# Medical Genetics

JORDE | CAREY | BAMSHAD

SIXTH EDITION

ELSEVIER

eBooks for Practicing Clinicians

ELSEVIER

# Any screen. Any time. Anywhere.

Activate the eBook version
of this title at no additional charge.

Elsevier eBooks for Practicing Clinicians gives you the power to browse and search content, view enhanced images, highlight and take notes—both online and offline.

## Unlock your eBook today.

1. Visit **expertconsult.inkling.com/redeem**
2. Scratch box below to reveal your code
3. Type code into "Enter Code" box
4. Click "Redeem"
5. Log in or Sign up
6. Go to "My Library"

## It's that easy!

Place Peel Off Sticker Here

For technical assistance:
email expertconsult.help@elsevier.com
call 1-800-401-9962 (inside the US)
call +1-314-447-8300 (outside the US)

# Medical Genetics

SIXTH EDITION

**Lynn B. Jorde, PhD**
Mark and Kathie Miller Presidential Professor and Chair
Department of Human Genetics
University of Utah Health Sciences Center
Salt Lake City, Utah

**John C. Carey, MD, MPH**
Professor
Division of Medical Genetics
Department of Pediatrics
University of Utah Health Sciences Center
Salt Lake City, Utah

**Michael J. Bamshad, MD**
Professor and Chief, Division of Genetic Medicine
Allan and Phyllis Treuer Endowed Chair in Genetics and Development
Department of Pediatrics
University of Washington School of Medicine
Seattle Children's Hospital
Seattle, Washington

1600 John F. Kennedy Blvd.
Ste 1800
Philadelphia, PA 19103-2899

MEDICAL GENETICS, SIXTH EDITION ISBN: **978-0-323-59737-1**


**Notice**

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds or experiments described herein. Because of rapid advances in the medical sciences, in particular, independent verification of diagnoses and drug dosages should be made. To the fullest extent of the law, no responsibility is assumed by Elsevier, authors, editors or contributors for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

**Library of Congress Control Number: 2019946145**

*Content Strategist:* Alexandra Mortimer
*Content Development Specialist:* Laura Klein and Humayra Khan
*Publishing Services Manager:* Deepthi Unni
*Senior Project Manager:* Manchu Mohan
*Design Direction:* Bridget Hoette

Printed in China

Last digit is the print number: 9 8 7 6 5 4 3 2 1

To Our Families

*Debra, Eileen,* and *Alton Jorde*
*Leslie, Patrick,* and *Andrew Carey*
*Dana, Marlowe, Jerry,* and *Joanne Bamshad*

# Foreword

J.B.S. Haldane (1951) titled an anthology of some of his more dyspeptic writings, "Everything Has a History," an idea not entirely original with Haldane but clearly applicable to the field of medical genetics. More than 200 years ago scientists such as Buffon, Lamarck, Goethe, and Kielmeyer reflected on how the developmental history of each organism related to the history of life on Earth. Based on these ideas, the discipline of biology was born in 18th century Europe, enjoyed adolescence as morphology and comparative anatomy in the 19th century, and reached adulthood in the 20th century as the field of genetics. However, the late 19th century definition of genetics (heredity) as the science of variation (and its causes) is still valid. Thus, human genetics is the science of human variation, medical genetics the science of abnormal human variation, and clinical genetics that branch of medicine that cares for individuals and families with abnormal variation of structure and function.

Variability and its pertinence to evolution became the central pre-occupation of biology during the 19th century, lasting unabated till now, beginning with Galton (1877, 1889) and *continuous* variation and Bateson (1894) and *discontinuous* variation. After 1900 the term variation or variability was effortlessly incorporated into phenotype analysis. But long before the quantitative genetics of the anthropologists and population geneticists, physicians struggled with the familial occurrence or transmission of evidently hereditary conditions such as hemophilia, red-headedness, or polydactyly, and familial occurrence of malformations such as cleft palate or spina bifida. Now, almost 120 years after the rediscovery of Mendel's work, we have available single-cell, type-specific transcriptomics to test phylogenetic hypotheses, and to investigate changes in splicing and transcription over the life of the organism. Medical genetics likewise has become effective, productive, even enjoyable, and diagnostically more reliable in virtually all respects from metabolic-biochemical, and cytogenetic to molecular levels.

In 1945 the University of Utah established the Laboratory for the Study of Hereditary and Metabolic Disorders (later called the Laboratory of Human Genetics). Here, an outstanding group of scientists performed pioneering studies on cleft lip and palate, muscular dystrophy, albinism, deafness, hereditary polyposis of the colon (Gardner syndrome), and familial breast cancer. These predecessors would be enormously proud of their current peers at the University of Utah, whose successes have advanced knowledge in every aspect of the field of genetics.

In their attempts to synthesize the story of genetics and its applications to human variability, health and disease, development, and cancer, the authors of this text have succeeded admirably. This concise, well-written and -illustrated, carefully edited and indexed book is highly recommended to undergraduate students, new graduate students, medical students, genetic counseling students, nursing students, and students in the allied health sciences. Importantly, it is also a wonderful text for practicing physicians (primary care providers and specialists) who want an authoritative introduction to the basis and principles of modern genetics as applied to human health and development. This text, by distinguished and internationally respected colleagues and friends who love to teach, is a joy to read in its expression of enthusiasm and of wonder, which Aristotle said was the beginning of all knowledge.

Einstein once said, "the most incomprehensible thing about the world is that it is comprehensible." When I began to work in the field of medical genetics, the gene was widely viewed as incomprehensible. Indeed, some scientists, such as Goldschmidt, cast doubt on the very existence of the gene, although the great American biologist E.B. Wilson had predicted its chemical nature more than 100 years previously. During my lifetime, genetics, the core of medicine, entered more or less effectively all of its disciplines, enabling students of Hippocrates to also become students of Mendel. In this text, genes and their function in health and disease are made comprehensible in a manner that should have wide appeal. It is a pleasure to send it on its way in this 6th edition.

***John Opitz, MD***
***Salt Lake City, Utah***

# Preface

Medical genetics is a rapidly progressing field. No textbook can remain factually current for long, so we have attempted to emphasize the central principles of genetics and their clinical application. In particular, this textbook integrates recent developments in molecular genetics and genomics with clinical practice.

This new edition maintains the format and presentation that were well received in five previous editions. Basic principles of molecular biology are introduced early in the book so that they can be discussed and applied in subsequent chapters. The chapters on autosomal and X-linked disorders include updated discussions of topics such as genomic imprinting, anticipation, and germline and somatic mosaicism. The chapter on cytogenetics highlights important advances in this area, including cytogenomic microarrays and newly described microdeletion syndromes. Whole-genome DNA sequencing and disease-gene identification, which constitute a central focus of modern medical genetics, are treated at length. Chapters are included on the rapidly developing fields of immunogenetics and cancer genetics. Considerable discussion is devoted to the genetics of common adult diseases, such as heart disease, diabetes, stroke, and hypertension. The estimation of polygenic risk scores for common diseases is discussed, as is the potential for better delineating underlying causal genes. The book concludes with chapters on genetic diagnosis (again emphasizing current approaches such as noninvasive prenatal testing and whole-exome and whole-genome sequencing), gene therapy, personalized medicine, and clinical genetics and genetic counseling.

As in previous editions, this book can also be accessed electronically through the studentconsult website (https://studentconsult.inkling.com) and there is a pin code in the front of the book to allow you access to this version of the book. Along with the digital version of the textbook you will find hyperlinks to relevant sites, and a battery of text questions and answers.

Several pedagogical aids are incorporated in this book:

- Clinical Commentary boxes present detailed coverage of the most important genetic diseases and provide examples of modern clinical management.
- Mini-summaries highlighted in gold are placed on nearly every page to help the reader understand and summarize important concepts.
- Study questions provided at the end of each chapter assist the reader in review and comprehension.
- A detailed glossary is included at the end of the book.
- Key terms are emphasized in boldface.
- Important up-to-date suggested readings are listed at the end of each chapter.

In addition, the sixth edition retains these important features:

- All chapters have been thoroughly updated, with special attention given to rapidly changing topics such as high-throughput DNA sequencing, genetic diagnosis, gene therapy, cancer genetics, and the genetics of common diseases.
- To facilitate the creation of illustrations for teaching purposes, all images on the website (including line drawings from the textbook) can be downloaded via the evolve book site (http://evolve.elsevier.com/Jorde/).
- An expanded comprehensive index includes all text citations of all diseases.

This textbook evolved from courses we teach for medical students, nursing students, genetic counseling students, and graduate and undergraduate students in human genetics. These students are the primary audience for this book, but it should also be useful for house staff, physicians, and other health-care professionals who wish to become more familiar with medical genetics.

# Acknowledgments

Many of our colleagues have generously donated their time and expertise in reading and commenting on portions of this book. We extend our sincere gratitude to Erica Andersen, PhD; Diane Bonner, PhD; Arthur Brothman, PhD; Peter Byers, MD; William Carroll, MD; Karin Chen, MD; Jessica Chong, PhD; Debbie Dubler, MS; Ruth Foltz, MS; Ron Gibson, MD, PhD; Susan Hodge, PhD; Rajendra Kumar-Singh, PhD; James Kushner, MD; Christina Lam, MD; Claire Leonard, MD; William McMahon, MD; Lawrence Merritt, MD; Dan Miller, MD, PhD; Sampath Prahalad, MD; Gary Schoenwolf, PhD; Leslie R. Schover, PhD; Sarah South, PhD; Maxine J. Sutcliffe, PhD; Carl Thummel, PhD; Thérèse Tuohy, PhD; Scott Watkins, MS; John Weis, PhD; and H. Joseph Yost, PhD. In addition, a number of colleagues provided photographs; they are acknowledged individually in the figure captions. We wish to thank Peeches Cedarholm, RN; Karin Dent, MS; Bridget Kramer, RN; and Ann Rutherford, BS, for their help in obtaining and organizing the photographs. The karyotypes in Chapter 6 were provided by Arthur Brothman, PhD, Erica Andersen, PhD, and Bonnie Issa, BS. We would also like to thank Tara Wenger for her work on the cases included in the online version of this text.

Our editor at Elsevier, Humayra Khan, offered ample encouragement and understanding.

Finally, we wish to acknowledge the thousands of students with whom we have interacted during the past three decades. Teaching involves communication in both directions, and we have undoubtedly learned as much from our students as they have learned from us.

***Lynn B. Jorde***
***John C. Carey***
***Michael J. Bamshad***

# Contents

# 1 Background and History

Genetics is playing an increasingly important role in the practice of clinical medicine. Medical genetics, once largely confined to relatively rare conditions seen by only a few specialists, is now becoming a central component of our understanding of most major diseases. These include not only the pediatric diseases, but also common adult diseases such as heart disease, diabetes, many cancers, and many psychiatric disorders. Because all components of the human body are influenced by genes, genetic disease is relevant to all medical specialties. Today's health-care practitioners must understand the science of medical genetics.

## What Is Medical Genetics?

Medical genetics involves any application of genetics to medical practice. It thus includes studies of the inheritance of diseases in families, mapping of disease genes to specific locations on chromosomes, analyses of the molecular mechanisms through which genes cause disease, and the diagnosis and treatment of genetic disease. As a result of rapid progress in molecular genetics, DNA-based diagnosis is available for several thousand inherited conditions, and gene therapy—the alteration of genes to correct genetic disease—is now effective for some conditions. Medical genetics also includes genetic counseling, in which information regarding risks, prognoses, and treatments is communicated to patients and their families.

## Why Is a Knowledge of Medical Genetics Important for Today's Health-Care Practitioner?

There are several reasons health-care practitioners must understand medical genetics. Genetic diseases make up a large percentage of the total disease burden in pediatric and adult populations (Table 1.1). This percentage will continue to grow as our understanding of the genetic basis of disease grows. In addition, modern medicine is placing increasing emphasis on prevention. Because genetics provides a basis for understanding the fundamental biological makeup of the organism, it naturally leads to a better understanding of the disease process. In some cases, this knowledge can lead to prevention of the disorder. It also leads to more effective disease treatment. Prevention and effective treatment are among the highest goals of medicine. The chapters that follow provide many examples of the ways genetics contributes to these goals. But first, this chapter reviews the foundations upon which current practice is built.

## A Brief History

The inheritance of physical traits has been a subject of curiosity and interest for thousands of years. The ancient Hebrews and Greeks, as well as later medieval scholars, described many genetic phenomena and proposed theories to account for them. Many of these theories were incorrect. Gregor Mendel (Fig. 1.1), an Austrian monk who is usually considered the father of genetics, advanced the field significantly by performing a series of cleverly designed experiments on living organisms (garden peas). He then used this experimental information to formulate a series of fundamental principles of heredity.

Mendel published the results of his experiments in 1865 in a relatively obscure journal. It is one of the ironies of biological science that his discoveries, which still form the foundation of genetics, received little recognition for 35 years. At about the same time, Charles Darwin formulated his theories of evolution, and Darwin's cousin, Francis Galton, performed an extensive series of family studies (concentrating especially on twins) in an effort to understand the influence of heredity on various human traits. Neither scientist was aware of Mendel's work.

Genetics as it is known today is largely the result of research performed during the 20th century. Mendel's principles were independently rediscovered in 1900 by three different scientists working in three different countries. This was also the year in which Landsteiner discovered the ABO blood group system. In 1902, Archibald Garrod described alkaptonuria as the first "inborn error of metabolism." In 1909, Johannsen coined the term **gene** to denote the basic unit of heredity.

The next several decades were a period of considerable experimental and theoretical work. Several organisms, including *Drosophila melanogaster* (fruit flies) and *Neurospora crassa* (bread mold) served as useful experimental systems in which to study the actions and interactions of genes. For example, H. J. Muller demonstrated the genetic consequences of ionizing radiation in the fruit fly. During this period, much of the theoretical basis of population genetics was developed by three central figures: Ronald Fisher, J. B. S. Haldane, and Sewall Wright. In addition, the modes of inheritance of several important genetic diseases, including phenylketonuria, sickle cell disease, Huntington disease, and cystic fibrosis, were established. In 1944, Oswald Avery showed that genes are composed of **deoxyribonucleic acid (DNA).**

Probably the most significant achievement of the 1950s was the specification of the physical structure of DNA by James Watson and Francis Crick in 1953. Their seminal paper, which was only one page long, formed the basis for

what is now known as **molecular genetics** (the study of the structure and function of genes at the molecular level). Another significant accomplishment in that decade was the correct specification of the number of human chromosomes. Since the early 1920s it had been thought that humans had 48 chromosomes in each cell. Only in 1956 was the correct number, 46, finally determined. The ability to count and identify chromosomes led to a flurry of new findings in cytogenetics, including the discovery in 1959 that Down syndrome is caused by an extra copy of chromosome 21.

Technological developments since 1960 have brought about significant achievements at an ever-increasing rate. The most spectacular advances have occurred in the field of molecular genetics. Thousands of genes have been mapped to specific chromosome locations. The Human Genome Project, a large collaborative venture begun in 1990, provided a nearly complete human DNA sequence in 2003 (the term **genome** refers to all of the DNA in an organism). Important developments in computer technology have helped to decipher the barrage of data being generated by this and related projects. In addition to mapping genes,

**TABLE 1.1** A Partial List of Some Important Genetic Diseases

| Disease | Approximate Prevalence |
|---|---|
| **CHROMOSOME ABNORMALITIES** | |
| Down syndrome | 1/700 to 1/1000 |
| Klinefelter syndrome | 1/1000 males |
| Trisomy 13 | 1/10,000 |
| Trisomy 18 | 1/6000 |
| Turner syndrome | 1/2500 to 1/10,000 females |
| **SINGLE-GENE DISORDERS** | |
| Adenomatous polyposis coli | 1/6000 |
| Adult polycystic kidney disease | 1/1000 |
| α-1 Antitrypsin deficiency | 1/2500 to 1/10,000 (whites)* |
| Cystic fibrosis | 1/2000 to 1/4000 (whites) |
| Duchenne muscular dystrophy | 1/3500 males |
| Familial hypercholesterolemia | 1/500 |
| Fragile X syndrome | 1/4000 males; 1/8000 females |
| Hemochromatosis (hereditary) | 1/300 whites are homozygotes; approximately 1/1000 to 1/2000 are affected |
| Hemophilia A | 1/5000 to 1/10,000 males |
| Hereditary nonpolyposis colorectal cancer | Up to 1/200 |
| Huntington disease | 1/20,000 (whites) |
| Marfan syndrome | 1/10,000 to 1/20,000 |
| Myotonic dystrophy | 1/7000 to 1/20,000 (whites) |
| Neurofibromatosis type 1 | 1/3000 to 1/5000 |
| Osteogenesis imperfecta | 1/5000 to 1/10,000 |
| Phenylketonuria | 1/10,000 to 1/15,000 (whites) |
| Retinoblastoma | 1/20,000 |
| Sickle cell disease | 1/400 to 1/600 blacks* in America; up to 1/50 in central Africa |
| Tay-Sachs disease | 1/3000 Ashkenazi Jews |
| Thalassemia | 1/50 to 1/100 (South Asian and circum-Mediterranean populations) |
| **MULTIFACTORIAL DISORDERS** | |
| **Congenital Malformations** | |
| Cleft lip with or without cleft palate | 1/500 to 1/1000 |
| Club foot (talipes equinovarus) | 1/1000 |
| Congenital heart defects | 1/200 to 1/500 |
| Neural tube defects (spina bifida, anencephaly) | 1/200 to 1/1000 |
| Pyloric stenosis | 1/300 |
| **Adult Diseases** | |
| Alcoholism | 1/10 to 1/20 |
| Alzheimer disease | 1/10 (Americans older than 65 years) |
| Bipolar disorder | 1/100 to 1/200 |
| Cancer (all types) | 1/3 |
| Diabetes (types 1 and 2) | 1/10 |
| Heart disease or stroke | 1/3 to 1/5 |
| Schizophrenia | 1/100 |
| **MITOCHONDRIAL DISEASES** | |
| Kearns-Sayre syndrome | Rare |
| Leber hereditary optic neuropathy (LHON) | Rare |
| Mitochondrial encephalopathy, lactic acidosis, and stroke-like episodes (MELAS) | Rare |
| Myoclonic epilepsy and ragged red fiber disease (MERRF) | Rare |

*The term "white" refers to individuals of predominantly European descent; the term "black" refers to individuals of predominantly sub-Saharan African descent. These terms are used for convenience; some of the challenges in accurately describing human populations are discussed in Chapter 14.

geneticists have pinpointed the molecular defects underlying several thousand genetic diseases. This research has contributed greatly to our understanding of the ways gene defects can cause disease, opening paths to more effective treatment and potential cures.

## Types of Genetic Diseases

Humans are estimated to have approximately 21,000 protein-coding genes. Alterations in these genes, or in combinations of them, can produce genetic disorders. These disorders are classified into several major groups:

- **Chromosome disorders,** in which entire chromosomes (or large segments of them) are missing, duplicated, or otherwise altered. These disorders include diseases such as Down syndrome and Turner syndrome
- Disorders in which single genes are altered; these are often termed *mendelian conditions,* or **single-gene disorders.** Well-known examples include cystic fibrosis, sickle cell disease, and hemophilia
- **Multifactorial disorders,** which result from a combination of multiple genetic and environmental causes. Many birth defects, such as cleft lip and cleft palate, as well as many adult disorders, including heart disease and diabetes, belong in this category
- **Mitochondrial disorders,** a relatively small number of diseases caused by alterations in the small cytoplasmic mitochondrial chromosome.

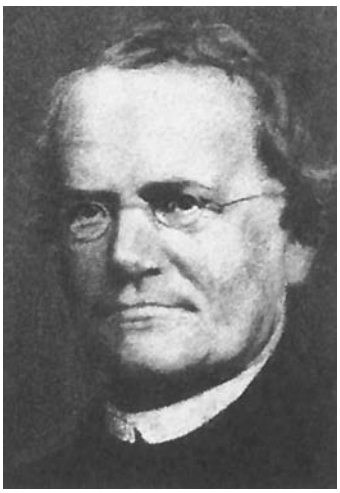

**Fig. 1.1** Gregor Johann Mendel. (From Raven PH, Johnson GB. *Biology.* 3rd ed. St Louis: Mosby; 1992.)

Table 1.1 provides some examples of each of these types of diseases.

Of these major classes of diseases, the single-gene disorders have probably received the greatest amount of attention. These disorders are classified according to the way they are inherited in families: autosomal dominant, autosomal recessive, or X-linked. These modes of inheritance are discussed extensively in Chapters 4 and 5. The first edition of McKusick's *Mendelian Inheritance in Man,* published in 1966, listed only 1368 autosomal traits and 119 X-linked traits. Today the online version of McKusick's compendium lists more than 24,000 genes and traits, of which more than 23,000 are autosomal, more than 1200 are X-linked, 60 are Y-linked, and 68 are in the mitochondrial genome. DNA variants responsible for more than 6000 traits, most of which are inherited diseases, have been identified. With continued advances, these numbers are certain to increase.

Although some genetic disorders, particularly the single-gene conditions, are strongly determined by genes, many others are the result of multiple genetic and nongenetic factors. One can therefore think of genetic diseases as lying along a continuum (Fig. 1.2), with disorders such as cystic fibrosis and Duchenne muscular dystrophy situated at one end (strongly determined by genes) and conditions such as measles situated at the other end (strongly determined by environment). Many of the most prevalent disorders, including many birth defects and many common diseases such as diabetes, hypertension, heart disease, and cancer, lie somewhere in the middle of the continuum. These diseases are the products of varying degrees of genetic and environmental influences.

## The Clinical Impact of Genetic Disease

Less than a century ago, diseases of largely nongenetic causation (i.e., those caused by malnutrition, unsanitary conditions, and pathogens) accounted for the great majority of deaths in children. During the 20th century, however, public health vastly improved in much of the world. As a result, genetic diseases have come to account for an ever-increasing percentage of deaths among children in developed countries. For example, the percentage of pediatric deaths due to genetic causes in various hospitals in the United Kingdom increased from 16.5% in 1914 to 50% in 1976 (Table 1.2).

In addition to contributing to a large fraction of childhood deaths, genetic diseases also account for a large share of admissions to pediatric hospitals. Several surveys have


**Fig. 1.2** Continuum of disease causation. Some diseases (e.g., cystic fibrosis) are strongly determined by genes, whereas others (e.g., infectious diseases) are strongly determined by environment.

**TABLE 1.2** Percentages of Childhood Deaths in United Kingdom Hospitals Attributable to Nongenetic and Genetic Causes

| Cause | London 1914 | London 1954 | Newcastle 1966 | Edinburgh 1976 |
|---|---|---|---|---|
| **NONGENETIC*** | | | | |
| All causes | 83.5 | 62.5 | 58.0 | 50.0 |
| **GENETIC** | | | | |
| Single gene | 2.0 | 12.0 | 8.5 | 8.9 |
| Chromosomal | — | — | 2.5 | 2.9 |
| Multifactorial | 14.5 | 25.5 | 31.0 | 38.2 |

Data from Rimoin DL, Connor JM, Pyeritz RE, Korf BR. *Emery and Rimoin's Principles and Practice of Medical Genetics*. London: Churchill Livingstone; 2007.
*Infections, for example.

shown that between 30% and 70% of hospital admissions involve conditions with substantial genetic components (e.g., Mendelian conditions, chromosome abnormalities, or congenital malformations). It is estimated that Mendelian diseases alone account for ≈20% of infant mortality and ≈10% of pediatric hospitalizations.

Another way to assess the importance of genetic diseases is to ask, "What fraction of persons in the population will be found to have a genetic disorder?" A variety of factors can influence the answer to this question. For example, some diseases are found more frequently in certain ethnic groups. Cystic fibrosis is especially common among those of European heritage, whereas sickle cell disease is especially common among those of African descent. Some diseases are more common in older persons. For example, colon cancer, breast cancer, and Alzheimer disease are caused by dominant genes in a small fraction (5–10%) of cases but are not usually manifested until later in life. The prevalence estimates for these genetic diseases would be higher in an older population. Variations in diagnostic and recording practices can also cause prevalence estimates to vary. Accordingly, the prevalence figures shown in Table 1.3 are given as rather broad ranges. Keeping these sources of variation in mind, it is notable that a recognizable genetic disease will be diagnosed in at least 3% to 7% of the population at some point. This tabulation does not include most cases of the more common adult diseases, such as heart disease, diabetes, and cancer, although it is known that these diseases also have genetic components. If such diseases are included, the clinical impact of genetic disease is considerable indeed.

**TABLE 1.3** Approximate Prevalence of Genetic Disease in the General Population

| Type of Genetic Disease | Lifetime Prevalence per 1000 Persons |
|---|---|
| Autosomal dominant | 3–9.5 |
| Autosomal recessive | 2–2.5 |
| X-linked | 0.5–2 |
| Chromosome disorder | 6–9 |
| Congenital malformation* | 20–50 |
| Total | 31.5–73 |

*Congenital means "present at birth." Most congenital malformations are thought to be multifactorial and therefore probably have both genetic and environmental components.

## Suggested Readings


Baird PA, Anderson TW, Newcombe HB, Lowry RB. Genetic disorders in children and young adults: a population study. *Am J Hum Genet*. 1988;42:677–693.

Bell CJ, Dinwiddie DL, Miller NA, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med*. 2011;3(65):65ra4.

Dunn LC. *A Short History of Genetics*. New York: McGraw-Hill; 1965.

McCandless SE, Brunger JW, Cassidy SB. The burden of genetic disease on inpatient care in a children's hospital. *Am J Hum Genet*. 2004;74:121–127.

McKusick VA, Harper PS. History of medical genetics. In: Rimoin DL, Pyeritz RE, Korf BR, eds. *Emery and Rimoin's Principles and Practice of Medical Genetics*. 6th ed. Academic Press; 2013.

Passarge E. *Color Atlas of Genetics*. 3rd ed. Stuttgart: Georg Thieme Verlag; 2007.

Rimoin DL, Pyeritz RE, Korf BR. *Emery and Rimoin's Principles and Practice of Medical Genetics*. 7th ed. Academic Press; 2013.

Scriver CR, Sly WS, Childs G, et al. *The Metabolic and Molecular Bases of Inherited Disease*. 8th ed. New York: McGraw-Hill; 2001.

Stevenson DA, Carey JC. Contribution of malformations and genetic disorders to mortality in a children's hospital. *Am J Med Genet A*. 2004;126A:393–397.

Watson JD, Crick FHC. Molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature*. 1953;171:737.


### Internet Resources


Dolan DNA Learning Center, Cold Spring Harbor Laboratory (a useful online resource for learning and reviewing basic principles): https://www.dnalc.org/

Genetic Science Learning Center (another useful resource for learning and reviewing basic genetic principles): http://learn.genetics.utah.edu/

Landmarks in the History of Genetics: http://cogweb.ucla.edu/EP/DNA_history.html

National Human Genome Research Institute Educational Resources: http://www.genome.gov/Education

Online Mendelian Inheritance in Man (OMIM) (a comprehensive catalog and description of single-gene conditions): http://www.ncbi.nlm.nih.gov/Omim/

University of Kansas Medical Center Genetics Education Center (a large number of links to useful genetics education sites): http://www.kumc.edu/gec/

# 2 Basic Cell Biology: Structure and Function of Genes and Chromosomes

All genetic diseases involve defects at the level of the cell. For this reason, one must understand basic cell biology to understand genetic disease. Errors can occur in the replication of genetic material or in the translation of genes into proteins. Such errors commonly produce single-gene disorders. In addition, errors that occur during cell division can lead to disorders involving entire chromosomes. To provide the basis for understanding these errors and their consequences, this chapter focuses on the processes through which genes are replicated and translated into proteins, as well as the process of cell division.

In the 19th century, microscopic studies of cells led scientists to suspect that the nucleus of the cell (Fig. 2.1) contains the important mechanisms of inheritance. They found that **chromatin,** the substance that gives the nucleus a granular appearance, is observable in the nuclei of nondividing cells. Just before a cell undergoes division, the chromatin condenses to form microscopically observable, threadlike structures called **chromosomes** (from the Greek words for "colored bodies"). With the rediscovery of Mendel's breeding experiments at the beginning of the 20th century, it soon became apparent that chromosomes contain **genes.** Genes are transmitted from parent to offspring and are considered the basic unit of inheritance. It is through the transmission of genes that physical traits such as eye color are inherited in families. Diseases can also be transmitted through genetic inheritance.

