

Basic genetics

Bruce R. Korf, MD, PhD

*Department of Genetics, University of Alabama at Birmingham,
1530 3rd Avenue, South Kaul 230, Birmingham, AL 35294, USA*

It has been recognized for approximately a century that genetic factors play a role in human disease, but until recently genetics was perceived as focusing only on rare disorders. Despite major advances such as chromosomal analysis, prenatal diagnosis, and newborn screening, genetics has played a minor role in day-to-day primary care. Recent advances, especially the sequencing of the human *genome*, are changing this picture rapidly. It is expected that genetics will play an increasingly central role in all areas of medicine and particularly in primary care, as the genetic contributions to common disorders come into focus. Most practicing physicians have had little training in genetics, and rapid advances in the field make it increasingly difficult to keep up. This article reviews some of the basic principles of genetics to provide a foundation for understanding the application of genetics to medical practice. Definitions of terms that appear here in *italics* can be found in the on-line, illustrated glossary at www.genetests.org.

Genes and the human genome

Our understanding of the concept of the *gene* evolved over most of the twentieth century, beginning with the recognition that genes function to encode proteins. The structure of *DNA* came to light in the middle of the century, ushering in a period during which the mechanisms of gene *replication* and *expression* came under study. The last quarter of the century saw the introduction of methods of DNA sequence analysis, culminating in the determination of the human genome sequence just at the turn of the century. This section reviews the current understanding of the structure and function of genes and their organization into the genome.

E-mail address: bkorf@uab.edu

Structure and function of genes

The basic unit of genetic function is the *gene*, the chemical basis for which is the *DNA molecule*. DNA consists of a pair of strands of a sugar-phosphate backbone attached to a set of pyrimidine and purine bases (Fig. 1). The strands are held together by hydrogen bonds between adenine and thymine bases and between guanine and cytosine bases. Together these strands form a double helix. The strands separate during DNA replication, and the base sequence of the newly synthesized strand is dictated by the complementarity of adenine with thymine and guanine with cytosine. DNA therefore contains within its structure the information necessary for its replication.

The sequence of bases in DNA also provides the code that determines the structure of proteins. Proteins consist of chains of amino acids. The specific ordering of amino acids determines the unique properties of each protein. The amino acid sequence of a protein is determined by the sequence of bases

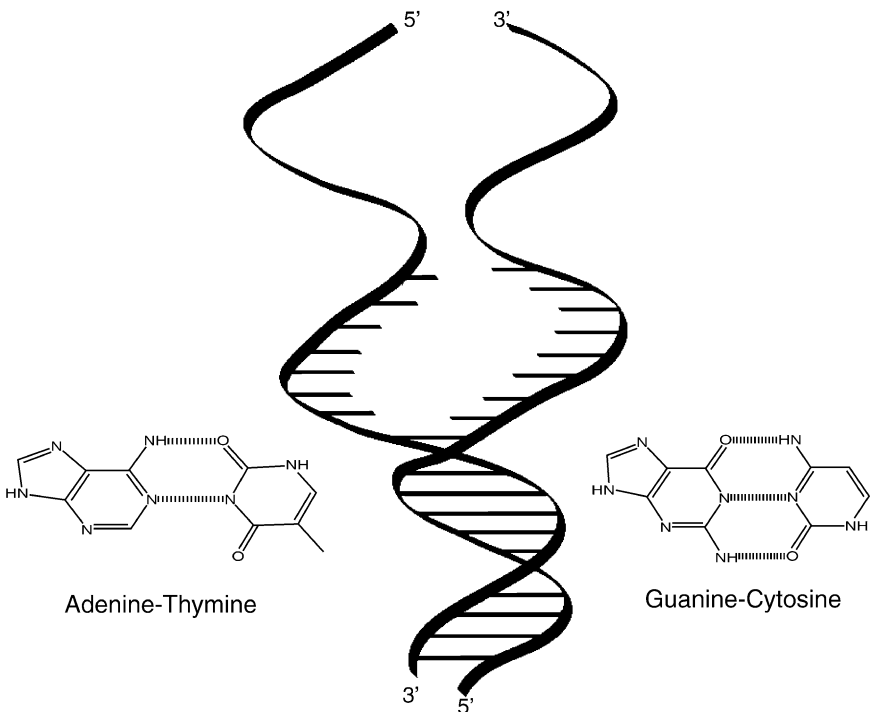


Fig. 1. DNA consists of a double helical structure, with the two sugar-phosphate strands linked together by hydrogen bonds between cytosine–guanine or adenine–thymine bases. The two strands are oriented in an antiparallel manner, with the free ends identified by the 5' or 3' ribose sites. DNA replication involves a partial unwinding of the strands and assembly of a new strand using the other as a template.

in the stretch of DNA that encodes the protein. Each amino acid is represented in DNA by a triplet of bases (*codon*). This genetic code is more or less universal to all organisms. The base sequence of one strand of DNA is copied into a complementary RNA (*mRNA*), which is in turn translated on the ribosome into protein (Fig. 2).

The process of copying the DNA sequence of a gene into messenger RNA (mRNA) is referred to as *transcription*. Gene expression is tightly controlled, with particular genes being turned on or off in particular cells at specific times in development or in response to physiologic signals. Transcriptional control is effected by the binding of repressor or activator proteins to a region just before the coding sequence of the gene, the *promoter sequence*. Other DNA sequences, called enhancers, function at longer range to decompact the DNA in a region to increase the rate of transcription.

