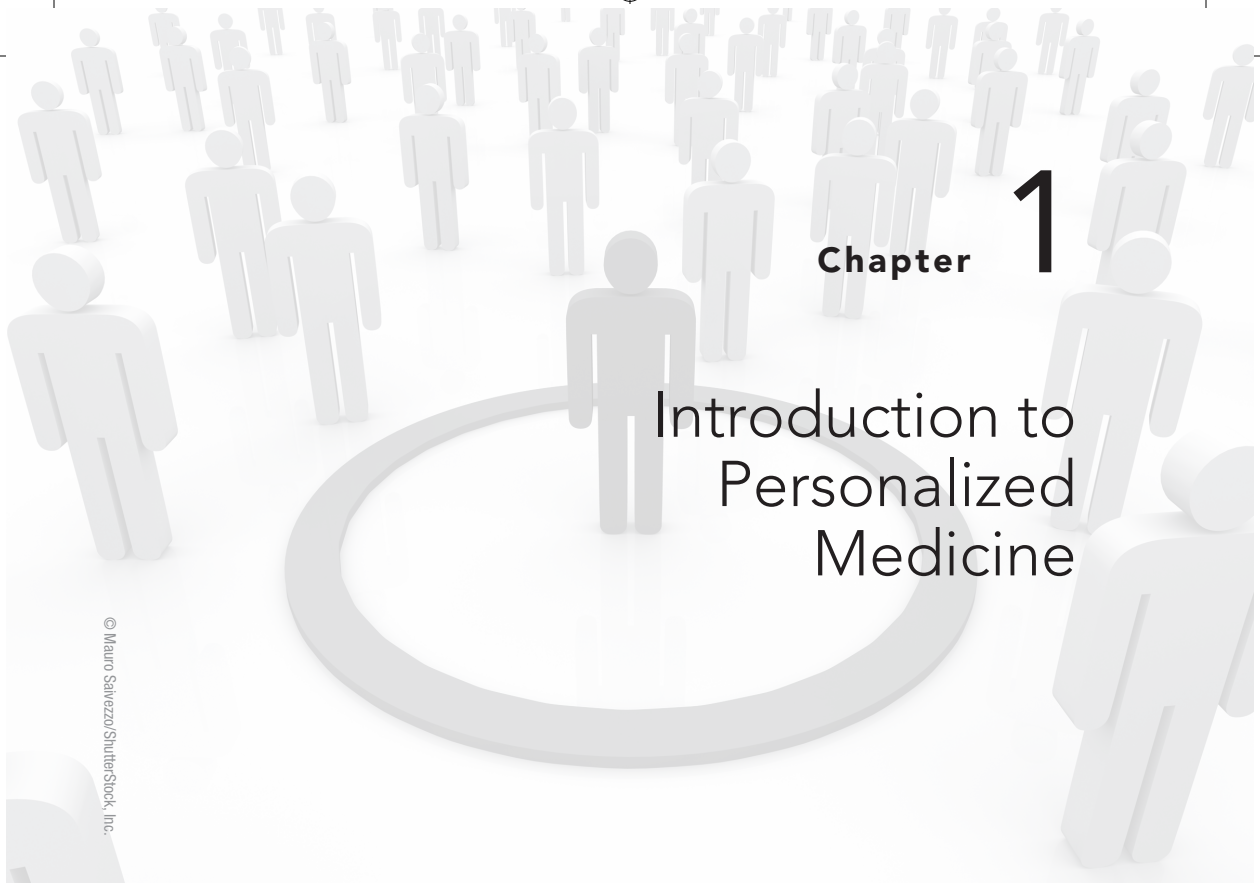


Chapter 1

Introduction to Personalized Medicine

Learning Objectives

Upon completion of this chapter, the student will be able to:

1. Recognize basic mechanisms of the expression of genetic information as traits—from the deoxyribonucleic acid (DNA) sequence to transcribed ribonucleic acid (RNA), to translated proteins, to phenotype.
2. Differentiate among the major types of genetic variation, including non-synonymous, synonymous, nonsense single nucleotide polymorphisms (SNPs), and insertion/deletion (indel) polymorphisms in both genotypic and phenotypic terms.
3. Provide specific examples that establish the relationship between altered drug disposition and polymorphic cytochrome P450 enzymes.
4. Describe how polymorphic genetic variation can be utilized to predict individualized responses to drug therapy.

Key Terms	Definition
allele	Alternate sequences or versions of the same gene inherited from each parent.
biomarker (genomic)	A measurable DNA and/or RNA characteristic that is an indicator of normal biologic processes, pathogenic processes, and/or response to therapeutic or other interventions.
codon	Three adjacent nucleotide bases that ultimately encodes a specific amino acid.
exon	A nucleotide sequence that codes information for protein synthesis.
gene	Regions of the genome (DNA) that contain the instructions to make proteins.
genome	The entire DNA of an organism.
genotype	The specific set of alleles inherited at a locus on a given gene.
haplotype	A series of polymorphisms that are inherited together.
heterozygous	Possessing two different alleles for the same trait.
histone	A protein around which DNA coils to form chromatin, thus "packaging" DNA.
homozygous	Possessing identical alleles for the same trait.
indel	Insertion or deletion of DNA either as single nucleotides or spanning regions of DNA involving many nucleotides.
intron	A nucleotide sequence in DNA that does not code information for protein synthesis and is removed before translation of messenger RNA.
monogenic trait	Characteristics derived from a single gene.
multigenic trait	Characteristics derived from multiple genes.
mutation	A change in DNA sequence between individuals.
nucleotide	One of the structural components, or building blocks, of DNA, including adenine (A), cytosine (C), guanine (G), and thymine (T), and of RNA, including adenine (A), cytosine (C), guanine (G), and uracil (U).
personalized medicine	The use of patient-specific information and biomarkers to make more informed choices regarding the optimal therapeutic treatment regimen for a given patient.
pharmacodynamics (PD)	The relationship between drug exposure and pharmacologic response.
pharmacogenetics (PGt)	The study of a gene involved in response to a drug.
pharmacogenomics (PGx)	The study of many genes, in some cases the entire genome, involved in response to a drug.
pharmacokinetics (PK)	The relationship of time and drug absorption, distribution, metabolism, and excretion.
phenotype	An individual's expression of a physical trait or physiologic function due to genetic makeup and environmental and other factors.
polymorphism	A mutation in DNA in a given population that may be observed at greater than 1% frequency.
reference sequence number (refSNP, rs#, rs)	A unique and consistent identifier of a given single nucleotide polymorphism (SNP).
single nucleotide polymorphism (SNP)	A variant DNA sequence in which a single nucleotide has been replaced by another base.
topoisomerase	A class of enzymes that alter the supercoiling of double-stranded DNA.
wild-type	The typical or normally occurring genotype of an organism.
xenobiotics	Substances (often drugs) introduced into the body but not produced by it.

Introduction

In its simplest terms, **personalized medicine** is the use of patient-specific information and **biomarkers** to make more informed choices regarding the optimal therapeutic treatment regimen for that patient, rather than reliance on population-based therapeutic trends. **Pharmacogenetics (PGt)** is the aspect of personalized medicine whereby patient-specific genomic biomarkers are used to choose the optimal drug and/or dose for the patient, with the goal of assuring drug efficacy in the patient while minimizing or avoiding the risk of an adverse drug reaction. The successful implementation of pharmacogenetics in the clinic is dependent upon a number of different processes and data, including *a priori* knowledge about a specific **allele** in the genome and its linkage to altered **pharmacokinetics (PK)** and/or **pharmacodynamics (PD)** (compared to the statistical norm in the population), the ability to accurately test a patient for the presence of a specific allele in his or her genome, and the ability to offer the patient more effective alternatives than would be typically offered to a patient in the statistical norm of the population. Key to this process is the utilization of prior discoveries and clinical findings (e.g., data) regarding a specific genomic allele relevant to the pharmacokinetics and/or pharmacodynamics of the prescribed or intended drug, and then *predicting* how the patient will respond to the drug. Finally, the utilization of pharmacogenetics and personalized medicine must add value to healthcare. In other words, the costs associated with the implementation and utilization of pharmacogenetics and personalized medicine must be ethically and economically justified by reducing the negative (adverse) effects and costs associated with adverse drug reactions as well as the assurance of more effective drug therapy outcomes for the healthcare consumer population.

Here, we distinguish the term *pharmacogenetics* from the now more commonly used **pharmacogenomics (PGx)**. In its purest sense, pharmacogenetics refers to the study of a gene involved in response to a drug, whereas pharmacogenomics (PGx) refers to the study of all genes in the genome involved in response to a drug.[1] However, the vernacular that has emerged in recent years often uses the term *pharmacogenomics* to reference the entirety of the science and the methods that study the interface of genomics, genetics, and drugs used in clinical therapeutics.

Living Systems and the Genome

The adult human body contains trillions of different cells, each performing different functions to sustain life. Some of these are muscle cells, some make up our skin, some are blood cells, some form bone,

some are brain or liver cells, and so on. Each of these cells has developed within a specific tissue in the body to perform a specific function. For example, our red blood cells are capable of transporting molecular oxygen from our lungs to organs and tissues, and then transporting carbon dioxide back to the lungs to be removed from the body. This unique cellular capability is due to the presence of a specific protein in the red blood cell called hemoglobin. More specifically, what we commonly refer to as hemoglobin is actually a multimolecular structure that contains a heterocyclic organic molecule called heme, which is bound to an atom of iron, as well as two specific globular proteins. These globular proteins are alpha-globulin and beta-globulin, which are each derived from a specific gene in our genome. Thus, the genes are used by the cellular machinery as a "blueprint" or "instruction set" on how to make these proteins. Hence, **genes** are the regions of our genome that contain the instructions to make proteins, and proteins are the functional components of living systems. In simpler terms, if the genes are the "blueprints" for life, then proteins are the "bricks and mortar" of living systems. Proteins are the inherited, functional components of living organisms—inherited because they are derived from our genome, which we inherit from our parents and ancestors. Interestingly, less than 2% of our genome is actually used as template genes to make proteins.[2] We will discuss aspects of proteins later; let us first take a closer look at our genome.

The **genome** of an organism is the instruction set for that organism, or, more specifically, the instruction set for the development of its cells and tissues, as well as the maintenance of these cells and tissues throughout the life of the organism. The functional molecule that makes up our genome is DNA (deoxyribonucleic acid). Our genome is made up of four different DNA **nucleotide** bases (adenine-A, cytosine-C, guanine-G, and thymine-T), which are somewhat equivalent to a written language with four different letters. The human genome contains about three billion nucleotide bases (or letters; A, C, G, T) in the genome, and essentially each cell (that contains a nucleus) contains a copy of the entire genome. To appreciate the size of the human genome, consider that, if our genome was printed in paperback novel form, it would contain over four million pages. By no means is the human genome considered large within the spectrum of living organisms on earth. The onion (yes, the one you eat) has a genome that is more than six times the size of the human genome, and it has been estimated that certain lilies (flowering plants) have a genome that is 30 times bigger than our own.[3,4] We will not be discussing the complexities of plant genomes in this textbook, as scientists are only beginning to understand the different complex genomes in living organisms.