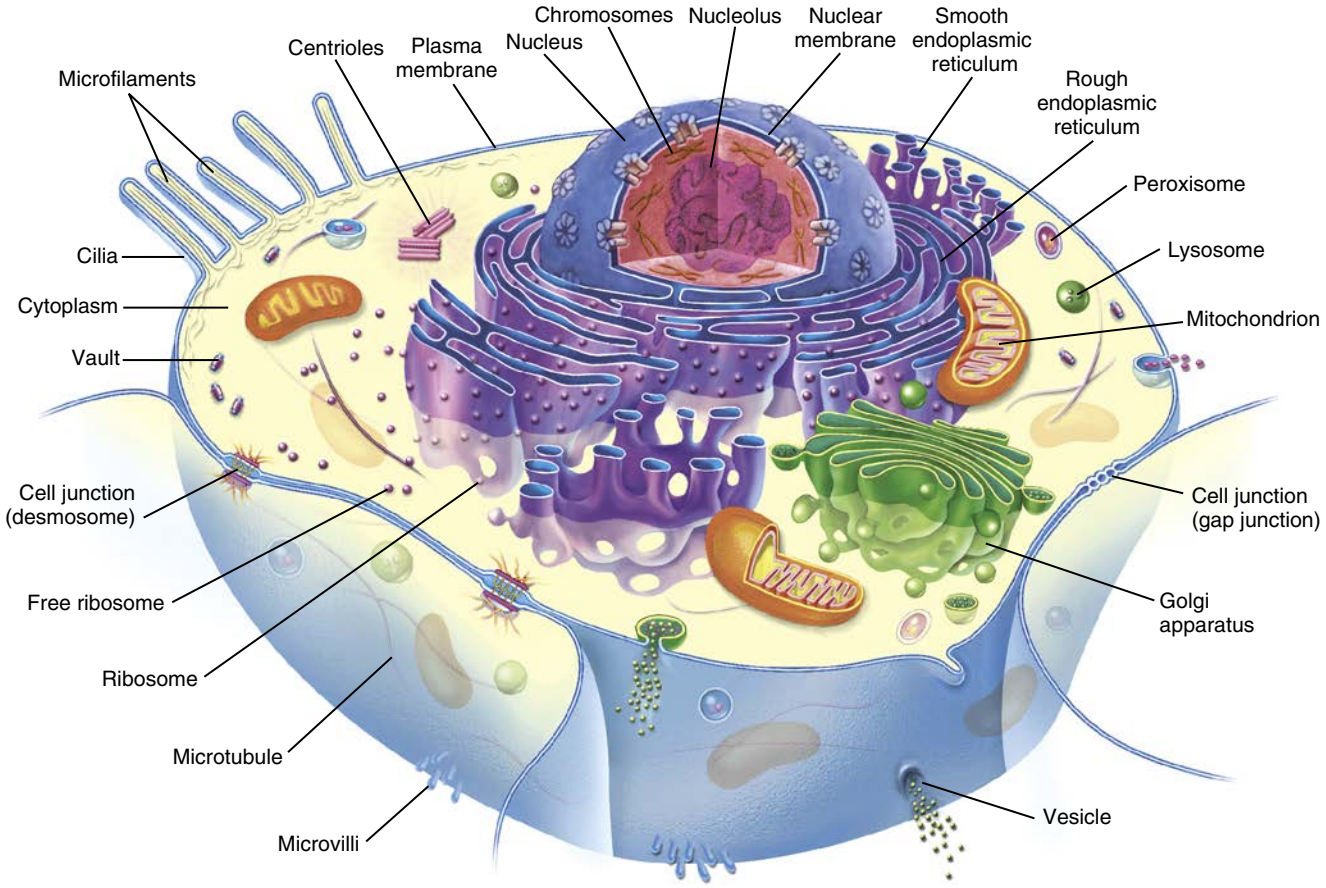


**Fig. 2.1 The anatomy of the cell.** (From McCance KL, Huether SE. *Pathophysiology: The Biologic Basis for Disease in Adults and Children*. 8th ed. St. Louis: Elsevier Mosby; 2019.)

Physically, genes are composed of **deoxyribonucleic acid (DNA).** DNA provides the genetic blueprint for all proteins in the body. Thus genes ultimately influence all aspects of body structure and function. Humans are estimated to have approximately 21,000 genes (sequences of DNA that code for proteins). An error (or **mutation**) in one of these genes often leads to a recognizable genetic disease. In addition, there are thousands of genes that encode a ribonucleic acid (RNA) product but not a protein.

Genes, the basic unit of inheritance, are contained in chromosomes and consist of DNA.

Each human **somatic cell** (cells other than the **gametes,** or sperm and egg cells) contains 23 pairs of different chromosomes, for a total of 46. One member of each pair is derived from the individual's father, and the other member is derived from the mother. One of the chromosome pairs consists of the **sex chromosomes.** In normal males, the sex chromosomes are a Y chromosome inherited from the father and an X chromosome inherited from the mother. Two X chromosomes are found in normal females, one inherited from each parent. The other 22 pairs of chromosomes are **autosomes.** The members of each pair of autosomes are said to be **homologs,** or **homologous,** because their DNA is very similar. The X and Y chromosomes are not homologs of each other.

Somatic cells, having two of each chromosome, are **diploid** cells. Human gametes have the **haploid** number of chromosomes, 23. The diploid number of chromosomes is maintained in successive generations of somatic cells by the process of **mitosis,** whereas the haploid number is obtained through the process of **meiosis.** Both of these processes are discussed in detail later in this chapter.

Somatic cells are diploid, having 23 pairs of chromosomes (22 pairs of autosomes and one pair of sex chromosomes). Gametes are haploid and have a total of 23 chromosomes.

# DNA, RNA, and Proteins: Heredity at the Molecular Level

## DNA

### Composition and Structure of DNA

The DNA molecule has three basic components: the pentose sugar, deoxyribose; a phosphate group; and four types of nitrogenous **bases** (so named because they can combine with hydrogen ions in acidic solutions). Two of the bases, **cytosine** and **thymine,** are single carbon–nitrogen rings called **pyrimidines.** The other two bases, **adenine** and **guanine,** are double carbon–nitrogen rings called **purines** (Fig. 2.2). The four bases are commonly represented by their first letters: C, T, A, and G.

One of the contributions of Watson and Crick in the mid-20th century was to demonstrate how these three components are physically assembled to form DNA. They proposed the now-famous **double helix** model, in which DNA can be envisioned as a twisted ladder with chemical bonds as its rungs (Fig. 2.3). The two sides of the ladder are composed of the sugar and phosphate components, held together by strong phosphodiester bonds. Projecting from each side of the ladder, at regular intervals, are the nitrogenous bases. The base projecting from one side is bound to the base projecting from the other side by relatively weak hydrogen bonds. The paired nitrogenous bases therefore form the rungs of the ladder.

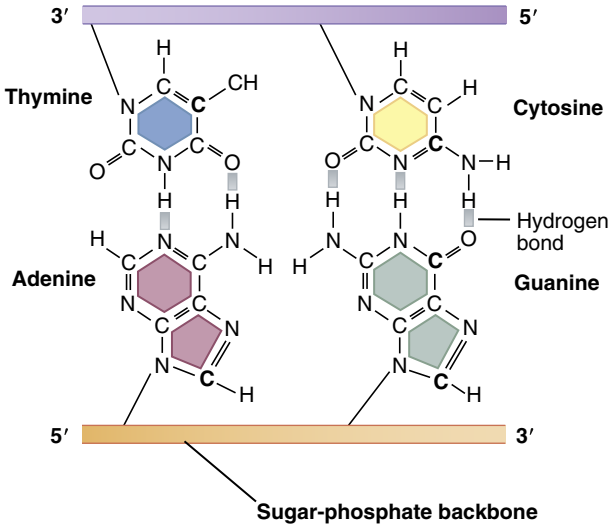


**Fig. 2.2** Chemical structure of the four bases, which shows hydrogen bonds between base pairs. Three hydrogen bonds are formed between cytosine–guanine pairs, and two bonds are formed between adenine–thymine pairs.

Fig. 2.2 illustrates the chemical bonds between bases and shows that the ends of the ladder terminate in either 3′ or 5′. These labels are derived from the order in which the five carbon atoms composing deoxyribose are numbered. Each DNA subunit, consisting of one deoxyribose, one phosphate group, and one base, is called a **nucleotide.**

Different sequences of nucleotide bases (e.g., ACCAAGTGC) specify different proteins. Specification of the body's many proteins must require a great deal of genetic information. Indeed, each haploid human cell contains approximately 3 billion nucleotide pairs, more than enough information to specify the composition of all human proteins.

The most important constituents of DNA are the four nucleotide bases: adenine, thymine, cytosine, and guanine. DNA has a double helix structure.

### DNA Coiling

Textbook illustrations usually depict DNA as a double helix molecule that continues in a long, straight line. However, if the DNA in a cell were actually stretched out in this way, it would be about 2 meters long. To package all of this DNA into a tiny cell nucleus, it is coiled at several levels. First, the DNA is wound around a **histone** protein core to form a **nucleosome** (Fig. 2.4). About 140 to 150 DNA bases are wound around each histone core, and then 20 to 60 bases form a spacer element before the next nucleosome complex. The nucleosomes in turn form

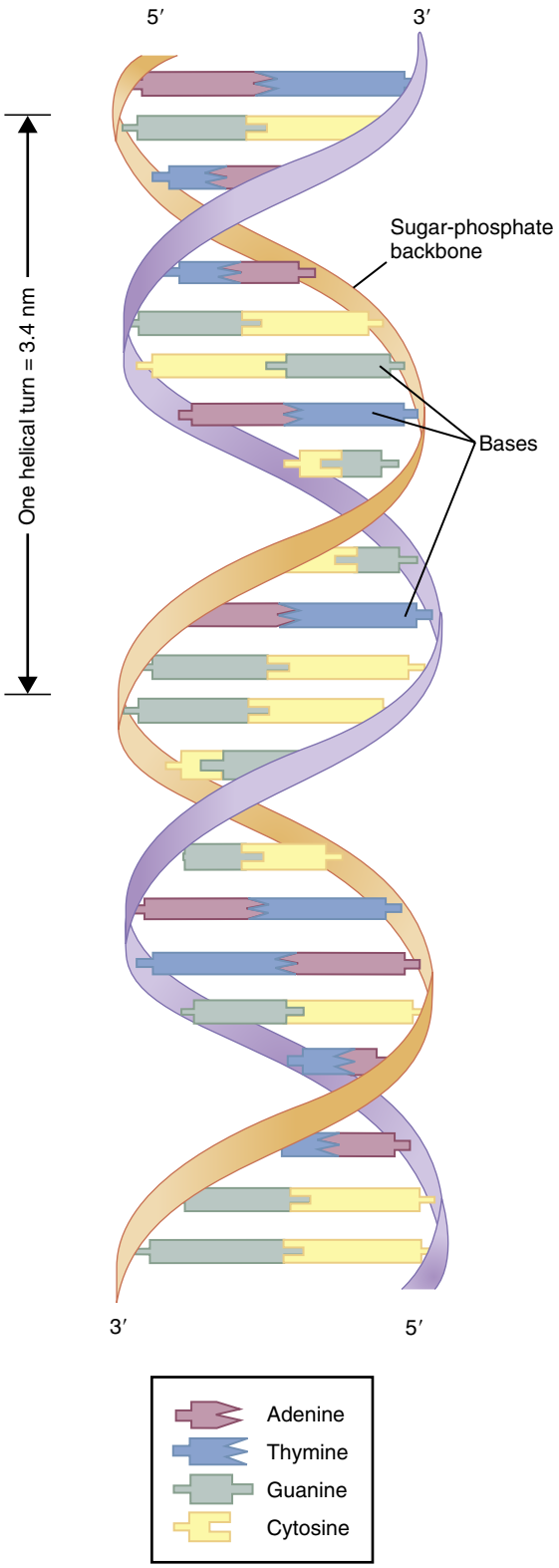


**Fig. 2.3** The DNA double helix, with sugar-phosphate backbone and nitrogenous bases.

a helical **solenoid;** each turn of the solenoid includes about six nucleosomes. The solenoids themselves are organized into **chromatin loops,** which are attached to a protein scaffold. Each of these loops contains approximately 100,000 base pairs (bp), or 100 **kilobases (kb)** of DNA. The end result of this coiling and looping is that the DNA, at its maximum stage of condensation, is only about 1/10,000 as long as it would be if it were fully stretched out.

DNA is a tightly coiled structure. This coiling occurs at several levels: the nucleosome, the solenoid, and 100-kb loops.

## Replication of DNA

As cells divide to make copies of themselves, identical copies of DNA must be made and incorporated into the new cells. This is essential if DNA is to serve as the fundamental genetic material. DNA **replication** begins as the weak hydrogen bonds between bases break, producing single DNA strands with unpaired bases. The consistent pairing of adenine with thymine and guanine with cytosine, known as **complementary base pairing,** is the key to accurate replication. The principle of complementary base pairing dictates that the unpaired base will attract a free nucleotide only if that nucleotide has the proper complementary base. For example, a portion of a single strand with the base sequence ATTGCT will bond with a series of free nucleotides with the bases TAACGA. The single strand is said to be a **template** upon which the complementary strand is built. When replication is complete, a new double-stranded molecule identical to the original is formed (Fig. 2.5).

Several different enzymes are involved in DNA replication. One enzyme unwinds the double helix and another holds the strands apart. Still another enzyme, **DNA polymerase,** travels along the single DNA strand, adding free nucleotides to the 3′ end of the new strand. Nucleotides can be added only to the 3′ end of the strand, so replication always proceeds from the 5′ to the 3′ end. When referring to the orientation of sequences along a gene, the 5′ direction is termed **upstream,** and the 3′ direction is termed **downstream.**

In addition to adding new nucleotides, DNA polymerase performs part of a **proofreading** procedure, in which a newly added nucleotide is checked to make certain that it is in fact complementary to the template base. If it is not, the nucleotide is excised and replaced with a correct complementary nucleotide base. This process substantially enhances the accuracy of DNA replication. When a DNA replication error is not successfully repaired, a mutation has occurred. As will be seen in Chapter 3, many such mutations cause genetic diseases.

DNA replication is critically dependent on the principle of complementary base pairing (A with T; C with G). This allows a single strand of the double-stranded DNA molecule to form a template for the synthesis of a new, complementary strand.

The rate of DNA replication in humans, about 40 to 50 nucleotides per second, is comparatively slow. In bacteria the rate is much higher, reaching 500 to 1000 nucleotides per second. Given that some human chromosomes have as many as 250 million nucleotides, replication would be an extraordinarily time-consuming process if it proceeded linearly from one end of the chromosome to the other: For

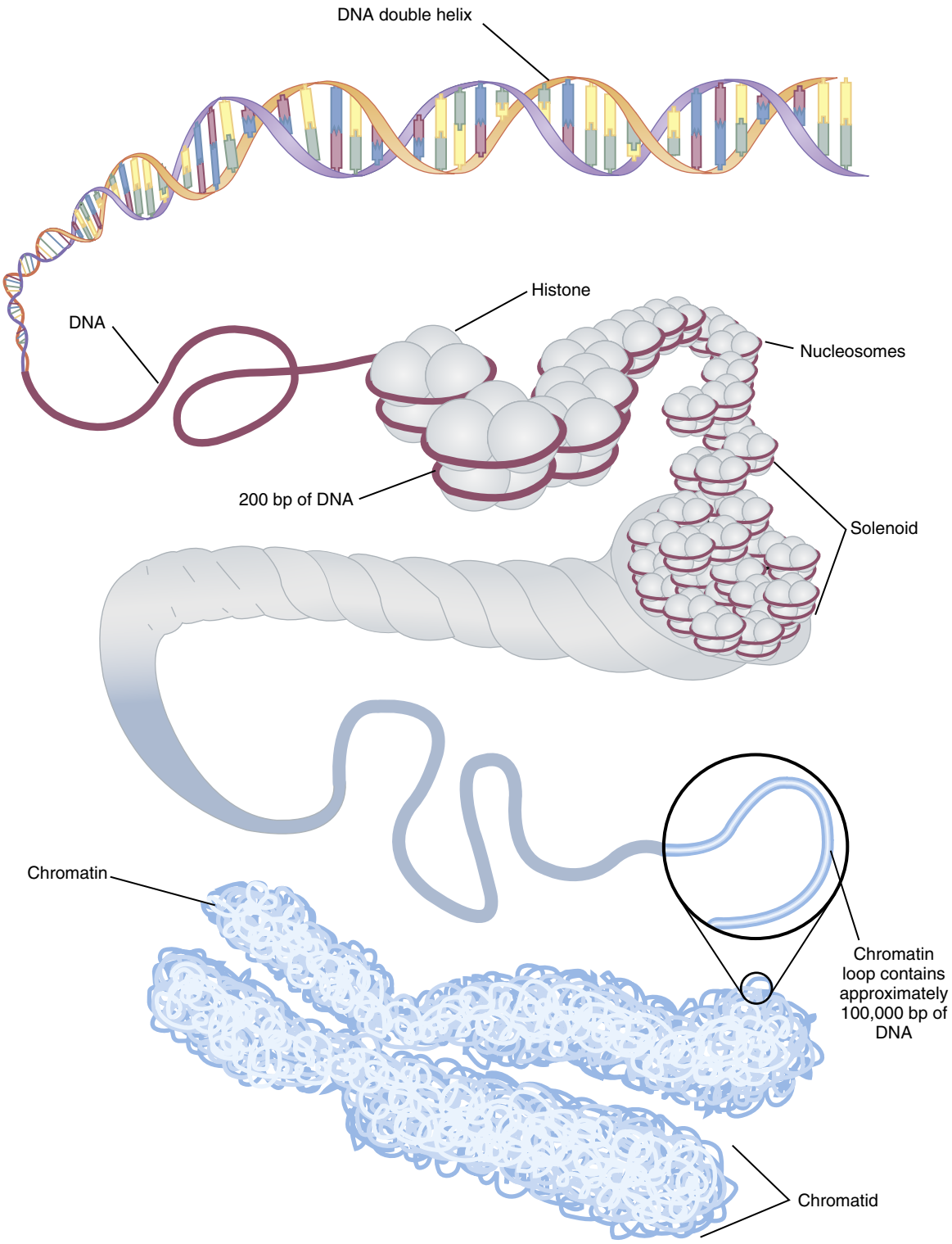


**Fig. 2.4** Patterns of DNA coiling. DNA is wound around histones to form nucleosomes. These are organized into solenoids, which in turn make up the chromatin loops.

a chromosome of this size, a single round of replication would take almost 2 months. Instead, replication begins at many different points along the chromosome, termed **replication origins.** The resulting multiple separations of the DNA strands are called **replication bubbles** (Fig. 2.6). By occurring simultaneously at many different sites along the chromosome, the replication process can proceed much more quickly.

Replication bubbles allow DNA replication to take place at multiple locations on the chromosome, greatly speeding the replication process.

## FROM GENES TO PROTEINS

While DNA is formed and replicated in the cell nucleus, protein synthesis takes place in the cytoplasm. The information

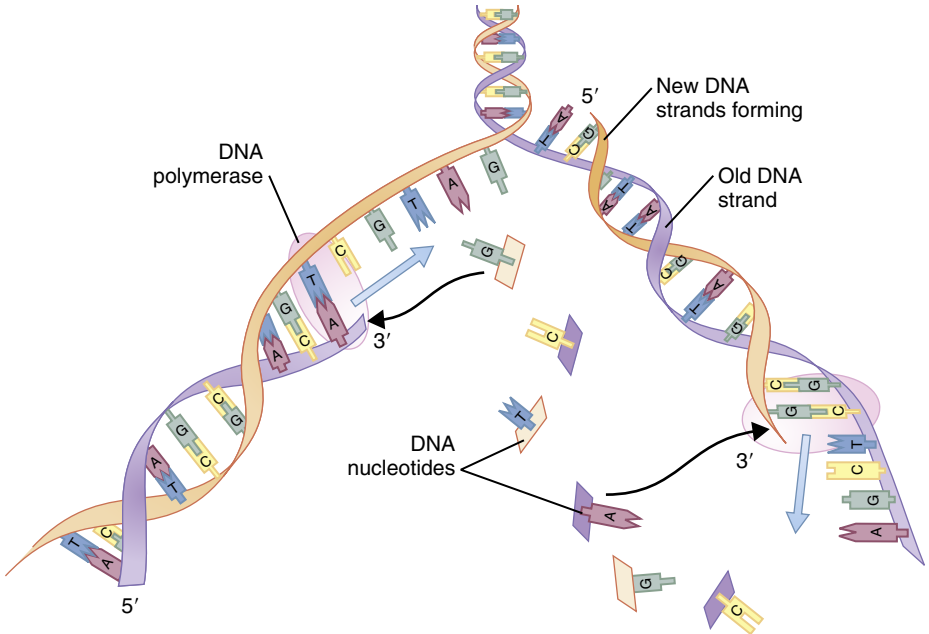


**Fig. 2.5** DNA replication. The hydrogen bonds between the two original strands are broken, allowing the bases in each strand to undergo complementary base pairing with free bases. This process, which proceeds in the 5′ to 3′ direction on each strand, forms two new double strands of DNA.

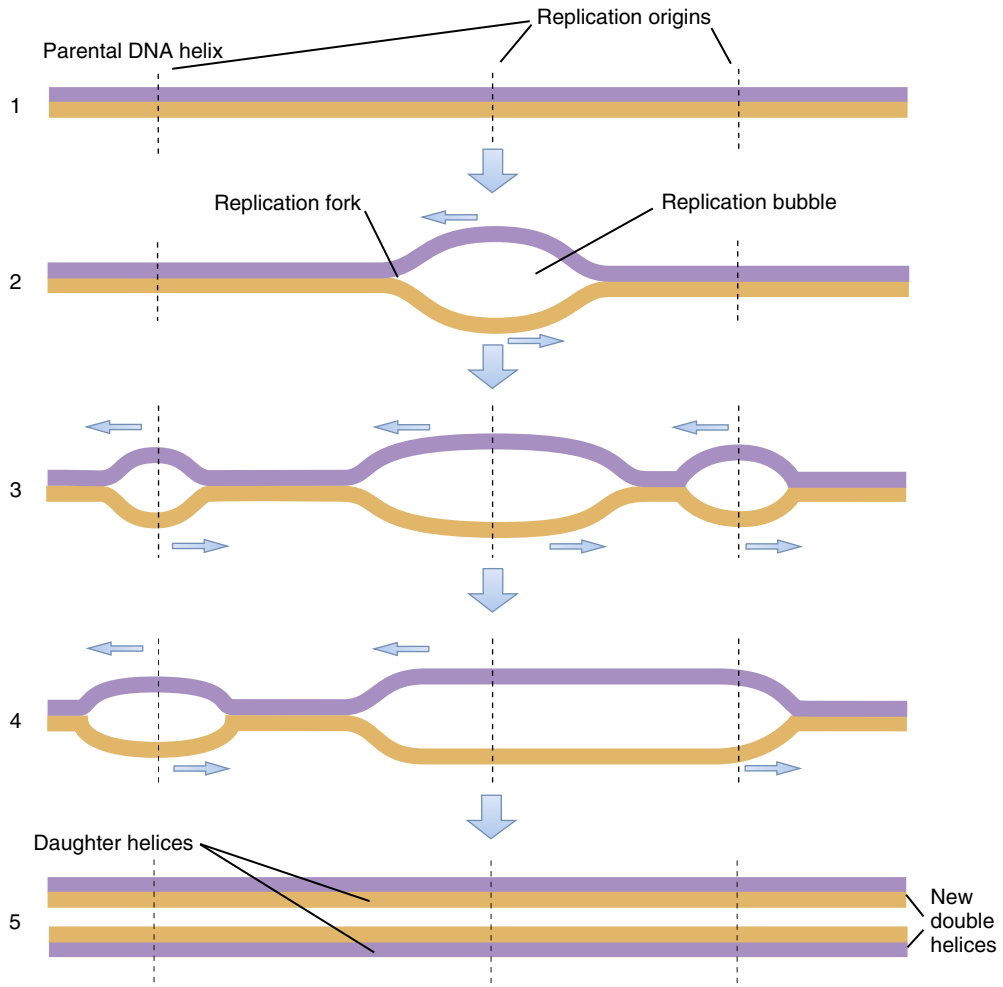


**Fig. 2.6** Replication bubbles form at multiple points along the DNA strand, allowing replication to proceed more rapidly.

contained in DNA must be transported to the cytoplasm and then used to dictate the composition of proteins. This involves two processes, transcription and translation. Briefly, the DNA code is **transcribed** into messenger RNA, which then leaves the nucleus to be **translated** into proteins. These processes, summarized in Fig. 2.7, are discussed at length later in this chapter. Transcription and translation are both mediated by **ribonucleic acid (RNA),** a type of nucleic acid that is chemically similar to DNA. Like DNA, RNA is composed of sugars, phosphate groups, and nitrogenous bases. It differs from DNA in that the sugar is ribose instead of deoxyribose, and uracil rather than thymine is one of the four bases. Uracil is structurally similar to thymine, so like thymine, it can pair with adenine. Another difference between RNA and DNA is that whereas DNA usually occurs as a double strand, RNA usually occurs as a single strand.

DNA sequences encode proteins through the processes of transcription and translation. These processes both involve RNA, a single-stranded molecule that is similar to DNA except that it has a ribose sugar rather than deoxyribose and a uracil base rather than thymine.

## Transcription

**Transcription** is the process by which an RNA sequence is formed from a DNA template (Fig. 2.8). The type of RNA produced by the transcription process is **messenger RNA (mRNA).** To initiate mRNA transcription, one of the **RNA polymerase** enzymes (RNA polymerase II) binds to a **promoter** site on the DNA (a promoter is a nucleotide sequence that lies just upstream of a gene). The RNA polymerase then pulls a portion of the DNA strands apart from each other, exposing unattached DNA bases. One of the two DNA

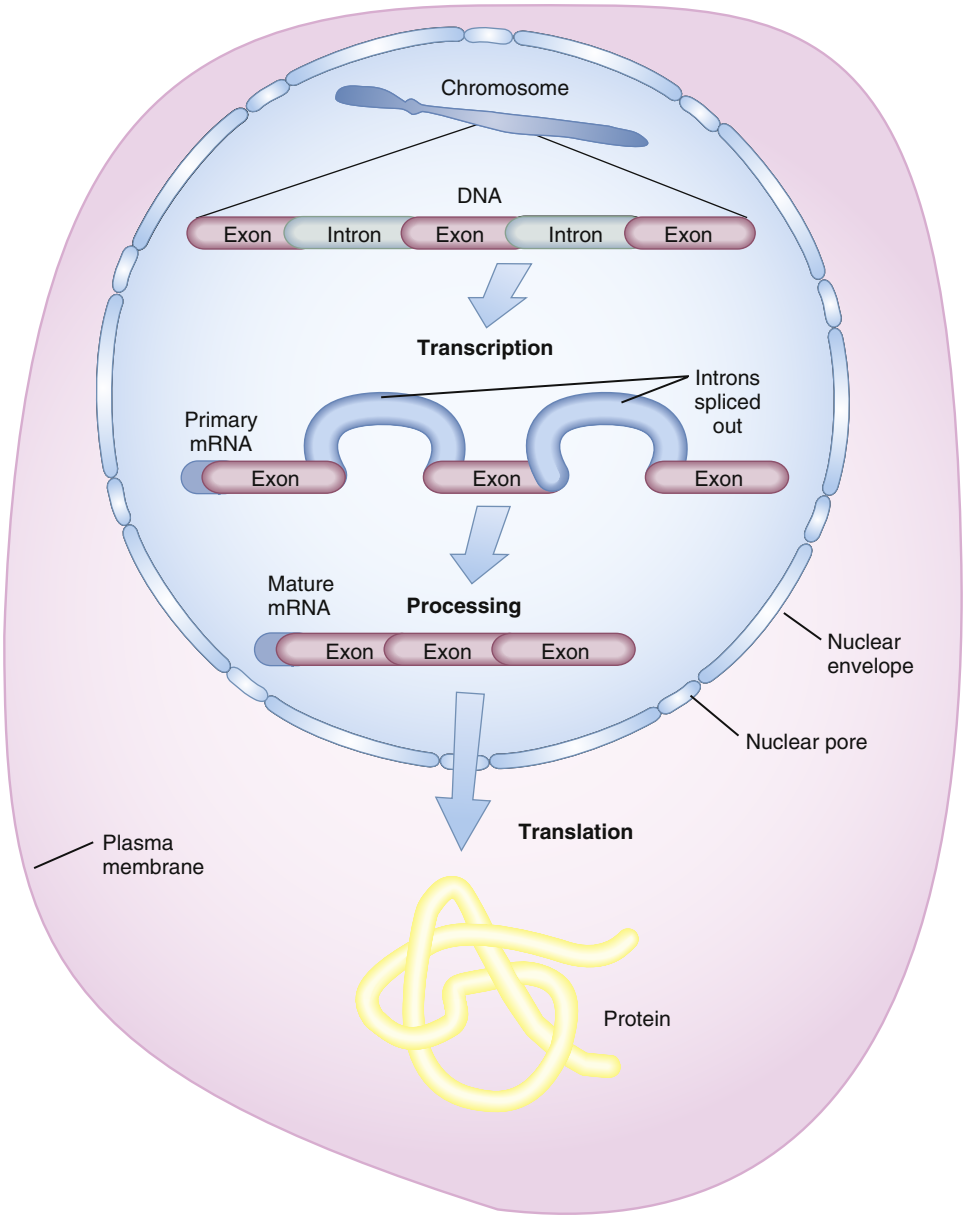


**Fig. 2.7** A summary of the steps leading from DNA to proteins. Replication and transcription occur in the cell nucleus. The mRNA is then transported to the cytoplasm, where translation of the mRNA into amino acid sequences composing a protein occurs.