Binding of an appropriate transcriptional activator results in partial opening of the DNA template at the start of the gene and attachment of the enzyme RNA polymerase. The polymerase reads the sequence of the template strand, copying a complementary RNA molecule that grows from the 5' to the 3' direction. The resulting mRNA is an exact copy of the DNA sequence, except that uridine takes the place of thymidine in RNA. Soon after transcription begins, a 7-methyl guanine residue is attached to the 5'-most base, forming the "cap." Transcription proceeds through the entire

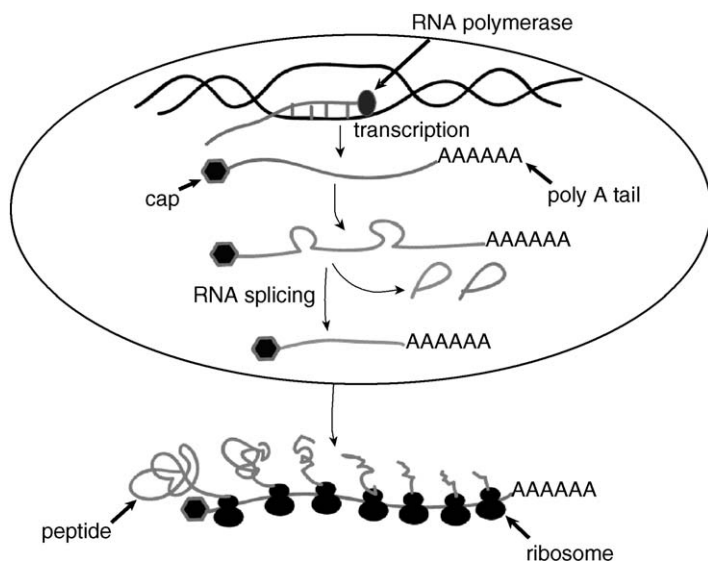


Fig. 2. Genetic information in the DNA is copied into a strand of mRNA. A 7 methylguanosine cap is added to the 5' end of the growing mRNA, and for most messages a poly A tail is added to the 3' end. Introns are removed from the message and exons are spliced together, and then the mature mRNA is exported to the cytoplasm. The peptide then is synthesized on ribosomes that bind to the mRNA.

coding sequence. Some genes include a sequence near the 3' end that signals cleavage at that site and enzymatic addition of 100–200 adenine bases, the poly-A tail. Polyadenylation is characteristic of so-called “housekeeping genes” that are expressed in most cell types. The 5' cap and the poly-A tail seem to function to stabilize the mRNA molecule and to facilitate its export to the cytoplasm.

The DNA sequence of a gene usually far exceeds the length required to encode its corresponding protein. This is accounted for by the fact the coding sequence is broken up into segments called *exons*, which are interrupted by other segments called *introns* (Fig. 3). Some exons may be only a hundred or so bases long, whereas introns can be several thousand bases in length. Much of the length of a gene therefore may be devoted to noncoding introns. The number of exons in a gene may be as few as one or two or may number in the dozens. The processing of the RNA transcript into mature mRNA requires the removal of the introns and *splicing* together of the exons. This is performed by an enzymatic process that occurs in the nucleus. The 5' end of an intron always consists of the two bases GU, following by a consensus sequence that is similar but not identical in all introns. This is the splice donor. The 3' end, the splice acceptor, ends in AG, preceded by a consensus sequence. The splice is initiated for each intron at the acceptor site, where the RNA molecule is cut and the 3' end of the intron bound to an internal A residue to form a lariat-shaped structure. A cut then is made at the splice donor and the exons flanking the intron then are ligated to one another.

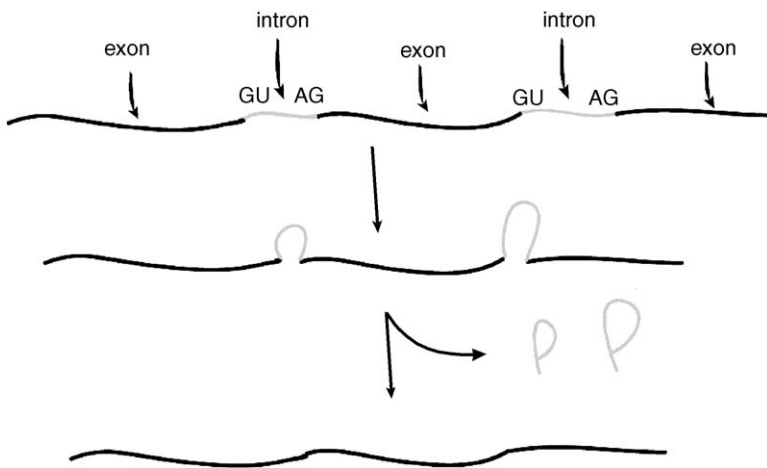


Fig. 3. This message consists of three exons and two introns. The 5' end of an intron includes a GU sequence (the splice donor) and the 3' end an AG (the splice acceptor). Introns are excised from the message to form lariat-like structures, and the exons are ligated together to form the mature mRNA.

The RNA splicing process offers another point of control of gene expression. Under the influence of control molecules present in particular cells, specific exons may be included or not included in the mRNA because of differential splicing. This results in the potential to produce multiple different proteins from the same gene, adding greatly to the diversity of proteins encoded in the genome. Specific exons may correspond with particular functional domains of proteins, leading to the production of multiple proteins with diverse functions from the same gene.