If we take the perspective of our genome representing information, then we must recognize that each cell has the instructions

(genes) for all the proteins that the organism can ever make, even though the cell may not use all this information. In other words, each cell (that has a nucleus) has all of the chromosomes of the genome, and therefore all the genes that we have inherited from our parents, yet each cell only uses a subset of these genes to make the proteins it needs to thrive and carry out its various functions. To conceptualize this phenomenon, imagine that each person in your college or organization was a cell, and each person has a computer that contains all the programs needed to make the entire organization run successfully. A person working in the accounting office would use the computer programs (i.e., genes) that are used to manage the organization's resources and inventory but would not use the programs used for personnel management (even though those programs are stored on the computer). Similarly, a muscle cell uses the genes to make the proteins used for mechanical contraction, but not the genes that make the proteins for detecting light that are used in the retinal cells of the eye.

Genetic Evolution and the Evolution of Genetics

Sexual reproduction is the fundamental process for enabling genetic diversity and the propagation of life on earth. It involves the passing of genetic information from viable parent organisms to their offspring. Thus, the offspring inherit the genetic information that allowed their parent organisms to thrive and survive in the environment. Furthermore, it allows two different successful organisms (the biological mother and father) with different genomic content to create variations of their respective genomes in their offspring, thereby creating genetically varied offspring. The life of these offspring represents a test of the content of their genome, and reaching sexual maturity and successfully reproducing reinforces the rigor of their genomic content (i.e., the viability of the organism was sufficient to endure its environment), which is passed on to subsequent generations. This is the basis of natural selection, or "survival of the fittest," from the perspective of the inherited genome. The genetic variability among the population (of a given species) appears extremely important for the evolutionary success of the species because it allows the species to adapt to changes in its environment over generations by reinforcing the traits that confer viability. In other words, as changes in the environment emerge and exert selective pressure on the species, members of the population that harbor the genetic content (that encode physical or behavioral traits) to overcome these environmental changes survive and shape the genetic content and physical traits of subsequent generations. This process is a fundamental tenet of evolution on earth.

We are all familiar with the genetic diversity and variation in the human population, as evidenced by obvious physical traits, such as eye color, hair color, and so on. These physical traits are the result of inherited genes from our biological parents, and dictate aspects of our physical appearance. This genetic variation extends into many aspects of our genome and cell biology that are not always as obvious, such as those that affect behavior or aspects of cellular biology. These variations in the human population are very important to the perpetuation of our species. As our regional and planetary environments change over long periods of time, those individuals that are best suited to survive in the changing environment will thrive and continue to bear offspring. The makeup of our genome, and therefore our physiology, is the result of millions of years of evolution under the selective pressures of our environment as well as competition for survival. In other words, the physical and behavioral traits that provided our ancestors with a competitive advantage allowed those individuals to thrive and bear offspring, whereas individuals who lacked a given competitive advantage were much less likely to thrive and bear offspring. Therefore, the individuals that harbored advantageous traits passed their genes onto their offspring, and after thousands of generations of human evolution the content of our genome is the result of this evolutionary process. Thus, our genome contains the genes that conferred the beneficial traits needed for our ancestors to thrive.

These traits may be very subtle yet important, such as the ability to digest lactose into adulthood and therefore derive sustenance from the milk of the beasts of burden that our ancestors domesticated over the last 10,000 years. Our ancestors who experienced this evolutionary adaptation could better survive periods of famine and drought, and this trait was retained in our recent evolution. It should become obvious that the changes in our genetic makeup that resulted in specific competitive advantages would be passed down through generations, whereas changes in our genome that did not serve a beneficial purpose, or in many cases even reduced the viability of an individual, are not seen in the modern human genome. Hence, we are the modern beneficiaries of this genetic "arms race" of inherited traits that has been ongoing for millions of years.

When we consider the past influences that have shaped our biology and the content of our genome, we can begin to understand why exposure to certain chemicals and compounds that we may ingest pose a threat to our survival whereas other substances are safe. For example, certain mushrooms (e.g., the death cap mushroom) and frogs (e.g., the poison dart frog of South America) synthesize compounds that are toxic to organisms that would otherwise consume them as a nutrient source. The death cap mushroom synthesizes a compound called amanitin, and the poisonous dart frog synthesizes epibatidine

(among other alkaloids). It is well known that these compounds are highly toxic to humans and many other organisms.[5]

No evidence suggests that there are variations in the sensitivity to these poisons among humans, and thus we all avoid consuming these mushrooms and frogs, an adaptation that benefits the mushroom and the frog. Because these organisms, and their poisons, have existed in nature for millions of years alongside our ancestors, it is not necessary to consider variations in the toxicity of the poisonous compounds across the modern human population. Building upon our evolutionary theory, we can imagine a fictionalized paradigm of selective pressure 50,000 years ago where these mushrooms were abundant and only a subpopulation of humans harbored the ability to detoxify the poison in the mushroom and therefore safely consume the mushroom as a nutrient source. If this were the case and the mushrooms were an abundant source of nutrients, the humans that could safely consume the mushrooms would thrive, whereas the humans that were sensitive to the toxin would be less likely to thrive. In this fictitious example, it is likely that all humans living today would harbor the ability to detoxify and safely consume the poison simply due to selective pressure on our ancestors. In genomic terms, modern humans would harbor a gene in their genome that encoded an enzyme capable of breaking down the poison. Although this is not true for the poisons mentioned in this example, it is true for many other substances in nature that our ancestors encountered.

The example above is presented to demonstrate a fundamental difference between naturally occurring substances and modern pharmaceutical products. When we look at modern pharmaceutical compounds, we see a much more varied response among humans to both the safety and efficacy of these compounds, even though the process of drug development attempts to provide drugs that are efficacious and safe for the entire population. One reason for the varied responses to drugs is that these pharmaceutical compounds did *not* exist in nature and were not available for consumption by our ancestors. Therefore, no evolutionary selective pressures have been experienced in humans with respect to exposure to these pharmaceutical compounds, and the outcome of evolutionary selective pressure has not been manifested in the genome through thousands of generations. Thus, we expect a much more varied response in the population to these modern medicinal chemicals, compared to naturally occurring substances.

The development of safe and effective medicinal compounds is a challenge because there can be a spectrum of responses in the population regarding the safety and efficacy of a drug, and this can complicate the management of pharmacy and therapeutics in our modern healthcare system. In other words, due to the methods used to assess and approve new drug entities, modern drug approval requires that it be safe

and effective in a large majority of the population. Thus, drugs under development that have shown large variation as to their safety and/or efficacy have not gained marketing approval. It should be obvious that if the genetic basis for variations within the population to the safety and/or efficacy of a drug are studied and understood, then a drug that is effective in a known subpopulation could be approved, if that subpopulation can be identified through genomic testing (i.e., pharmacogenetics or pharmacogenomics). In fact, this movement in pharmacotherapy will result in safer and more effective drug use within subpopulations in our society and enable healthcare professionals to use genomic screening to predict how a patient will respond to a specific drug and therefore inform healthcare professionals as to which drug and/or dose is optimal for the patient. This is a principal tenet for the adoption of personalized medicine.

We have discussed how selective pressure and evolution have shaped the content of our genome, now let us look at our genome from a completely different perspective: How have advances in modern healthcare and disease management or, more specifically, extending the length of human life, inadvertently revealed (or invented) a new type of genetic predisposition to disease?

At some time in our recent history, pre-modern humans lived together in groups composed primarily of three generations (i.e., children, parents, and grandparents). In a simple version of anthropological theory, the parental generation (in the physical prime of their life) worked to search for resources, gather food, and defend the group, while the elders helped oversee the young of the group. It is important to note that larger groups of individuals consume, and therefore require, more resources (e.g., food, water) than smaller groups. Thus, larger groups of early humans were at a disadvantage in times of limited resources (e.g., drought, famine), compared to smaller groups. Therefore, it was not beneficial for pre-modern humans to have a long lifespan, as this would result in large groups that were at a competitive disadvantage; for this reason human life expectancy was much shorter than it is today.

Australopithecines appear to have had an average life expectancy of only 15 to 20 years and survived for about 300,000 generations, ending about two million years ago. More recently, early agriculturalists and nomadic pre-modern humans had an average life span of about 25 years and survived only about 500 generations. These ancestral life spans suggest that age-related declines in function after the age of 25 were due to the forces of natural selection. In modern times, the last 200 years of human history (about 10 generations), the average life span has increased from 43 to 75 years of age.[6]

If we consider that the vast majority of our ancestors only had life expectancies of younger than 45 years, then certainly their genomic content and function was to maintain optimal health until this age, with no evolutionary advantage to extending life span. The increase in life expectancy between pre-modern and post-modern humans is nearly instantaneous as compared to the much longer timelines associated with pre-modern human evolution. For the purposes of illuminating this perspective, it can be assumed that our genome is essentially identical to our pre-modern ancestors' of 5,000 years ago.

In other words, the modern human genome has evolved to support an individual's life until they reach about 45 years old (this may even be a generous estimate), even if living in a modern society. Or, more accurately, any health problems that have a basis in genetics would not have been passed down from our ancestors if the health problem manifested itself early in life (i.e., before about 35 years). However, if the genetic-based health problem manifests itself after the age of 45 years, it would not have exerted selective pressure against individuals that harbored this genetic allele. Therefore, it would not have negatively impacted the survival of our pre-modern ancestors, and it would be expected to be present in our genome today. In other words, extending human life beyond the age of 50 years "reveals" new diseases in the human population, and the management of these age-related disorders becomes more dependent on modern healthcare methods, practices, and technology as we age. From this perspective, if it is determined that an individual has a genetic predisposition for a disease or disorder with an expected onset at 60 years of age, it is not a failure of human evolutionary processes but simply an artifact of extending human life.

Many examples of age-related disorders with genetic underpinnings can be found in humans, and it is certain that more discoveries will be made linking specific genetic markers with age-related disorders. An example of an age-related disorder with a known genetic link is Huntington's disease. Huntington's disease involves an inherited genetic defect where an expansion of a three nucleotide repeat (CAG) in the protein-coding region of a specific gene (named Huntington) causes the protein to self-aggregate. The deleterious effects (symptoms) of this genetic defect are usually first manifested at about 40 years of age. Thus, there were no selective pressures to eliminate this genetic defect from the population in pre-modern humans because it was not a genetic defect until our life expectancy increased beyond 40 years. This can be said for essentially all genetically linked diseases in humans older than 40 years of age.

As we consider age-related diseases, the ability to utilize genomic screening in the clinic is very important in identifying people who

are predisposed to a specific age-dependent disorder. Ideally, utilizing genomic screening in this paradigm allows the patient ample time to take measures to reduce or eliminate the risk of the disorder, such as changes in diet, exercise, prophylactic medicines, and so on. In this case, it is important to note that although there is currently no "cure" for Huntington's disease, the disease can be diagnosed using genetic screening methods prior to the appearance of any disease symptoms. Thus, the use of genetic screening methods for disease risk should be carried out with adequate genetic counseling because (1) the results of genetic screening must be interpreted correctly and (2) there can be significant psychological ramifications associated with the results of genetic screening for the patient and family.