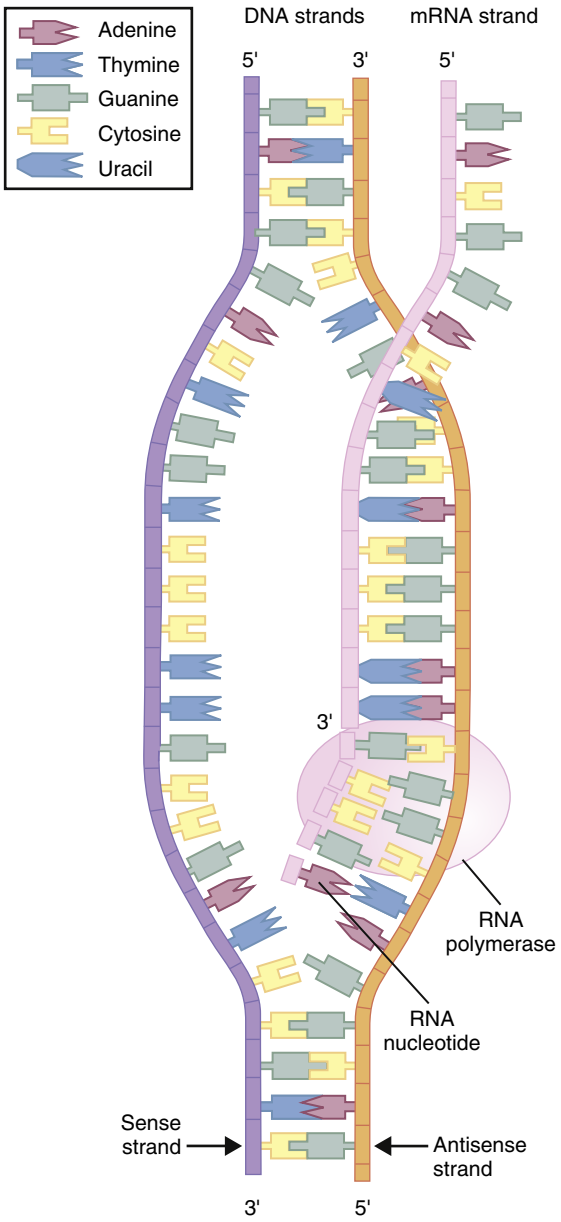


**Fig. 2.8** Transcription of DNA to mRNA. RNA polymerase II proceeds along the DNA strand in the 3′ to 5′ direction, assembling a strand of mRNA nucleotides that is complementary to the DNA template strand.

strands provides the template for the sequence of mRNA nucleotides. Although either DNA strand could in principle serve as the template for mRNA synthesis, only one is chosen to do so in a given region of the chromosome. This choice is determined by the promoter sequence, which orients the RNA polymerase in a specific direction along the DNA sequence. Because the mRNA molecule can be synthesized only in the 5′ to 3′ direction, the promoter, by specifying directionality, determines which DNA strand serves as the template. This template DNA strand is also known as the **antisense strand.** RNA polymerase moves in the 3′ to 5′ direction along the DNA template strand, assembling the complementary mRNA strand from 5′ to 3′ (see Fig. 2.8). Because of complementary base pairing, the mRNA nucleotide sequence is identical to that of the DNA strand that does not serve as the template—the **sense strand**—except for the substitution of uracil for thymine.

Soon after RNA synthesis begins, the 5′ end of the growing RNA molecule is capped by the addition of a chemically modified guanine nucleotide. This **5′ cap** appears to help prevent the RNA molecule from being degraded during synthesis, and later it helps to indicate the starting position for translation of the mRNA molecule into protein. Transcription continues until a group of bases called a **termination sequence** is reached. Near this point, a series of 100 to 200 adenine bases are added to the 3′ end of the RNA molecule. This structure, known as the **poly-A tail,** may be involved in stabilizing the mRNA molecule so that it is not degraded when it reaches the cytoplasm. RNA polymerase usually continues to transcribe DNA for several thousand additional bases, but the mRNA bases that are attached after the poly-A tail are eventually lost. Finally, the DNA strands and the RNA polymerase separate from the RNA strand, leaving a transcribed single mRNA strand. This mRNA molecule is termed the **primary transcript.**

In some human genes, such as the one that can cause Duchenne muscular dystrophy, several different promoters exist and are located in different parts of the gene. Thus transcription of the gene can start in different places, resulting in the production of somewhat different proteins. This allows the same gene sequence to code for variations of a protein in different tissues (e.g., muscle tissue versus brain tissue).

In the process of transcription, RNA polymerase II recognizes a promoter site near the 5′ end of a gene on the sense strand and, through complementary base pairing, helps to produce an mRNA strand from the antisense DNA strand.

## Transcription and the Regulation of Gene Expression

Some genes are transcribed in all cells of the body. These **housekeeping genes** encode products that are required for a cell's maintenance and metabolism. Most genes, however, are transcribed only in specific tissues at specific points in time. Therefore in most cells only a small fraction of genes are actively transcribed. This specificity explains why there is a large variety of different cell types making different protein products, even though almost all cells have exactly the same DNA sequence. For example, the globin genes are transcribed in the progenitors of red blood cells (where they help to form hemoglobin), and the low-density lipoprotein receptor genes are transcribed in liver cells.

Many different proteins participate in the process of transcription. Some of these are required for the transcription of all genes, and these are termed **general transcription factors.** Others, labeled **specific transcription factors,** have more specialized roles, activating only certain genes at certain stages of development. A key transcriptional element is RNA polymerase II, which was described previously. Although this enzyme plays a vital role in initiating transcription by binding to the promoter region, it cannot locate the promoter region on its own. Furthermore, it is incapable of producing significant quantities of mRNA by itself. Effective transcription requires the interaction of a large complex of approximately 50 different proteins.

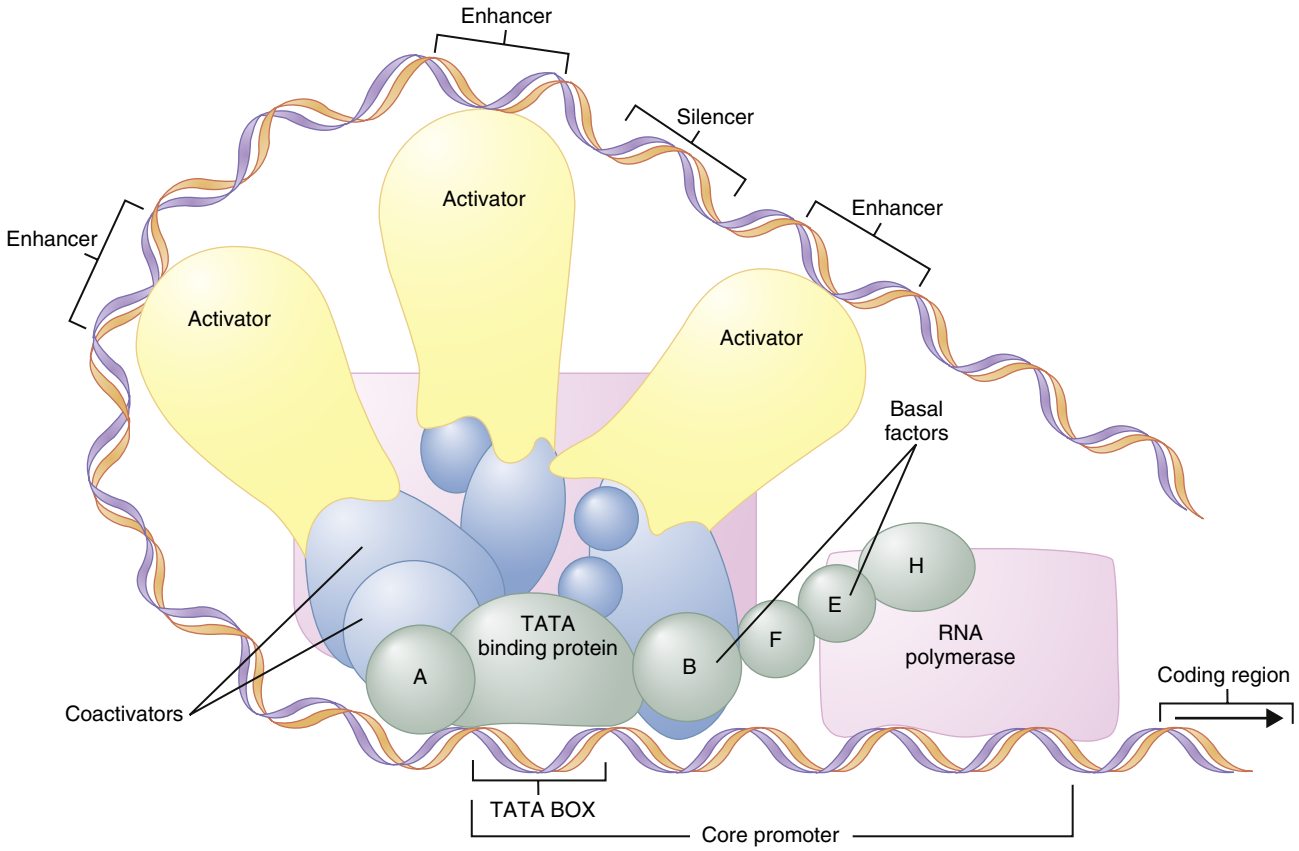


**Fig. 2.9** Key elements of transcription control include general (basal) transcription factors and specific enhancers and silencers. The activity of enhancers is mediated by activators and coactivators, which are specific transcription factors. (from Tjian R. Molecular machines that control genes. *Sci Am.* 1995;272[2]:54-61.)

These include general (basal) transcription factors, which bind to RNA polymerase and to specific DNA sequences in the promoter region (sequences such as TATA and others needed for initiating transcription). The general transcription factors allow RNA polymerase to bind to the promoter region so that it can function effectively in transcription (Fig. 2.9).

The transcriptional activity of specific genes can be greatly increased by interaction with sequences called **enhancers,** which may be located thousands of bases upstream or downstream of the gene. Enhancers do not interact directly with genes. Instead, they are bound by a class of specific transcription factors that are termed **activators.** Activators bind to a second class of specific transcription factors called **coactivators,** which in turn bind to the general transcription factor complex described previously (see Fig. 2.9). This chain of interactions, from enhancer to activator to coactivator to the general transcription complex and finally to the gene itself, increases the transcription of specific genes at specific points in time. Whereas enhancers help to increase the transcriptional activity of genes, other DNA sequences, known as **silencers,** help to repress the transcription of genes through a similar series of interactions.

Mutations in enhancer, silencer, or promoter sequences, as well as mutations in the genes that encode transcription factors, can lead to faulty expression of vital genes and consequently to genetic disease. Examples of such diseases are discussed in the following chapters.

Transcription factors are required for the transcription of DNA to mRNA. General transcription factors are used by all genes, and specific transcription factors help to initiate the transcription of genes in specific cell types at specific points in time. Transcription is also regulated by enhancer and silencer sequences, which may be located thousands of bases away from the transcribed gene.

The large number and complexity of transcription factors allow fine-tuned regulation of gene expression. But how do the transcription factors locate specific DNA sequences? This is achieved by **DNA-binding motifs:** configurations in the transcription-factor protein that allow it to fit snugly and stably into a unique portion of the DNA double helix. Several examples of these binding motifs are listed in Table 2.1, and Fig. 2.10 illustrates the binding of one such motif to DNA. Each major motif contains many variations that allow specificity in DNA binding.

An intriguing type of DNA-binding motif is contained in the high-mobility group (HMG) class of proteins. These proteins are capable of bending DNA and can facilitate interactions between distantly located enhancers and the appropriate basal factors and promoters (see Fig. 2.9).

Transcription factors contain DNA-binding motifs that allow them to interact with specific DNA sequences. In some cases, they bend DNA so that distant enhancer sequences can interact with target genes.

**TABLE 2.1** The Major Classes of DNA-Binding Motifs Found in Transcription Factors

| Motif | Description | Human Disease Examples |
|---|---|---|
| Helix–turn–helix | Two α helices are connected by a short chain of amino acids, which constitute the turn. The carboxyl-terminal helix is a recognition helix that binds to the DNA major groove. | Homeodomain proteins (HOX): mutations in human *HOXD13* and *HOXA13* cause synpolydactyly and hand–foot–genital syndrome, respectively. |
| Helix–loop–helix | Two α helices (one short and one long) are connected by a flexible loop. The loop allows the two helices to fold back and interact with one another. The helices can bind to DNA or to other helix–loop–helix structures. | Mutations in the *TWIST* gene cause Saethre–Chotzen syndrome (acrocephalosyndactyly type III) |
| Zinc finger | Zinc molecules are used to stabilize amino acid structures (e.g., α helices, β sheets), with binding of the α helix to the DNA major groove. | *BRCA1* (breast cancer gene); *WT1* (Wilms tumor gene); *GL13* (Greig syndrome gene); vitamin D receptor gene (mutations cause rickets) |
| Leucine zipper | Two leucine-rich α helices are held together by amino acid side chains. The α helices form a Y-shaped structure whose side chains bind to the DNA major groove. | *RB1* (retinoblastoma gene); *JUN* and *FOS* oncogenes |
| β Sheets | Side chains extend from the two-stranded β sheet to form contacts with the DNA helix. | TBX family of genes: *TBX5* (Holt–Oram syndrome); *TBX3* (ulnar–mammary syndrome) |

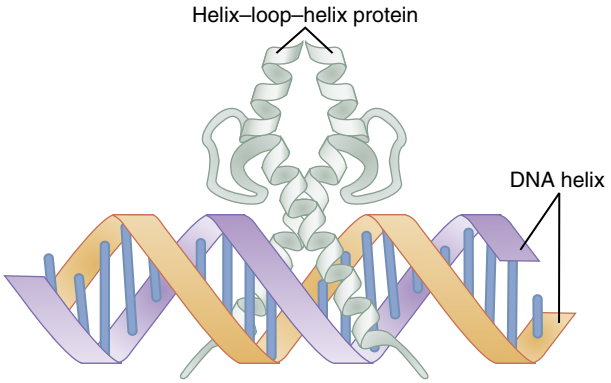


**Fig. 2.10** A helix–loop–helix motif binds tightly to a specific DNA sequence.

Gene activity can be related to patterns of chromatin coiling or condensation (**chromatin** is the combination of DNA and the histone proteins around which the DNA is wound). Decondensed or open chromatin regions, termed **euchromatin,** are typically characterized by histone **acetylation,** the attachment of acetyl groups to lysine residues in the histones. Acetylation of histones reduces their binding to DNA, helping to decondense the chromatin so that it is more accessible to transcription factors. Euchromatin is thus transcriptionally active. In contrast, **heterochromatin** is usually less acetylated, more condensed, and transcriptionally inactive.

Transcriptional silencing is also associated with **methylation** of promoter regions, in which methyl groups are attached to the DNA molecule. Methylation renders promoters less accessible to transcription factors. Factors such as methylation and histone modification, which alter the expression of genes but do not change the DNA sequence itself, are examples of **epigenetic** ("over" genetics) modification. Epigenetics, which plays key roles in development and cancer, is discussed further in Chapters 8 and 11.

Gene expression can also be influenced by **microRNAs (miRNA),** which are small RNA molecules (17–27 nucleotides) that are not translated into proteins but can bind to and downregulate mRNA. MicroRNAs, which are considered another form of epigenetic modifier, have been found to play important roles in gene regulation and in a number of types of cancer (see Chapter 11).

Heterochromatin, which is highly condensed and hypoacetylated, tends to be transcriptionally inactive, whereas euchromatin, which is acetylated and less condensed, tends to be transcriptionally active. Transcription is also regulated by methylation and by the attachment of microRNAs to mRNA. These factors are all examples of epigenetic modification.

## Gene Splicing

The primary mRNA transcript is exactly complementary to the base sequence of the DNA template. In **eukaryote**[*] an important step takes place before this RNA transcript leaves the nucleus. Sections of the RNA are removed by nuclear enzymes, and the remaining sections are spliced together to form the functional mRNA that will migrate to the cytoplasm. The excised sequences are called **introns,** and the sequences that are left to code for proteins are called **exons** (Fig. 2.11). Only after gene splicing is completed does the **mature transcript** move out of the nucleus into the cytoplasm. Some genes contain **alternative splice sites,** which allow the same primary transcript to be spliced in different ways, ultimately producing different protein products from the same gene. Errors in gene splicing, like replication errors, are a form of mutation that can lead to genetic disease.

Introns are spliced out of the primary mRNA transcript before the mature transcript leaves the nucleus. Exons contain the mRNA that specifies proteins.

## The Genetic Code

Proteins are composed of one or more **polypeptides**, which are in turn composed of sequences of **amino acids.** The body contains 20 different types of amino acids, and the amino acid sequences that make up polypeptides must in some way be designated by the DNA after transcription into mRNA.

Because there are 20 different amino acids and only four different RNA bases, a single base could not be specific for a single amino acid. Similarly, specific amino acids could not be defined by couplets of bases (e.g., adenine followed by

[*] Eukaryotes are organisms that have a defined cell nucleus, as opposed to prokaryotes, which lack a defined nucleus.

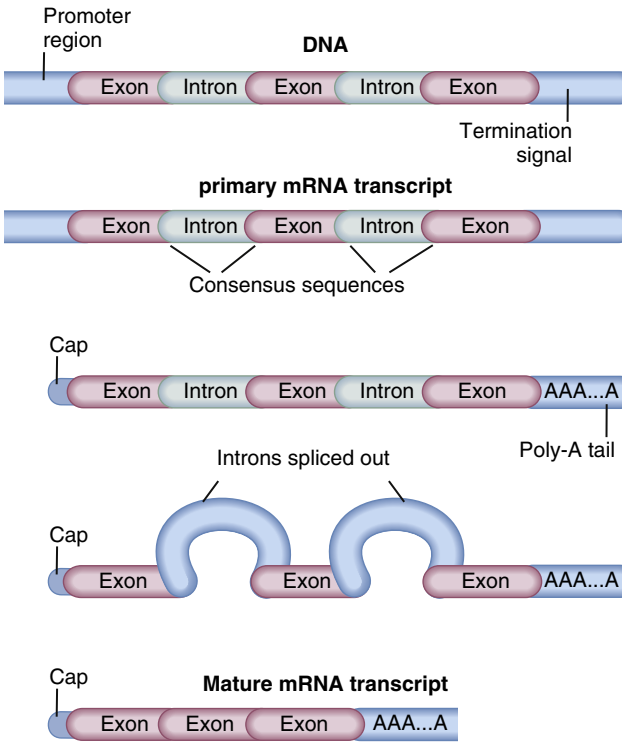


**Fig. 2.11** Gene splicing. Introns are precisely removed from the primary mRNA transcript to produce a mature mRNA transcript. Consensus sequences mark the sites at which splicing occurs.

guanine, or uracil followed by adenine) because only 16 (4 × 4) different couplets are possible. If triplet sets of bases are translated into amino acids, however, 64 (4 × 4 × 4) combinations can be achieved—more than enough to specify each amino acid. Conclusive proof that amino acids are specified by these triplets of bases, or **codons,** was obtained by manufacturing synthetic nucleotide sequences and allowing them to direct the formation of polypeptides in the laboratory. The correspondence between specific codons and amino acids, known as the **genetic code,** is shown in Table 2.2.

Of the 64 possible codons, three signal the end of a gene and are known as **stop codons.** These are UAA, UGA, and UAG. The remaining 61 codons all specify amino acids. This means that most amino acids can be specified by more than one codon, as Table 2.2 shows. The genetic code is thus said to be **degenerate.** Although a given amino acid may be specified by more than one codon, each codon can designate only one amino acid.

Individual amino acids, which compose proteins, are encoded by units of three mRNA bases, termed codons. There are 64 possible codons and only 20 amino acids, so the genetic code is degenerate.

A significant feature of the genetic code is that it is universal: the great majority of organisms use the same DNA codes to specify amino acids. An important exception to this rule occurs in **mitochondria,** cytoplasmic organelles that are the sites of cellular respiration (see Fig. 2.1). The mitochondria have their own extranuclear DNA molecules, and several codons of mitochondrial DNA encode different amino acids than do the same nuclear DNA codons.

**TABLE 2.2** The Genetic Code*

| First Position (5′ END) ↓ | Second Position U | Second Position C | Second Position A | Second Position G | Third Position (3′ END) ↓ |
|---|---|---|---|---|---|
| U | Phe | Ser | Tyr | Cys | U |
| U | Phe | Ser | Tyr | Cys | C |
| U | Leu | Ser | STOP | STOP | A |
| U | Leu | Ser | STOP | Trp | G |
| C | Leu | Pro | His | Arg | U |
| C | Leu | Pro | His | Arg | C |
| C | Leu | Pro | Gln | Arg | A |
| C | Leu | Pro | Gln | Arg | G |
| A | Ile | Thr | Asn | Ser | U |
| A | Ile | Thr | Asn | Ser | C |
| A | Ile | Thr | Lys | Arg | A |
| A | Met | Thr | Lys | Arg | G |
| G | Val | Ala | Asp | Gly | U |
| G | Val | Ala | Asp | Gly | C |
| G | Val | Ala | Glu | Gly | A |
| G | Val | Ala | Glu | Gly | G |

*Ala,* Alanine; *Arg,* arginine; *Asn,* asparagine; *Asp,* aspartic acid; *Cys,* cysteine; *Gln,* glutamine; *Glu,* glutamic acid; *Gly,* glycine; *His,* histidine; *Ile,* isoleucine; *Leu,* leucine; *Lys,* lysine; *Met,* methionine; *Phe,* phenylalanine; *Pro,* proline; *Ser,* serine; *Thr,* threonine; *Trp,* tryptophan; *Tyr,* tyrosine; *Val,* valine.

*Examples: UUG is translated into leucine; UAA is a stop codon; GGG is translated into glycine. Under some circumstances the UGA codon can specify an amino acid called selenocysteine, which is often called the 21st amino acid.

## Translation

**Translation** is the process in which mRNA provides a template for the synthesis of a polypeptide. However, mRNA cannot bind directly to amino acids. Instead it interacts with molecules of **transfer RNA (tRNA),** which are cloverleaf-shaped RNA strands of about 80 nucleotides. As Fig. 2.12 illustrates, each tRNA molecule has a site at the 3′ end for the attachment of a specific amino acid by a covalent bond. At the opposite end of the cloverleaf is a sequence of three nucleotides called the **anticodon,** which undergoes complementary base pairing with an appropriate codon in the mRNA. The attached amino acid is then transferred to the polypeptide chain being synthesized. In this way mRNA specifies the sequence of amino acids by acting through tRNA.

The cytoplasmic site of protein synthesis is the **ribosome,** which consists of roughly equal parts of enzymatic proteins and **ribosomal RNA (rRNA).** The function of rRNA is to help bind mRNA and tRNA to the ribosome. During translation, depicted in Fig. 2.13, the ribosome first binds to an initiation site on the mRNA sequence. This site consists of a specific codon, AUG, which specifies the amino acid methionine (this amino acid is usually removed from the polypeptide before the completion of polypeptide synthesis). The ribosome then binds the tRNA to its surface so that base pairing can occur between tRNA and mRNA. The ribosome moves along the mRNA sequence, codon by codon, in the 5′ to 3′ direction. As each codon is processed, an amino acid is translated by the interaction of mRNA and tRNA.

In this process, the ribosome provides an enzyme that catalyzes the formation of covalent peptide bonds between the adjacent amino acids, resulting in a growing polypeptide. When the ribosome arrives at a stop codon on the

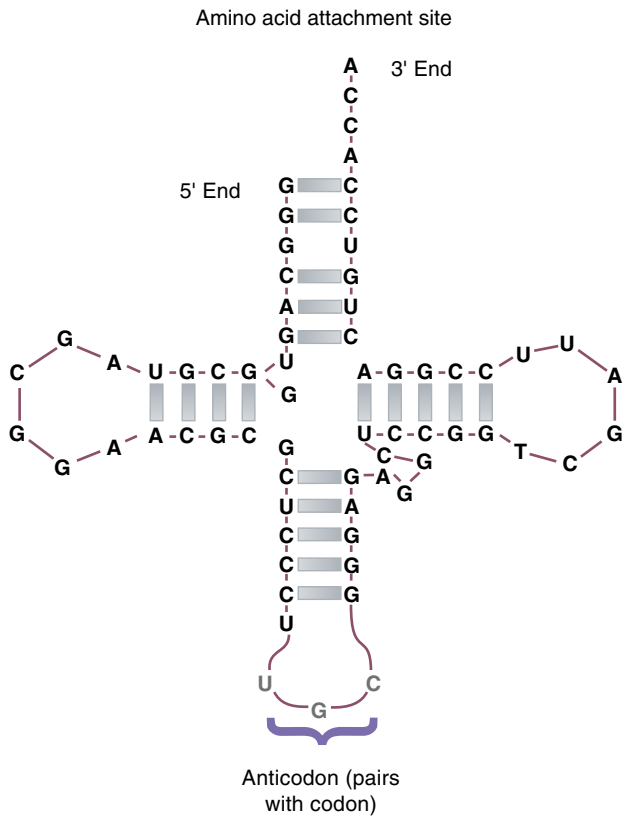


**Fig. 2.12** The structure of a tRNA molecule. In two dimensions, the tRNA has a cloverleaf shape. Note the 3′ site of attachment for an amino acid. The anticodon pairs with a complementary mRNA codon.

mRNA sequence, translation and polypeptide formation cease. The amino ($NH_2$) terminus of the polypeptide corresponds to the 5′ end of the mRNA strand, and the carboxyl (COOH) terminus corresponds to the 3′ end. After synthesis is completed, the mRNA, the ribosome, and the polypeptide separate from one another. The polypeptide is then released into the cytoplasm.

In the process of translation, the mRNA sequence serves as a template to specify sequences of amino acids. These sequences, which form polypeptides, are assembled by ribosomes. The tRNA and rRNA molecules interact with mRNA in the translation process.

Before a newly synthesized polypeptide can begin its existence as a functional protein, it often undergoes further processing, termed **posttranslational modification.** These modifications can take a variety of forms, including cleavage into smaller polypeptide units or combination with other polypeptides to form a larger protein. Other possible modifications include the addition of carbohydrate side chains to the polypeptide. Such modifications may be needed, for example, to produce proper folding of the mature protein or to stabilize its structure. An example of a clinically important protein that undergoes considerable posttranslational modification is type I collagen (Clinical Commentary 2.1).

Posttranslational modification consists of various chemical changes that occur in proteins shortly after they are translated.

### Noncoding RNA

The progression from DNA to RNA to proteins is considered part of the "central dogma" of molecular biology and genetics. However, several types of RNA are never translated into amino acids but may nevertheless have important functions. For example, **microRNAs** (**miRNAs**) are small RNA sequences, 17 to 27 nucleotides in length. They are complementary to specific mRNA sequences, and can bind to and down-regulate these mRNAs, thus lowering their expression levels. More than 1000 human miRNAs have been identified, and some of them play important roles in regulating mRNAs related to cancer (see Chapter 11). Each miRNA can bind to multiple mRNA targets and therefore can regulate many different gene transcripts. **Small interfering RNAs** (**siRNAs**), another type of short RNA sequence, bind single specific targets and are used in cancer therapy and some forms of gene therapy (see Chapter 13). Still another type of nontranslated RNA is termed **long noncoding RNA** (**lncRNA**) and is classified as having a length greater than 200 nucleotides. The human genome contains more than 10,000 lncRNAs. While the function of most lncRNAs remains unknown, some are involved in gene regulation. Perhaps the best example is the lncRNA encoded by the *XIST* gene, which inactivates one X chromosome in the genomes of female embryos (see Chapter 5).