The mature mRNA is exported to the cytoplasm for *translation* into protein. During translation, the mRNA sequence is read into the amino acid sequence of a protein. The translational machinery consists of a protein–RNA complex called the ribosome. The sequence is read in triplets, called *codons*, beginning at the 5' end of the mRNA, which always is AUG, encoding methionine (although often this methionine residue later is cleaved off). Each codon corresponds with a particular complementary anticodon that is part of another RNA molecule called transfer RNA (tRNA). tRNA molecules bind specific amino acids defined by their anticodon sequence. Protein translation therefore consists of binding of a specific tRNA to the appropriate codon, which juxtaposes the next amino acid in the growing peptide that is enzymatically linked by an amide bond to the peptide. The process ends when a stop codon is reached (UAA, UGA, or UAG). The peptide then is released from the ribosome for transport to the appropriate site within the cell or for secretion from the cell. Post-translational modification such as glycosylation begins during the translation process and continues after translation is complete. Specific amino acid sequences determine the trafficking of the peptide within the cell, directing it for further processing and transport.

The human genome

The past decade has seen the completion of an international effort to sequence the human genome. Analysis of the sequence data has yielded several important insights. The first is that there are fewer genes in the human genome than had been expected. Early estimates of the number of human genes hovered at approximately 100,000; this was a crude estimate based on the facts that there are 3 billion base pairs of DNA and an average gene is approximately 30,000 bases. The actual number, though still not precisely known, seems to be closer to 30,000—not dramatically more than many “lower organisms” such as the fruit fly or the flatworm. The next surprise is that protein-encoding sequences comprise only approximately 5% of the genome. Approximately 50% consists of repeated sequences that do not encode protein. Some genes encode RNA that does not get translated into protein. This includes transfer and ribosomal RNAs, RNA molecules that are involved in the mRNA splicing process, and RNAs involved in gene regulation. Third, it has become apparent that genes are not distributed

evenly along the chromosomes. Rather they are distributed in clusters, with some areas being gene-rich and others gene-poor.

It also has been found that variation in base sequence between individuals is common. Such variations are referred to as *polymorphisms*, a term that is defined formally later in this article. The most frequently occurring polymorphisms involve changes of a single base pair of DNA, which occur approximately once in several hundred base pairs. These “*single nucleotide polymorphisms*” (SNPs) are an especially valuable tool in gene mapping and in searching for associations of specific genes with multifactorial traits, as is elaborated later.

Even with the identification of all human genes we are a long way from understanding how their expression is regulated and how gene products function in relation to one another and the environment. Developing tools for approaching these questions is another goal of the human genome project. Efforts also are underway to sequence genomes of other organisms, including microorganisms (*Escherichia coli* and the yeast *Saccharomyces cerevisiae*), the worm *Caenorhabditis elegans* (which has served as a model for development of multicellular organisms), the fruit fly *Drosophila*, the mouse, and many others. It is expected that insights gained from the study of genome structure and function in these model organisms will enhance understanding of human biology.

Gene mutation

The properties of a protein are determined by its amino acid sequence, which in turn is determined by the base sequence of the gene that encodes it. Alterations in the DNA coding sequence are referred to as *mutations*. Mutations can lead to complete failure of expression of a gene, aberrant regulation, or abnormal function of protein products of the gene.

Mutations can be classified either by the nature of the change in the DNA coding sequence or by the impact of the change on gene expression. DNA sequence changes consist of *deletions*, *insertions*, *rearrangements*, or single base changes (Fig. 4). *Deletions* can stretch over large regions, encompassing an entire gene or even group of genes, or can be as small as a single base.

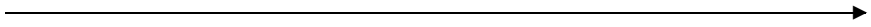
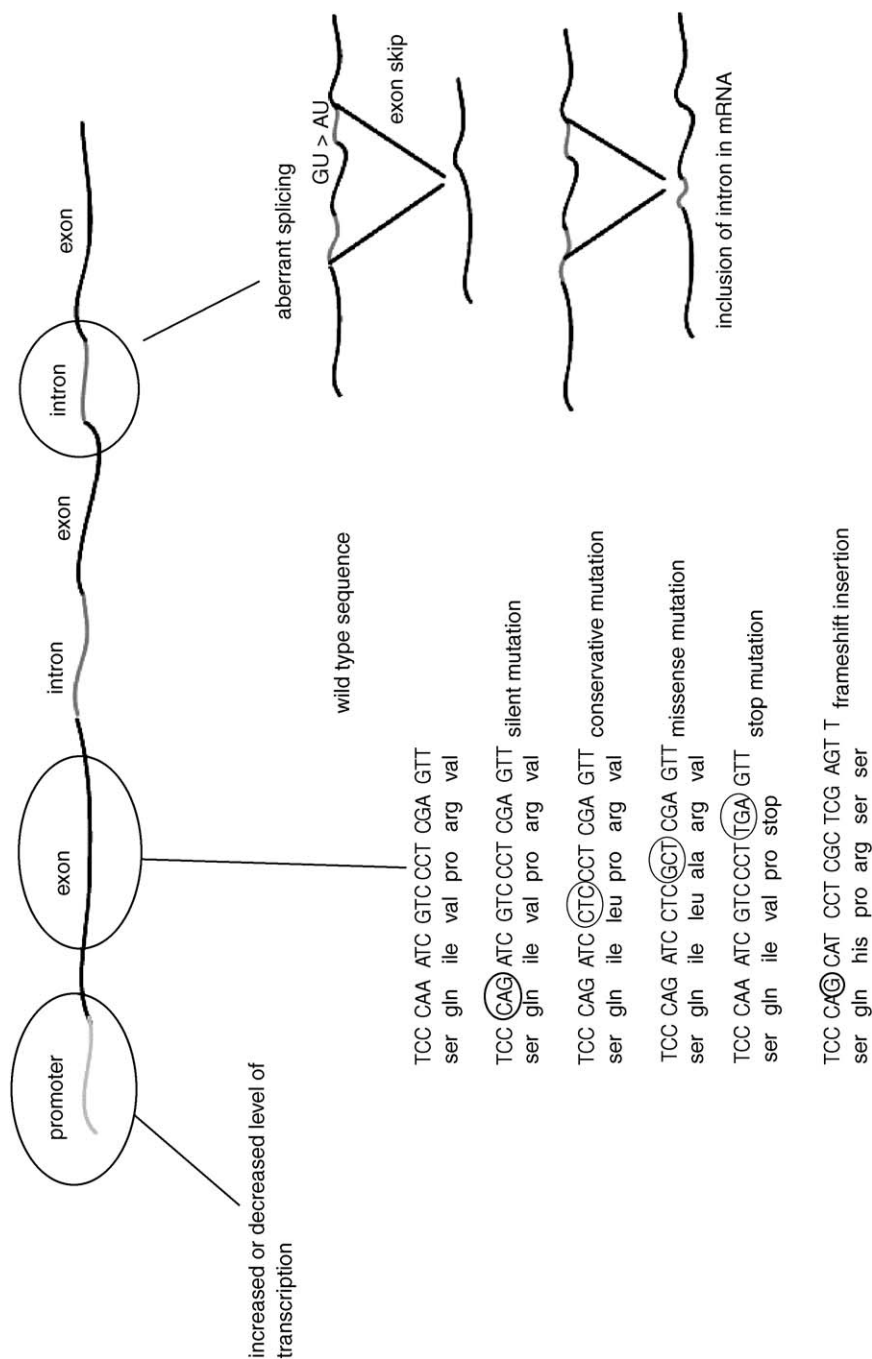