Genome Structure and Gene Regulation

Less than 2% of the human genome is made up of gene sequences that encode proteins, and these genes are distributed throughout the 23 chromosomes and mitochondrial DNA of the human genome. The remaining 98% of the genome exists between gene sequences (i.e., intergenic DNA sequence) and contains many important regions that are key elements to DNA replication and DNA regulatory machinery. For example, polymorphic variations in intergenic DNA sequence may influence DNA tertiary structure directly or alter binding sites of DNA regulatory machinery, including **histones** and **topoisomerases**, which exert profound influence on overall gene regulation, cellular signaling, and homeostatic responses to environmental stresses. Indeed, recent studies implicate mechanisms of DNA–histone binding in the potential underlying pathophysiology of mood disorders and drug addiction while pointing to potential therapeutic targets for novel antidepressant and antipsychotic therapies.[7] One insight that is gained when considering the size of plant and animal genomes, and the relatively small fraction of these genomes that actually encode proteins (i.e., genes), is that the retention of large noncoding regions in the genome over millions of years of generations does not appear to consume excessive cellular resources that place the organism at a disadvantage to survival and/or there is an evolutionary advantage to retaining these large noncoding regions, even those regions that do not appear to be critical for DNA replication. Note that the relevance of intergenic DNA sequence to pharmacogenetics is emerging. One potential example of these effects is that of the O^6-methylguanine DNA methyltransferase (*MGMT*) enzyme that repairs DNA damage induced by alkylating chemotherapeutic drugs such as temozolomide. Evidence suggests that hypermethylation of DNA regions upstream of MGMT suppress its expression in some types of B lymphoma cells,

causing increased susceptibility to the cytotoxic effects of anticancer medications used in treatment.[8] However, more research and linkage studies must be carried out to fully understand how specific allelic variations in these regions will be utilized to alter drug dose and/or drug choice in clinical practice.

The human genome is made up of approximately 25,000 distinct genes, each capable of coding a unique protein, and it is these proteins that enable our cells to carry out the many different molecular, enzymatic, and mechanical processes that enable life. Because most drugs interact with proteins, pharmacogenetics deals primarily with genetic variations that affect gene regulation (i.e., DNA sequence variations that alter how much of each protein is being synthesized in the cell) and protein function or activity (i.e., DNA sequence variations in the gene that alter the amino acid sequence of the protein). Pharmacogenetics involves an understanding of how individual genetic differences in a population are the cause of variable responses to a specific dose of a drug in a population. In order to effectively examine the interactions between pharmacokinetics, pharmacodynamics, and genetics, we must first understand how genes are regulated in the cell and how the gene sequence (coding sequence) defines the primary sequence of a protein.

The simplest description of a gene's structure can be divided into (1) a regulatory region, where the cellular machinery exerts its effect on if, and how much, the gene will be "activated" or used by the cell and (2) the coding region, where the DNA sequence directly correlates with the protein sequence (see **Figure 1-1**).

The regulatory region contains specific DNA sequences and motifs where transcription factors and other regulatory elements bind, thereby promoting or preventing the transcription of the gene. During gene transcription, ribonucleic acid (RNA) polymerase binds within the regulatory region and then moves along the coding sequence to create a direct copy of the gene sequence. This RNA copy will undergo further processing before leaving the nucleus of the cell, ultimately coupling with the ribosome to synthesize the protein from the gene. In **Figure 1-2**, the details of eukaryotic transcription are described.

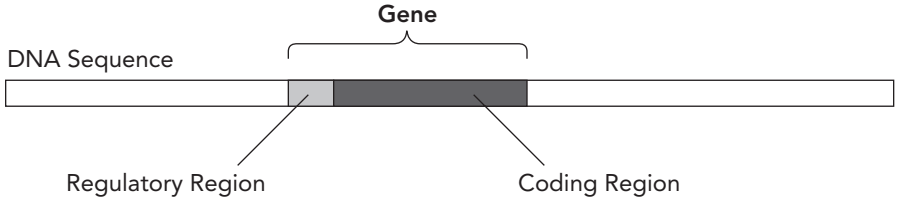


Figure 1-1 Simplified gene structure. The structure of a gene can be divided into the regulatory region, which is responsive to cellular machinery controlling its expression, and the coding region, where the DNA sequence directly correlates with the protein sequence.

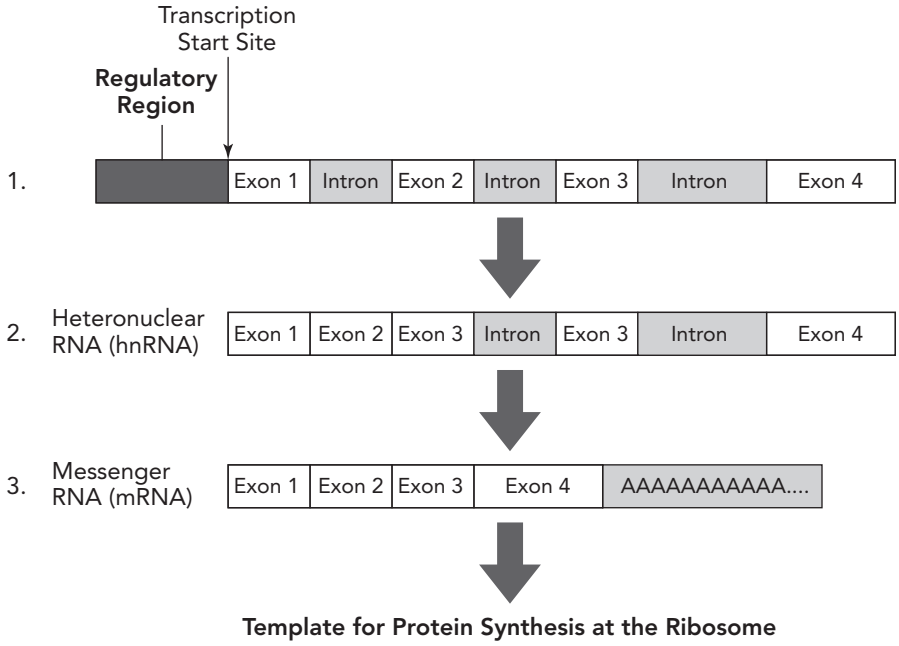


Figure 1-2 Eukaryotic transcription and translation. Genetic information derived from the DNA sequence is converted to functional proteins when: beginning at the transcription start site, the double-stranded DNA sequence is directly copied into a single-stranded heteronuclear RNA (hnRNA); messenger RNA (mRNA) is formed when introns are spliced or removed from the hnRNA and a poly-A tail is added; and mRNA is translated into a protein via protein synthesis at the ribosome.

As shown in the figure, the double-stranded DNA sequence (1) is directly copied into a single-stranded RNA sequence known as heteronuclear *RNA* (hnRNA), beginning at the transcription start site. Heteronuclear RNA is then (2) processed by removing the **intron** regions, a process termed splicing, and (3) a poly-A tail is added, resulting in messenger RNA (mRNA), which moves from the nucleus to the ribosome for protein synthesis.

Note that the removal of the intron sequence in the hnRNA results in the concatenation of the **exon** sequences in the mRNA, which represents the coding sequence for the protein. At the ribosome, the genetic coding sequence (nucleic acids) is converted to the protein sequence (amino acids). Each and every amino acid in a protein is coded by three nucleic acids, called a **codon** (see **Figure 1-3** for a codon key). For example, the nucleic acid codon "AUG" encodes for the amino acid methionine in a protein sequence. Note that the thymine (T) in DNA is replaced by uracil (U) in RNA. In addition to the codons that encode specific amino acids, three codons (UAA, UAG,

First Position	Second Position: U	Second Position: C	Second Position: A	Second Position: G	Third Position
U	UUU – Phe UUC – Phe UUA – Leu UUG – Leu	UCU – Ser UCC – Ser UCA – Ser UCG – Ser	UAU – Tyr UAC – Tyr UAA – Stop UAG – Stop	UGU – Cys UGC – Cys UGA – Stop UGG – Trp	U C A G
C	CUU – Leu CUC – Leu CUA – Leu CUG – Leu	CCU – Pro CCC – Pro CCA – Pro CCG – Pro	CAU – His CAC – His CAA – Gln CAG – Gln	CGU – Arg CGC – Arg CGA – Arg CGG – Arg	U C A G
A	AUU – Ile AUC – Ile AUA – Ile AUG – Met	ACU – Thr ACC – Thr ACA – Thr ACG – Thr	AAU – Asn AAC – Asn AAA – Lys AAG – Lys	AGU – Ser AGC – Ser AGA – Arg AGG – Arg	U C A G
G	GUU – Val GUC – Val GUA – Val GUG – Val	GCU – Ala GCC – Ala GCA – Ala GCG – Ala	GAU – Asp GAC – Asp GAA – Glu GAG – Glu	GGU – Gly GGU – Gly GGA – Gly GGG – Gly	U C A G

Ala	Alanine	Gly	Glysine	Pro	Proline
Arg	Arginine	His	Histidine	Ser	Serine
Asn	Asparagine	Ile	Isoleucine	Thr	Threonine
Asp	Aspartic Acid	Leu	Leucine	Trp	Tryptophan
Cys	Cysteine	Lys	Lysine	Tyr	Tyrosine
Gln	Glutamine	Met	Methionine	Val	Valine
Glu	Glutamic Acid	Phe	Phenylalanine		

Figure 1-3 The genetic code. In the expression of genetic information, the codon key describes the code for each amino acid in a protein based on three nucleic acids, termed a "codon."

and UGA) encode a "stop" command, thereby stopping the growth of the protein at that point in the sequence.

Cell Biology and the Human Genome

In living systems, the cell is the basic unit of life. Each cell that contains a nucleus contains the entire genome of the organism, although

the cell only utilizes a subset of genes to enable its viability and function within the organism. Within the nucleus of human cells are 23 pairs of chromosomes (46 chromosomes total). Chromosomes were originally discovered over 100 years ago using basic dyes and state-of-the-art microscopes (at that time), thus the name chromosome simply means "colored body" as a description of how they were first observed in the nucleus of the cell. One pair of chromosomes is associated with gender and is commonly referred to as the sex chromosomes. Females have two "X" sex chromosomes, whereas males have an "X" and a "Y" sex chromosome.

In simple terms, chromosomes are essentially unbroken polymers of double-stranded DNA. They often are associated with histone proteins that enable an efficient "packaging" of the DNA prior to cell division. The state of DNA in the cell correlates with the different phases of cell division (see **Figure 1-4**). It should be obvious that when a cell divides into two daughter cells, each cell must have a copy of the genome to remain viable. The cell goes through four phases to replicate itself, which includes replication of its genomic content. In the G1 phase, the activity of the cell is largely dedicated to growth and maintenance of the functions of the cell. As a cell prepares to undergo mitotic division, it enters the S phase, during which the entirety of the DNA (chromosomes) in the cell is duplicated (i.e., DNA synthesis = "S" phase), resulting in two copies of each chromosome. Completion of the DNA (chromosome) duplication leads to the G2 phase, and the

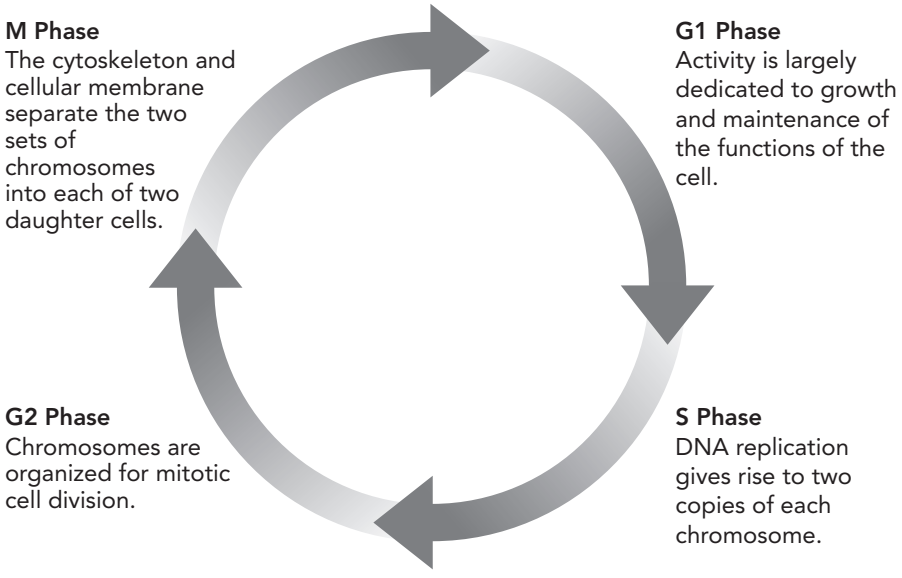


Figure 1-4 The four phases of the cell cycle.

chromosomes are organized in preparation for mitotic cell division. During the M phase of cell division, the cellular membrane separates the two sets of chromosomes into each of two daughter cells, and each daughter cell reenters the G1 phase, each with a complete copy of the genome within the chromosomes of its nucleus.

