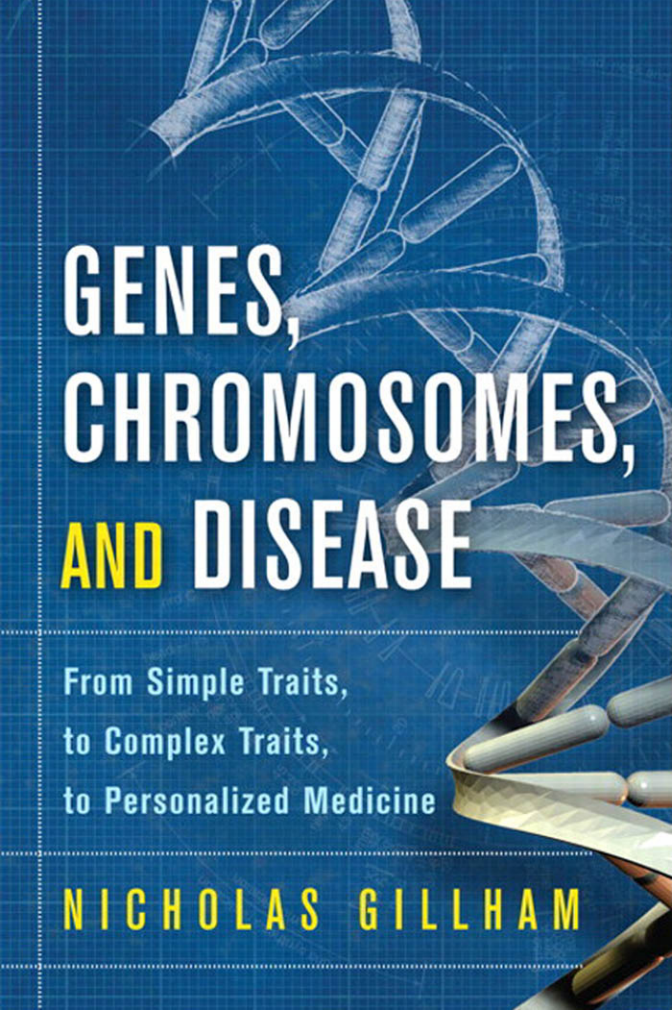
GENES,
CHROMOSOMES,
AND DISEASE
From Simple Traits,
to Complex Traits,
to Personalized Medicine
NICHOLAS GILLHAM

Genes, Chromosomes, and Disease

This page intentionally left blank

Genes, Chromosomes, and Disease

From Simple Traits, to Complex Traits, to Personalized Medicine

Nicholas Wright Gillham

Vice President, Publisher: Tim Moore
Associate Publisher and Director of Marketing: Amy Neidlinger
Acquisitions Editor: Kirk Jensen
Editorial Assistant: Pamela Boland
Operations Manager: Gina Kanouse
Senior Marketing Manager: Julie Phifer
Publicity Manager: Laura Czaja
Assistant Marketing Manager: Megan Colvin
Cover Designer: Chuti Prasertsith
Cover Illustration: Digital Art/Corbis
Managing Editor: Kristy Hart
Project Editor: Betsy Harris
Copy Editor: Karen Annett
Proofreader: Kathy Ruiz
Indexer: Rebecca Salerno
Senior Compositor: Gloria Schurick
Manufacturing Buyer: Dan Uhrig

Publishing as FT Press Science
Upper Saddle River, New Jersey 07458

Printed in the United States of America
First Printing June 2011

ISBN-10: 0-13-707544-8
ISBN-13: 978-0-13-707544-7

Pearson Education LTD.
Pearson Education Australia PTY, Limited.
Pearson Education Singapore, Pte. Ltd.
Pearson Education Asia, Ltd.
Pearson Education Canada, Ltd.
Pearson Educación de Mexico, S.A. de C.V.
Pearson Education—Japan
Pearson Education Malaysia, Pte. Ltd.