Fig. 4. Mutations can occur anywhere along the length of a gene. Mutations at the promoter sequence can lead to increased or decreased levels of transcription of the gene. Exon mutations include a variety of types of changes. A base sequence change that does not alter the amino acid encoded is a silent mutation. If the change leads to substitution of an amino acid with similar chemical properties, the mutation is conservative. A missense mutation leads to amino acid substitution that may change the chemical properties of the protein. If the change alters a codon to a chain termination signal, premature termination of protein translation occurs. Insertion or deletion of a number of bases that is not an integral multiple of 3 leads to frameshift and usually to premature termination of translation. Mutations in the splice donor or acceptor sequences can lead to exon skipping or inclusion of some intron sequence in the final processed message.



Insertions likewise can consist of large stretches of DNA (typically several hundred base pairs sometimes more) or can be as small as one base. Other rearrangements include *inversions* or *translocations* (attachment of one chromosomal region to another).

Deletion of an entire gene obviously results in lack of expression of that *allele*. Smaller intragenic deletions might result in lack of a particular exon or contiguous group of exons. Deletion of bases within an exon leads to loss of one or more amino acids from the protein if the deletion is of an integral multiple of three bases (remember that amino acids are encoded by triplets of bases). Deletion of any other number of bases causes a *frameshift*, altering the grouping of triplets of bases to form codons. The amino acid sequence is radically altered and usually a stop codon is encountered soon, leading to premature termination of translation. Often the truncated peptide is inactive and is degraded in the cell, so frameshift mutations usually result in lack of protein expression. The impact of insertions is similar to that of deletions: insertion of integral multiples of three bases leads to insertion of additional amino acids, whereas insertion of a nonintegral multiple of three leads to a frameshift.

Single base changes may affect the protein in subtle or in more radical ways. In some cases, the altered codon encodes the same amino acid as the original (many amino acids are encoded by multiple codons), called a silent mutation. Other mutations cause insertion of an incorrect amino acid into the protein. If the substituted amino acid has similar chemical properties to the original amino acid, the mutation is said to be conservative. The impact on the function of the protein may be negligible or the change may be compatible with some degree of function. Alternatively the alteration might result in a major disruption of protein function, including complete loss of function, or conversely, aberrant gain of function. Finally, a base change might convert an amino acid codon to a stop codon, resulting in premature termination of translation of the protein and consequent lack of expression.

In addition to involving the coding sequence, mutations also may occur in regulatory regions or inside introns, the latter disrupting the splicing process. Regulatory mutations may lead to loss of transcriptional control and failure of repression or inadequate activation of a gene. Mutations within the intron splice donor or acceptor regions lead either to failure of a splice to occur (leading to exon skipping) or to inclusion of some intron sequence (if a sequence similar to the donor or acceptor occurs within the intron), which has the functional impact of an insertion and usually results in a stop codon or frameshift.

Patterns of genetic transmission

An individual inherits two copies of every gene (except those carried on the sex chromosomes), one copy from each parent. The copies of a particular gene are called *alleles*. The specific pair of alleles at a *locus* constitutes the

genotype, and the resulting physical trait is the *phenotype*. Each human cell contains 46 *chromosomes*, including 22 pairs of non-sex chromosomes and two sex chromosomes (XX in females, XY in males).

Single gene transmission

A *recessive* trait is expressed only in individuals who inherit a mutant allele from both parents (Fig. 5). Such an individual is said to be *homozygous*, whereas the parents are *heterozygous* carriers. Carriers do not express the trait because the non-mutant allele is dominant. Two carrier parents face a one in four chance that any of their offspring will be homozygous. In contrast, a *dominant* trait is expressed in either the heterozygous or homozygous state (Fig. 6). A heterozygous affected individual has a 50% chance of passing the mutant allele to any offspring. If the mutation is rare, most who are affected will be heterozygous, not homozygous. Moreover, many of the medically important dominant traits are lethal when homozygous.