In addition to the chromosomal DNA found in the nucleus of the cell, a relatively small amount of DNA is found in the mitochondria. Mitochondria harbor about 16 kilobases of DNA (called mtDNA) in a circular form. In humans, the mtDNA contains 37 genes that encode proteins utilized by the mitochondria for energy production and protein synthesis. The mtDNA undergoes its own replication as mitochondria replicate within the cells of our body. Importantly, mtDNA is maternally inherited because the large female gamete (egg cell) contains hundreds of thousands of mitochondria, whereas the mtDNA in the much smaller male gamete (sperm cell) is not incorporated during fertilization of the egg. Thus, the fertilized egg only contains maternal mtDNA and is therefore used in genetic research for mapping maternal inheritance across generations.

Cells exist in the human body that do not harbor a nucleus and therefore lack a copy of the human genome. The red blood cells (RBCs, also known as erythrocytes) and platelets (also called thrombocytes) of the blood are derived from parent cells in the long bones of our bodies. Red blood cells are involved in oxygen transport in the blood and are derived from a process called erythropoiesis involving progenitor cells (e.g., proerythroblasts, polychromatic erythroblasts), whereas platelets are involved in blood clotting and are derived from megakaryocytes. These non-nucleated cells lack nuclear DNA, yet they harbor genetic information in the form of RNA, allowing the cells to synthesize proteins. The DNA that is obtained from a blood sample is actually derived from the white blood cells, which make up about 1% of blood volume in healthy adults.[9]

This is important when we consider using DNA genotyping to support advances in healthcare. If we use genotyping to screen for a neurological disease, we do not need to sample the human brain directly because almost all other cells in our body contain the complete genome. Therefore, we can carry out genotyping using cells that are easily obtainable (e.g., white blood cells, from a buccal swab to gather the cells from inside the mouth), thereby providing a noninvasive, nondestructive method for gaining access to our genomic information.

Categorically, genetic testing in humans is routinely carried out in four distinct areas: paternity and/or maternity, DNA forensics, disease predisposition, and pharmacogenetics (see **Table 1-1**). Paternity/maternity testing is used to establish a biological relationship between a parent and an offspring, whereas DNA forensics can determine the origin and/or identity of a biological sample. In both of

Table 1-1

Categories of Human Genetic Testing

	Paternity or Maternity Testing	DNA Forensics	Disease Predisposition	Pharmacogenetics
Utility	Determine biological parent.	Determine identity of crime scene DNA sample.	Determine cause of, or predisposition for, disease or disorder, or if the patient is a carrier for an inherited disease.	Predict optimal drug and/or dose for specific patient.
Sample source	Buccal swab	Varied	Buccal swab, saliva, or blood sample	Buccal swab, saliva, or blood sample
Target	Short tandem repeats (STR)	Short tandem repeats (STR)	Allelic variations linked to disease/disorder	Genes for drug metabolism enzymes, drug transporters, and drug receptors
Rapid testing turnaround required	Infrequently	Infrequently	No	Yes

these areas, the genomic biomarkers commonly tested are called short tandem repeats, or STRs, which are short repeated sequences of DNA. Another growing area of human genomics involves testing for specific genomic biomarkers associated with disease, where the genetic cause for a disease or disorder is established as a diagnostic tool or used to determine the risk of developing the disease.

Pharmacogenetics, however, points to important distinctions among these areas of genetic testing. Each has shown tremendous utility and societal value. Yet, in order to derive the full clinical potential of genetic testing in pharmacogenetics, information regarding genetic variation as it relates to the disposition and effect of medications must be immediately available to caregivers. Thus, the value of pharmacogenetics is more likely to be dependent on technologies and information systems/procedures that allow for rapid testing and provide clinicians with more real-time access to a patient's individual genetic data.

Genetic Variation and Personalized Medicine

The essence of personalized medicine is individual genetic variation. The most obvious and perhaps most basic examples of individual genetic variation are observed outwardly. Readily apparent physical traits, such as skin tone, eye color, hair color, height, and even shoe

size, are all dictated by genes that vary, in some cases dramatically, between individuals. In this sense, the gene–trait interface could be described in modern, colloquial terms as "designing" an avatar in a video game. Each player is offered choices that determine the appearance of the avatar. Analogous to a genetic menu of sorts, one can scroll through screens of options, ranging from body type to facial structure, where nuances such as the thickness of the eyebrows, shape of the nose, and distance between the eyes are presented. These choices allow an avatar to assume a uniqueness that, although immensely oversimplified, can be extrapolated to represent genetic variation and the direct relationship between genetic identity and physical traits (see **Figure 1-5**). Yet, as we move from electronic simplification to genetic reality, the avatar analogy quickly fades—the vast complexity of the human genome provides for a much deeper level of variation between individuals.

Analysis of the human genome following publication of its first complete sequence in 2003 only begins to describe this complexity. As described earlier, each human germ-line cell contains approximately three billion nucleotide base pairs of DNA comprising around 25,000 genes, and among this immense store of genetic code there is tremendous intraspecies homogeneity, a fact underscored by the discovery that all humans share roughly 99.9% of the DNA sequence.

Such uniformity makes perfect sense. Genes encode for proteins for which functions are nearly always precisely limited by their tertiary and quaternary structure, which dictates efficiency of enzymatic and/or biological processes. One dramatic example is actin, a type of cytoskeletal scaffold that owns the title of being the most abundant protein in nearly all human cells, comprising anywhere from 10–20% of total cell protein. In fact, the typical hepatocyte contains an estimated 500 million actin molecules, giving the cytosol a gel-like rather than fluid consistency.[10] As is implied by their abundance, actin proteins are essential for a variety of biological functions, such as structural integrity, cell shape, cell motility, chromosome morphology, and muscle contraction, as well as a host of intracellular events, including gene transcription and translation. Thus, it is of little surprise that the six human genes encoding for the three actin isoforms (α, β, and γ) are among the most highly conserved in the entire genome, being second only to the histone family of DNA-binding proteins. In fact, the DNA sequence of human actin is over 80% identical to that found in yeast, with a near 96% amino acid homology.[11]

This incredible degree of interspecies homogeneity means that biological activity in eukaryotes is extremely sensitive to changes in the DNA sequence. Indeed, entire clusters of genes exist with a sole recognized function of minimizing DNA mutations during

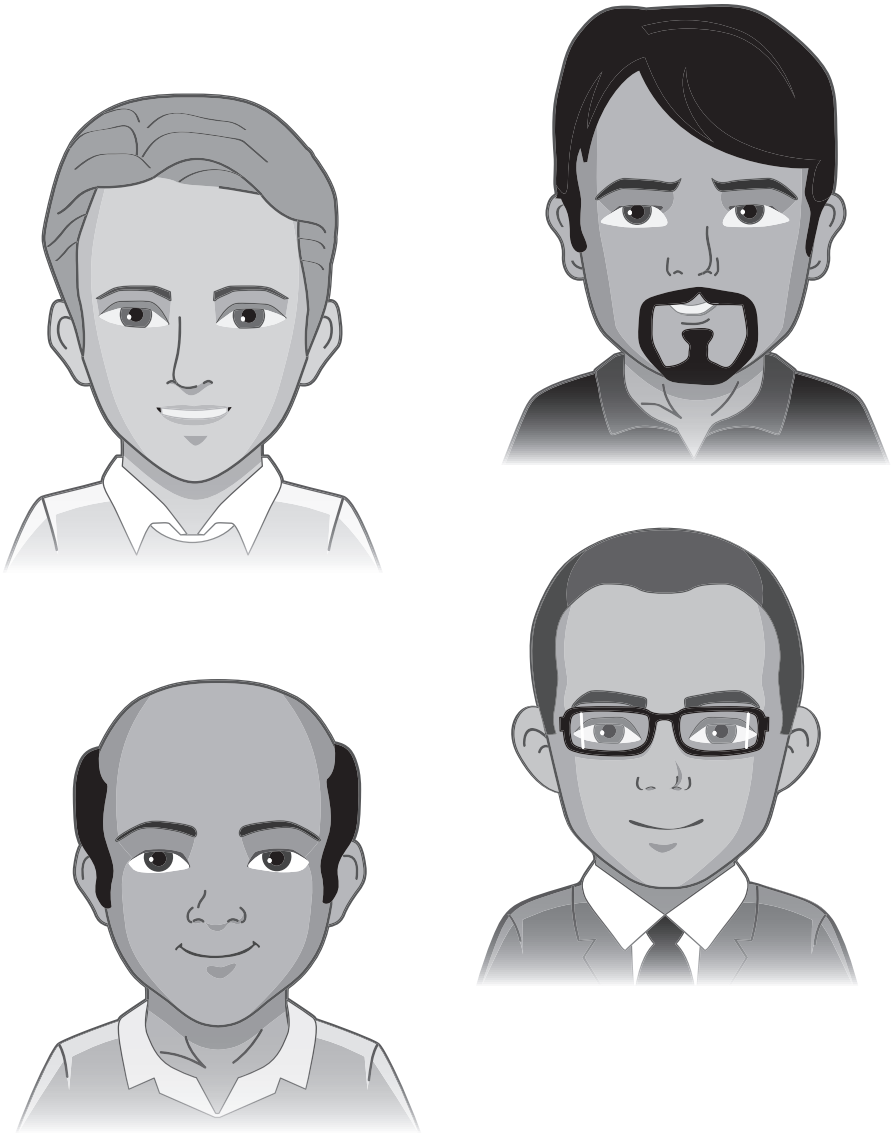


Figure 1-5 Avatars representing the authors of this text (generated by www.pickaface.net).
Source: Courtesy of Fredy Sujono from www.pickaface.net.

cell division. One such family of proteins is known as the mismatch repair genes.[12] Also highly conserved, this family of nine unique proteins "proofread" newly replicated daughter strands of DNA for relatively common errors in base incorporation by DNA polymerases, errors that would otherwise result in nearly one mutation for every 1,000 base pairs replicated. Instead, mismatch repair enzymes identify

"mismatched" bases, excise them from the newly replicated daughter strand, and finally reinsert the correct deoxyribonucleotide base. This effectively reduces the average mutation rate by six orders of magnitude, or to less than one base change per billion bases.