## The Structure of Genes and the Genome

Some aspects of gene structure, such as the existence of introns and exons, have already been touched on. Alterations of different parts of genes can have quite distinct consequences in terms of genetic disease. It is therefore necessary to describe more fully the details of gene structure. A schematic diagram of gene structure is given in Fig. 2.16.

### INTRONS AND EXONS

The intron–exon structure of genes, discovered in 1977, is one of the attributes that distinguishes eukaryotes from prokaryotes. Introns form the major portion of most eukaryotic genes. As noted previously, introns are spliced out of the mRNA before it leaves the nucleus, and this splicing must be under very precise control. Enzymes that carry out splicing are directed to the appropriate locations by DNA sequences known as **consensus sequences** (so named because they are common in all eukaryotic organisms), which are situated adjacent to each exon.

Because most eukaryotic genes are composed primarily of introns, it is natural to ask whether introns might have some function. One interesting hypothesis is that introns, by lengthening genes, encourage the shuffling of genes when homologous chromosomes exchange material during meiosis (see later discussion). It has also been suggested that introns evolved to modify the amount of time required for DNA replication and transcription.

Surprisingly, some introns contain transcribed genes that are apparently unrelated to the gene in which the introns are contained. For example, introns of the human neurofibromatosis type 1 *(NF1)* gene contain three genes

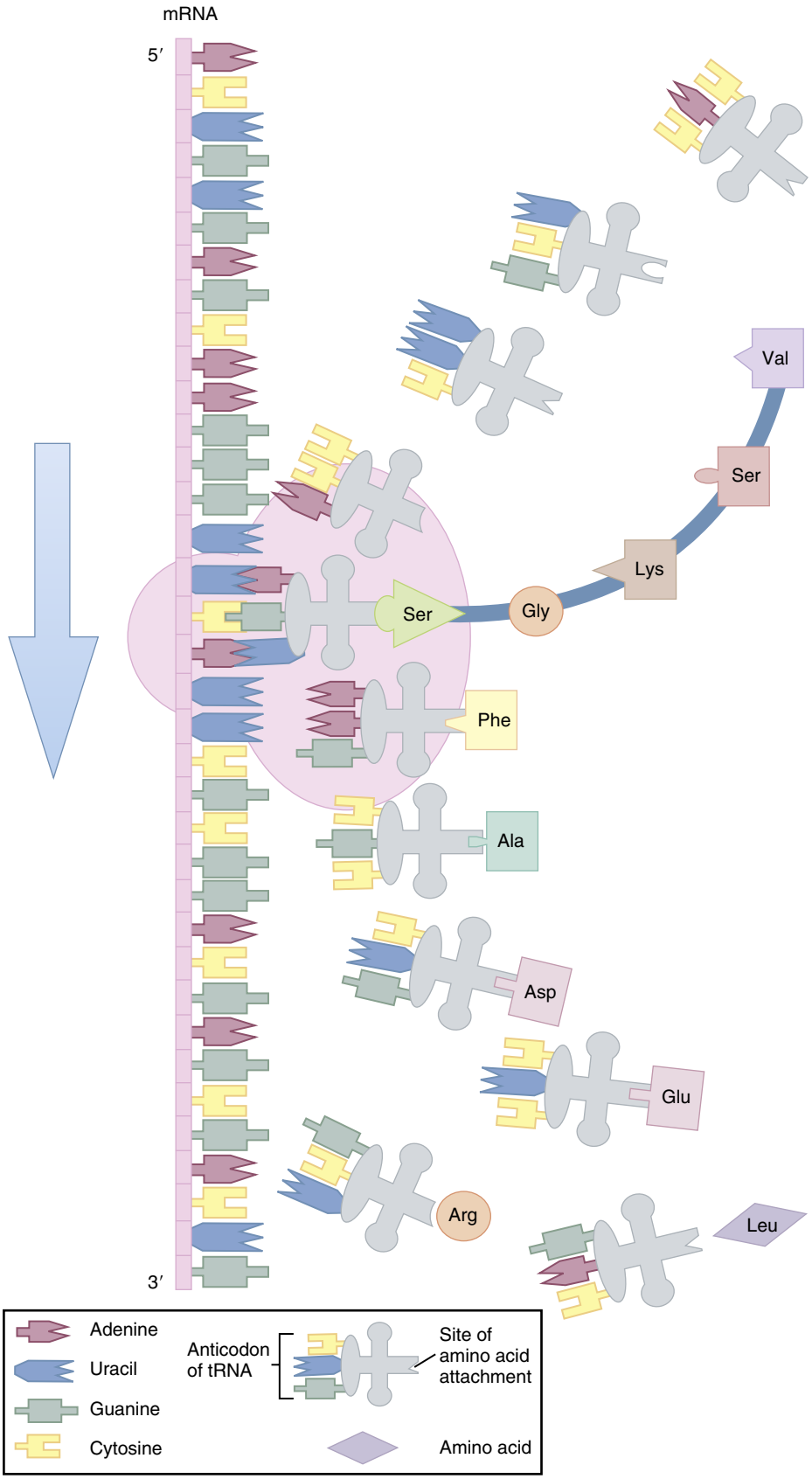


**Fig. 2.13** Translation of mRNA to amino acids. The ribosome moves along the mRNA strand in the 5′ to 3′ direction, assembling a growing polypeptide chain. In this example, the mRNA sequence GUG AGC AAG GGU UCA has assembled five amino acids (Val, Ser, Lys, Gly, and Ser, respectively) into a polypeptide.

## Clinical Commentary 2.1

### Osteogenesis Imperfecta, an Inherited Collagen Disorder

**TABLE 2.3** Subtypes of Osteogenesis Imperfecta

| Type | Disease Features |
|---|---|
| I | Mild bone fragility, blue sclerae, hearing loss in 60% of patients, normal or near-normal stature, few bone deformities, dentinogenesis imperfecta in some cases |
| II | Most severe form, with extreme bone fragility, long bone deformities, compressed femurs, gray-blue sclerae; lethal in the prenatal or perinatal periods (most die of respiratory failure during the first week of life, and survival beyond one year is highly unusual) |
| III | Severe bone fragility, very short stature, variably blue sclerae, progressive bone deformities, hearing loss in 80% of patients; dentinogenesis imperfecta is common |
| IV | Short stature, normal-to-gray sclerae, mild to moderate bone deformity, hearing loss in some patients, dentinogenesis imperfecta is common; bone fragility is variable |
| V | Similar to type IV but also includes calcification of interosseous membrane of forearm, radial head dislocation, and hyperplastic callus formation |
| VI | Similar to type IV but not caused by type I collagen mutations; calcification in interosseous membranes is not seen; no dentinogenesis imperfecta |
| VII | White sclerae, early lower limb deformities, shortening of all limbs, congenital fractures, osteopenia |
| VIII | Severe, often lethal phenotype similar to type II but not caused by type I collagen mutations; severe osteoporosis, shortened long bones, white sclerae |

Types I–IV are caused by mutations in the two genes that encode type I collagen protein; types V–VIII have been identified on the basis of distinct bone histology and are caused by mutations in genes involved in posttranslational processing of the gene product. Types VII and VIII have an autosomal recessive mode of inheritance, and several additional rare autosomal recessive and dominant forms of osteogenesis imperfecta have been identified (see Table 1 in Forlino and Marini, 2016). There is substantial phenotypic overlap among many of these categories of osteogenesis imperfecta.

As its name implies, osteogenesis imperfecta is a disease caused by defects in the formation of bone. This disorder, sometimes known as brittle bone disease, affects approximately 1 in 15,000 to 1 in 20,000 births in all ethnic groups.

Approximately 90% of osteogenesis imperfecta cases are caused by defects in type I collagen, a major component of bone that provides much of its structural stability. The function of collagen in bone is analogous to that of the steel bars incorporated in reinforced concrete.

When type I collagen is improperly formed, the bone loses much of its strength and fractures easily. Patients with osteogenesis imperfecta can suffer hundreds of bone fractures (Fig. 2.14), or they might experience only a few, making this disease highly variable in its expression (the reasons for this variability are discussed in Chapter 4). In addition to bone fractures, patients can have short stature, hearing loss, abnormal tooth development (dentinogenesis imperfecta), bluish sclerae, and various bone deformities, including scoliosis. Bone deformities, as well as abnormal collagen, result in decreased pulmonary function, which is the major cause of morbidity and mortality in this condition. Osteogenesis imperfecta was traditionally classified into four major types, all of which are caused by mutations in either of the two genes that encode type I collagen. A number of additional types, caused by mutations in other genes, have subsequently been added (Table 2.3). There is currently no cure for this disease, and management consists primarily of the repair of fractures, and in some cases the use of external or internal bone support (e.g., surgically implanted rods). Additional therapies include the administration of bisphosphonates to decrease bone resorption and human growth hormone to facilitate growth. The effectiveness of bisphosphonates for reducing the rate of bone fractures is a subject of debate. Physical rehabilitation also plays an important role in clinical management.

Type I collagen is a trimeric protein (i.e., having three subunits) with a triple helix structure. It is formed from a precursor protein, type 1 procollagen. Two of the three subunits of type 1 procollagen, labeled proα1(I) chains, are encoded by an 18-kb (kb = 1000 base pairs) gene on chromosome 17, and the third subunit, the pro-α2(I) chain, is encoded by a 38-kb gene on chromosome 7. Each of these genes contains more than 50 exons. After transcription and splicing, the mature mRNA formed from each gene is only 5 to 7 kb long. The mature mRNAs proceed to the cytoplasm, where they are translated into polypeptide chains by the ribosomal machinery of the cell.

At this point, the polypeptide chains undergo a series of posttranslational modifications. Many of the proline and lysine residues* are hydroxylated (i.e., hydroxyl groups are added) to form hydroxyproline and hydroxylysine, respectively. (Several rare forms of osteogenesis imperfecta, including type VII, are caused by mutations in genes involved in the hydroxylation process.) The three polypeptides, two proα1(I) chains and one proα2(I) chain, begin to associate with one another at their COOH termini. This association is stabilized by sulfide bonds that form between the chains near the COOH termini. The triple helix then forms, in zipper-like fashion, beginning at the COOH terminus and proceeding toward the $NH_2$ terminus. Some of the hydroxylysines are glycosylated (i.e., sugars are added), a modification that commonly occurs in the rough endoplasmic reticulum (see Fig. 2.1). The hydroxyl groups in the hydroxyprolines help to connect the three chains by forming hydrogen bonds, which stabilize the triple helix. Critical to proper folding of the helix is the presence of a glycine in every third position of each polypeptide.

Once the protein has folded into a triple helix, it moves from the endoplasmic reticulum to the Golgi apparatus (see Fig. 2.1) and is secreted from the cell. Yet another modification then takes place: the procollagen is cleaved by proteases near both the $NH_2$ and the COOH termini of the triple helix, removing some amino acids at each end. These amino acids performed essential functions earlier in the life of the protein (e.g., helping to form the triple helix structure, helping to thread the protein through the endoplasmic reticulum) but are no longer needed. This cleavage results in the mature protein, type I collagen. The collagen then assembles itself into fibrils, which react with adjacent molecules outside the cell to form the covalent cross-links that impart tensile strength to the fibrils.

* A residue is an amino acid that has been incorporated into a polypeptide chain.

*Continued*

## Clinical Commentary 2.1—cont'd

### Osteogenesis Imperfecta, an Inherited Collagen Disorder

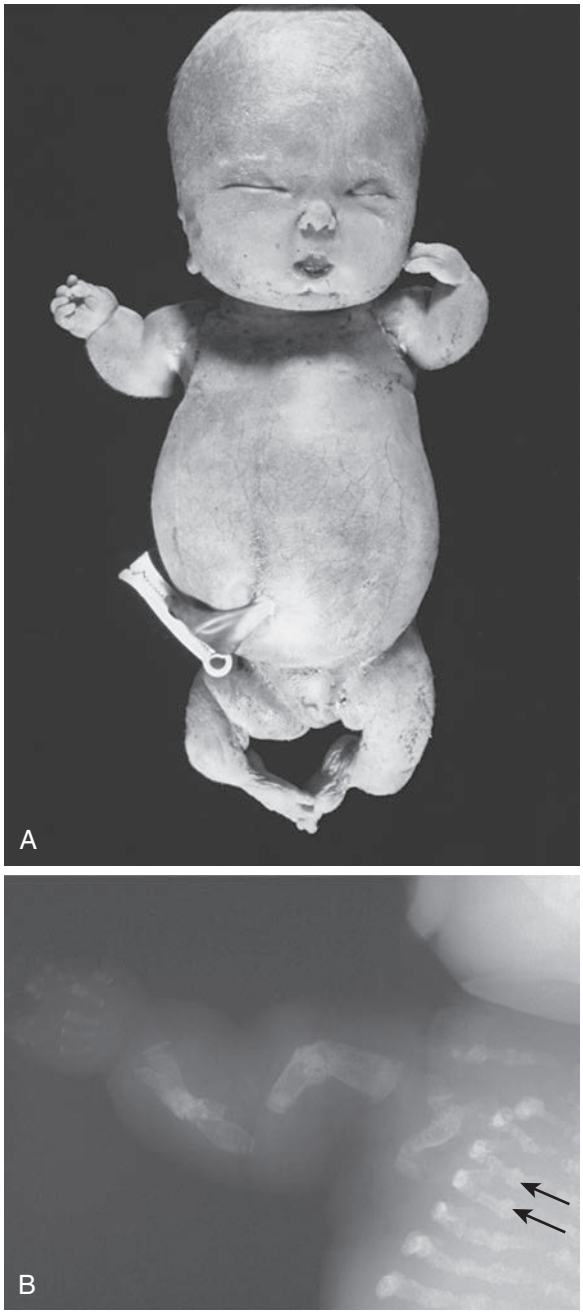


**Fig. 2.14 A,** A stillborn infant with type II osteogenesis imperfecta (the perinatal lethal form). The infant had a type I procollagen mutation and short, slightly twisted limbs. **B,** Radiograph of an infant with type II osteogenesis imperfecta. Note calluses from rib fractures, which are observable as "beads" on the ribs *(arrows).*

The path from the DNA sequence to the mature collagen protein involves many steps (Fig. 2.15). The complexity of this path provides many opportunities for mistakes (in replication, transcription, translation, or posttranslational modification) that can cause disease. Of the more than 1500 type I collagen mutations now known to cause osteogenesis imperfecta, the most common type produces a replacement of glycine with another amino acid. Because only glycine is small enough to be accommodated in the center of the triple helix structure, substitution of a different amino acid causes instability of the structure and thus poorly formed fibrils. This type of mutation is often seen in severe forms of osteogenesis imperfecta. Other mutations can cause excess posttranslational modification of the polypeptide chains, again producing abnormal fibrils. Other examples of disease-causing mutations are provided in the suggested readings at the end of this chapter.

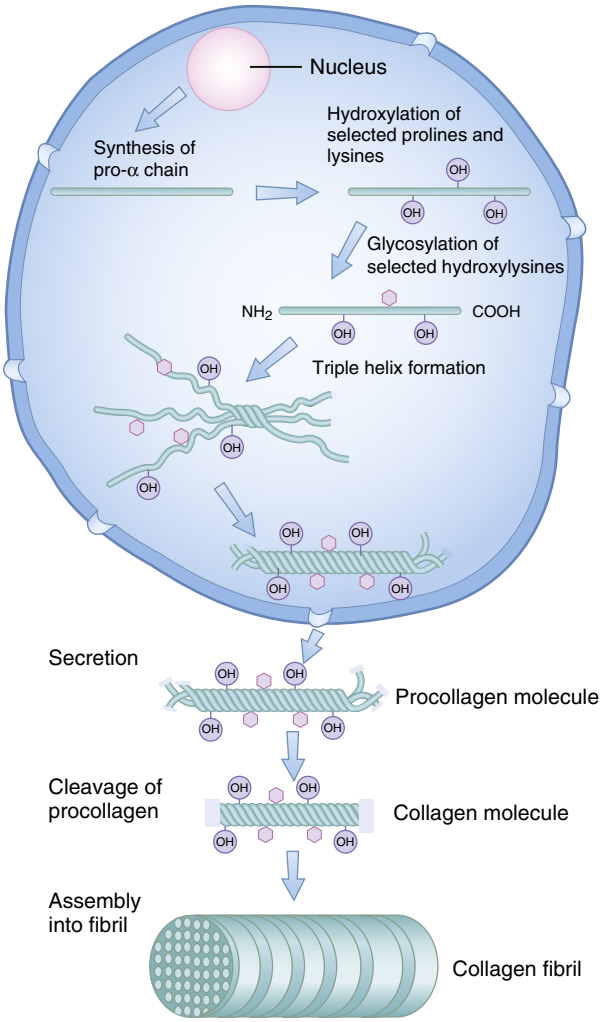


**Fig. 2.15** The process of collagen fibril formation. After the proα polypeptide chain is formed, a series of posttranslational modifications takes place, including hydroxylation and glycosylation. Three polypeptide chains assemble into a triple helix, which is secreted outside the cell. Portions of each end of the procollagen molecule are cleaved, resulting in the mature collagen molecule. These molecules then assemble into collagen fibrils.

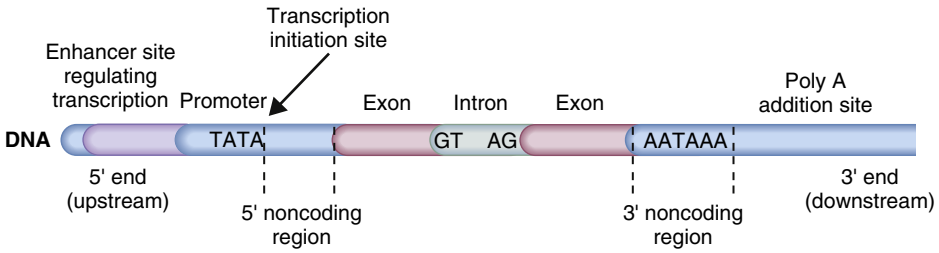


**Fig. 2.16** Details of gene structure, showing promoter and upstream regulation (enhancer) sequences and a poly A addition site.

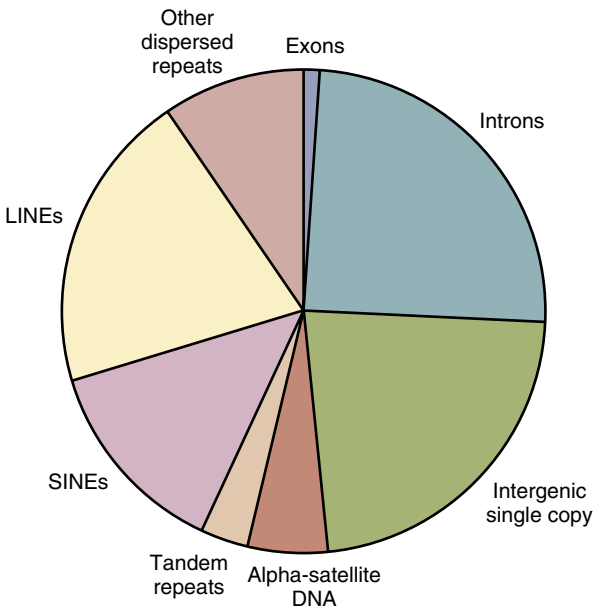


**Fig. 2.17** Structural composition of the human genome. Because of limitations in mapping repetitive sequences, these figures are approximate. In addition, there is some overlap among categories (e.g., repeat sequences are sometimes found in introns). The category "other dispersed repeats" includes DNA transposons, LTR (long terminal repeat) retrotransposons, and low-copy number duplications.

that are transcribed in the direction opposite that of the NF1 gene itself. These genes appear to have no functional relationship to the *NF1* gene. Similar gene inserts have been found within the factor VIII gene *(F8)* on the human X chromosome.

## TYPES OF DNA

Although most of the emphasis in genetics is given to the DNA that encodes proteins, only 1% to 2% of the 3 billion nucleotide pairs in the human genome actually perform this role. It has been estimated that about 75% of human DNA is transcribed into mRNA at some time in some cells, but the function of most of this mRNA (if any) is unknown. To better understand the nature of all types of DNA, we briefly review the several categories into which it is classified (Fig. 2.17).

The first and most important class of DNA is termed **single-copy DNA.** As the name implies, single-copy DNA sequences are seen only once (or possibly a few times) in the genome. Single-copy DNA accounts for about half of the genome and includes the protein-coding genes. However, protein-coding DNA (exons) represents only a small fraction of all single-copy DNA, most of which is found in introns or in DNA sequences that lie between genes.

The other half of the genome consists of **repetitive DNA,** sequences that are repeated over and over again in the genome, often thousands of times. There are two major classes of repetitive DNA: **dispersed repetitive DNA** and **satellite DNA.** Satellite repeats are clustered together in certain chromosome locations, where they occur as tandem repeats (i.e., the beginning of one repeat occurs immediately adjacent to the end of another). Dispersed repeats, as the name implies, tend to be scattered singly throughout the genome; they do not occur in tandem.

The term *satellite* is used because these sequences, owing to their composition, can easily be separated by centrifugation in a cesium chloride density gradient. The DNA appears as a satellite, separate from the other DNA in the gradient. This term is not to be confused with the satellites that can be observed microscopically on certain chromosomes (see Chapter 6). Satellite DNA accounts for approximately 8% to 10% of the genome and can be further subdivided into several categories. Alpha-Satellite DNA occurs as tandem repeats of a 171-bp sequence that can extend to several million base pairs or longer. This type of satellite DNA is found near the centromeres of chromosomes. **Minisatellites** are blocks of tandem repeats (each 11 bp long or greater) whose total length is much smaller, usually a few thousand base pairs. A final category, **microsatellites,** are smaller still; the repeat units are 1 to 10 bp long, and the total length of the array is usually less than a few hundred base pairs. Minisatellites and microsatellites are of special interest in human genetics because they vary in length among individuals, making them highly useful for forensic identification and gene mapping (see Chapter 8). A minisatellite or microsatellite is found at an average frequency of one per 2 kb in the human genome; altogether they account for about 3% of the genome.

Dispersed repetitive DNA makes up about 45% of the genome, and these repeats fall into several major categories (Fig. 2.18). The two most common categories are short interspersed elements **(SINEs)** and long interspersed elements **(LINEs).** Individual SINEs range in size from 90 to 500 bp, and individual LINEs can be as large as 7000 bp. One of the most important types of SINEs in humans is the 300-bp *Alu* element, so named because each repeat contains a DNA sequence that can be cut by the *Alu* restriction enzyme (see Chapter 3 for further discussion). The *Alu* repeats are a **family** of DNA elements, meaning that all of them have highly similar DNA sequences. About 1 million *Alu* repeats are scattered throughout the genome; they thus constitute approximately 10% of all human DNA. LINE repeats constitute another 20% of the human genome. A

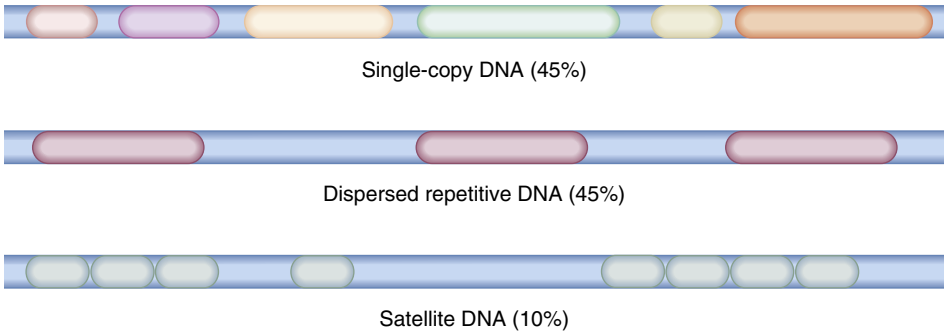


**Fig. 2.18** Single-copy DNA sequences are unique and are dispersed throughout the genome. Satellite DNA sequences are repetitive elements that occur together in clusters. Dispersed repeats are similar to one another but do not cluster together.

remarkable feature of *Alu* sequences, as well as LINEs, is that some of them can generate copies of themselves, which can then be inserted into other parts of the genome. This is how they have achieved such remarkable numbers. These insertions can sometimes interrupt a protein-coding gene, causing genetic disease (examples are discussed in Chapter 4).

There are several major types of DNA, including single-copy DNA, satellite DNA, and dispersed repetitive DNA. The latter two categories are both classes of repeated DNA sequences. Only about 1% to 2% of human DNA actually encodes proteins.

## The Cell Cycle

During the course of development, each human progresses from a single-cell **zygote** (an egg cell fertilized by a sperm cell) to a marvelously complex organism containing approximately 30 trillion ($3x10^{13}$) individual cells.[a] Because few cells last for a person's entire lifetime, new ones must be generated to replace those that die. Both of these processes—development and replacement—require the manufacture of new cells. The cell division processes that are responsible for the creation of new diploid cells from existing ones are **mitosis** (nuclear division) and **cytokinesis** (cytoplasmic division). Before dividing, a cell must duplicate its contents, including its DNA; this occurs during **interphase.** The alternation of mitosis and interphase is referred to as the **cell cycle.**

As Fig. 2.19 shows, a typical cell spends most of its life in interphase. This portion of the cell cycle is divided into three phases, **G1, S,** and **G2.** During **G1** (gap 1, the interval between mitosis and the onset of DNA replication), synthesis of RNA and proteins takes place. DNA replication occurs during the S (synthesis) phase. During G2 (the interval between the S phase and the next mitosis), some DNA repair takes place, and the cell prepares for mitosis. By the time G2 has been reached, the cell contains two identical copies of each of the 46 chromosomes. These identical chromosomes are referred to as **sister chromatids.** Sister chromatids often exchange material during interphase, a process known as **sister chromatid exchange.**

[a] Our bodies also contain at least the same number of microbial cells (e.g., bacteria and fungi) and non-cellular genetic material such as viruses, the genomes of which are collectively known as the **microbiome.**

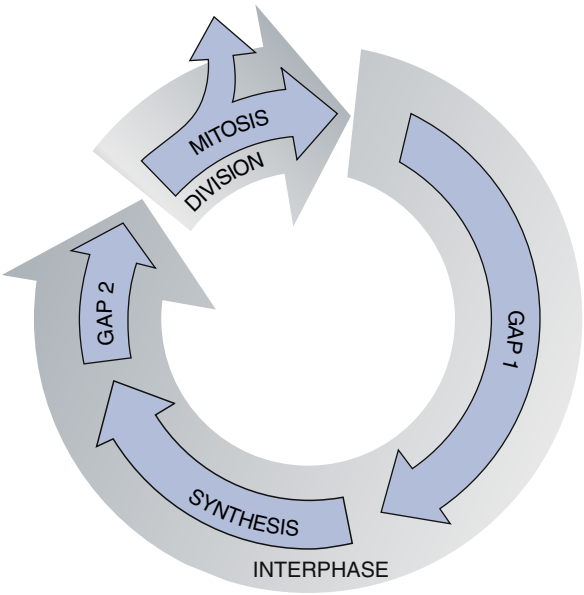


**Fig. 2.19** Major phases of the mitotic cell cycle, showing the alternation of interphase and mitosis (division).

The cell cycle consists of the alternation of cell division (mitosis and cytokinesis) and interphase. DNA replication and protein synthesis take place during interphase.

The length of the cell cycle varies considerably from one cell type to another. In rapidly dividing cells such as those of epithelial tissue (found, for example, in the lining of the intestines and in the lungs), the cycle may be completed in less than 10 hours. Other cells, such as those of the liver, might divide only once each year or so. Some cell types, such as skeletal muscle cells and neurons, largely lose their ability to divide and replicate in adults. Although all stages of the cell cycle have some variation in length, the great majority of variation is due to differences in the length of the G1 phase. When cells stop dividing for a long period, they are often said to be in the G0 stage.