An *X-linked trait* is transmitted from females to half their offspring, but a male transmits the gene to all of his daughters but none of his sons (Fig. 7). X-linked recessive traits tend to be expressed in males, who have only one X chromosome. In females, most genes on one of the two X chromosomes are inactivated in each cell early in development, so a female is a sort of “mosaic,” having cells expressing one or the other X. Although the choice of X to inactivate is random in any cell, there are rare instances in which nonrandom inactivation occurs, making it possible for a female to express an X-linked recessive trait. This also can occur if the female has only one X chromosome (Turner syndrome). X-linked dominant traits can affect males and females; some are lethal in males and therefore are seen only in females.

Mutations that cause recessive disorders are most often loss of function changes, such as stop mutations, frameshifts, or missense mutations that

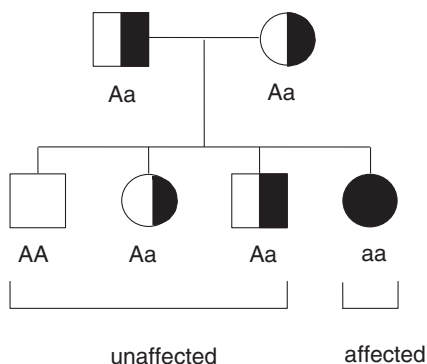


Fig. 5. For an autosomal recessive trait, an affected individual must be homozygous for the mutant gene (designated here as an “aa” genotype). Parents are heterozygous carriers (“Aa”) but do not express the phenotype.

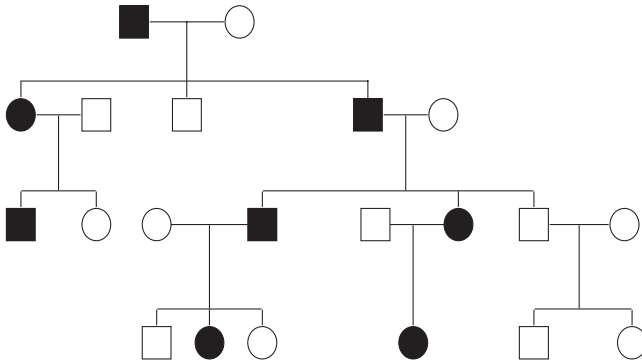


Fig. 6. An autosomal dominant trait is transmitted from generation to generation. Males or females can transmit the trait to either sex. An affected individual has a 50% chance of transmitting the mutant or nonmutant allele to any offspring.

significantly impair protein function. This pattern of inheritance is especially common for enzyme deficiencies such as *phenylketonuria* (PKU). Because of the catalytic nature of enzymes, loss of 50% of enzyme activity in a carrier is well tolerated. Complete loss of function, however, stops the reaction and leads to substrate accumulation and product deficiency. The mutations that cause dominant traits tend to be more diverse. Some are caused by loss of function, indicating that the system is sensitive to a 50% reduction in the quantity of gene product. Others are caused by gain of function mutations, in which the protein is active at times and places it should not be or has increased levels of activity. Only a single allele need be activated in this way, explaining the dominance.

Individuals with a mutant genotype do not always express the phenotype. Lack of expression is referred to as nonpenetrance. Sometimes the disorder

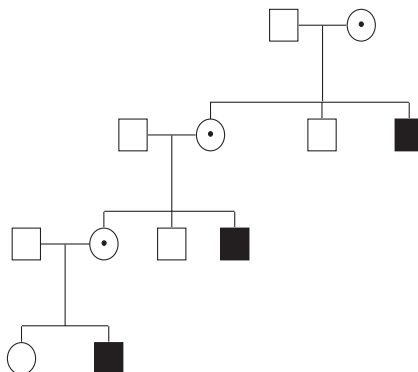


Fig. 7. An X-linked recessive trait is transmitted from a carrier female to approximately 50% of her offspring, but only hemizygous males express the trait.

appears only with the passage of time, in which case *penetrance* is said to be age-dependent. Nonpenetrance can be inferred when a dominant disorder skips a generation—an affected child has an affected grandparent, but the parent is not affected (and is therefore nonpenetrant). Genetic traits also can display *variable expressivity*, defined as differences in the degree or quality of expression of a specific phenotype. Some traits display a tendency to become more severe with each generation, a phenomenon referred to as *anticipation*. These traits tend to be neurologic disorders caused by triplet repeat expansions. The genes for these disorders include a run of multiple repeats of a specific triplet of DNA bases (eg, CAGCAGCAG...). The mutations consist of a larger than normal number of repeats. Long, repeating sequences are unstable during DNA replication and tend to expand, so that offspring have a longer series of repeats than the parent. Anticipation is explained by the larger the repeat size, the more severe the disorder and the greater the likelihood that the repeat will expand still more in the next generation.

Two additional recent discoveries regarding genetic transmission are *imprinting* and *mitochondrial inheritance*. *Imprinting* refers to the differential expression of maternal and paternal alleles for a particular gene. A small number of genes are subject to imprinting. Whether a mutation in an imprinted gene produces a phenotypic change depends on whether it occurs on a copy of an imprinted gene that was inherited from the parent whose allele is expressed in the offspring.

