearly to produce a sheet of residues). There are also structures composed from these secondary structures which have been documented (referred to as *tertiary structures*). The 3 dimensional structure of a protein is important for its functioning since proteins are the working parts of cells, hence the structure of a protein is generally conserved in evolution. Since glycine is not as constrained in its angles, it is often in tight turns in the 3 dimensional structure, hence its location needs to be conserved in random mutations if the protein is to continue its function. Other residues are important for forming certain other 3-dimensional structures, hence they need to be preferentially conserved in mutation too.

Many proteins have an effect on the cell by binding to other molecules, such as DNA or other proteins, and the ability of a protein to bind to another molecule is greatly influenced by the 3 dimensional shape of the protein. The object to which a protein binds is called its *ligand*. After a protein binds to another molecule, the other molecule (and the protein itself) may undergo changes. We often refer to this process as one protein activating the other since frequently proteins exist freely in the cell until encountering an enzyme which promotes a chemical reaction by lowering the level of energy needed for the reaction to occur.

Proteins are the primary chemicals that are involved in the biochemical reactions described in Section 2. For a multicellular organism to survive, the component cells need to coordinate their activities. Such coordination is achieved by intracellular signaling. A cell will respond to certain stimuli (such as heat or lack of nutrients) by producing signaling molecules and releasing these molecules through its cell membrane. This is achieved by a sequence of chemical reactions that involves converting certain segments of DNA into RNA which then allows certain protein products to be produced. These proteins then allow the signaling molecules to bud off the surface of the cell. Animal cells are covered with molecules that selectively bind to specific signaling molecules. If the ligand of a signaling molecule is encountered on the surface of another cell, the signaling molecule will bind to that cell and the cell to which it binds will then undergo a set of changes. Other signaling molecules are able to penetrate the cell membrane. These changes are brought about by the physical binding of the signaling molecule to the surface of the cell as such physical binding can lead the cell to alter the set of proteins that the cell is producing. In this way, changes in gene expression in one cell can lead to changes in the gene expression of many cells. Moreover, as gene expression changes in the second set

of cells, they alter the set of proteins they secrete which can lead to other outcomes further down stream in a connected sequence of events. In a given cell, many pathways will operate simultaneously (some involving the same genes) and the outcome of all these forces determines what the cell does. In this way, a cell acts as a certain cell type in response to the genes that it is expressing. Hence it is the action and use of genes via the process of gene expression that determines the phenotype of a cell, and the collective action of cells, which is also coordinated by gene expression, determines the phenotype of an organism.

Proteins are frequently organized into families of proteins based on function. While a variety of methods to construct such families have been proposed, there are typically thousands of families and frequently there are subfamilies distinguished with these families. Databases have been constructed to allow researchers to ascertain the family of a given protein (for example the PANTHER data base, Mi H, Lazareva-Ulitsky B, Loo R et al. (2006)). For example, Eukaryotic protein kinases is one of the largest families of proteins. All proteins in this family have nearly the same sequence, structure and function (they are used by cells to communicate with each other). Protein families typically extend across organisms and frequently organisms will have multiple versions of proteins from a family. Another commonly studied gene family is the major histocompatibility complex. Three distinct classes of this protein family are commonly distinguished, and they are called class 1, 2, and 3. This family of genes is central to immune control. Humans have 3 proteins that belong to MHC class 1, and they are called the human leukocyte antigens, which, like most proteins and genes, is usually referred to by its abbreviated name HLA (the 3 different types in humans are called HLA-A, HLA-B and HLA-C). These genes produce receptors that cover cell surfaces and allow cells from the immune system to determine if other molecules are foreign or produced by itself.

A major initiative aimed at classifying proteins (and thereby the genes that produce these proteins) is the *Gene Ontology Consortium*. This is a group of researchers attempting to develop a common set of procedures for classifying proteins, and the classification system is referred to as the gene ontology (or just GO). A distinctive feature of GO is that it classifies genes by 3 separate criteria: biological process, cellular component and molecular function. These ontologies attempt to place genes into a directed acyclical graph so that as one moves further down the graph structure there are fewer genes in any category and the process, component or function is more precise, i.e. the classifications get more coarse as one moves up the graphical structure. A gene in one group at a fine level of separation can point to multiple coarser classifications since a gene may play a different role in a different cellular context. Proteins that are involved in the same cellular process are typically referred to as being in a common pathway. Note that there is not a simple relationship between gene families and pathway membership as genes from the same family may or may not be in a common pathway.

As an example of a well understood pathway, in the growth factor-regulated MAP kinase pathway (Robinson and Cobb, 1977), once the signaling molecule (a growth factor) binds to the cell surface receptor for a tyrosine kinase (an enzyme used for intracellular communication and a member of the Eukaryotic protein kinase family), the receptor and other proteins nearby undergo post-translational modifications that lead to certain other proteins coming to the location where the receptor is located. These other proteins include Ras (an important protein in the development of cancer) which then becomes activated. Once activated, Ras can recruit the protein Raf to its location which then activates the protein MEK, which in turn activates MAP kinase. After MAP kinase is activated, it can activate transcription factors which then alter the pattern of gene expression (activated MEK also activates these other transcription factors). For this reason, researchers often think of genes as working in groups, or as part of a biological pathway. In actuality, many receptors have their ligands bind to a cell, and the response of the cell to the collection of these signals depends on the current levels of signals and the levels of all proteins in the cell. For this reason, it makes sense to frame many questions about the operation of proteins in cells in terms of the interactions between pathways of genes.