Cells divide in response to important internal and external cues. Before the cell enters mitosis, for example, DNA replication must be accurate and complete and the cell must have achieved an appropriate size. The cell must respond to extracellular stimuli that require increased or decreased rates of division. Complex molecular interactions mediate this regulation. Among the most important of the

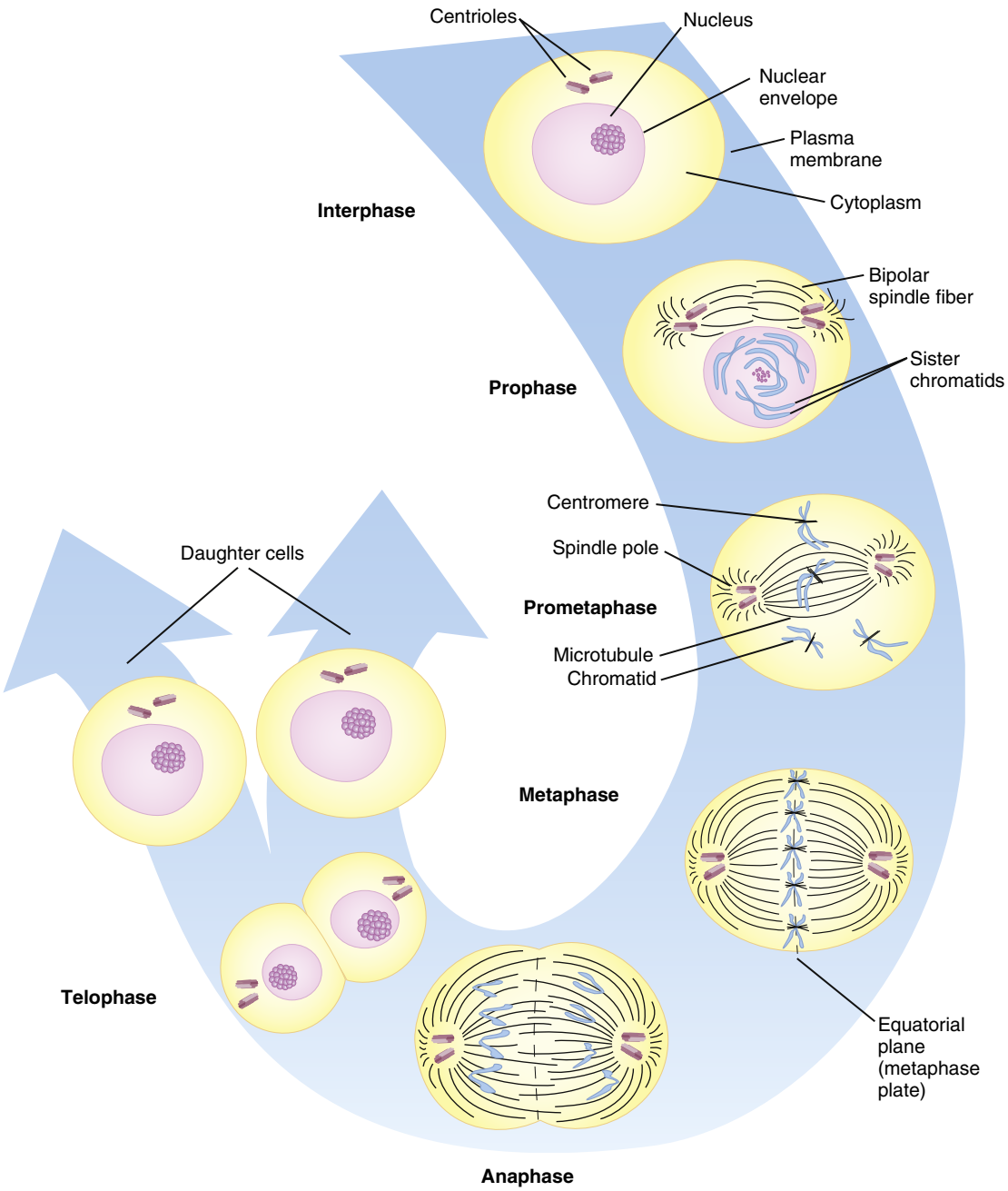


**Fig. 2.20** The stages of mitosis, during which two identical diploid cells are formed from one original diploid cell.

molecules involved are the **cyclin-dependent kinases (CDKs),** a family of kinases that phosphorylate other regulatory proteins at key stages of the cell cycle. To carry out this function, the CDKs must form complexes with various **cyclins,** proteins that are synthesized at specific cell-cycle stages and are then degraded when CDK action is no longer needed. The cyclins and CDKs, as well as the many proteins that interact with them, are subjects of intense study because of their vital role in the cell cycle and because their malfunction can lead to cancer (see Chapter 11).

The length of the cell cycle varies in different cell types. Critical to regulation of the cell cycle are CDKs, which phosphorylate other proteins and cyclins that form complexes with CDKs. Faulty regulation of the cell cycle can lead to cancer.

## MITOSIS

Although mitosis usually requires only 1 to 2 hours to complete, this portion of the cell cycle involves many critical and complex processes. Mitosis is divided into several phases (Fig. 2.20). During **prophase,** the first mitotic stage, the chromosomes become visible under a light microscope as they condense and coil (chromosomes are not clearly visible during interphase). The two sister chromatids of each chromosome lie together, attached at a point called the **centromere.** The nuclear membrane, which surrounds the nucleus, disappears during this stage. **Spindle fibers** begin to form, radiating from two **centrioles** located on opposite sides of the cell. The spindle fibers become attached to the centromeres of each chromosome and eventually pull the two sister chromatids in opposite directions.

The chromosomes reach their most highly condensed state during **metaphase,** the next stage of mitosis. Because they are highly condensed, they are easiest to visualize microscopically during this phase. For this reason, clinical diagnosis of chromosome disorders is usually based on metaphase chromosomes. During metaphase, the spindle fibers begin to contract and pull the centromeres of the chromosomes, which are now arranged along the middle of the spindle (the **equatorial plane** of the cell).

During **anaphase,** the next mitotic stage, the centromere of each chromosome splits, allowing the sister chromatids to separate. The chromatids are then pulled by the spindle fibers, centromere first, toward opposite sides of the cell. At the end of anaphase, the cell contains 92 separate chromosomes, half lying near one side of the cell and half near the other side. If all has proceeded correctly, the two sets of chromosomes are identical.

**Telophase,** the final stage of mitosis, is characterized by the formation of new nuclear membranes around each of the two sets of 46 chromosomes. Also the spindle fibers disappear, and the chromosomes begin to decondense. Cytokinesis usually occurs after nuclear division and results in a roughly equal division of the cytoplasm into two parts. With the completion of telophase, two diploid **daughter cells,** both identical to the original cell, have been formed.

Mitosis is the process through which two identical diploid daughter cells are formed from a single diploid cell.

## MEIOSIS

When an egg cell and a sperm cell unite to form a zygote, their chromosomes are combined into a single cell. Because humans are diploid organisms, there must be a mechanism to reduce the number of chromosomes in gametes to the haploid state. Otherwise the zygote would have 92, instead of the normal 46, chromosomes. The primary mechanism by which haploid gametes are formed from diploid precursor cells is **meiosis.**

Two cell divisions occur during meiosis. Each meiotic division has been divided into stages with the same names as those of mitosis, but the processes involved in some of the stages are quite different (Fig. 2.21). During meiosis I, often called the **reduction division stage,** two haploid cells are formed from a diploid cell. These diploid cells are the **oogonia** in females and the **spermatogonia** in males. After meiosis I, a second meiosis, the **equational division,** takes place, during which each haploid cell is replicated.

The first stage of the meiotic cell cycle is **interphase I,** during which important processes such as replication of chromosomal DNA take place. The second phase of meiosis I, **prophase I,** is quite complex and includes many of the key events that distinguish meiosis from mitosis. Prophase I begins as the chromatin strands coil and condense, causing them to become visible as chromosomes. During the process of **synapsis,** the homologous chromosomes pair up, side by side, lying together in perfect alignment (in males, the X and Y chromosomes, being mostly nonhomologous, line up end to end). This pairing of homologous chromosomes is an important part of the cell cycle that does not occur in mitosis. As prophase I continues, the chromatids of the two chromosomes intertwine. Each pair of intertwined homologous chromosomes is either **bivalent** (indicating two chromosomes in the unit) or **tetrad** (indicating four chromatids in the unit).

A second key feature of prophase I is the formation of **chiasmata** (plural of **chiasma**), cross-shaped structures that mark attachments between the homologous chromosomes (Fig. 2.22). Each chiasma indicates a point at which the homologous chromosomes exchange genetic material. This process, called **crossing over,** produces chromosomes that consist of combinations of parts of the original chromosomes. This chromosomal shuffling is important because it greatly increases the possible combinations of genes in each gamete and thereby increases the number of possible combinations of human traits. Also, as discussed in Chapter 8, this phenomenon is critically important in inferring the order of genes along chromosomes. At the end of prophase I, the bivalents begin to move toward the equatorial plane, a spindle apparatus begins to form in the cytoplasm, and the nuclear membrane dissipates.

**Metaphase I** is the next phase. As in mitotic metaphase, this stage is characterized by the completion of spindle formation and alignment of the bivalents, which are still attached at the chiasmata, in the equatorial plane. The two centromeres of each bivalent now lie on opposite sides of the equatorial plane.

During **anaphase I,** the chiasmata disappear and the homologous chromosomes are pulled by the spindle fibers toward opposite poles of the cell. The key feature of this phase is that unlike the corresponding phase of mitosis, the centromeres do not duplicate and divide, so that only half of the original number of chromosomes migrate toward each pole. The chromosomes migrating toward each pole thus consist of one member of each pair of autosomes and one of the sex chromosomes.

The next stage, **telophase I,** begins when the chromosomes reach opposite sides of the cell. The chromosomes uncoil slightly, and a new nuclear membrane begins to form. The two daughter cells each contain the haploid number of chromosomes, and each chromosome has two sister chromatids. In humans cytokinesis also occurs during this phase. The cytoplasm is divided approximately equally between the two daughter cells in the gametes formed in males. In those formed in females, nearly all of the cytoplasm goes into one daughter cell, which will later form the egg. The other daughter cell becomes a **polar body,** a small, nonfunctional cell that eventually degenerates.

Meiosis I (reduction division) includes a prophase I stage in which homologous chromosomes line up and exchange material (crossing over). During anaphase I the centromeres do not duplicate and divide. Consequently only one member of each pair of chromosomes migrates to each daughter cell.

The equational division, meiosis II, then begins with **interphase II.** This is a very brief phase. The important feature of interphase II is that, unlike interphase I, no replication of DNA occurs. **Prophase II,** the next stage, is quite similar to mitotic prophase, except that the cell nucleus contains only the haploid number of chromosomes. During prophase II, the chromosomes thicken as they coil, the nuclear membrane disappears, and new spindle fibers are formed. This is followed by **metaphase II,** during which

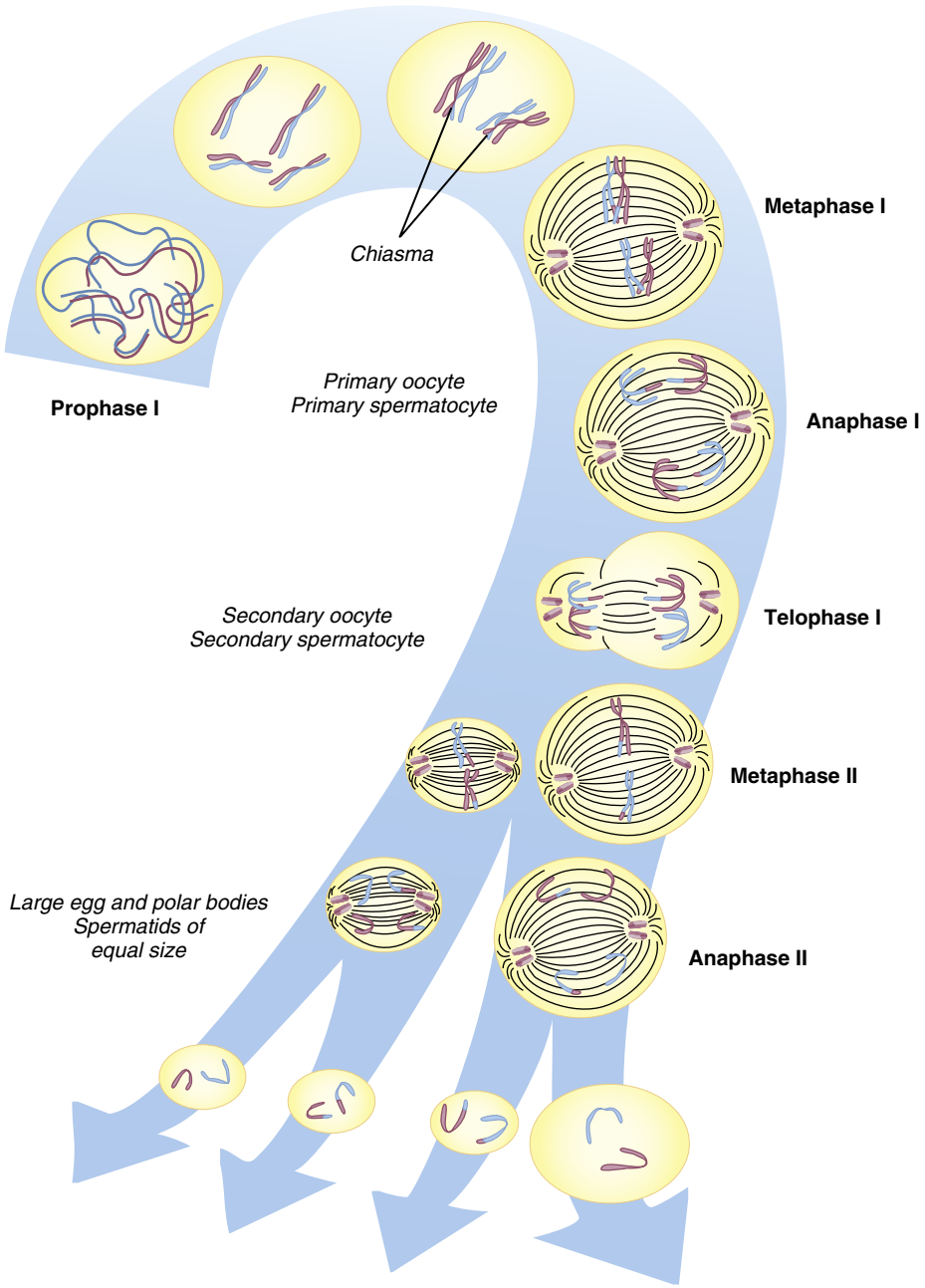


**Fig. 2.21** The stages of meiosis, during which haploid gametes are formed from a diploid cell. For brevity, prophase II and telophase II are not shown. Note the relationship between meiosis and spermatogenesis and oogenesis.

the spindle fibers pull the chromosomes into alignment at the equatorial plane.

**Anaphase II** then follows. This stage resembles mitotic anaphase in that the centromeres split and each carries a single chromatid toward a pole of the cell. The chromatids have now separated, but because of chiasma formation and crossing over, the newly separated sister chromatids might not be identical (see Fig. 2.21).

**Telophase II,** like telophase I, begins when the chromosomes reach opposite poles of the cell. There they begin to uncoil. New nuclear membranes are formed around each group of chromosomes, and cytokinesis occurs. In gametes formed in males, the cytoplasm is again divided equally between the two daughter cells. The end result of male meiosis is thus four functional daughter cells, each of which has an equal amount of cytoplasm. In female gametes, unequal division of the cytoplasm again occurs, forming the egg cell and another polar body. The polar body formed during meiosis I sometimes undergoes a second division, so three polar bodies may be present when the second stage of meiosis is completed.

Meiosis is a specialized cell division process in which a diploid cell gives rise to haploid gametes. This is accomplished by combining two rounds of division with only one round of DNA replication.

Most chromosome disorders are caused by errors that occur during meiosis. Gametes can be created that contain

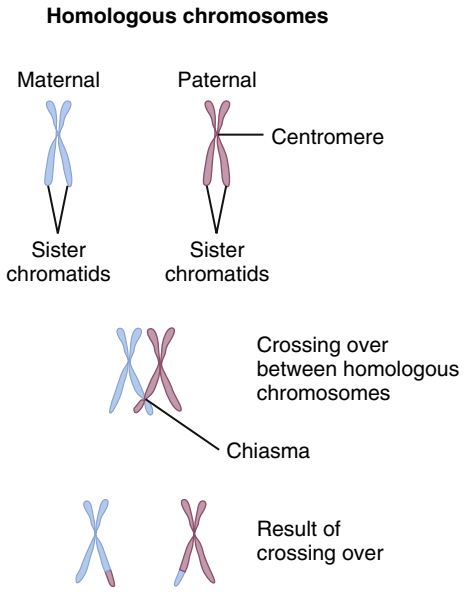


**Fig. 2.22** The process of chiasma formation and crossing over results in the exchange of genetic material between homologous chromosomes.

missing or additional chromosomes or chromosomes with altered structures. In addition, mitotic errors that occur early in the life of the embryo can affect enough of the body's cells to produce clinically significant disease. Mitotic errors occurring at any point in one's lifetime can, under some circumstances, cause cancer. Cancer genetics is discussed in Chapter 11, and chromosome disorders are the subject of Chapter 6.

## THE RELATIONSHIP BETWEEN MEIOSIS AND GAMETOGENESIS

The stages of meiosis can be related directly to stages in **gametogenesis,** the formation of gametes (see Fig. 2.21). In mature males, the seminiferous tubules of the testes are populated by spermatogonia, which are diploid cells. After going through several mitotic divisions, the spermatogonia produce **primary spermatocytes.** Each primary spermatocyte, which is also diploid, undergoes meiosis I to produce a pair of **secondary spermatocytes,** each of which contains 23 double-stranded chromosomes. These undergo meiosis II, and each produces a pair of **spermatids** that contain 23 single-stranded chromosomes. The spermatids then lose most of their cytoplasm and develop tails for swimming as they become mature **sperm** cells. This process, known as **spermatogenesis,** continues throughout the life of the mature male.

In spermatogenesis, each diploid spermatogonium produces four haploid sperm cells.

**Oogenesis,** the process by which female gametes are formed, differs in several important ways from spermatogenesis. Whereas the cycle of spermatogenesis is constantly recurring, much of female oogenesis is completed before birth. Diploid oogonia divide mitotically to produce **primary oocytes** by the third month of fetal development. More than 6 million primary oocytes are formed during gestation, and they are suspended in prophase I by the time of birth. Meiosis continues only when a mature primary oocyte is ovulated. In meiosis I, the primary oocyte produces one secondary oocyte (containing the cytoplasm) and one polar body. The **secondary oocyte** then emerges from the follicle and proceeds down the fallopian tube, with the polar body attached to it. Meiosis II begins only if the secondary oocyte is fertilized by a sperm cell. If this occurs, one haploid mature **ovum,** containing the cytoplasm, and another haploid polar body are produced. The polar bodies eventually disintegrate. About 1 hour after fertilization, the nuclei of the sperm cell and ovum fuse, forming a diploid zygote. The zygote then begins its development into an embryo through a series of mitotic divisions.

In oogenesis, one haploid ovum and three haploid polar bodies are produced meiotically from a diploid oogonium. In contrast to spermatogenesis, which continues throughout the life of the mature male, the first phase of oogenesis is completed before the female is born; oogenesis is then halted until ovulation occurs.

## Study Questions

1. Consider the following double-stranded DNA sequence: 5′-CAG AAG AAA ATT AAC ATG TAA-3′ 3′-GTC TTC TTT TAA TTG TAC ATT-5′. If the bottom strand serves as the template, what is the mRNA sequence produced by transcription of this DNA sequence? What is the amino acid sequence produced by translation of the mRNA sequence?

2. Arrange the following terms according to their hierarchical relationship to one another: genes, chromosomes, exons, codons, nucleotides, and genome.

3. Less than 2% of human DNA encodes proteins. Furthermore, in a given cell type only 10% of the coding DNA actively encodes proteins. Explain these statements.

4. What are the major differences between mitosis and meiosis?

5. The human body contains approximately $3x10^{13}$ cells. Starting with a single-cell zygote, how many mitotic cell divisions, on average, would be required to produce this number of cells?

6. How many mature sperm cells will be produced by 100 primary spermatocytes? How many mature egg cells will be produced by 100 primary oocytes?

### Suggested Readings

Alberts B, Johnson A, Lewis J, et al. *Molecular Biology of the Cell*. 6th ed. New York: Garland Science; 2015.
Forlino A, Marini JC. Osteogenesis imperfecta. *Lancet*. 2016;387:1657–1671.
Krebs JE, Goldstein ES, Kilpatrick ST. *Lewin's Genes XII*. Burlington, MA: Jones & Bartlett Learning; 2018.
Strachan T, Read AP. *Human Molecular Genetics*. 4th ed. London: Garland Science; 2011.

**Internet Resources**

Mitosis and meiosis animation: https://www.youtube.com/watch?v=5kVVaRcEI1Y
Tutorials and animations of DNA replication, transcription, and translation: http://www.hhmi.org/biointeractive/

# 3 Genetic Variation: Its Origin and Detection

Humans display a substantial amount of genetic variation. This is reflected in traits such as height, blood pressure, and skin color. Included in the spectrum of genetic variation are disease states, such as cystic fibrosis or type 1 neurofibromatosis (see Chapter 4). This aspect of genetic variation is the focus of medical genetics.

All genetic variation originates from the process known as **mutation,** which is defined as a change in DNA sequence. Mutations can affect either **germline** cells (cells that produce gametes) or **somatic cells** (all cells other than germline cells). Mutations in somatic cells can lead to cancer and are thus of significant concern. However, this chapter is focused primarily on germline mutations, because they can be transmitted from one generation to the next.

As a result of mutations, a gene can differ among individuals in terms of its DNA sequence. The differing sequences are referred to as **alleles.** A gene's location on a chromosome is termed a **locus** (from the Latin word for "place"). For example, it might be said that a person has a certain allele at the β-globin locus on chromosome 11. If a person has the same allele on both members of a chromosome pair, they are said to be a **homozygote.** If the alleles differ in DNA sequence, they are a **heterozygote.** The combination of alleles that is present at a given locus is termed the **genotype.**

In human genetics, the term *mutation* has often been reserved for DNA sequence changes that cause genetic diseases and are consequently relatively rare, with a population frequency less than 1%. DNA sequence variants that are more common in populations are conventionally described as **polymorphisms** ("many forms," describing multiple alleles at a locus). **Loci** (plural of locus) that contain multiple alleles are termed **polymorphic.** Nowadays, however, alleles that have a frequency less than 1% are sometimes called polymorphisms as well. In addition, many common polymorphisms are now known to influence the risks for complex, common diseases such as diabetes and heart disease (see Chapter 12), so the traditional distinction between mutation (rare and disease-causing) and polymorphism (common and benign) has become increasingly blurred.

One of Gregor Mendel's important contributions to genetics was to show that the effects of one allele at a locus can mask those of another allele at the same locus. He performed **crosses** (matings) between pea plants homozygous for a "tall" allele (i.e., having two identical copies of an allele that we will label *H*) and plants homozygous for a "short" allele (having two copies of an allele labeled *h*). This cross, which can produce only heterozygous *(Hh)* offspring, is illustrated in the **Punnett square** shown in Fig. 3.1. Mendel found that the offspring of these crosses were all tall, even though they were heterozygotes. This is because the *H* allele is **dominant,** and the *h* allele is **recessive.** (It is conventional to label the dominant allele in upper case and the recessive allele in lower case.) The term *recessive* comes from a Latin root meaning "to hide." This describes the behavior of recessive alleles well: in heterozygotes, the consequences of a recessive allele are hidden. A dominant allele exerts its effect in both the homozygote *(HH)* and the heterozygote *(Hh),* whereas the presence of the recessive allele is physically observable only when it occurs in homozygous form *(hh).* Thus short pea plants can be created only by crossing parent plants that each carry at least one *h* allele. An example is a heterozygote × heterozygote cross, shown in Fig. 3.2.

In this chapter we examine mutation as the source of genetic variation. We discuss the types of mutation, the causes and consequences of mutation, and the biochemical and molecular techniques that are now used to detect genetic variation in human populations.

## Mutation: The Source of Genetic Variation

### TYPES OF MUTATION

Some mutations consist of an alteration of the number or structure of chromosomes in a cell. These major chromosome abnormalities can be observed microscopically and are the subject of Chapter 6. Here the focus is on mutations that affect only single genes and are not microscopically observable. Most of our discussion centers on mutations that take place in coding DNA or in regulatory sequences, because mutations that occur in other parts of the genome usually have no clinical consequences.

One important type of single-gene mutation is the **base-pair substitution,** in which one base pair is replaced by another.* This can result in a change in the amino acid sequence. However, because of the redundancy of the genetic code, many of these mutations do not change the amino acid sequence and thus typically have no effect. Such mutations are called **silent substitutions.**

*In molecular genetics, base-pair substitutions are also termed **point mutations.** However, "point mutation" was used in classical genetics to denote any mutation small enough to be unobservable under a microscope.

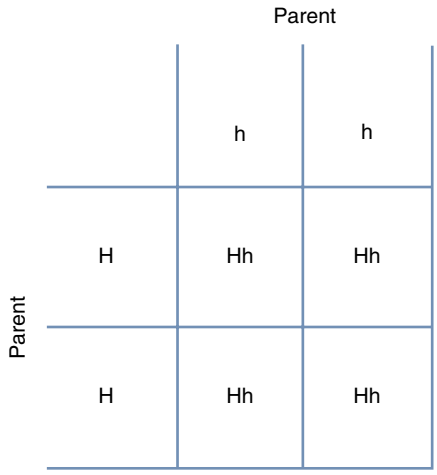


**Fig. 3.1** Punnett square illustrating a cross between *HH* and *hh* homozygote parents.

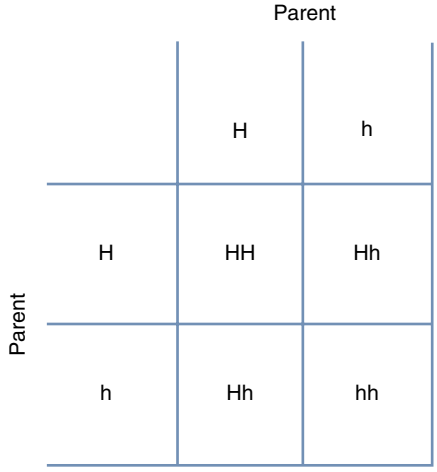


**Fig. 3.2** Punnett square illustrating a cross between two *Hh* heterozygotes.

Base-pair substitutions that alter amino acids consist of two basic types: **missense mutations,** which produce a change in a single amino acid, and **nonsense mutations,** which produce one of the three stop codons (UAA, UAG, or UGA) in the messenger RNA (mRNA) (Fig. 3.3). Because stop codons terminate translation of the mRNA, a premature nonsense mutation, also termed a **stop-gain**, results in early termination of the polypeptide chain or in destruction of the transcript through a process known as **nonsense-mediated mRNA decay**. Conversely, if a stop codon is altered so that it encodes an amino acid (a **stop-loss**), an abnormally elongated polypeptide can be produced. Alterations of amino acid sequences can have profound consequences, and many of the serious genetic diseases discussed later are the result of such alterations.

A second major type of mutation consists of **deletions** or **insertions** of one or more base pairs. These mutations, which can result in extra or missing amino acids in a protein, are often detrimental. An example of such a mutation is the 3-bp deletion that is found in most persons with cystic fibrosis (see Chapter 4). Deletions and insertions tend to be especially harmful when the number of missing or extra base pairs is not a multiple of three. Because codons consist of groups of three base pairs, such insertions or deletions can alter the downstream codons. This is a **frameshift mutation** (Fig. 3.4). For example, the insertion of a single base (an A in the second codon) converts a DNA sequence read as 5′-ACT GAT TGC GTT-3′ to 5′-ACT GAA TTG CGT-3′. This changes the amino acid sequence from Thr-Asp-Cys-Val to Thr-Glu-Leu-Arg. Often a frameshift mutation produces a stop codon downstream of the insertion or deletion, resulting in a truncated polypeptide.

On a larger scale, **duplications** of whole genes can also lead to genetic disease. A good example is given by Charcot–Marie–Tooth disease. This disorder, named after the three physicians who described it more than a century ago, is a peripheral nervous system condition that leads to progressive atrophy of the distal limb muscles. It affects approximately 1 in 2500 persons and exists in several different forms. About 70% of patients who have the most common form (type 1A) display a 1.5 million-bp (base pair) duplication on one copy of chromosome 17. As a result, they have three, rather than two, copies of the genes in this region. One of these genes, *PMP22*, encodes a component of peripheral myelin. The increased dosage of the gene product contributes to the demyelination that is characteristic of this form of the disorder. Interestingly, a deletion of this same region produces a different disease, hereditary neuropathy with liability to pressure palsies (paralysis). Because either a reduction (to 50%) or an increase (to 150%) in the gene product produces disease, this gene is said to display **dosage sensitivity.** Base-pair substitutions in *PMP22* itself can produce yet another disease, Dejerine–Sottas syndrome, which is characterized by distal muscle weakness, sensory alterations, muscular atrophy, and enlarged spinal nerve roots.

Other types of mutation can alter the regulation of transcription or translation. A **promoter mutation** can decrease the affinity of RNA polymerase for a promoter site, often resulting in reduced production of mRNA and thus decreased production of a protein. Mutations of transcription factor genes or enhancer sequences can have similar effects.