The basis for *mitochondrial inheritance* is that each mitochondrion contains hundreds of copies of a double-stranded circular DNA molecule. This DNA encodes 13 of the proteins involved in oxidative phosphorylation and a full set of tRNAs and ribosomal RNAs. Mutations in mitochondrial DNA present as failure in energy metabolism. Virtually all of the mitochondria in the zygote are of maternal origin, so mutations in mitochondrial DNA tend to be maternally-inherited (Fig. 8). Also there

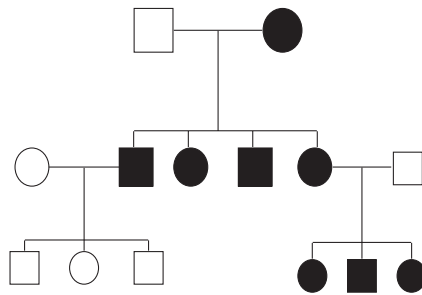


Fig. 8. A female transmits a mitochondrial trait to all of her offspring, whereas a male does not transmit the trait at all. If the trait is caused by mutation of a mitochondrial gene, there is likely variable expression among members of the family, because each has a different proportion of mutant or nonmutant mitochondrial DNA because of heteroplasmy.

may be a mixture of normal and mutant mitochondria DNAs in the same cell, a phenomenon known as *heteroplasmy*, and mitochondria segregate passively to the daughter cells at mitosis and meiosis. This can lead to multiple cell populations with various proportions of mutant and wild-type mitochondrial DNA, leading to differences in expression of a mitochondrial trait in different tissues or in individual members of the same family.

Multifactorial inheritance

Most of the disorders caused by changes in single genes are rare. Genetic factors also contribute to the etiology of common disorders, but here multiple genes interacting with one another and with environmental factors are required to cause the phenotype. This is referred to as *multifactorial inheritance*. To some extent the distinction between single gene and multifactorial disorders is artificial; even single gene traits are affected by modifying genes and the environment. Nevertheless the distinction is useful in that it acknowledges that, for a large number of common traits, genetic factors play a role, but no single gene is likely to be overwhelmingly important.

Multifactorial traits include a wide array of rare and common disorders with larger or smaller contribution of genetic factors. There are several ways in which the genetic contribution can be recognized. First is clustering within families. Risk for a multifactorial trait is higher in relatives of an affected individual than in the general population. This familial clustering, however, does not occur in accordance with Mendelian dominant or recessive inheritance. Multifactorial traits also display a higher rate of concordance in identical twins as opposed to full siblings. For quantitative traits such as height or blood pressure, the genetic contribution can be estimated from statistical analysis of correlations among relatives.

Various models of multifactorial inheritance have been used in studies of common traits. The additive polygenic model is based on the notion that alleles at multiple genes exert an additive effect to determine a quantitative phenotype. The threshold model postulates that there is a distribution of “liability” toward a particular trait in the population, but the trait itself only occurs when a threshold of liability is crossed. This can explain “all or none” phenotypes such as cleft palate. In some cases there are one or more genes of major effect, whereas in others the genetic contribution is more complex. As genomic knowledge develops, the identification of genes that contribute to common multifactorial disorders will have significant impact in helping to develop diagnostic tests and treatments based on understanding of pathophysiology.

Genetic epidemiology

The application of genetics in medicine is based largely on the concept of genetic variation. Genetic variants include extreme changes that are

responsible for single gene disorders, such as sickle cell anemia or cystic fibrosis. At the other extreme are genetic variations that are of no medical significance but that account for differences, such as hair or eye color. In between are the variants that account for risk for common disorders, for example hypertension or diabetes. These variants occur with different frequencies in different populations, accounting in part for population-based predilections for certain disorders.

Allele frequency and polymorphism

The relationship between allele and genotype frequency is given by the Hardy-Weinberg equation. For a two-allele system, A and a, if the frequency of A = p and of a = q, then the frequency of the genotype AA = p^2 , Aa = 2pq, and aa = q^2 . The Hardy-Weinberg equation can be used to calculate allele frequency from disease frequency, which is useful in genetic counseling. For example, cystic fibrosis has a disease frequency of 1:2,500 in individuals of northern European ancestry. The frequency of the disease-causing allele, q, is therefore the square root, 1/50, and the carrier frequency is 2pq, or approximately 1/25.

As has been noted already, sequence variation is common in the human genome. A locus with multiple alleles in which the frequency of the least common is 1% or greater is said to be *polymorphic*. There are many different types of polymorphisms in the human genome. Variations of an individual DNA base, single nucleotide polymorphisms (SNPs), have been mentioned already. Other types of polymorphisms include variations in the number of repeated sequences: either blocks of DNA tens or hundreds bases in length or simple repeats of 2–4 base units. The latter have been useful in genetic linkage studies, but the polymorphisms usually are not of functional significance. SNPs on the other hand may indeed have functional significance when they reside within the coding sequences of genes.

Mutation and selection

Genetic variation is the outcome of the gradual accumulation of mutations in the human genome over the course of evolution. Some of these changes have no impact, especially if they occur in regions of the genome that do not encode protein. At the opposite extreme, some are lethal and therefore are not passed to the next generation. It might seem that this fate would await most changes that have a deleterious effect, yet clearly many unfavorable changes persist within the genome. There are several reasons for this.

First is the striking of a balance between mutation and selection. Although mutation rates tend to be low, at low allele frequencies the loss caused by selection may be counterbalanced by replacement of mutant alleles by new mutation. Second, under certain circumstances the deleterious allele might confer an advantage. In some cases the advantage might have occurred sometime in the past, in a different era and under different

environmental conditions than exist today. This may have been the case, for example, for the mutations that cause hemochromatosis. At some time and place in the past there may have been an advantage to increased levels of iron absorption, but today this leads to iron overload and tissue toxicity. For recessive traits, the selective advantage might apply to heterozygotes, with the homozygotes for either genotype being at a relative disadvantage. This is the case, for example, with hemoglobin disorders such as sickle cell anemia or thalassemia. Although homozygotes for the globin mutation suffer the effects of severe anemia, compared with heterozygotes, wild-type homozygotes are at increased risk for severe malaria.