7 Some basic laboratory techniques

There are a number of fairly standard techniques that can be used to measure the level of a DNA, RNA or protein molecule from a collection of cells. A basic distinction in these methods is between cloning and hybridization. In DNA cloning, a particular DNA fragment is amplified from a sample to produce many copies of the DNA fragment. After this cloning, the amplified products are separated from the rest and properties of the fragment can be studied. The most commonly used method for DNA cloning is called polymerase chain reaction (PCR). This technique can also be used to measure the quantity of RNA by first using the RNA to make cDNA (i.e. a sequence of DNA that is the complement of the DNA sequence used to generate the RNA molecule, called complementary DNA), then using PCR to amplify this cDNA (this is called reverse transcriptase PCR because the enzyme reverse transcriptase is used to generate the cDNA from the RNA). In contrast, with molecular hybridization, a particular DNA or RNA fragment is detected by another molecule. An example of the latter technique, referred to as a southern blot hybridization, can be used to determine aspects of a sequence of DNA at specific locations. In particular this method can be used to determine what alleles an individual has at a particular locus. Northern blots are a variation of southern blots that allow one to measure the quantity of RNA in a collection of cells. Finally, a slightly different technique, called a western blot, can be used to quantify the level of a protein in a particular tissue sample. All of these techniques require one to specify the exact molecule (i.e. DNA fragment or protein) one wants to measure and typically only a few types of molecules will be quantified in a study that uses these techniques. In the final chapters we will discuss microarray technology, a molecular hybridization technique, which allows one to measure the level of mRNA for thousands of genes simultaneously. Consideration of thousands of genes instead of just a handful has changed the nature of investigations in molecular biology to the extent that researchers refer to considerations of just a few genes as genetics, whereas consideration of thousands of genes is referred to as *genomics*. Nonetheless, these techniques can be used to investigate if a gene or protein is expressed at different levels in different patient populations or in different tissue types. Since we generally think of DNA as being the same in every cell in the body (which isn't quite true since there are mutations that occur during cell division, and some of these are extremely important since such mutations may induce cancer), usually DNA is isolated from white blood cells (i.e. cells that function as part of the immune system and are easily found in a blood sample) and in most studies involving DNA, samples of white blood cells are stored in special freezers so one can further investigate the DNA in the future. While red blood cells may appear to be a more natural choice (since they are the most common cell in the blood) they don't have a nucleus since they are terminally differentiated cells (i.e. mature red blood cells don't divide).

There are also a number of organisms that are frequently studied in genetics. The use of another animal to understand human genetic processes constitutes a model for that genetic process. Studying these organisms can lead to insights not only regarding genetic mechanisms, but also provides a way to develop technology in a more cost effective way than using human samples. The prokaryote *E. coli* is commonly used to understand some basic genetic mechanisms. Even though humans are distantly related to this bacterium, a number of genetic mechanisms are common. A simple eukaryote that has received a great deal of attention is *Saccharomyces cerevisiae*, a single celled yeast. Surprisingly, a large fraction of yeast genes have a similar gene found in mammals. Other extensively studied organisms include *C. elegans* (a 1 mm long roundworm), and *Brachydanio rerio* (the zebrafish) for insights regarding development and *Drosophila melanogaster* (the fruit fly). All these organisms have a quick developmental cycles and are easy to genetically manipulate. Finally, *mus musculus* (i.e. the mouse) has been used in many investigations since mice have short life spans and almost every gene has a corresponding human gene (in fact large chromosomal regions are common between humans and mice). The entire genome is available for all of these important models for human

genetics.

Exercises

1. If 98% of human DNA is not used for producing proteins (i.e. non-coding DNA), what is the mean length of a gene in terms of bases? What is the mean length of a protein in terms of amino acids?
2. In humans, what is the mean number of gene alleles on a chromosome pair if all chromosomes have the same number of genes?
3. If a gene has 4 exons each with 200 codons, what is the length of its associated mRNA molecule? How many introns does it have?
4. Suppose nucleotides arise in a DNA sequence independently of each other and all are equally likely. How often does one expect to encounter a start codon just by chance? How often will one encounter a stop codon by chance? What is the mean length of the genes one would detect in such random DNA? How many would you detect in the a sequence the length of the human genome?
5. Translate the following mRNA sequence into a sequence of amino acids:
AUGUUGUGCGGUAGCUGUUAU.
6. Show that 2 unique proteins are encoded by the following DNA sequence:
UUAGAUGUGUGGUCGAGUCAUACUGACUUGA.

Note that the gene doesn't necessarily start at the first letter in this sequence, the orientation of the sequence is unknown (i.e. we don't know which end is the 3' or 5' end) and a gene may be found on the complementary sequence. Also, give the protein sequences.