Mutations can also interfere with the splicing of introns as mature mRNA is formed from the primary mRNA transcript. **Splice-site mutations,** those that occur at intron–exon boundaries, alter the splicing signal that is necessary for proper excision of an intron. Splice-site mutations can occur at the GT sequence that defines the 5′ splice site (the **donor site**) or at the AG sequence that defines the 3′ splice site (the **acceptor site**). They can also take place in the sequences that lie near the donor and acceptor sites. When such mutations occur, the excision is often made within the next exon, at a splice site located in the exon. These splice sites, whose DNA sequences differ slightly from those of normal splice sites, are ordinarily unused and hidden within the exon. They are thus termed **cryptic splice sites.** The use of a cryptic site for splicing results in partial deletion of the exon, or in other cases, the deletion of an entire exon. As Fig. 3.5 shows, splice-site mutations can also result in the abnormal inclusion of part or all of an intron in the mature mRNA. Finally, a mutation can occur at a cryptic splice site, causing it to appear as a normal splice site and thus to compete with the normal splice site.

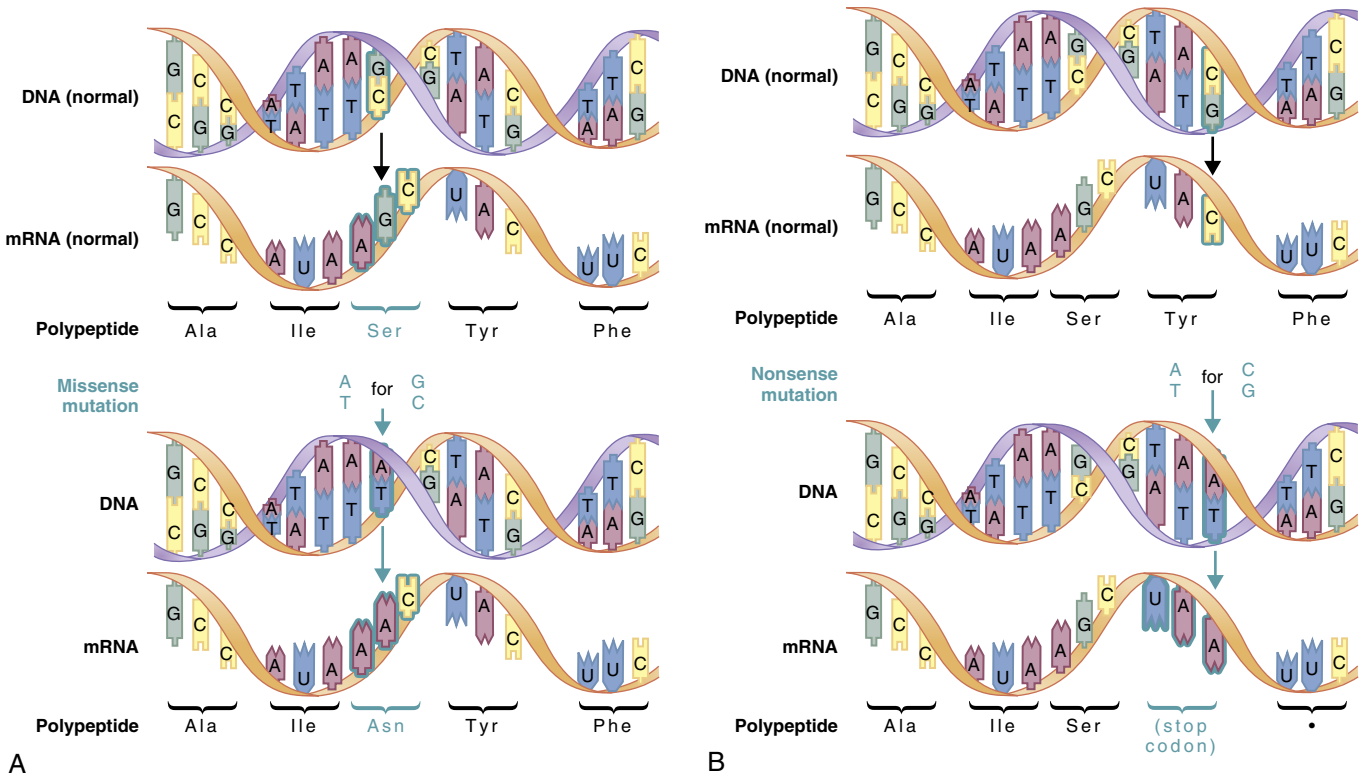


**Fig. 3.3** Base-pair substitution. Missense mutations **A,** produce a single amino acid change, whereas nonsense mutations **B,** produce a stop codon in the mRNA. Stop codons terminate translation of the polypeptide.

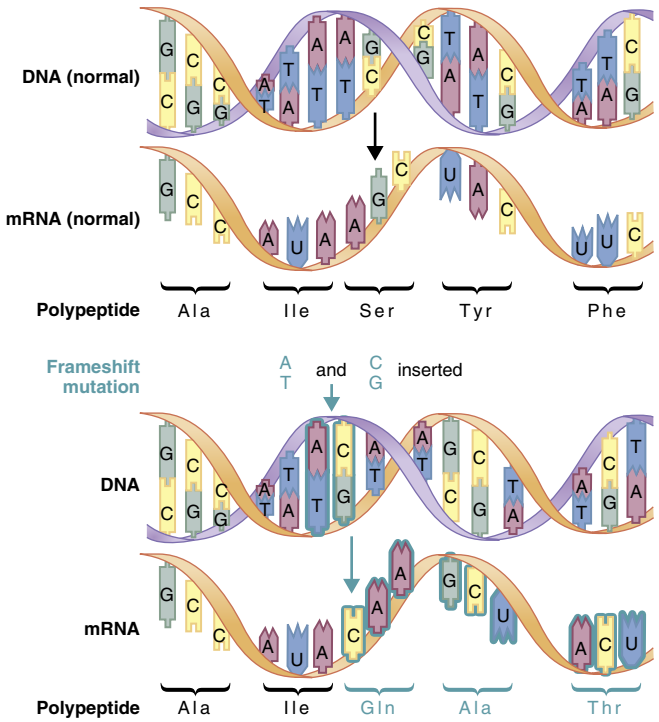


**Fig. 3.4** Frameshift mutations result from the addition or deletion of a number of bases that is not a multiple of three. This alters all of the codons downstream from the site of insertion or deletion.

Several types of DNA sequences are capable of propagating copies of themselves; these copies are then inserted in other locations on chromosomes (examples include the LINE and *Alu* repeats, discussed in Chapter 2). Such insertions can cause frameshift mutations. The insertion of **mobile elements** has been shown to cause isolated cases of type 1 neurofibromatosis, Duchenne muscular dystrophy, β-thalassemia, familial breast cancer, familial polyposis (colon cancer), and hemophilia A and B (clotting disorders) in humans.

The final type of mutation to be considered here affects tandem repeated DNA sequences (see Chapter 2) that occur within or near certain disease-related genes. The repeat units are usually 3 bp long, so a typical example would be CAGCAGCAG. A normal person has a relatively small number of these tandem repeats (e.g., 10 to 30 CAG consecutive elements) at a specific chromosome location. Occasionally the number of repeats increases during meiosis or possibly during early fetal development, so that a newborn might have hundreds or even thousands of tandem repeats. When this occurs in certain regions of the genome, it causes genetic disease. Like other mutations, these **expanded repeats** can be transmitted to the patient's offspring. More than 20 genetic diseases are now known to be caused by expanded repeats (see Chapters 4 and 5).

Mutations are the ultimate source of genetic variation. Some mutations result in genetic disease, but most have no physical effects. The principal types of mutation are missense, nonsense, frameshift, promoter, and splice-site mutations. Mutations can also be caused by the random insertion of mobile elements, and some genetic diseases are known to be caused by expanded repeats.

## MOLECULAR CONSEQUENCES OF MUTATION

It is useful to think of mutations in terms of their effects on the protein product. Broadly speaking, mutations can produce either a **gain of function** or a **loss of function** of

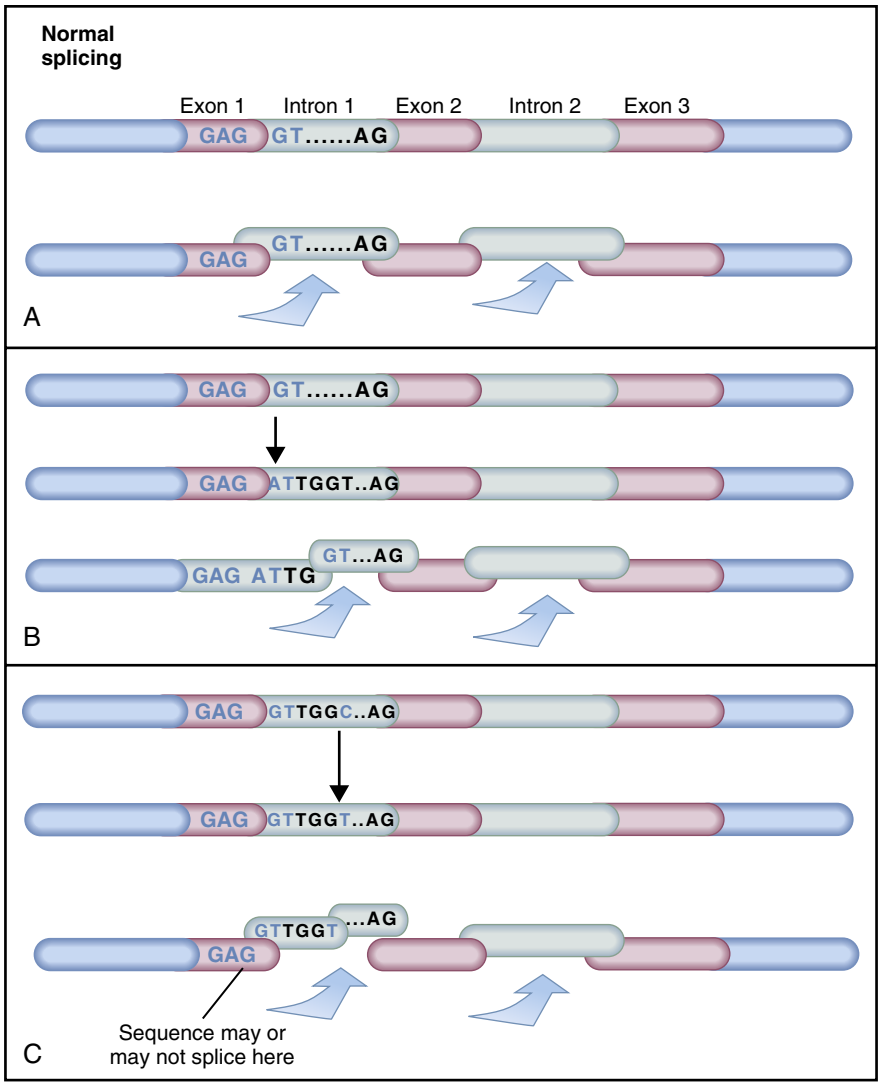


**Fig. 3.5** **A,** Normal splicing. **B,** Splice-site mutation. The donor sequence, GT, is replaced with AT. This results in an incorrect splice that leaves part of the intron in the mature mRNA transcript. In another example of splice-site mutation **C**, a second GT donor site is created within the first intron, resulting in a combination of abnormally and normally spliced mRNA products.

the protein product (Fig. 3.6). Gain-of-function mutations occasionally result in a completely novel protein product. More commonly, they result in overexpression of the product or inappropriate expression (i.e., in the wrong tissue or in the wrong stage of development). Gain-of-function mutations produce dominant disorders. Charcot–Marie–Tooth disease can result from overexpression of the protein product, which is considered a gain-of-function mutation. Huntington disease, discussed in Chapter 4, is another example.

Loss-of-function mutations are often seen in recessive diseases, where the mutation results in the loss of 50% of the protein product (e.g., a metabolic enzyme), but the 50% that remains is sufficient for normal function. The heterozygote is thus unaffected, but the homozygote, having little or no protein product, is affected. In some cases, however, 50% of the gene's protein product is not sufficient for normal function **(haploinsufficiency),** and a dominant disorder can result. Haploinsufficiency is seen, for example, in the autosomal dominant disorder known as familial hypercholesterolemia (see Chapter 12). In this disease, a single copy of a mutation (heterozygosity) reduces the number of low-density lipoprotein (LDL) receptors by 50%. Cholesterol levels in heterozygotes are approximately double those of normal homozygotes, resulting in a substantial increase in the risk of heart disease. As with most disorders involving haploinsufficiency, the disease is more serious in affected homozygotes (who have few or no functional LDL receptors) than in heterozygotes.

A **dominant negative** mutation results in a protein product that not only is abnormal but also inhibits the function of the protein produced by the normal allele in the heterozygote. Typically, dominant negative mutations are seen in genes that encode multimeric proteins (i.e., proteins composed of two or more subunits). Type I collagen (see Chapter 2), which is composed of three helical subunits, is an example of such a protein. An abnormal helix created by a single mutation can combine with the other helices, distorting them and producing a seriously compromised triple-helix protein.

Mutations can result in either a gain of function or a loss of function of the protein product. Gain-of-function mutations are sometimes seen in dominant diseases. Loss of function is seen in recessive diseases and in diseases involving haploinsufficiency, in which 50% of the gene product is insufficient for normal function. In dominant negative mutations, the abnormal protein product interferes with the normal protein product.

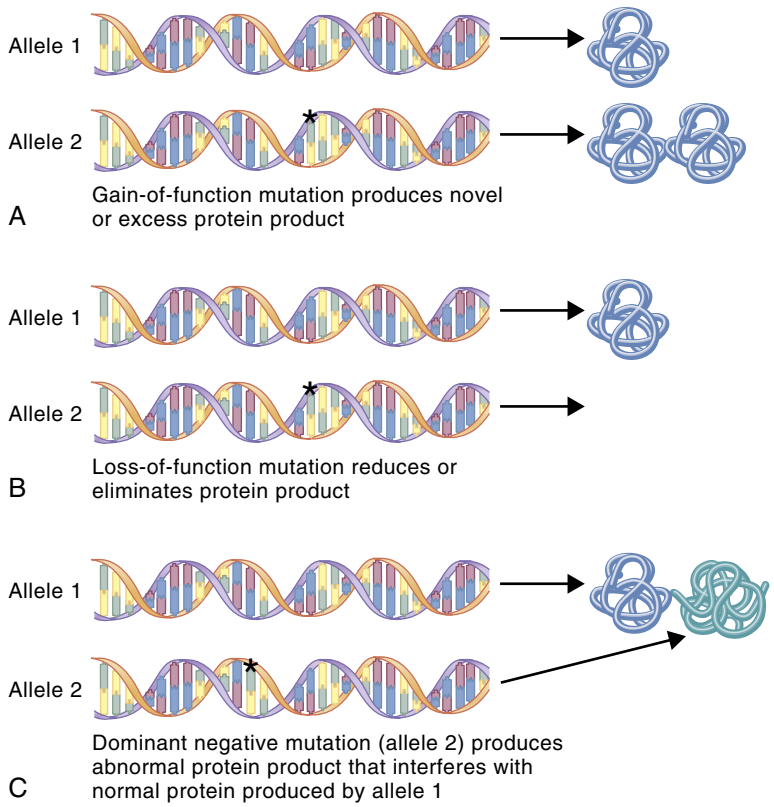


**Fig. 3.6 A,** Gain-of-function mutations produce a novel protein product or an increased amount of protein product. **B,** Loss-of-function mutations decrease the amount of protein product. **C,** Dominant negative mutations produce an abnormal protein product that interferes with the otherwise normal protein product of the normal allele in a heterozygote.

## CLINICAL CONSEQUENCES OF MUTATION: THE HEMOGLOBIN DISORDERS

Genetic disorders of human hemoglobin are the most common group of single-gene diseases; an estimated 7% of the world's population carries one or more mutations of the genes involved in hemoglobin synthesis. Because almost all of the types of mutation described in this chapter have been observed in the hemoglobin disorders, these diseases serve as an important illustration of the clinical consequences of mutation.

The hemoglobin molecule is a tetramer composed of four polypeptide chains, two labeled α and two labeled β. The β chains are encoded by a gene on chromosome 11, and the α chains are encoded by two genes on chromosome 16 that are very similar to each other. Typically, an individual has two normal β genes and four normal α genes (Fig. 3.7). Ordinarily tight regulation of these genes ensures that roughly equal numbers of α and β chains are produced. Each of these **globin** chains is associated with a **heme** group, which contains an iron atom and binds with oxygen. This property allows hemoglobin to perform the vital function of transporting oxygen in erythrocytes (red blood cells).

The hemoglobin disorders can be classified into two broad categories: structural abnormalities, in which the hemoglobin molecule is altered, and thalassemias, a group of conditions in which either the α- or the β-globin chain is structurally normal but reduced in quantity. Another condition, hereditary persistence of fetal hemoglobin (HPFH), occurs when fetal hemoglobin, encoded by the α-globin genes and by two β-globin–like genes called $^{A}\gamma$ and $^{G}\gamma$ (see Fig. 3.7), continues to be produced after birth (normally

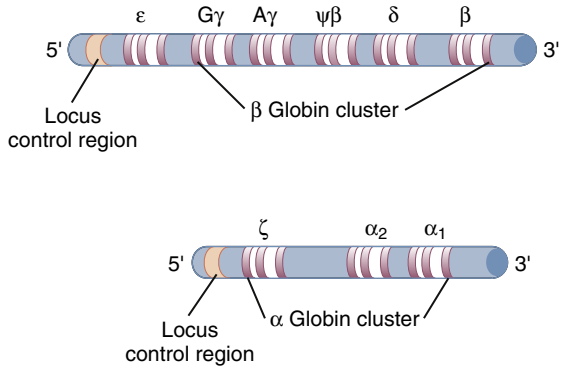


**Fig. 3.7** The α-globin gene cluster on chromosome 16 and the β-globin gene cluster on chromosome 11. The β-globin cluster includes the ε-globin gene, which encodes embryonic globin, and the γ-globin genes, which encode fetal globin. The ψβ gene is not expressed. The α-globin cluster includes the ζ-globin gene, which encodes embryonic α-globin.

γ-chain production ceases and β-chain production begins at the time of birth). HPFH does not cause disease but instead can compensate for a lack of normal adult hemoglobin.

A large array of different hemoglobin disorders have been identified. The discussion that follows is a greatly simplified presentation of the major forms of these disorders. The hemoglobin disorders, the mutations that cause them, and their major features are summarized in Table 3.1.

### Sickle Cell Disease

Sickle cell disease, which results from an abnormality of hemoglobin structure, is seen in approximately 1 in 400 to 1 in 600

**TABLE 3.1** Summary of the Major Hemoglobin Disorders

| Disease | Mutation Type | Major Disease Features |
|---|---|---|
| Sickle cell disease | β-globin missense mutation | Anemia, tissue infarctions, infections |
| HbH disease | Deletion or abnormality of three of the four α-globin genes | Moderately severe anemia, splenomegaly |
| Hydrops fetalis (Hb Barts) | Deletion or abnormality of all four α-globin genes | Severe anemia or hypoxemia, congestive heart failure; stillbirth or neonatal death |
| $\beta^0$-Thalassemia | Usually nonsense, frameshift, or splice-site donor or acceptor mutations; no β-globin produced | Severe anemia, splenomegaly, skeletal abnormalities, infections; often fatal during first decade if untreated |
| $\beta^+$-Thalassemia | Usually missense, regulatory, or splice-site consensus sequence or cryptic splice-site mutations; small amount of β-globin produced | Features similar to those of $\beta^0$-thalassemia but often somewhat milder |

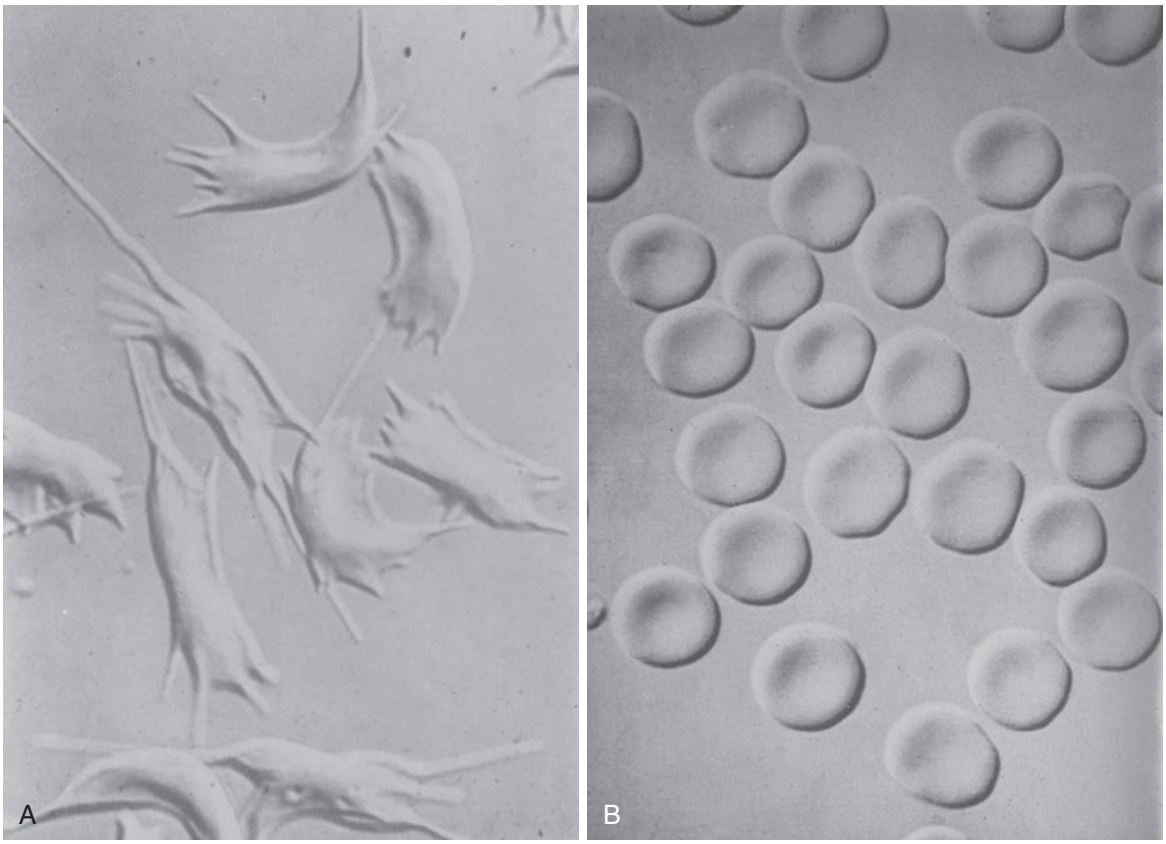


**Fig. 3.8 A,** Erythrocytes from patients with sickle cell disease assume a characteristic shape under conditions of low oxygen tension. **B,** Compare with normal erythrocytes.

African American births. It is even more common in parts of Africa, where it can affect up to 1 in 50 births, and it is also seen in Mediterranean, Middle Eastern, and South Asian populations. Sickle cell disease is typically caused by a single missense mutation that affects a substitution of valine for glutamic acid at position 6 of the β-globin polypeptide chain (Fig. 3.8). In homozygotes, this amino acid substitution alters the structure of hemoglobin molecules such that they form aggregates, causing erythrocytes to assume a characteristic sickle shape under conditions of low oxygen tension (see Fig. 3.8, *A*). These conditions are experienced in capillaries, the tiny vessels whose diameter is smaller than that of the erythrocyte. Normal erythrocytes (see Fig. 3.8, *B*) can squeeze through capillaries, but sickled erythrocytes are less flexible and are unable to do so. In addition, the abnormal erythrocytes tend to stick to the vascular endothelium (the innermost lining of blood vessels).

The resultant vascular obstruction produces localized hypoxemia (lack of oxygen), painful sickling crises, and infarctions of various tissues, including bone, spleen, kidneys, brain, and lungs (an infarction is tissue death due to hypoxemia). Premature destruction of the sickled erythrocytes decreases the number of circulating erythrocytes and the hemoglobin level, producing **anemia.** The spleen becomes enlarged (splenomegaly), but infarctions eventually destroy this organ, producing some loss of immune function. This contributes to the recurrent and sometimes fatal bacterial infections (especially pneumonia) that are commonly seen in persons with sickle cell disease. About 10% of persons with sickle cell disease experience a stroke before age 20 years. In North America, it is estimated that the life expectancy of persons with sickle cell disease is reduced by about 30 years.

Sickle cell disease, which causes anemia, tissue infarctions, and multiple infections, is the result of a single missense mutation that produces an amino acid substitution in the β-globin chain.

## Thalassemia

The term *thalassemia* is derived from the Greek word *thalassa* ("sea"). Thalassemia was first described in populations living near the Mediterranean Sea, although it is also common in portions of Africa, the Mideast, India, and Southeast Asia. In contrast to sickle cell disease, in which a mutation alters the structure of the hemoglobin molecule, the mutations that cause thalassemia reduce the quantity of either α globin or β globin. Thalassemia can be divided into two major groups, α-thalassemia and β-thalassemia, depending on the globin chain that is reduced in quantity. When one type of chain is decreased in number, the other chain type, unable to participate in normal tetramer formation, tends to form molecules consisting of four chains of the excess type only. These are termed **homotetramers,** in contrast to the **heterotetramers** normally formed by α and β chains. In α-thalassemia, the α-globin chains are deficient, so the β chains (or γ chains in the fetus) are found in excess. They form homotetramers that have a greatly reduced oxygen-binding capacity, producing hypoxemia. In β-thalassemia, the excess α chains form homotetramers that precipitate and damage the cell membranes of red blood cell precursors (i.e., the cells that form erythrocytes). This leads to premature erythrocyte destruction and anemia.

Most cases of α-thalassemia are caused by deletions of the α-globin genes. The loss of one or two of these genes has no clinical effect. The loss or abnormality of three of the α genes produces moderately severe anemia and splenomegaly (hemoglobin H [HbH] disease). Loss of all four α genes, a condition seen primarily among Southeast Asians, produces hypoxemia in the fetus and *hydrops fetalis* (a condition in which there is a massive buildup of fluid). Severe hydrops fetalis often causes stillbirth or neonatal death.

The α-thalassemia conditions are usually caused by deletions of α-globin genes. The loss of three of these genes leads to moderately severe anemia, and the loss of all four is fatal.

Persons with a β-globin mutation in one copy of chromosome 11 (heterozygotes) are said to have β-thalassemia minor, a condition that involves little or no anemia and does not ordinarily require clinical management. Those in whom both copies of the chromosome carry a β-globin mutation develop either β-thalassemia major (also called Cooley's anemia) or a less serious condition, β-thalassemia intermedia. Beta-globin may be completely absent ($\beta^0$-thalassemia), or it may be reduced to about 10% to 30% of the normal amount ($\beta^+$-thalassemia). Typically, $\beta^0$-thalassemia produces a more severe disease phenotype, but because disease features are caused by an excess of α-globin chains, patients with $\beta^0$-thalassemia are less severely affected when they also have α-globin mutations that reduce the quantity of α-globin chains.

Beta-globin is not produced until after birth, so the effects of β-thalassemia major are not seen clinically until the age of 2 to 6 months. These patients develop severe anemia. If the condition is left untreated, substantial growth retardation can occur. The anemia causes bone marrow expansion, which in turn produces skeletal changes, including a protuberant upper jaw and cheekbones and thinning of the long bones (making them susceptible to fracture). Splenomegaly (Fig. 3.9) and infections are common, and patients with untreated β-thalassemia major often die during the first decade of life. Beta-thalassemia can vary considerably in severity, depending on the precise nature of the responsible mutation.

**Fig. 3.9** A child with β-thalassemia major who has severe splenomegaly.

In contrast to α-thalassemia, gene deletions are relatively rare in β-thalassemia. Instead, most cases are caused by single-base mutations. Nonsense mutations, which result in premature termination of translation of the β-globin chain, usually produce $\beta^0$-thalassemia. Frameshift mutations also typically produce the $\beta^0$ form. In addition to mutations in the β-globin gene itself, alterations in regulatory sequences can occur. Beta-globin transcription is regulated by a promoter, two enhancers, and an upstream region known as the **locus control region (LCR)** (see Fig. 3.7). Mutations in these regulatory regions usually result in reduced synthesis of mRNA and a reduction, but not complete absence, of β-globin ($\beta^+$-thalassemia). Several types of splice-site mutations have also been observed. If a point mutation occurs at a donor or acceptor site, normal splicing is destroyed completely, producing $\beta^0$-thalassemia. Mutations in the surrounding consensus sequences usually produce $\beta^+$-thalassemia. Mutations also occur in the cryptic splice sites found in introns or exons of the β-globin gene, which cause these sites to be available to the splicing mechanism. These additional splice sites then compete with the normal splice sites, producing some normal and some abnormal β-globin chains. The result is usually $\beta^+$-thalassemia.