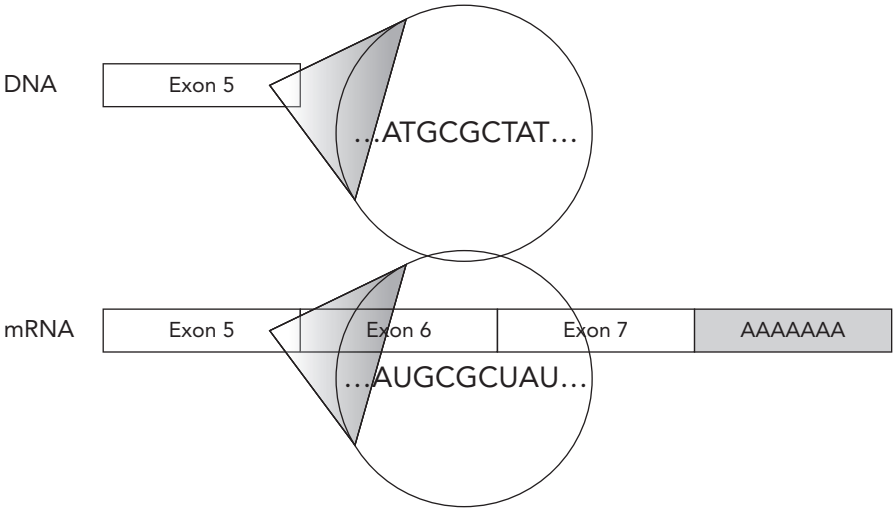
The importance of mutation-reducing enzymatic activity is obvious. DNA sequence fidelity transmitted from parent cell to daughter and from parent organism to offspring allows for continuity of gene sequence, which provides for continuity of inherited traits. Moreover, evolutionary pressures of selection work toward maintaining individuals with as little genetic diversity as possible, at least with respect to the many thousands of genes, like those for actin, whose activity is required for sustainable life.

In spite of these Herculean cellular efforts and the constant evolutionary pressures that favor DNA fidelity across generations, genetic variation persists. Small changes in genetic code continue to arise, and these often more subtle mutations, known as polymorphisms, give rise to a deeper, and in some ways more defining, characteristic of genetic variation among individuals.

Polymorphic Genetic Variation

In the most basic sense, changes in the genetic code are observed as differences in DNA sequence called **mutations**. These changes in sequence may or may not produce observable differences in traits either in an individual or in its offspring. Mutations that occur in genomic DNA between individuals gives rise to genetic variation—that one person's DNA sequence differs from another at specific bases. Some mutations are more common than others in a population. When a particular mutation occurs in at least 1% of individuals in the population, it is commonly referred to as a **polymorphism**, which is derived literally from the Greek word meaning "many forms." For example, if at a given location in the genome 4% of individuals contain adenine (A) but the other 96% contain a cytosine (C), the A represents a polymorphism. In this way, the term polymorphism is used to help describe the prevalence of a specific genetic variation between individuals within a population.

Variants are incredibly common. Individuals differ in their DNA on average by one base pair for every 100 to 300 base pairs throughout the genome, although their frequency can be much greater within a given gene. It has been estimated that as many as 9 to 10 million polymorphisms may reside in the human genome, yet it is highly unlikely that any one individual will carry all possible polymorphic variations.[13] However, because of their frequency, polymorphisms are particularly useful in describing genetic differences between individuals, especially differences that define discrete subpopulations within the population as a whole.



		Second Position					
		U	**C**	**A**	**G**		
First Position	**U**	UUU – Phe UUC – Phe UUA – Leu UUG – Leu	UCU – Ser UCC – Ser UCA – Ser UCG – Ser	UAU – Tyr UAC – Tyr UAA – Stop UAG – Stop	UGU – Cys UGC – Cys UGA – Stop UGG – Trp	**U** **C** **A** **G**	**Third Position**
	C	CUU – Leu CUC – Leu CUA – Leu CUG – Leu	CCU – Pro CCC – Pro CCA – Pro CCG – Pro	CAU – His CAC – His CAA – Gin CAG – Gin	CGU – Arg CGC – Arg CGA – Arg CGG – Arg	**U** **C** **A** **G**	
	A	AUU – Ile AUC – Ile AUA – Ile AUG – Met	ACU – Thr ACC – Thr ACA – Thr ACG – Thr	AAU – Asn AAC – Asn AAA – Lys AAG – Lys	AGU – Ser AGC – Ser AGA – Arg AGG – Arg	**U** **C** **A** **G**	
	G	GUU – Val GUC – Val GUA – Val GUG – Val	GCU – Ala GCC – Ala GCA – Ala GCG – Ala	GAU – Asp GAC – Asp GAA – Glu GAG – Glu	GGU – Gly GGU – Gly GGA – Gly GGG – Gly	**U** **C** **A** **G**	

	Wild-Type Sequence	**Genetic Variation**
mRNA	…ACC—GCC—UAU…	…ACC—UCC—UAU…
Protein Activity	…Thr—Ala—Tyr…	…Thr—Ser—Tyr…
Clinical Outcome	Normal binding to β-myosin Normal cardiac muscle contraction	Reduced β-myosin binding Decreased cardiac muscle contraction and clinical symptoms of left ventricular hypertrophy

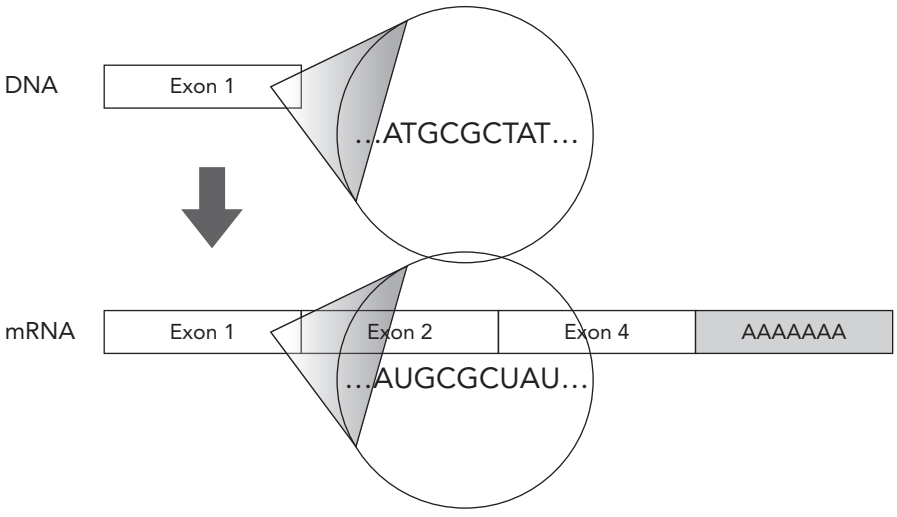
The manifestation of variation in the genetic code can be dramatic. One such example is found in the human α-actin gene. Familial hypertrophic cardiomyopathy (FHC) is an autosomal dominant congenital disease that leads to compromised cardiac function (syncope, angina, arrhythmias, and heart failure) and is the leading cause of sudden death in young people.[14] At least nine different mutations in α-actin have been directly linked to FHC, including a guanine to thymine (G→T) mutation at base 253 of exon 5 in the actin gene. This change, where the **wild-type** or typical sequence found in "normal" individuals is altered, results in the substitution of the amino acid serine for alanine at position 295 within the actin protein and is denoted as Ala295Ser. The simple G→T variation results in an actin molecule whose binding affinity for β-myosin is diminished, which reduces the strength of cardiac muscle contraction and can contribute to potentially fatal hypertrophy of the left ventricle (see **Figure 1-6**).[15]

Understandably, potentially serious physiological consequences that can be expressed at a young age make the actin Ala295Ser variation less likely to be transmitted generationally. However, a far greater degree of genetic variation is interspersed throughout the genome. Remember that nearly 99% of the genome is contained within regions of DNA considered noncoding or intergenic that do not directly encode for protein. Thus, the vast majority of variations are likely to be neither harmful nor beneficial *per se*. Yet, there is a growing appreciation for the potential role of polymorphisms in directly causing, or indirectly associating with, characteristics and traits that vary between groups within a population, especially as it pertains to individual responses to drugs.

In general, polymorphisms can be categorized into two main types: **single nucleotide polymorphisms**, commonly referred to as **SNPs** (pronounced "snips"), and insertions or deletions, commonly referred to as **indels**, with each category further differentiated into subcategories based on the nature, location, and effect of the polymorphism.

The most common type of polymorphism in pharmacogenetics is the SNP (see **Figure 1-7**). Single nucleotide polymorphisms are polymorphisms that occur at a single nucleotide where any one of the four bases of DNA (A, C, G, and T) may be substituted for another. An estimated 90% of all genetic variation in the human genome is thought to

Figure 1-6 Variation in the human cardiac α-actin gene associated with familial hypertrophic cardiomyopathy (FHC). A guanine to thymine (G→T) mutation in exon 5 of the human cardiac α-actin gene results in variation in the mRNA codon sequence and subsequent mistranslation of serine at amino acid position 295 rather than alanine. The resulting actin molecule exhibits reduced binding affinity for β-myosin, resulting in diminished cardiac muscle contraction and clinical symptoms associated with hypertrophy of the left ventricle.



First Position		Second Position: U	Second Position: C	Second Position: A	Second Position: G	Third Position	
	U	UUU – Phe UUC – Phe UUA – Leu UUG – Leu	UCU – Ser UCC – Ser UCA – Ser UCG – Ser	UAU – Tyr UAC – Tyr UAA – Stop UAG – Stop	UGU – Cys UGC – Cys UGA – Stop UGG – Trp	U C A G	
	C	CUU – Leu CUC – Leu CUA – Leu CUG – Leu	CCU – Pro CCC – Pro CCA – Pro CCG – Pro	CAU – His CAC – His CAA – Gln CAG – Gln	CGU – Arg CGC – Arg CGA – Arg CGG – Arg	U C A G	
	A	AUU – Ile AUC – Ile AUA – Ile AUG – Met	ACU – Thr ACC – Thr ACA – Thr ACG – Thr	AAU – Asn AAC – Asn AAA – Lys AAG – Lys	AGU – Ser AGC – Ser AGA – Arg AGG – Arg	U C A G	
	G	GUU – Val GUC – Val GUA – Val GUG – Val	GCU – Ala GCC – Ala GCA – Ala GCG – Ala	GAU – Asp GAC – Asp GAA – Glu GAG – Glu	GGU – Gly GGU – Gly GGA – Gly GGG – Gly	U C A G	

	Wild-Type Sequence	Synonymous SNP	Nonsynonymous SNP
mRNA Codon Sequence	…AUG—CGC—UAU…	…AUG—CGA—UAU…	…AUG—AGC—UAU…
Protein Sequence	…Met—Arg—Tyr…	…Met—Arg—Tyr…	…Met—Ser—Tyr…

Figure 1-7 Synonymous and nonsynonymous SNPs. Synonymous, or sense, SNPs are changes to a single nucleotide that alter the mRNA codon sequence without changes to the translated protein. In this hypothetical example, a cytosine to adenine (C → A) polymorphism changes the codon from CGC to CGA, but both codons are translated to arginine. In contrast, nonsynonymous, or missense, SNPs are changes to a single nucleotide that result in altered mRNA codon sequence and subsequent mistranslation of the protein. In this case, a cytosine to adenine (C → A) polymorphism results in the translation of a serine rather than arginine.

be derived from SNPs. Interestingly, the substitution of C → T constitutes roughly two out of every three SNPs.[13] Single nucleotide polymorphisms can be located in either coding or noncoding regions of DNA. Recall that coding regions contained within the genes make up less than 2% of the total DNA in the genome. As a result of the relative paucity of bases that make up this region, SNPs in coding regions occur less frequently than SNPs in noncoding regions but have a far greater potential to influence the phenotype of an individual. In this sense, the old colloquialism "location, location, location" certainly applies to SNPs.