Some genetic variants occur more commonly in specific populations. This may be caused by local environmental conditions that favor particular alleles. Another reason might be a phenomenon known as the *founder effect* (Fig. 9). If a mutant allele is introduced into a small population in which breeding occurs only among members of that population for many generations, that mutation will have a higher frequency than in other groups. This accounts for a high frequency of particular mutations in groups such as Ashkenazi Jews, Amish, northern European groups, for example.

Genetic linkage and association

Relating the 30,000 or more genes in the human genome to specific effects on health and disease remains a significant task. Nevertheless the information available from the human genome sequence has provided a critical resource. Two major approaches used are *genetic linkage* and association.

Linkage studies are based on the fact genes are arrayed in a linear order on the chromosomes, an order that is the same from individual to individual. A specific pair of alleles in two closely linked genes tends to segregate together from generation to generation unless a *crossover* occurs at meiosis between the loci (Fig. 10). The frequency of such crossover events is a function of the distance between the loci, so measuring *recombination rates* between adjacent loci provides a kind of genetic map. The human gene map is now densely populated with polymorphic loci whose precise position is known. These polymorphisms can be used to identify a locus that segregates in a family together with a genetic disorder caused by mutation in a specific gene. Such an analysis results in *mapping* the disease gene, from which information the disease gene itself can be identified. When this approach began to be used in the mid-1980s, the process of searching for a disease gene even after it had been mapped was laborious, but now the effort is substantially easier using the resources of the human genome sequence. A trickle of reports of newly discovered genes responsible for disease has increased now to a flood.

The linkage approach works well if a single gene contributes in a major way to a particular disorder, but it is less well suited to discovery of genes

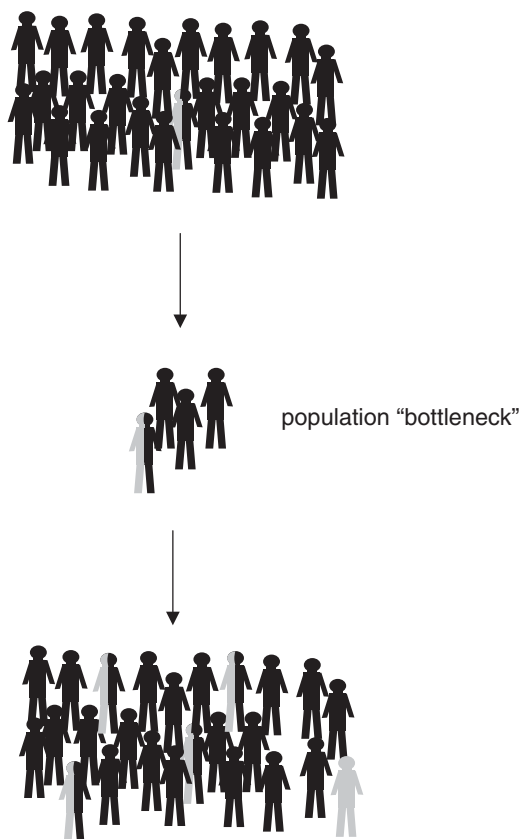


Fig. 9. The founder effect involves the relative increase in frequency of a rare gene that occurs when the population goes through a “bottleneck.” Among the small number of “founders,” by chance the mutation is represented with disproportionately high frequency.

that make more modest contributions to multifactorial disorders. Here, the approach of genetic association analysis is used (Fig. 11). At present this involves testing specific “candidate” polymorphisms—variants in genes that are believed to be likely candidates for playing a role in a particular disease process. The testing often involves a case-control study in which the frequency of a specific allele is compared between populations of affected and nonaffected individuals. If the allele is found more often in affected individuals, it is implicated as being associated with the disease. This could mean that the allele itself contributes to the cause of disease or that it is located nearby another allele that is the causative factor. This is referred to as *linkage disequilibrium*. Association studies have the potential to dissect the genetic contributions to common disorders, but the process is a challenging one. Individual alleles may make small contributions, requiring large study groups, and there may be differences in the genetic contributions