Many different types of mutations can produce β-thalassemia. Nonsense, frameshift, and splice-site donor and acceptor mutations tend to produce more severe disease. Regulatory mutations and those involving splice-site consensus sequences and cryptic splice sites tend to produce less severe disease.

Hundreds of different β-globin mutations have been reported. Consequently most patients with β-thalassemia are not homozygotes in the strict sense; they usually have a *different* β-globin mutation on each copy of chromosome 11 and are termed **compound heterozygotes** (Fig. 3.10). Even

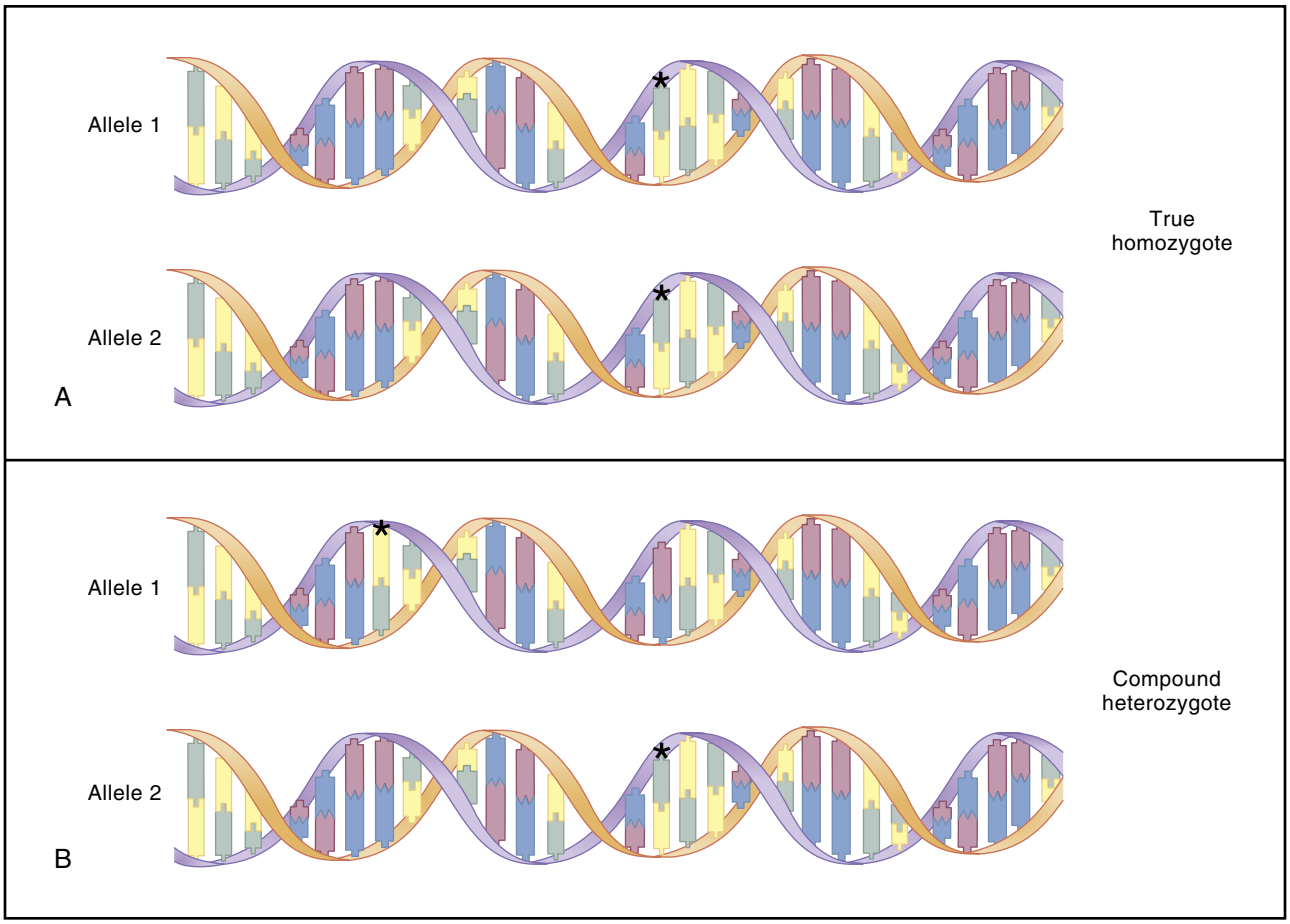


**Fig. 3.10 A,** True homozygotes have two alleles that are identical in DNA sequence. Here the homozygote has two copies of a single-base mutation, shown by the *asterisk* in the same position in the DNA sequence. Both mutations (alleles 1 and 2) have a loss-of-function effect, giving rise to a recessive disease. **B,** The same effect is seen in a compound heterozygote, which has two different mutations *(asterisks)* in two different locations in the gene's DNA sequence. Each allele has a loss-of-function effect, again causing a recessive disease.

though the mutations differ, each of the two β-globin genes is altered, producing a disease state. It is common to apply the term *homozygote* loosely to compound heterozygotes.

Patients with sickle cell disease or β-thalassemia major are sometimes treated with blood transfusions and with chelating agents that remove excess iron introduced by the transfusions. Prophylactic administration of antibiotics and antipneumococcal vaccine help to prevent bacterial infections in patients with sickle cell disease, and analgesics are administered for pain relief during sickling crises. Bone marrow transplantation, which provides donor stem cells that produce genetically normal erythrocytes, has been performed on patients with severe β-thalassemia or sickle cell disease. However, it is often impossible to find a suitably matched donor, and the mortality rate from this procedure is still fairly high (approximately 5% to 30%, depending on the severity of disease and the age of the patient). A lack of normal adult β-globin can be compensated for by reactivating the genes that encode fetal β-globin (the γ-globin genes, discussed previously). Agents such as hydroxycarbamide and butyrate can reactivate these genes and are being investigated. In addition, the transcription factor encoded by *BCL11A*, which ordinarily silences γ-globin expression after birth, is being explored as a target of drug and/or gene therapy for sickle cell disease. Beta-thalassemia is also a strong candidate for gene therapy (see Chapter 13).

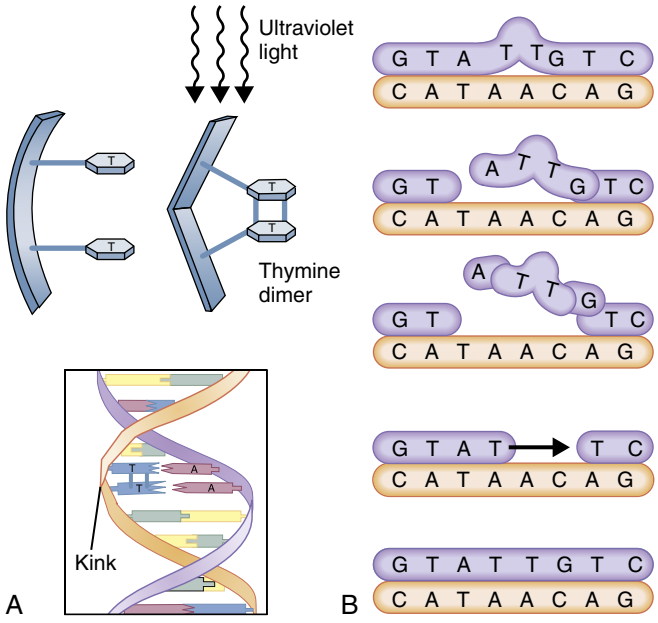


**Fig. 3.11 A,** Pyrimidine dimers originate when covalent bonds form between adjacent pyrimidine (cytosine or thymine) bases. This deforms the DNA, interfering with normal base pairing. **B,** The defect is repaired by removal and replacement of the dimer and bases on either side of it, with the complementary DNA strand used as a template.

## Clinical Commentary 3.1 The Effects of Radiation on Mutation Rates

Because mutation is a rare event, it is difficult to measure directly in humans. The relationship between radiation exposure and mutation is similarly difficult to assess. For a person living in a developed country, a typical lifetime exposure to ionizing radiation is about 6 to 7 rem.* About one-third to one-half of this amount is thought to originate from medical and dental x-ray procedures.

Unfortunately, a few human populations have received much larger radiation doses. The most thoroughly studied such population is the survivors of the atomic bomb blasts that occurred in Hiroshima and Nagasaki, Japan, at the close of World War II. Many of those who were exposed to high doses of radiation died from radiation sickness. Others survived, and many of the survivors produced offspring.

To study the effects of radiation exposure in this population, a large team of Japanese and American scientists conducted medical and genetic investigations of some of the survivors. A significant number developed cancers and chromosome abnormalities in their somatic cells, probably as a consequence of radiation exposure. To assess the effects of radiation exposure on the subjects' germlines, the scientists compared the offspring of those who suffered substantial radiation exposure with the offspring of those who did not. Although it is difficult to establish radiation doses with precision, there is no doubt that, in general, those who were situated closer to the blasts suffered much higher exposure levels. It is estimated that the exposed group received roughly 30 to 60 rem of radiation, many times the average lifetime radiation exposure.

In a sample of 77,000 offspring of these survivors, researchers assessed a large number of factors, including stillbirths, chromosome abnormalities, birth defects, cancer before 20 years of age, death before 26 years of age, and various measures of growth and development (e.g., intelligence quotient). There were no statistically significant differences between the offspring of persons who were exposed to radiation and the offspring of those who were not exposed. In addition, direct genetic studies of mutations have been carried out using minisatellite polymorphisms and protein electrophoresis, a technique that detects mutations that lead to amino acid changes (discussed elsewhere in this chapter). Parents and offspring were compared to determine whether germline mutations had occurred at various loci. The numbers of mutations detected in the exposed and unexposed groups were statistically equivalent.

More recently, studies of those who were exposed to radiation from the Chernobyl nuclear power plant accident have demonstrated a significant increase in thyroid cancers among children exposed to radiation. This reflects the effects of somatic mutations. The evidence for increased frequencies of germline mutations in protein-coding DNA, however, remains unclear. A number of other studies of the effects of radiation on humans have been reported, including investigations of those who live near nuclear power plants. The radiation doses received by these persons are substantially smaller than those of the populations discussed previously, and the results of these studies are equivocal.

It is remarkable that even though there was substantial evidence for radiation effects on somatic cells in the Hiroshima and Nagasaki studies, no detectable effect could be seen for germline cells. What could account for this? Because large doses of radiation are lethal, many of those who would have been most strongly affected would not be included in these studies. Furthermore, because germline mutation rates are very small, even relatively large samples of radiation-exposed persons may be insufficient to detect increases in mutation rates. It is also possible that DNA repair compensated for some radiation-induced germline damage.

These results argue that radiation exposure, which is clearly associated with somatic mutations, should not be taken lightly. Above-ground nuclear testing in the American Southwest has produced increased rates of leukemia and thyroid cancer in a segment of the population. Radon, a radioactive gas that is produced by the decay of naturally occurring uranium, can be found at dangerously high levels in some homes and poses a risk for lung cancer. Any unnecessary exposure to radiation, particularly to the gonads or to developing fetuses, should be avoided.

*A rem is a traditional unit for measuring radiation exposure. It is roughly equal to 0.01 joule of absorbed energy per kilogram of tissue. Parental exposures have also been estimated in Gray (Gy), which is another measure of the deposition of energy in tissue. In the minisatellite mutation study, the average gonadal dosage among "exposed" parents was 1.47 Gy, which is many times higher than lifetime exposure in the general population.

## CAUSES OF MUTATION

A large number of agents are known to cause **induced mutations.** These mutations, which are attributed to known environmental causes, can be contrasted with **spontaneous mutations,** which arise naturally in cells, for example during DNA replication. Agents that cause induced mutations are known collectively as **mutagens.** Animal studies have shown that **radiation** is an important class of mutagen (Clinical Commentary 3.1). **Ionizing radiation,** such as that produced by x-rays and nuclear fallout, can eject electrons from atoms, forming electrically charged ions. When these ions are situated within or near the DNA molecule, they can promote chemical reactions that change DNA bases. Ionizing radiation can also break the bonds of double-stranded DNA. This form of radiation can reach all cells of the body, including the germline cells.

**Nonionizing radiation** does not form charged ions but can move electrons from inner to outer orbits within an atom. The atom becomes chemically unstable. Ultraviolet (UV) radiation, which occurs naturally in sunlight, is an example of nonionizing radiation. UV radiation causes the formation of covalent bonds between adjacent pyrimidine bases (i.e., cytosine or thymine). These **pyrimidine dimers** (a dimer is a molecule having two subunits) are unable to pair properly with purines during DNA replication; this results in a base-pair substitution (Fig. 3.11). Because UV radiation is absorbed by the skin, it does not reach the germline, but it can cause skin cancer (Clinical Commentary 3.2).

A variety of chemicals can also induce mutations, sometimes because of their chemical similarity to DNA bases. Because of this similarity, these **base analogs,** such as 5-bromouracil, can be substituted for a true DNA base during replication. The analog is not exactly the same as the

## Clinical Commentary 3.2 Xeroderma Pigmentosum: A Disease of Faulty DNA Repair

An inevitable consequence of exposure to UV radiation is the formation of potentially dangerous pyrimidine dimers in the DNA of skin cells. Fortunately the highly efficient nucleotide excision repair (NER) system removes these dimers in normal persons. Among those affected with the rare autosomal recessive disease xeroderma pigmentosum (XP) this system does not work properly, and the resulting DNA replication errors lead to base-pair substitutions in skin cells. XP varies substantially in severity, but early symptoms are usually seen in the first 1 to 2 years of life. Patients develop dry, scaly skin (xeroderma) along with extensive freckling and abnormal skin pigmentation (pigmentosum). Skin tumors, which can be numerous, typically appear by 10 years of age. It is estimated that the risk of skin tumors in persons with XP is elevated approximately 1000-fold. These cancers are concentrated primarily in sun-exposed parts of the body. Patients are advised to avoid sources of UV light (e.g., sunlight), and cancerous growths are removed surgically. Neurological abnormalities are seen in about 30% of persons with XP. Severe, potentially lethal malignancies can occur before 20 years of age.

The NER system is encoded by at least 28 different genes, and inherited mutations in any of 7 of these genes can give rise to XP. These genes encode helicases that unwind the double-stranded DNA helix: an endonuclease that cuts the DNA at the site of the dimer; an exonuclease that removes the dimer and nearby nucleotides; a polymerase that fills the gap with DNA bases (using the complementary DNA strand as a template); and a ligase that rejoins the corrected portion of DNA to the original strand.

It should be emphasized that the expression of XP requires inherited germline mutations of an NER gene as well as subsequent uncorrected somatic mutations of genes in skin cells. Some of these somatic mutations can affect genes that promote cancer (see Chapter 11), resulting in tumor formation. The skin-cell mutations themselves are somatic and thus are not transmitted to future generations.

NER is but one type of DNA repair. The table below provides examples of a number of other diseases that result from defects in various types of DNA repair mechanisms (Fig. 3.12) (Table 3.2).

**Fig. 3.12** Xeroderma pigmentosum. This patient's skin has multiple hyperpigmented lesions, and skin tumors on the forehead have been marked for excision.

**TABLE 3.2** Examples of Diseases That Are Caused by a Defect in DNA Repair

| Disease | Features | Type of Repair Defect |
|---|---|---|
| Xeroderma pigmentosum | Skin tumors, photosensitivity, cataracts, neurological abnormalities | Nucleotide excision repair defects, including mutations in helicase and endonuclease genes |
| Cockayne syndrome | Reduced stature, skeletal abnormalities, optic atrophy, deafness, photosensitivity, intellectual disability | Defective repair of UV-induced damage in transcriptionally active DNA; considerable etiological and symptomatic overlap with xeroderma pigmentosum and trichothiodystrophy |
| Fanconi anemia | Anemia; leukemia susceptibility; limb, kidney, and heart malformations; chromosome instability | As many as eight different genes may be involved, but their exact role in DNA repair is not yet known |
| Bloom syndrome | Growth deficiency, immunodeficiency, chromosome instability, increased cancer incidence | Mutations in the reqQ helicase family |
| Werner syndrome | Cataracts, osteoporosis, atherosclerosis, loss of skin elasticity, short stature, diabetes, increased cancer incidence; sometimes described as "premature aging" | Mutations in the reqQ helicase family |
| Ataxia-telangiectasia | Cerebellar ataxia, telangiectases,* immune deficiency, increased cancer incidence, chromosome instability | Normal gene product is likely to be involved in halting the cell cycle after DNA damage occurs |
| Hereditary nonpolyposis colorectal cancer | Proximal bowel tumors, increased susceptibility to several other types of cancer | Mutations in any of six DNA mismatch-repair genes |

*Telangiectases are vascular lesions caused by the dilatation of small blood vessels. This typically produces discoloration of the skin.

base it replaces, so it can cause pairing errors during subsequent replications. Other chemical mutagens, such as acridine dyes, can physically insert themselves between existing bases, distorting the DNA helix and causing frameshift mutations. Still other mutagens can directly alter DNA bases, causing replication errors. An example of the latter is nitrous acid, which removes an amino group from cytosine, converting it to uracil. Although uracil is normally found in RNA, it mimics the pairing action of thymine in DNA. Thus it pairs with adenine instead of guanine, as the original cytosine would have done. The end result is a base-pair substitution.

Hundreds of chemicals are now known to be mutagenic in laboratory animals. Among these are nitrogen mustard, vinyl chloride, alkylating agents, formaldehyde, sodium nitrite, and saccharin. Some of these chemicals are much more potent mutagens than others. Nitrogen mustard, for example, is a powerful mutagen, whereas saccharin is a relatively weak one. Although some mutagenic chemicals are produced by humans, many occur naturally in the environment (e.g., aflatoxin $B_1$, a common contaminant of foods).

Many substances in the environment are known to be mutagenic, including ionizing and nonionizing radiation and hundreds of different chemicals. These mutagens are capable of causing base substitutions, deletions, and frameshifts. Ionizing radiation can induce double-stranded DNA breaks. Some mutagens occur naturally, and others are generated by humans.

## DNA REPAIR

Considering that 3 billion DNA base pairs must be replicated in each cell division, and considering the large number of mutagens to which we are exposed, DNA replication is surprisingly accurate. A primary reason for this accuracy is the process of **DNA repair,** which takes place in all normal cells of higher organisms. Several dozen enzymes are involved in the repair of damaged DNA. They collectively recognize an altered base, excise it by cutting the DNA strand, replace it with the correct base (determined from the complementary strand), and reseal the DNA. These repair mechanisms are estimated to correct at least 99.9% of initial errors.

Because DNA repair is essential for the accurate replication of DNA, defects in DNA repair systems can lead to many types of disease. For example, inherited mutations in genes responsible for **DNA mismatch repair** result in the persistence of cells with replication errors (i.e., **mismatches**) and can lead to some types of cancer (see Chapter 11). A diminished capacity to repair **double-stranded DNA breaks** can lead to ovarian and/or breast cancer. **Nucleotide excision repair** is necessary for the removal of larger changes in the DNA helix (e.g., pyrimidine dimers); defects in excision repair lead to a number of diseases, of which xeroderma pigmentosum is an important example (see Clinical Commentary 3.2).

DNA repair helps to ensure the accuracy of the DNA sequence by correcting replication errors (mismatches), repairing double-stranded DNA breaks, and excising damaged nucleotides.

## MUTATION RATES

How often do spontaneous mutations occur? At the nucleotide level, the mutation rate is estimated to be about $1.3 \times 10^{-8}$ per base pair per generation (this figure represents mutations that have escaped the process of DNA repair). Thus each gamete contains approximately 35 new mutations, the great majority of which occur in noncoding DNA. At the level of the gene, the mutation rate is quite variable, ranging from $10^{-4}$ to $10^{-7}$ per locus per cell division. There are at least two reasons for this large range of variation: the size of the gene and the susceptibility of certain nucleotide sequences.

First, genes vary tremendously in size. The somatostatin gene, for example, is quite small, containing 1480 bp. In contrast, the gene responsible for Duchenne muscular dystrophy *(DMD)* spans more than 2 million bp. As might be expected, larger genes present larger targets for mutation and usually experience mutation more often than do smaller genes. The *DMD* gene, as well as the genes responsible for hemophilia A and type 1 neurofibromatosis, are all very large and have high mutation rates.

Second, it is well established that certain nucleotide sequences are especially susceptible to mutation. These are termed **mutation hot spots.** The best-known example is the two-base **(dinucleotide)** sequence CG. In mammals, about 80% of CG dinucleotides are **methylated;** a methyl group is attached to the cytosine base (these dinucleotide sequences are also labeled CpG [cytosine-phosphate-guanine], to distinguish the two-base DNA sequence from a single pair of complementary bases, C and G). A methylated cytosine, 5-methylcytosine, easily loses an amino group, converting it to thymine. The end result is a mutation from cytosine to thymine (Fig. 3.13). Surveys of mutations in human genetic diseases have shown that the mutation rate at CG dinucleotides is about 12 times higher than at other dinucleotide sequences. Mutation hot spots, in the form of CG dinucleotides, have been identified in a number of important human disease genes, including the procollagen genes

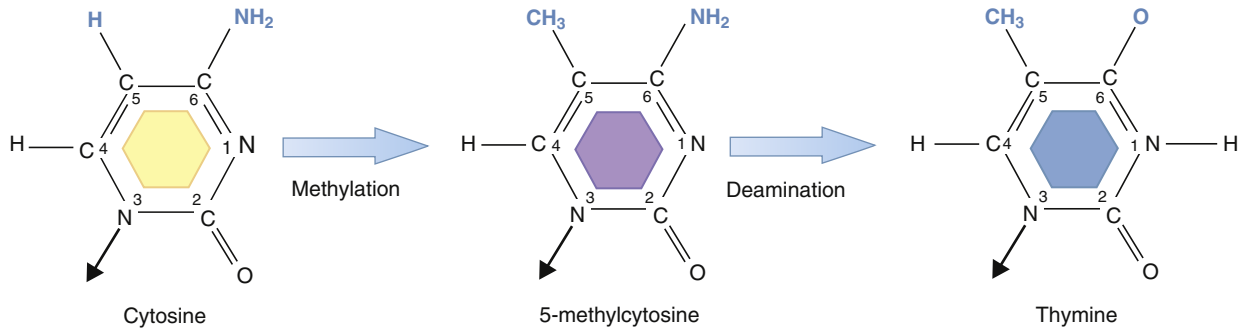


**Fig. 3.13** Cytosine methylation. The addition of a methyl group ($CH_3$) to a cytosine base forms 5-methylcytosine. The subsequent loss of an amino group (deamination) forms thymine. The result is a cytosine → thymine substitution.

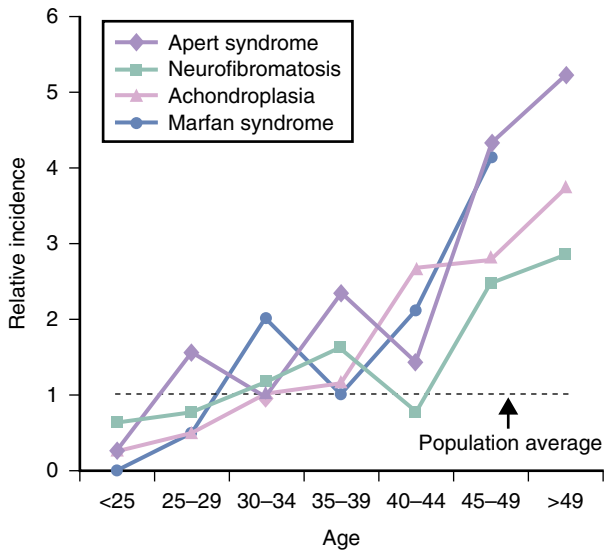


**Fig. 3.14** Paternal age effect. For some single-gene disorders, the risk of producing a child with the condition *(y-axis)* increases with the father's age *(x-axis)*.

responsible for osteogenesis imperfecta (see Chapter 2). Other disease examples are discussed in Chapters 4 and 5.

Parental age is strongly correlated with the probability of transmitting a mutation to one's offspring. Some chromosome abnormalities increase dramatically with maternal age (see Chapter 6), and single-base mutations increase with paternal age. The latter is seen in several single-gene disorders, including Marfan syndrome and achondroplasia. As Fig. 3.14 shows, the risk of producing a child with Marfan syndrome is several times higher for a father older than 40 years than for a father in his 20s. This paternal age effect is usually attributed to the fact that the stem cells giving rise to sperm cells continue to divide throughout life, which allows a progressive buildup of DNA replication errors. Recent comparisons of whole genome sequences in parents and offspring estimate that approximately one to two additional mutations are transmitted with each additional year of paternal age. Most studies show a smaller but significant effect of maternal age on the single-base mutation rate, on the order of approximately 0.5 additional mutations per year of maternal age.

Large genes, because of their size, are generally more likely to experience mutations than are small genes. Mutation hot spots, particularly methylated CG dinucleotides, experience elevated mutation rates. For many single-gene disorders there is a substantial increase in mutation risk with advanced paternal age.

# Detection and Measurement of Genetic Variation

For centuries humans have been intrigued by the differences that can be seen among individuals. Attention was long focused on observable differences such as skin color or body shape and size. Only in the 20th century did it become possible to examine variation in genes, the consequence of mutations accumulated through time. The evaluation and measurement of this variation in populations and families are important for mapping genes to specific locations on chromosomes, a key step in determining gene function (see Chapter 8). The evaluation of genetic variation also provides the basis for genetic diagnosis, and it is highly useful in forensics. In this section, several key approaches to detecting genetic variation in humans are discussed in historical sequence.

## BLOOD GROUPS

Several dozen **blood group** systems have been defined on the basis of antigens located on the surfaces of erythrocytes. Some are involved in determining whether a person can receive a blood transfusion from a specific donor. Because individuals differ extensively in terms of blood groups, these systems provided an important early means of assessing genetic variation.

Each of the blood group systems is determined by a different gene or set of genes. The various antigens that can be expressed within a system are the result of different DNA sequences in these genes. Two blood-group systems that have special medical significance—the ABO and Rh blood groups—are discussed here. The ABO and Rh systems are both important in determining the compatibility of blood transfusions and tissue grafts. Some combinations of these blood groups can produce maternal-fetal incompatibility, sometimes with serious results for the fetus. These issues are discussed in detail in Chapter 9.

### The ABO Blood Group

Human blood transfusions were carried out as early as 1818, but they were often unsuccessful. After transfusion, some recipients suffered a massive, sometimes fatal, hemolytic reaction. In 1900 Karl Landsteiner discovered that this reaction was caused by the ABO antigens located on erythrocyte surfaces. The ABO system consists of two major antigens, labeled A and B. A person can have one of four major blood types: people with blood type A carry the A antigen on their erythrocytes, those with type B carry the B antigen, those with type AB carry both A and B, and those with type O carry neither antigen. Each individual has antibodies that react against any antigens that are not found on their own red blood cell surfaces. For example, a person with type A blood has anti-B antibodies, and transfusing type B blood into this person provokes a severe antibody reaction. It is straightforward to determine ABO blood type in the laboratory by mixing a small sample of a person's blood with solutions containing different antibodies and observing which combinations cause the observable clumping that is characteristic of an antibody–antigen interaction.

The ABO system, which is encoded by a single gene on chromosome 9, consists of three primary alleles, labeled $I^A$, $I^B$, and $I^O$. (There are also subtypes of both the $I^A$ and $I^B$ alleles, but they are not addressed here.) Persons with the $I^A$ allele have the A antigen on their erythrocyte surfaces (blood type A), and those with $I^B$ have the B antigen on their cell surfaces (blood type B). Those with both alleles express both antigens (blood type AB), and those with two copies of the $I^O$ allele have neither antigen (type O blood). Because the $I^O$ allele produces no antigen, persons who are $I^A\,I^O$ or $I^B\,I^O$ heterozygotes have blood types A and B, respectively (Table 3.3).

Because populations vary substantially in terms of the frequency with which the ABO alleles occur, the ABO locus was the first blood group system to be used extensively in

**TABLE 3.3** Relationship between ABO Genotype and Blood Type

| Genotype | Blood Type | Antibodies Present |
|---|---|---|
| $I^AI^A$ | A | Anti-B |
| $I^AI^O$ | A | Anti-B |
| $I^BI^B$ | B | Anti-A |
| $I^BI^O$ | B | Anti-A |
| $I^AI^B$ | AB | None |
| $I^OI^O$ | O | Anti-A and anti-B |

studies of genetic variation among individuals and populations. For example, early studies showed that the A antigen is relatively common in western European populations, and the B antigen is especially common among Asians. Neither antigen is common among native South American populations, the great majority of whom have blood type O.