Single nucleotide polymorphisms with the most direct genetic influence are located within the coding region of DNA. These polymorphisms are classified as either synonymous (also called sense mutations), which result in translation of the same amino acid, or nonsynonymous (also called missense mutations), which result in translation of a different amino acid. Another type of coding SNP can be classified as a nonsense mutation in that the polymorphism results in the inappropriate insertion of a stop codon in the growing mRNA, ultimately leading to a truncated protein product. In these ways, SNPs may cause important differences in gene function and/or expression. For example, mRNA transcripts used for translation can be directly altered by SNPs, leading to compromised transcript stability or altered RNA splicing. Likewise, coding nonsynonymous or nonsense SNPs may influence protein structure, stability, substrate affinities, and so on.

Apolipoprotein E (ApoE), a gene associated with Alzheimer's disease, can serve as an example of the effects of nonsynonymous SNPs located in the coding region of a gene.[16] Apolipoprotein E is a member of a family of proteins whose function is to bind to and assist in the transport of lipids in the circulatory system and is the predominant lipoprotein in the brain. Two SNPs, both thymine to cytosine (T→C) substitutions, are located within ApoE that result in the translation of more basic arginine residues at amino acid positions 112 and 158 instead of neutral cysteines. These changes, when found together, are known as the ApoE ε4 allele and transform ApoE into an isoform that exhibits increased binding affinity to amyloid β, a small protein involved in the pathology of Alzheimer's disease.[17] Apolipoprotein ε4 is found in high abundance in neurofibrillary tangles characteristic of Alzheimer's disease.[17] In fact, the SNPs associated with the ApoE ε4 allele, which occur in 5% of the population, are now considered to be the single greatest genetic risk factor for the development of Alzheimer's disease, which is the leading cause of senile dementia in the elderly and effects nearly 25 million adults worldwide.

Importantly, the influence of SNPs is not limited to those found directly in coding regions. At least one important function of noncoding DNA is to regulate the expression of mRNA transcripts. Thus, noncoding polymorphisms located in regulatory regions, including promoters, areas of DNA that respond to cellular machinery that

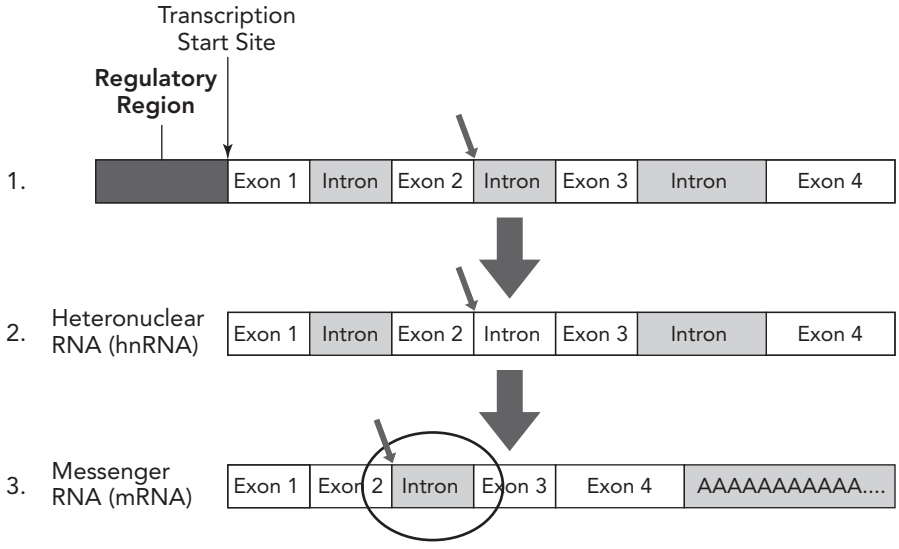


Figure 1-8 The potential impact of noncoding SNPs. SNPs located in noncoding regions of DNA, such as promoters, introns, and the boundary between exons and introns, can result in altered splicing and/or expression of mRNA transcripts. In this example, the SNP located at the proximal intronic boundary between exons 2 and 3 (indicated by arrows) alters the DNA sequence recognized by splicing machinery within the cell, eliminating the splice site. The resulting mRNA transcript erroneously retains the intron, leading to translation of the intron sequence into an altered protein product.

control gene expression, introns, and the boundary between exons and introns, lead to potential changes in transcription factor binding, mRNA transcript stability, or RNA splicing (see **Figure 1-8**).[18]

It is worth noting, not without irony, that there is considerable variation in the nomenclature used to describe genetic variation. Frequently, the same polymorphisms are described by different names in various basic science and clinical sources in the literature. For instance, a hypothetical single-base variation from adenine to thymine could be designated as A→T, A/T, A>T, or even A123T or 123A>T to denote base position within the gene. Making matters even more confusing, early studies of polymorphisms did not benefit from the standardized DNA sequence databases that exist today, such as the National Center for Biotechnology Information, or NCBI (www.ncbi.nlm.nih.gov). Rather, investigators studying identical regions of DNA frequently used sequences or fragments of DNA with different starting points relative to the actual genomic sequence. Thus, studies of our hypothetical polymorphism at position 123 could appear in the literature as A123T in one study and A323T in another if the sequence used in the latter began 200 bases upstream relative to that used in the former.

Although it will likely take some time for standardized nomenclature to take hold in the literature, recent efforts have produced several proposals for a systematic methodology of SNP nomenclature.

One prominent example is from the Human Genome Variation Society.[19] Its recommendations for the naming of human sequence variation promote a basic system focusing on first naming the gene of interest followed by designating the level of sequence variation: at the level of DNA, located in either coding regions designated as "c," genomic or noncoding regions as "g," or mitochondrial regions as "m." This nomenclature is not to be confused with the molecular biological term complementary DNA, which also is designated cDNA, and is likewise derived from reverse transcribing messenger RNA, or mRNA, so that only exons are included in the sequence. Thus, a coding reference sequence represents only DNA information contained in processed mRNA, whereas gDNA sequences represent DNA information identical to how it exists in the genome, containing DNA from introns, exons, and intergenic regions. Ribonucleic acid and protein sequence variation are respectively designated by "r." or "p." Actual variation in a sequence is described by listing first the reference or wild-type sequence/base followed by the sequence variation. Thus, applying this nomenclature system, the α-actin variation already described would be named c.253G>T to indicate the variation of sequence at position 253 in the coding reference sequence in the α-actin gene where the reference base guanine has been replaced by the variant thymine. This naming system could further be applied to describe the resulting change in terms of base substitution at the protein level using p.295S>A or p.295Ser>Ala where the serine at amino acid position 295 in the α-actin protein is changed to alanine.

Another SNP nomenclature system that is widely used is the **reference sequence number**, or the **refSNP**, **rs#**, or **rs**. Developed for use in the Single Nucleotide Polymorphism Database (dbSNP) hosted by the NCBI, this system is designed to reference genetic variation such as SNPs according to more precise locations within the genome rather than the arbitrary and varying segments of DNA frequently used in individual studies.[20] This is akin to providing each SNP with an exact chromosomal street address, where possible, that is used to define the SNP. For instance, rs113513162 is the specific, consistent identifier in dbSNP for the c.253G>T actin mutation in exon 5 of the ACTC gene located on chromosome 15. Efforts such as these that normalize the nomenclature and referencing of variation in the genetic code have proven valuable in decreasing the incidence of ambiguous or misleading literature references to SNPs.

Genetic variation can also be described at the whole-gene level. Perhaps the most relevant example for the purposes of this text is that of the human cytochrome (CYP) P450 genes for which gene-wide

variation is defined by well-accepted nomenclature.[21] In this system, the superfamily designation of "CYP" precedes that given for family (indicated by number), subfamily (indicated by letter), and individual subfamily member (again indicated by number). Importantly, allelic differences are defined by a number or a number and a letter following an asterisk (*) designation. It is important to note that in this nomenclature system the "*1" designation most commonly refers to the wild-type gene, whereas integers of "2" or greater denote polymorphic alleles typically numbered in order of their discovery and validation. For some genes, the nomenclature also includes the designation of "*1A" as the wild-type and "*1B," "*1C," "*1D," and so on as variants.

All told, this system allows for genotypic variation, in some cases involving multiple SNPs, to be described in phenotypic terms by referencing differences in an allele rather than a nucleotide. For example, CYP2C9 is a primary metabolizing enzyme of drugs, including the antiseizure medication phenytoin, the anticoagulant warfarin, and many nonsteroidal anti-inflammatory drugs, such as naproxen. A SNP that occurs within the CYP2C9 gene resulting in a cytosine to thymine (C→T; rs1799853) conversion leads to decreased enzymatic function. This allelic polymorphism is designated by CYP2C9*2 and is used to denote individuals susceptible to elevated drug levels following administration of typical doses of these medications (see **Figure 1-9**).

The other major category of polymorphism is indels. This genetic variation involves the insertion or deletion of DNA either as single nucleotides or as two or more nucleotides, in some cases spanning regions of DNA encompassing an entire gene. One of the best characterized forms of indels is the duplication of the cytochrome P450 drug metabolizing enzyme CYP2D6, where individuals have been

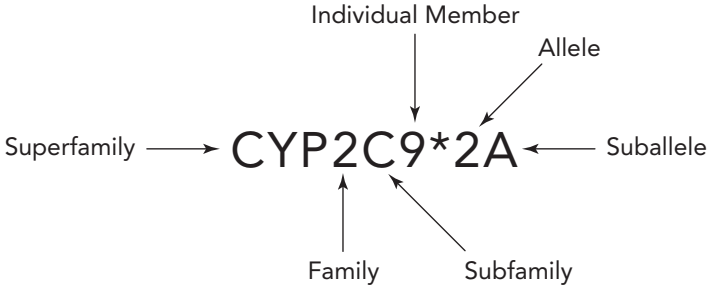


Figure 1-9 Nomenclature for the cytochrome P-450 (CYP) alleles. The established nomenclature system for alleles of the cytochrome P-450 (CYP) superfamily designates "CYP" followed by family number, subfamily letter, and individual subfamily number. Allelic differences are defined by number or a number and letter following an asterisk (*).

found to possess as many as 13 copies of the gene. In contrast, GSTT1, a gene encoding for the glutathione-conjugating enzyme glutathione S-transferase theta-1, is entirely deleted in some individuals, which sometimes leads to reduced metabolism of **xenobiotics**, particularly those with electrophillic and hydrophobic properties.[22] In this case, the existence of alternative metabolic pathways for some compounds means that the phenotypic effect of this gene deletion may not be observed.

A classic example of the pharmacogenetic consequences of polymorphic variation can be found in a family member of the cytochrome P450 enzymes, CYP2D6. Located on chromosome 22, CYP2D6 is a primary mechanism for the metabolism of nearly 100 drugs, including many antidepressants, such as fluoxetine; many neuroleptics, such as haloperidol; beta blockers, such as propanolol; and analgesics, such as codeine. Individuals carrying the wild-type alleles for CYP2D6 (CYP2D6*1) are phenotypically considered extensive metabolizers (EM) in that substrates at CYP2D6, such as the drugs aforementioned, are metabolized efficiently.[23] If one were to compare CYP2D6 metabolic activity to the volume dial on a stereo, the CYP2D6*1 allele would be analogous to a normal setting (see **Figure 1-10**).