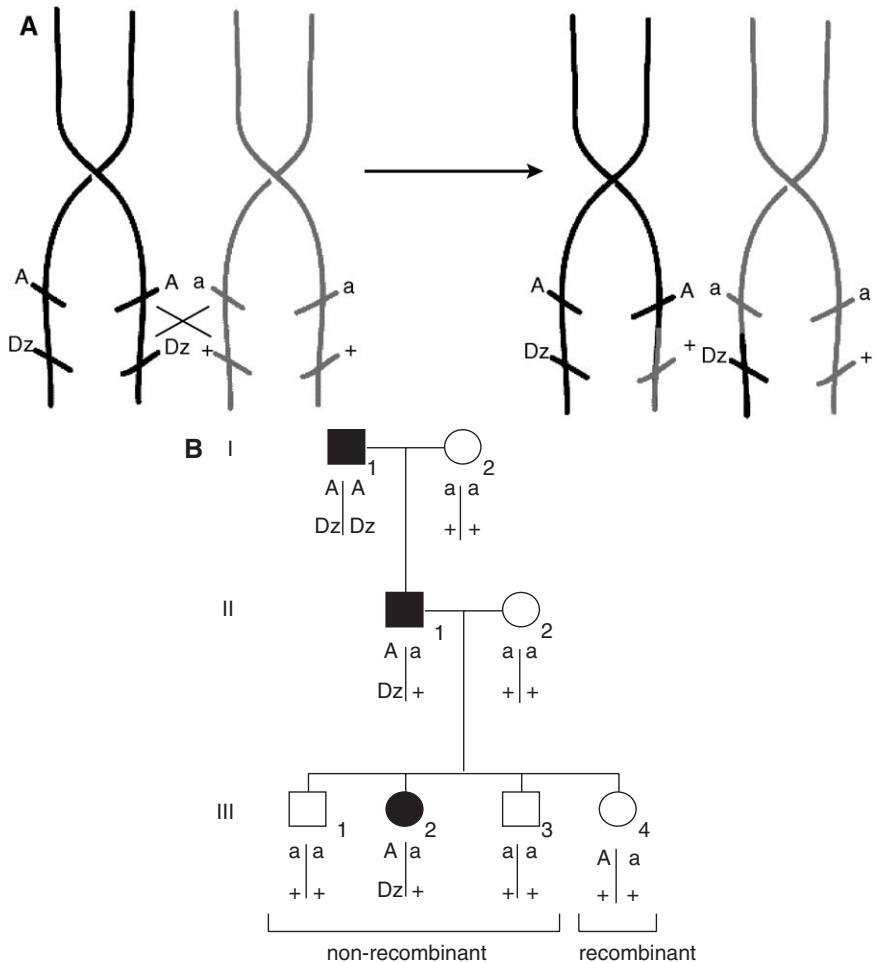


Fig. 10. (A) A pair of closely linked genes (marker A and a disease gene, designated Dz for the mutant allele and + for the normal wild-type) may undergo recombination if a crossover occurs between the two during meiosis. (B) Genetic linkage analysis involves inspection of families for evidence of recombination. In this family, individual II-1 inherited the A allele and the disease from his father. Offspring III-4 is recombinant, because she inherited the A allele from her father together with the + allele. The frequency with which such recombination events occurs provides a measure of the genetic distance between the loci.

to disease in different groups. Also, at present the approach requires making an educated guess as to the alleles most likely to be associated with disease. It would be ideal if there were a way to search the entire genome for association of any allele with a disease, in which case it would be possible to identify previously unsuspected associations. For now the requirements for genotyping the large numbers of SNPs in a large population are prohibitively expensive, but efforts are underway to reduce the cost of this

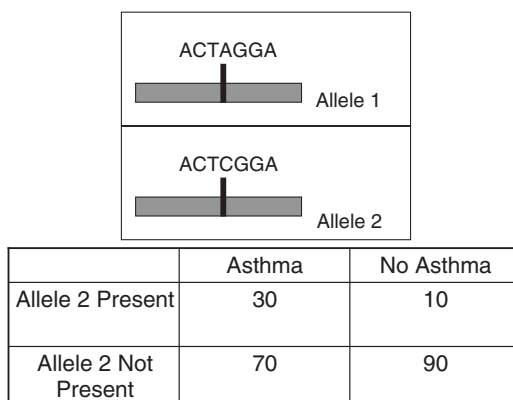


Fig. 11. In an association study, one measures the frequency with which a particular single nucleotide polymorphism (the two alleles, as shown, differ in a single base, either A or C on one strand) occurs in affected or nonaffected individuals. In this example, allele 2 is disproportionately represented in individuals with asthma, supporting the hypothesis of an association of allele 2 with this disorder.

analysis, with the potential of major payoffs in understanding the genetics of common disorders.

Sources of information

The genome is in essence a cache of information, and therefore genomics can be conceptualized as a kind of information science. It is clear that the 30,000 or so genes, the 100,000 or more proteins, and the myriad interactions of genes, proteins, metabolites, and the like comprise a data set of staggering complexity. Putting all this information together to understand the structure and function of cells, tissue, organs, and organisms has required launching a new discipline of bioinformatics. From the clinical point of view things are no simpler. Keeping track of clinical information and using the resources of genetics and genomics for clinical decision making requires access to genetic information at the point of care. Some of the major internet resources that provide public access to genetic information are described in this section.

OMIM

OMIM is an acronym for *Online Mendelian Inheritance in Man*. OMIM began in the 1960s as a database of human genetic traits and genes and for many years was published in book form. It is now freely available on the internet at <http://www.ncbi.nlm.nih.gov/omim/> and contains more than 10,000 entries. Entries include human genetic traits, including single gene disorders, and individual genes. There is an extensive referenced review of

essential data, together with a clinical synopsis for medical disorders. OMIM also includes links to databases of genetic sequence information, protein structure, and so on, for any gene that has been identified. Major mutations also are listed. OMIM is searchable by specific clinical features, so it is possible to obtain a genetic differential diagnosis using the search feature.

GeneReviews

GeneReviews began as a database of genetic testing laboratories but has expanded to include a set of reviews of specific genetic disorders. The laboratory directory also can be reached at www.genetests.org. It is searchable according to particular disorders or laboratories and is best used to identify a clinical or research laboratory that offers testing for a particular genetic disorder. The GeneReviews component provides a set of peer-reviewed articles describing diagnosis, testing, management, and counseling for a wide variety of genetic disorders. It also provides an illustrated glossary. Terms *italicized* in this article are defined in the glossary at www.genetests.org.

Summary

Genetics is moving rapidly from a position at the fringes of medical practice toward the center. Although the transformation will be gradual, physicians will be using the tools of genetics and genomics at virtually every phase of the medical decision-making process. Some familiarity with the basic concepts of genetics is therefore essential to insure that these tools are used well and wisely. There is also a need for access to point-of-care information and decision-making support. Although genetics is likely to be integrated into medical education for the current generation of students, there remains a significant challenge in providing up-to-date information and understanding to physicians currently in practice who will be witness to major advances during the balance of their careers.