### The Rh System

Like the ABO system, the Rh system is defined on the basis of antigens that are present on erythrocyte surfaces. This system is named after the rhesus monkey, the experimental animal in which it was first isolated by Landsteiner in the late 1930s. It is typed in the laboratory by a procedure similar to the one described for the ABO system. Rh alleles vary considerably among individuals and populations and thus have been another highly useful tool for assessing genetic variation. The molecular basis of variation in both the ABO and the Rh systems has been elucidated (for further details, see the suggested readings at the end of this chapter), and it is becoming increasingly common to type these systems by directly examining an individual's DNA sequence rather than by assessing an antibody-antigen reaction.

The blood groups, of which the ABO and Rh systems are examples, have provided an important means of studying human genetic variation. Blood group variation is the result of antigens that occur on the surface of erythrocytes.

## PROTEIN ELECTROPHORESIS

**Protein electrophoresis,** developed first in the 1930s and applied widely to humans in the 1950s and 1960s, increased the number of detectable polymorphic loci considerably. This technique makes use of the fact that a single amino acid difference in a protein (the result of a mutation in the corresponding DNA sequence) can cause a slight difference in the electrical charge of the protein.

An example is the common sickle cell disease mutation discussed earlier. The replacement of glutamic acid with valine in the β-globin chain produces a difference in electrical charge because glutamic acid has two carboxyl groups, whereas valine has only one carboxyl group. Electrophoresis can be used to determine whether a person has normal hemoglobin (HbA) or the mutation that causes sickle cell disease (HbS) (Fig. 3.15). The hemoglobin is placed in an electrically charged gel composed of starch, agarose, or polyacrylamide (see Fig. 3.15, *A*). The slight difference in charge resulting from the amino acid difference causes the HbA and HbS forms to migrate at different rates through the gel. The protein molecules are allowed to migrate for several hours and are then stained with chemical solutions so that their positions can be seen (see Fig. 3.15, *B*). From the resulting pattern it can be determined whether the person is an HbA homozygote, an HbS homozygote, or a heterozygote having HbA on one chromosome copy and HbS on the other.

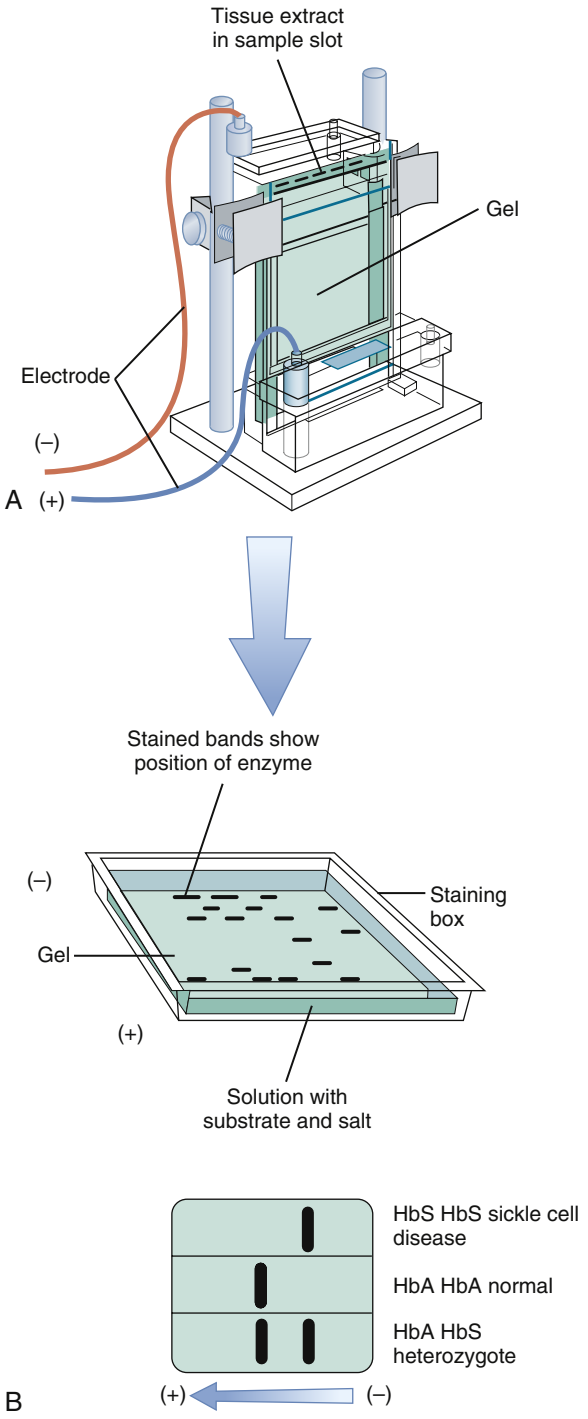


**Fig. 3.15** The process of protein electrophoresis. **A,** A tissue sample is loaded in the slot at the top of the gel, and an electrical current is run though the gel. After staining, distinct bands, representing molecules with different electrical charges and therefore different amino acid sequences, are visible. **B,** HbA homozygotes show a single band closer to the positive pole, whereas HbS homozygotes show a single band closer to the negative pole. Heterozygotes, having both alleles, show two bands.

Protein electrophoresis has been used to detect amino acid variation in hundreds of human proteins. However, silent substitutions, which do not alter amino acids, cannot be detected by this approach. In addition, some amino acid substitutions do not alter the electrical charge of the protein molecule. For these reasons, protein electrophoresis detects only about one-third of the mutations that occur in coding DNA. In addition, single-base substitutions in noncoding DNA are not usually detected by protein electrophoresis.

Protein electrophoresis detects variations in genes that encode certain serum proteins. These variations are observable because proteins with slight differences in their amino acid sequence migrate at different rates through electrically charged gels.

## DETECTING VARIATION AT THE DNA LEVEL

Each human haploid DNA sequence (i.e., the sequence inherited from one parent) differs from any other human haploid sequence by at least 3 to 4 million DNA base pairs, which amounts to one single-base difference every 1000 base pairs. Because there are only a few hundred or so blood group and protein electrophoretic polymorphisms, these approaches have detected only a tiny fraction of human DNA variation. Yet the assessment of this variation is critical to gene identification and genetic diagnosis (see Chapters 8 and 13). Fortunately, molecular techniques developed since the 1980s have enabled the detection of millions of new polymorphisms at the DNA level. These types of variants, and the techniques used to detect them, have revolutionized both the practice and the potential of medical genetics.

### Single Nucleotide Polymorphisms

The most numerous type of polymorphism in the human genome consists of variants at single nucleotide positions on a chromosome, or **single nucleotide polymorphisms (SNPs).** For example, one individual might have an A-T base pair at a given position, while another would have a G-C base pair. Increasingly the term **single nucleotide variant (SNV)** is used, reflecting the fact that many of these variants are rare (< 1% frequency) in populations. SNPs or SNVs, when they occur in functional DNA sequences, can cause inherited diseases, although most are harmless. Increasingly they are being detected by microarray and direct sequencing methods, which are discussed later in this chapter.

### Tandem Repeat Polymorphisms

The great majority of SNPs have only two possible alleles, placing a limit on the amount of genetic diversity they can reveal. More diversity at a single locus could be observed if it had many alleles, rather than just two. This type of diversity is seen in microsatellite and minisatellite DNA. As discussed in Chapter 2, microsatellites and minisatellites are regions in which the same DNA sequence is repeated over and over in tandem (Fig. 3.16). Microsatellites are composed of units that are only 1 to 10 bp long, which are termed **short tandem repeats (STRs)**, whereas minisatellites contain longer repeat units. The number of repeat units in a given region varies substantially from individual to individual; a specific region could have as few as two or three repeats or as many as 20 or more. These polymorphisms can therefore reveal a high degree of genetic variation. STRs are easy to assay using polymerase chain reaction (PCR) (see below), and more than 1 million of them are distributed throughout the human genome. These properties make them useful for mapping genes by the process of linkage analysis, discussed in Chapter 8. They are also useful in forensic applications, such as paternity testing and the identification of criminal suspects (see Box 3.1) (Fig. 3.17).

Single nucleotide polymorphisms (SNPs), or single nucleotide variants (SNVs), represent nucleotide positions whose DNA bases vary among individuals. STRs are a type of polymorphism that results from varying numbers of microsatellite repeats in a specific DNA region. Because they can have many different alleles in a population, they are highly useful in forensics and in gene mapping.

### Insertions, Deletions, and Copy Number Variants

Throughout the human genome, there are sections of DNA that vary in their number of copies from one individual to another. Variants whose size is less than 50 bp are caused by insertions or deletions and are termed **indels** (unlike STRs, indels do not occur in multiple tandem copies). On average, each human genome contains approximately 600,000 indels, most of which are less than 10 bp in size. Variants larger than 50 bp are termed **structural variants** and can be as large as entire chromosomes (see Chapter 6). An important class of structural variants are **copy number variants (CNVs),** typically defined as DNA sequences larger than 500–1,000 base pairs (definitions of the sizes of indels, structural variants, and CNVs vary somewhat and have not yet been standardized). CNVs may be present in zero to more than a dozen copies in a haploid genome, and each human is heterozygous for at least 100 CNVs (i.e., they inherited a different number of copies from the mother than from the father). Although CNVs are much less numerous than SNPs, their large individual size means that they account for at least several million total base pair differences between any pair of haploid DNA sequences (roughly the same amount as SNPs). Some CNVs have been shown to be associated with inherited diseases, and some

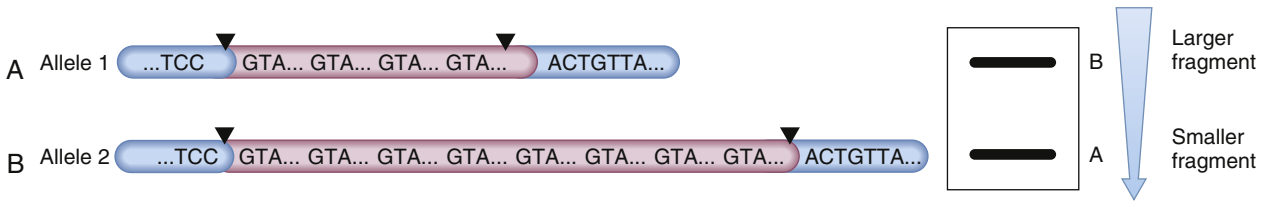


**Fig. 3.16** Tandem repeat polymorphisms. Bands of differing length **(A** and **B)** are created by different numbers of tandem repeats in the DNA on the two copies of a chromosome. Following amplification and labeling of the region that contains the polymorphism, different fragment lengths are separated by electrophoresis and visualized on an autoradiogram.

## BOX 3.1 DNA Profiles in the Forensic Setting

Because of the large number of polymorphisms observed in the human genome, it is virtually certain that each of us is genetically unique (with the exception of identical twins, whose DNA sequences are nearly always identical). It follows that genetic variation could be used to identify individuals, much as a conventional fingerprint does. Because DNA can be found in any tissue sample, including blood, semen, and hair,* genetic variation has substantial potential in forensic applications (e.g., criminal cases, paternity suits, identification of accident victims). STRs, with their many alleles, are very useful in establishing a highly specific **DNA profile.**

The principle underlying a DNA profile is quite simple. If we examine enough polymorphisms in a given individual, the probability that any other individual in the population has the same allele at each examined locus becomes extremely small. DNA left at the scene of a crime in the form of blood or semen, for example, can be typed for a series of STRs. Because of the extreme sensitivity of the PCR approach, even a tiny sample several years old can yield enough DNA for laboratory analysis (although extreme care must be taken to avoid contamination when using PCR with such samples). The detected alleles are then compared with the alleles of a suspect. If the alleles in the two samples match, the suspect is implicated (see Fig. 3.17).

A key question is whether another person in the general population might have the same alleles as the suspect. Could the DNA profile then falsely implicate the wrong person? In criminal cases, the probability of obtaining an allele match with a random member of the population is calculated. Because of the high degree of allelic variation in STRs, this probability is usually very small. The use of 13 or more STRs, which is now common practice, yields random match probabilities in the neighborhood of 1 in 1 trillion. Provided that a large enough number of loci are used under well-controlled laboratory conditions, and provided that the data are collected and evaluated carefully, DNA profiles can furnish highly useful forensic evidence. DNA profiles are now used in many thousands of criminal court cases each year.

Although we tend to think of such evidence in terms of identifying the guilty party, it should be kept in mind that when a match is not obtained, a suspect may be exonerated. In addition, postconviction DNA testing has resulted in the release of hundreds of persons who were wrongly imprisoned. Thus DNA profiles can also benefit the innocent.

In recent years, panels of SNPs have been developed that can predict an individual's genetic ancestry, eye color, and hair color with a fair degree of accuracy. Methylation profiles can estimate age within several years. Some predictions about the identity of a perpetrator can be made on the basis of the DNA sample alone. In addition, a number of criminal suspects have been identified because their DNA profile, or that of one or more relatives, is already stored in a national database and matches the DNA profile from an evidentiary sample. Although these developments increase the potential for forensic identification, they also pose new legal and ethical questions.

*Even fingerprints left at a crime scene sometimes contain enough DNA for PCR amplification and DNA profiling.

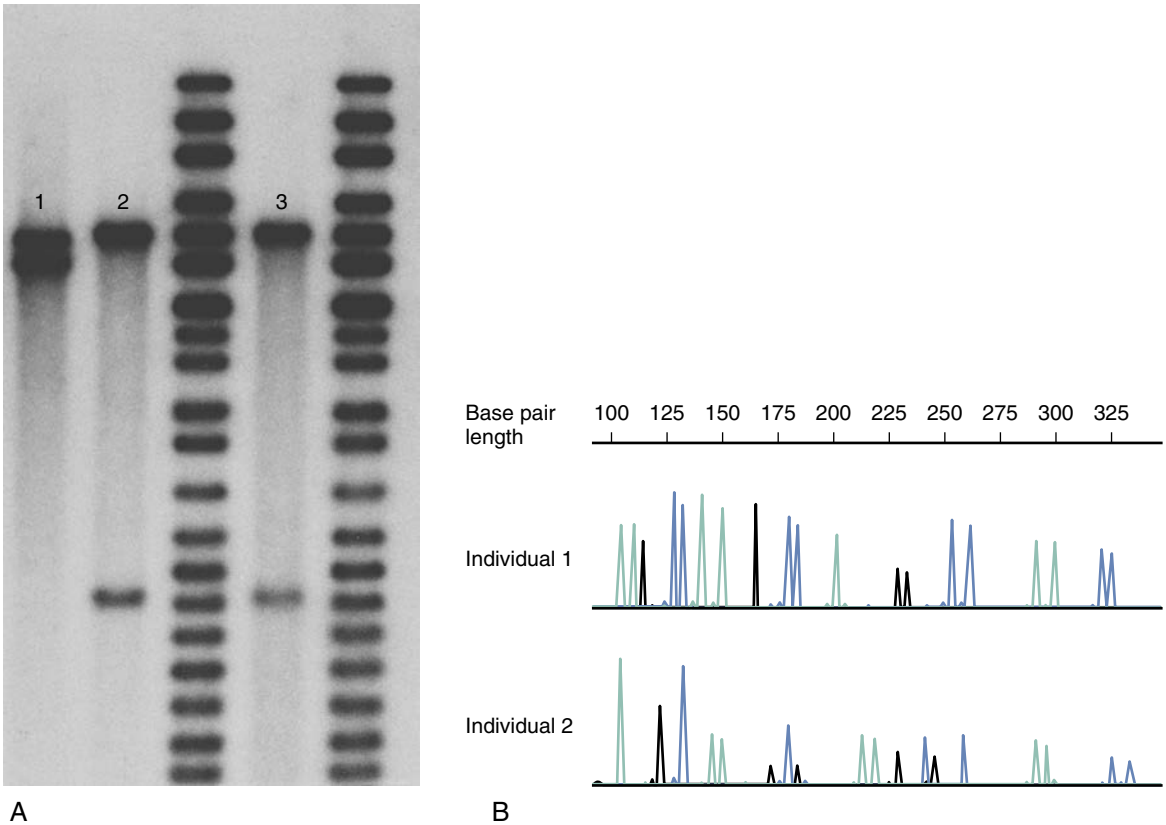


**Fig. 3.17** DNA profiles. **A,** An autoradiogram shows that the band pattern of DNA from suspect 1 does not match the DNA taken from the crime scene (3), whereas the band pattern from suspect 2 does match. In practice, multiple STRs are assayed to reduce the possibility of a false match. **B,** STRs are now commonly assayed using a capillary gel apparatus. The resulting STR profile is displayed as an electropherogram, in which the locations of peaks indicate the lengths (in base pairs) of each STR allele. (Panel **A** courtesy Jay Henry, Criminalistics Laboratory, Department of Public Safety, State of Utah.)

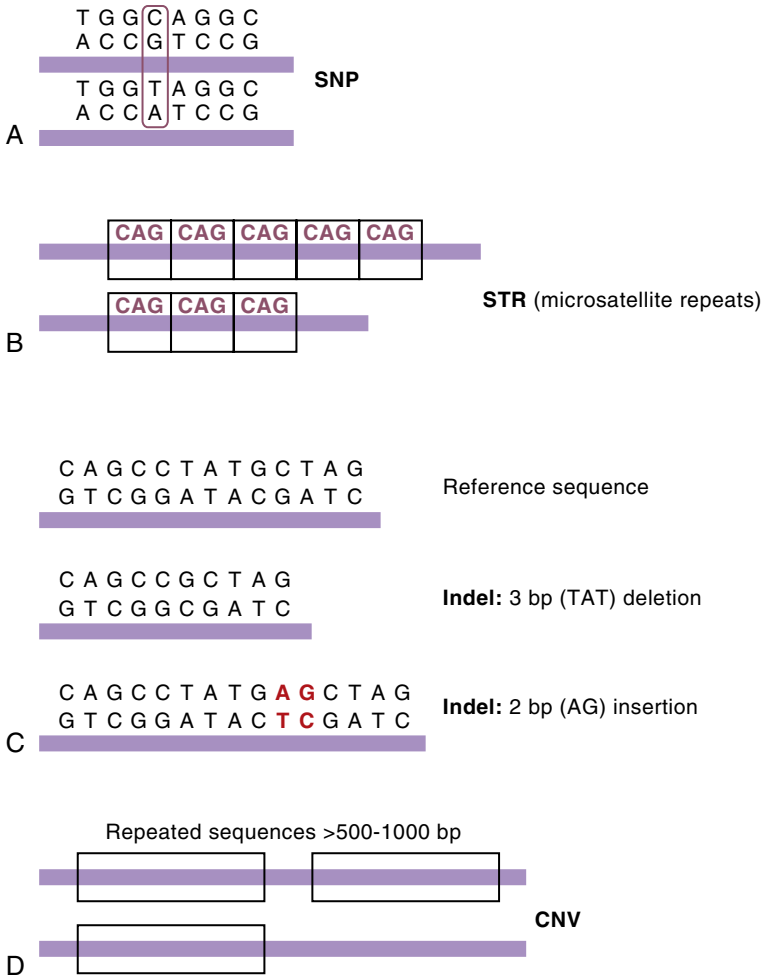


**Fig. 3.18** **A,** Single nucleotide polymorphisms (SNPs) are single-base variations in the genome. **B,** Tandem repeats consist of short segments of DNA, such as microsatellites, that are repeated over and over, in tandem. **C,** Indels are single, small (< 50 bp) deletions or insertions of DNA, relative to the reference sequence. **D,** Copy number variants (CNVs) represent differences in the numbers of larger repeated segments of DNA (>500 bp).

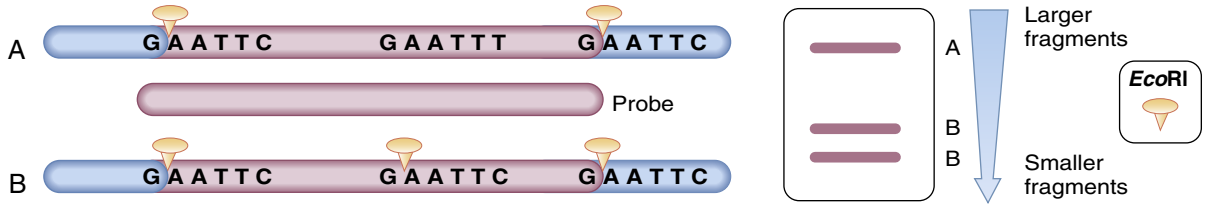


**Fig. 3.19** Cleavage of DNA by the *Eco*RI restriction enzyme. In **B,** the enzyme cleaves the three GAATTC recognition sequences, producing two smaller fragments. In **A,** the middle sequence is GAATTT instead of GAATTC, so it cannot be cleaved by the enzyme. The result is a single, longer fragment.

are associated with response to specific therapeutic drugs. Fig. 3.18 highlights the differences among SNPs, tandem repeats, indels, and CNVs.

Indels are defined as insertions or deletions of segments of DNA smaller than 50 bp. CNVs are a type of structural variant that consists of differences in the number of repeated DNA sequences longer than 500–1000 bp.

## Southern Blotting and Restriction Fragment Analysis

An early approach to the detection of genetic variation at the DNA level took advantage of the existence of bacterial enzymes known as **restriction endonucleases,** or **restriction enzymes.** These enzymes cleave human DNA at specific sequences, termed **restriction sites.** For example, the intestinal bacterium *Escherichia coli* produces a restriction enzyme called *Eco*RI, which recognizes the DNA sequence GAATTC. Each time this sequence is encountered, the enzyme cleaves the sequence between the G and the A (Fig. 3.19). A **restriction digest** of human DNA using this enzyme will produce more than 1 million DNA fragments **(restriction fragments).** These fragments are then subjected to gel electrophoresis, in which the smaller ones migrate more quickly through the gel than do the larger ones (Fig. 3.20). The DNA is denatured (i.e., converted from a double-stranded to a single-stranded form) by exposing it to alkaline chemical solutions. To fix their positions permanently, the DNA fragments are transferred from the gel to a solid membrane, such

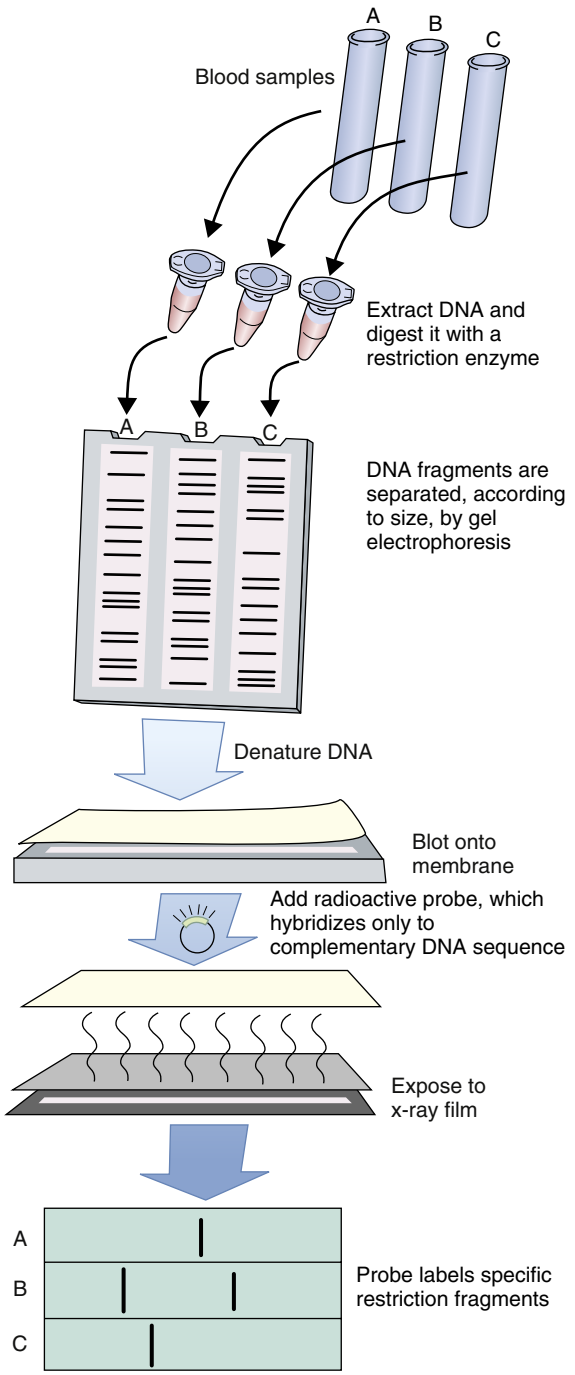


**Fig. 3.20** Restriction enzyme digestion and Southern blotting. DNA is extracted from blood samples from subjects **A, B,** and **C.** The DNA is digested by a restriction enzyme and then loaded on a gel. Electrophoresis separates the DNA fragments according to their size. The DNA is denatured and transferred to a solid membrane (Southern blot), where it is hybridized with a radioactive probe. Exposure to x-ray film (autoradiography) reveals specific DNA fragments (bands) of different sizes in individuals **A, B,** and **C.**

as nitrocellulose (this is a **Southern transfer,** named after the man who invented the process in the mid-1970s). At this point, the solid membrane, often called a **Southern blot,** contains many thousands of fragments arrayed according to their size. Because of their large number, the fragments are indistinguishable from one another.

To visualize only the fragments corresponding to a specific region of DNA, a **probe,** consisting of a small piece of single-stranded human DNA (a few kilobases [kb] in length), is constructed using recombinant DNA techniques (see Box 3.2) (Fig. 3.21). The probe is labeled, often with a radioactive isotope, and then exposed to the Southern blot. The

## BOX 3.2 Genetic Engineering, Recombinant DNA, and Cloning

In the last two decades most of the lay public has acquired at least a passing familiarity with the terms "recombinant DNA," "cloning," and "genetic engineering." Indeed, these techniques lie at the heart of what is often called the "new genetics."

**Genetic engineering** refers to the laboratory alteration of genes. An alteration that is of special importance in medical genetics is the creation of **clones.** Briefly, a clone is an identical copy of a DNA sequence. The following description outlines one approach to the cloning of human genes.

Our goal is to insert a human DNA sequence into a rapidly reproducing organism so that copies (clones) of the DNA can be made quickly. One system commonly used for this purpose is the **plasmid,** which is a small, circular, self-replicating piece of DNA that resides in many bacteria. Plasmids can be removed from bacteria or inserted into them without seriously disrupting bacterial growth or reproduction.

To insert human DNA into the plasmid, we need a way to cut DNA into pieces so that it can be manipulated. Restriction enzymes, discussed earlier in the text, perform this function efficiently. The DNA sequence recognized by the restriction enzyme *Eco*RI, GAATTC, has the convenient property that its complementary sequence, CTTAAG, is the same sequence, except backward. Such sequences are called **palindromes.** When plasmid or human DNA is cleaved with *Eco*RI, the resulting fragments have sticky ends. If human DNA and plasmid DNA are both cut with this enzyme, both types of DNA fragments contain exposed ends that can undergo complementary base pairing with each other. When the human and plasmid DNA are mixed together they recombine (hence the term **recombinant DNA**). The resulting plasmids contain human DNA **inserts.** The plasmids are inserted back into bacteria, where they reproduce rapidly through natural cell division. The human DNA sequence, which is reproduced along with the other plasmid DNA, is thus cloned (see Fig. 3.21).

The plasmid is referred to as a **vector.** Several other types of vectors may also be used as cloning vehicles, including **bacteriophages** (viruses that infect bacteria), **cosmids** (phage–plasmid hybrids capable of carrying relatively large DNA inserts), **yeast artificial chromosomes** (**YACs;** vectors that are inserted into yeast cells and that behave much like ordinary yeast chromosomes), **bacterial artificial chromosomes (BACs),** and **human artificial chromosomes** (see Chapters 8 and 13). Although plasmids and bacteriophages can accommodate only relatively small inserts (about 10 and 20 kb, respectively), cosmids can carry inserts of approximately 50 kb, and YACs can carry inserts up to 1000 kb in length.

Cloning can be used to create the thousands of copies of human DNA needed for Southern blotting and other experimental applications. In addition, this approach is now used to produce genetically engineered therapeutic products, such as insulin, interferon, human growth hormone, clotting factor VIII (used in the treatment of hemophilia A, a coagulation disorder), and tissue plasminogen activator (a blood clot–dissolving protein that helps to prevent heart attacks and strokes). When these genes are cloned into bacteria or other organisms, the organism produces the human gene product along with its own gene products. In the past, these products were obtained from donor blood or from other animals. The processes of obtaining and purifying them were slow and costly, and the resulting products sometimes contained contaminants. Genetically engineered gene products are rapidly becoming a cheaper, purer, and more efficient alternative.