Most of the clinically relevant CYP2D6 SNPs identified thus far result in diminished enzymatic activity associated with poor metabolizer (PM) or intermediate metabolizer (IM) phenotypes. For example, the CYP2D6*4 allele containing the 1846G>A polymorphism is a splicing defect in CYP2D6 that results in a truncated, nonfunctional protein product.[24] It is among the most common CYP2D6 SNPs found in Caucasian populations accounting for a significant percentage of mutant alleles. Another example of the PM phenotype is the CYP2D6*10 allele containing the 100C>T SNP, which results in diminished enzymatic activity via enhanced protein degradation.[25] It is the most common CYP-related polymorphism found in Asian populations (nearly 50% of individuals), whereas the CYP2D6*4 allele is seen at a much lower frequency in this group. Among individuals of African ethnicity, the CYP2D6*17 allele containing the 1023C>T polymorphism is most common, resulting in a deficiency of hydrolase activity due to reduced substrate-binding affinity.[26]

Evidence also suggests indel polymorphic expression of CYP2D6. Repetition of a 42-kilobase DNA fragment containing CYP2D6*2 results in CYP2D6 duplication that is phenotypically expressed as an ultrarapid metabolizer (UM) phenotype.[27] In fact, as many as 13 copies of the enzyme have been indentified in one individual's genome. Interestingly, this phenomenon is thought to have arisen from selective pressures associated with specific geographic regions. The incidence of CYP2D6 duplication has been reported with a frequency of less than 2% in Asians and less than 5% in Western Europeans but as much as 16% in Ethiopians.[28,29] Thus, the frequency of individuals possessing CYP2D6

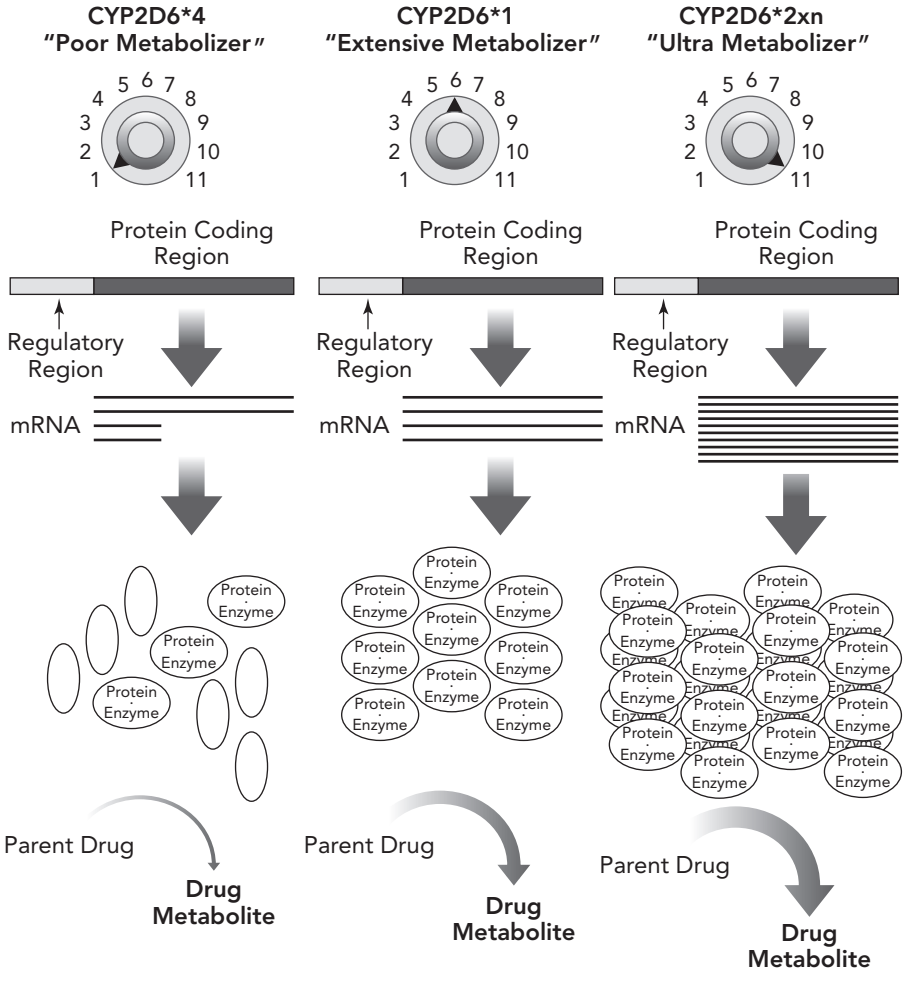


Figure 1-10 Genotypic and phenotypic differences in CYP2D6 metabolism. Individuals carrying the wild-type alleles for CYP2D6 (CYP2D6*1) are phenotypically considered "extensive metabolizers." Carriers of the CYP2D6*4 allele containing the 1846G>A polymorphism produce a truncated, nonfunctional protein product and therefore exhibit a "poor metabolizer" phenotype. The CYP2D6*2xn allele indicates repetition of a 42-kilobase DNA fragment, resulting in CYP2D6 duplication that is phenotypically expressed as an "ultrarapid metabolizer" phenotype.

duplication suggests a geographical gradient, possibly resulting from dietary pressures where, historically speaking, the detoxification capacity afforded by CYP2D6 duplication may have been essential for African diets relative to more European-based diets.

Fascinating though they may be from purely anthropological and genetic viewpoints, these observations have profound clinical

implications. First, for each of these groups, individuals possessing PM CYP2D6 polymorphisms may require reduced dosing of substrate drugs in order to avoid toxicities associated with decreased drug metabolism, which in many cases can be severe or even fatal. In contrast, individuals possessing CYP2D6 UM polymorphisms may require the polar opposite therapeutic course, that of increased rather than decreased dosing, in order to avoid symptoms associated with drug inefficacy. Second, these examples highlight the potential for pharmacogenetics to provide a mechanistic basis as to why individuals belonging to specific ethnic groups may respond very differently to standard drug therapy and eventually may provide a means for personalized dosing of those medications in advance. At the same time, these findings should provide ample caution against making assumptions based on ethnic background when treating individual patients. Remember, personalized medicine deals with using individual genetic information to support clinical decision making for optimal patient care. Ideally, increased prevalence of a pharmacogenetically relevant SNP in an ethnic population affords valuable consideration, but not a conclusion, at least not without genetic data specific to the individual patient.

Consider the therapeutic challenge of treating a patient with a needed medication whose primary metabolism occurs via CYP2D6, all while facing the unknown possibility that the patient's metabolic capacity could range anywhere from PM status to UM status. With this perspective, it is hardly surprising that according to the U.S. Food and Drug Administration nearly one million adverse drug reactions are reported each year in the United States, half of which lead to serious patient outcomes such as hospitalization or disability and almost 100,000 directly result in death.[30] Clearly, not all of these adverse events are attributable to pharmacogenetic influences. Many are undoubtedly the result of human error, such as administering the incorrect dose contrary to a correctly prescribed regimen. However, it ought to give one pause to realize that many adverse events are attributable not to human error, but to errors in humans. Or, in other words, adverse events arise not just when incorrect medications and/or doses are administered, but also when they are correctly prescribed and administered to individuals whose pharmacogenetic profiles may contraindicate such therapy.

In spite of these examples of dramatic phenotypes of polymorphic variation between individuals, it is more common that a pharmacogenetic trait cannot be clearly associated with a single SNP or indel. In this case, haplotypes can sometimes be used to associate **genotype** with phenotype. In true genetic terms, a **haplotype** refers to regions of DNA, such as a combination of alleles, that are inherited but that may or may not determine phenotype traits. Haplotypes are relatively

common. It has been estimated that most genes contain between 2 and 53 haplotypes, with an average of 14. A haplotype has been frequently used to describe groups of SNPs that are inherited together. Haplotypes themselves may not have a direct effect on drug response, but their proximity to an unidentified causative mutation may allow them to act as a marker for a particular drug response.

One example of the use of haplotypes in predicting individual drug responses is found in the β_2 adrenergic receptor (β_2AR).[31] Twelve haplotypes have been identified in the 5′ untranslated region (UTR) and in the coding region of the *ADRB2* gene that encodes for the β_2AR receptor. Several of these haplotypes have been associated with a greater than two fold increase in response to the β_2AR agonist albuterol, which is the prototypical agent in the class of sympathomimetic drugs used as first-line bronchodilators in treating symptoms of both asthma and chronic obstructive pulmonary disease (COPD). Importantly, no individual SNPs located within the haplotypes were able to be causatively linked to improved β_2AR-mediated bronchodilation. Thus, both SNPs and haplotypes can be used to map genetic changes that are associated with an individualized drug response.

The examples provided thus far show a direct link between genotype and phenotype—between the specific genetic makeup of an individual and the response of the individual to a drug. However, establishing an association between a genetic polymorphism and a specific drug response is more complicated when multiple polymorphisms within a gene and/or multiple genes are involved.

This is most easily discussed when considering traits that are **monogenic**, or those derived from a single gene. For example, each individual inherits two alleles of CYP2C9 (one from mom and one from dad). Therefore, the overall activity of CYP2C9 results from the combined contribution of both alleles. By definition, most individuals inherit two wild-type copies of CYP2C9, which means that most of us exhibit "normal" metabolic activity of the enzyme. However, what if an individual inherits the wild-type CYP2C9 allele from one parent but the CYP2C9*2 polymorphism from the other parent? In this case, the individual would be considered **heterozygous** for CYP2C9 (written CYP2C9*1/*2) in that he or she possesses two different alleles for the same gene, one fully functional and the other with compromised enzymatic activity. If both alleles were to contain the CYP2C9*2 polymorphism, the individual would be considered **homozygous** (CYP2C9*2/*2), resulting in greatly diminished metabolism by the CYP2C9 enzyme. Thus, one would expect to see a graded loss of metabolism across individuals who are wild-type (CYP2C9*1/*1), heterozygous (CYP2C9*1/*2), and homozygous (CYP2C9*2/*2) for the CYP2C9*2 polymorphism. This is referred to as a gene-dose response relationship (see **Figure 1-11**).

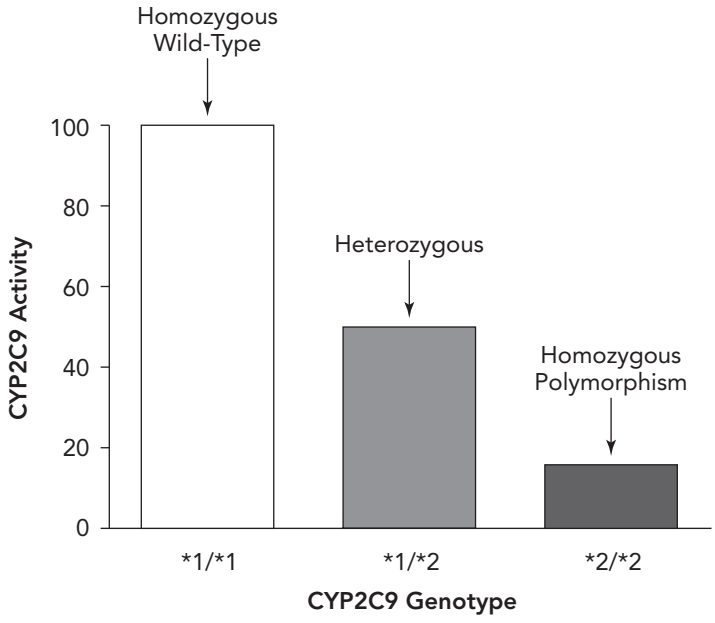


Figure 1-11 Allelic expression of CYP2C9*2 polymorphisms as an example of a monogenic trait. CYP2C9 activity results from the combined contribution of both alleles. Most individuals carry two alleles of the wild-type CYP2C9*1 and therefore exhibit full metabolic activity of the enzyme. Individuals who inherit one wild-type CYP2C9*1 allele and one polymorphic CYP2C9*2 allele are considered heterozygous for CYP2C9 and exhibit diminished enzymatic activity. If both alleles are the CYP2C9*2 polymorphism, the individual would be considered homozygous for the polymorphism, resulting in greatly diminished metabolism via CYP2C9.

In this simplistic example, the phenotype of CYP2C9 activity can be explained by the direct relationship between trait and genotype. But what if we expand our analysis to consider not just CYP2C9 activity but the overall response to a drug metabolized by the enzyme?

As an anticoagulant, warfarin has been used extensively to prevent thromboembolism but is limited in use by a narrow therapeutic index. Inadequate drug therapy increases the likelihood of potentially fatal thrombolytic events, whereas toxicity may result in life-threatening hemorrhaging. The anticoagulant effects of warfarin are mediated by inhibition of vitamin K epoxide reductase complex subunit 1 (VKORC1), a key factor in the clotting process.[32,33] Thus, the warfarin response is dependent on the function of its drug target, VKORC1, and its metabolizing enzyme, CYP2C9. Importantly, polymorphisms have been identified in VKORC1, including a guanine to adenine conversion (−1639G>A; rs9923231) that increases an individual's sensitivity to warfarin.[34,35] This means there exist subpopulations of patients

that carry the CYP2C9*2 polymorphism, the −1639G>A VKORC1 polymorphism, or both. Moreover, each individual will be either heterozygous or homozygous for each polymorphism, with each polymorphism potentially altering the anticoagulant response to warfarin.

This more intricate scenario describes a **multigenic trait** where the phenotypic expression of the trait (in this case the anticoagulant response to warfarin) is dependent upon the function of several genes rather than just one. As complicated as this may appear, this gene–gene interaction still greatly oversimplifies the actual clinical condition. Consider that the warfarin response is influenced by not just two genes (gene–drug interactions) but rather the confluence of many more factors, such as age, weight, and sex,which are further compounded by other environmental variables (gene–environment interactions), such as concurrent drug therapy, and behavioral choices, such as smoking or diet. Merely attempting to approximate such complexity helps to highlight the complicated relationship that can exist between a drug response and genotype.

Review Questions

1. Pharmacogenomics is the study of the relationship between genetic variation and drug response.
   a. True
   b. False
2. Genetic variation in the nucleotide sequence of DNA necessarily results in changes in amino acid sequence and protein functionality.
   a. True
   b. False
3. In describing genetic variation, mutations and polymorphisms can be differentiated by which of the following?
   a. Frequency of the variation
   b. Functional effects of the variation
   c. Location of the variation within the genome
   d. Mutations and polymorphisms are indistinguishable
4. A coding synonymous single nucleotide polymorphism is most likely to induce a change in which of the following?
   a. Enzyme–substrate affinity
   b. Receptor–ligand binding
   c. RNA splicing
   d. Transcription

5. Which of the following metabolic enzymes is associated with both poor metabolizer (PM) and ultrarapid metabolizer (UM) phenotypes?
   a. NADP
   b. CYP2D6
   c. TPMT
   d. VKORC1

6. Which of the following CYP2D6 polymorphisms is an example of an indel?
   a. CYP2D6*4
   b. CYP2D6*2
   c. CYP2D6*10
   d. CYP2D6*17

7. Which of the following is the best description of a haplotype?
   a. A common mutation in DNA in a given population observed at greater than 1% frequency.
   b. An observable characteristic or trait.
   c. A series of polymorphisms that are inherited together.
   d. Possessing two different alleles for the same trait.

8. Polymorphisms such as those found in CYP2C9 result in heterozygous individuals who often display intermediate enzyme activity and wild-type and homozygous individuals who display either fully functional or nonfunctional enzyme activity, respectively. This trimodal phenotype is indicative of which of the following?
   a. A monogenic trait
   b. A multigenic trait
   c. Neither A nor B
   d. It is not possible to tell.

9. A patient who recently started taking the antipsychotic medication haloperidol presents with dry mouth, restlessness, spasms of the neck muscles, and weight gain, all of which are adverse effects associated with haloperidol toxicity. Based on your knowledge of the pharmacogenetic influence of CYP polymorphisms, you speculate that this patient is:
   a. homozygous for the CYP2D6*4 allele.
   b. homozygous for the CYP2D6*2xn allele.
   c. homozygous for the VKCOR1 (AA) allele.
   d. homozygous for the CYP2C9*1 allele.

References

1. PharmGKB. Questions about pharmacogenomics: What is the difference between pharmacogenetics and pharmacogenomics? Available at: www.pharmgkb.org/resources/faqs.jsp. Accessed August 21, 2012.
2. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*. 2001;409(6822):860–921.
3. Greilhuber J, Volleth M, Loidl J. Genome size of man and animals relative to the plant *Allium cepa*. *Can J Genet Cytol*. 1983;25(6):554–560.
4. Lim K-B, Chung J-D, van Kronenburg B, et al. Introgression of *Lilium rubellum* Baker chromosomes into *L. longiflorum* Thunb: Agenome painting study of the F1 hybrid, BC1 and BC2 progenies. *Chromosome Res*. 2000;8:119–125.
5. Hallen HE, Luo H, Scott-Craig JS, Walton JD. Gene family encoding the major toxins of lethal *Amanita* mushrooms. *Proc Natl Acad Sci*. 2007;104(48):19097–19101.
6. Sauvain-Dugerdil C, Léridon H, Mascie-Taylor CGN. *Human clocks: The bio-cultural meanings of age*. New York: Peter Lang; 2006.
7. Covington HE, Maze I, LaPlant QC, et al. Antidepressant actions of histone deacetylase inhibitors. *J Neurosci*. 2009;29(37):11451–11460.
8. Rabik CA, Njoku MC, Dolan ME. Inactivation of O^6-alkylguanine DNA alkyltransferase as a means to enhance chemotherapy. *Cancer Treat Rev*. 2006;32(4):261–276.
9. Alberts B, Johnson A, Lewis J. Leukocyte functions and percentage breakdown. In: Alberts B (ed.). *Molecular biology of the cell*. 4th ed. New York: Garland Science; 2002.
10. Lodish H, Berk A, Matsudaira P, et al. Microfilaments and intermediate filaments. In: Lodish H, Berk A, Zipursky SL, Matsudaira P, Baltimore D, Darnell J (eds.). *Molecular cell biology*. New York: W.H. Freeman; 2004.
11. Vandekerckhove J, Weber K. At least six different actins are expressed in a higher mammal: An analysis based on the amino acid sequence of the amino-terminal tryptic peptide. *J Mol Biol*. 1978;126:783–802.
12. Iyer RR, Pluciennik A, Burdett V, Modrich PL. DNA mismatch repair: Functions and mechanisms. *Chem Rev*. 2006;106:302–323.
13. Chakravarti A. To a future of genetic medicine. *Nature*. 2001;409:822–823.
14. Maron BJ. Hypertrophic cardiomyopathy: A systematic review. *JAMA*. 2002;287:1308–1320.
15. Mogensen J, Klausen IC, Pedersen AK, et al. Alpha-cardiac actin is a novel disease gene in familial hypertrophic cardiomyopathy. *J Clin Invest*. 1999;103:R39–R43.
16. Cacabelos R, Martinez R, Fernandez-Novoa L, et al. Genomics of dementia. *Int J Alzheimers Dis*. 2012;2012:1–37.
17. Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science*. 1993;261:921–923.
18. Zhang W, Huang RS, Dolan ME. Integrating epigenomics into pharmacogenomic studies. *Pharmgenomics Pers Med*. 2008;2008(1):7–14.
19. den Dunnen JT, Antonarakis SE. Human Genome Variation Society. Recommendations for the description of sequence variants. *Hum Mutat*. 2000;15:7–12.

20. National Center for Biotechnology Information. dbSNP Short Genetic Variations. Available at: www.ncbi.nlm.nih.gov/projects/SNP. Accessed August 21, 2012.
21. Ingelman-Sundberg M, Daly AK, Oscarson M, Nebert DW. Human cytochrome P450 (CYP) genes: Recommendations for the nomenclature of alleles. *Pharmacogenetics*. 2000;10:91–93.
22. Sprenger R, Schlagenhaufer R, Kerb R, et al. Characterization of the glutathione S-transferase GSTT1 deletion: Discrimination of all genotypes by polymerase chain reaction indicates a trimodular genotype-phenotype correlation. *Pharmacogenetics*. 2000;10:557–565.
23. Kimura S, Umeno M, Skoda RC, et al. The human debrisoquine 4-hydroxylase (CYP2D) locus: Sequence and identification of the polymorphic CYP2D6 gene, a related gene, and a pseudogene. *Am J Hum Genet*. 1989;45:889–904.
24. Kagimoto M, Heim M, Kagimoto K, et al. Multiple mutations of the human cytochrome P450IID6 gene (CYP2D6) in poor metabolizers of debrisoquine: Study of the functional significance of individual mutations by expression of chimeric genes. *J Biol Chem*. 1990;265:17209–17214.
25. Yokota H, Tamura S, Furuya H, et al. Evidence for a new variant CYP2D6 allele CYP2D6J in a Japanese population associated with lower in vivo rates of sparteine metabolism. *Pharmacogenetics*. 1993;3:256–263.
26. Masimirembwa C, Persson I, Bertilsson L, et al. A novel mutant variant of the CYP2D6 gene (CYP2D6*17) common in a black African population: Association with diminished debrisoquine hydroxylase activity. *Br J Clin Pharmacol*. 1996;42:713–719.
27. Johansson I, Lundqvist E, Bertilsson L, et al. Inherited amplification of an active gene in the cytochrome P450 CYP2D locus as a cause of ultrarapid metabolism of debrisoquine. *Proc Natl Acad Sci*. 1993;90:11825–11829.
28. Bradford LD. CYP2D6 allele frequency in European Caucasians, Asians, Africans and their descendants. *Pharmacogenomics*. 2002;3:229–243.
29. Ingelman-Sundberg M. Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): Clinical consequences, evolutionary aspects, and functional diversity. *Pharmacogenomics J*. 2005;5:6–13.
30. U.S. Food and Drug Administration. Adverse Event Reporting System (AERS). Available at: www.fda.gov/Drugs/GuidanceComplianceRegulatoryInformation/Surveillance/AdverseDrugEffects/default.htm. Accessed August 20, 2012.
31. Drysdale CM, McGraw DW, Stack CB, et al. Complex promoter and coding region beta 2-adrenergic receptor haplotypes alter receptor expression and predict in vivo responsiveness. *Proc Natl Acad Sci*. 2000;97:10483–10488.
32. Higashi MK, Veenstra DL, Kondo LM, et al. Association between CYP2C9 genetic variants and anticoagulation-related outcomes during warfarin therapy. *JAMA*. 2002;287:1690–1698.
33. Li T, Chang CY, Jin DY, et al. Identification of the gene for vitamin K epoxide reductase. *Nature*. 2004;427:541–544.
34. Veenstra DL, Blough DK, Higashi MK, et al. CYP2C9 haplotype structure in European American warfarin patients and association with clinical outcomes. *Clin Pharmacol Ther*. 2005;77:353–364.
35. Wadelius M, Chen LY, Downes K, et al. Common VKORC1 and GGCX polymorphisms associated with warfarin dose. *Pharmacogenomics J*. 2005; 5:262–270.