Library of Congress Cataloging-in-Publication Data
Gillham, Nicholas W. (Nicholas Wright), 1932- author.
Genes, chromosomes, and disease : from simple traits, to complex traits, to personalized medicine/ Nicholas Gillham.
p. ; cm.
Includes bibliographical references.
ISBN-13: 978-0-13-707544-7 (hardback : alk. paper)
ISBN-10: 0-13-707544-8 (hardback : alk. paper)
1. Genetic disorders—Susceptibility. 2. Human chromosomes. I. Title.
[DNLM: 1. Genetic Disease, Inborn—genetics. 2. Chromosomes, Human. 3. Genetic Disease, Inborn—therapy. 4. Genetic Predisposition to Disease—genetics. 5. Genetic Techniques. QZ 50]
RB155.5.G55 2011
616'.042—dc22

2010051297

In memory of my brother Oliver

This page intentionally left blank

Contents

This page intentionally left blank

Preface

The science of genetics began in 1900 with the independent rediscovery of Mendel's 1866 paper by Carl Correns and Hugo de Vries. Until the middle of the nineteenth century, blending theories of inheritance prevailed, but it became clear to Charles Darwin and his cousin Francis Galton that the hereditary elements must be particulate to provide the kind of variation upon which natural selection could work. Each of them proposed a particulate theory of inheritance, but the particles had to be hypothetical as the architecture of the cell and its different components were only beginning to reveal themselves to the curious eye. By 1900, a great deal was known about cell structure. In particular, chromosomes had been identified and Walther Flemming, a German scientist, had characterized their behavior in cell division (mitosis). Another German scientist, Theodor Boveri, provided evidence that chromosomes of the germ cell lineage provided continuity between generations. And in 1902, an American graduate student, Walter Sutton, connected chromosomes with genes, in a classic paper. Thomas Hunt Morgan and his associates obtained experimental proof of the chromosome theory using the fruit fly *Drosophila* as a model. Working with *Drosophila* in his Fly Room at Columbia University, Morgan and his colleagues would elucidate many of the most important principles of Mendelian genetics.

In England, William Bateson became Mendel's great advocate. One would have thought such advocacy unnecessary except that, just about the time of Mendel's rediscovery, Francis Galton had come up with a model of inheritance, which he called his Ancestral Theory. Particularly in Great Britain, there was much controversy in the first decade of the twentieth century between Galton's supporters and Bateson. The Mendelians finally won out. In the course of these heated exchanges, Bateson became aware of the work of an English doctor, Archibald Garrod. Garrod was studying a disease called alkaptoneuria that caused the urine to blacken. His results suggested to Bateson that a recessive gene mutation might be involved. Bateson entered into a correspondence with Garrod, who in 1902 published a paper titled "The Incidence of Alkaptoneuria: A Study in Chemical Individuality." And with that paper, Garrod made the first connection between a human disease and a gene.

The aim of this book is to provide an overview of the relationship between genes and disease, what can be done about these diseases, and the prospects for the future as we enter the era of personalized medicine. The first three chapters deal with diseases that are simple in the sense that they result because of single gene mutations. Chapter 1, "Hunting for disease genes," considers the pedigree and its use in deciphering human genetic diseases and, at the end, the question of how many genetic diseases there are in the context of the structure of the human genome and the genes it contains. Chapter 2, "How genetic diseases arise," is about how the process of mutation gives rise to genetic defects, but also about how this same process has produced millions of tiny genomic changes called single nucleotide polymorphisms (SNPs). Most SNPs have little or no effect on the individual, but they are of major importance to those who desire to investigate genetic diseases, particularly complex ones. People with and without a genetic disease can be compared to see if any of these SNPs can be associated with specific diseases. The chapter also considers what happens when mistakes occur in partitioning chromosomes properly to sperm and eggs. Chapter 3, "Ethnicity and genetic disease," examines the reasons why some diseases are more prevalent in some races and ethnic groups than others and explains why this has nothing to do with race or ethnicity per se.

The second group of three chapters considers genetically complex diseases. Chapter 4, "Susceptibility genes and risk factors," is about genetic risk factors and diseases like type 2 diabetes, coronary disease, and asthma, where the environment also plays an important role. In each case, there are single gene mutations that can cause the disease. These disease mutations are considered in some detail as they show how certain single gene changes can lead to complex diseases. However, people with these single gene changes only represent a small fraction of those suffering from the disease. In most people who suffer from asthma, have type 2 diabetes, or are susceptible to coronary disease, there is a complex interplay between a variety of genetic risk factors and the environment. Unraveling these interactions is a work in progress.

Chapter 5, "Genes and cancer," discusses cancer, a large collection of different genetic diseases. What they all have in common is the propensity for uncontrolled growth. It has only been possible to work out the many different genetic pathways that lead to cancer because of basic research in cell biology. This has provided the necessary background

information on how the normal pathways themselves are organized. The topic of cancer genetics is so vast that select examples have been chosen to illustrate several different points concerning the disease. For example, cervical cancer shows how viruses sometimes act as causative agents of cancer. The greatly increased frequency of lung cancer in recent years illustrates that decades can elapse between the exposure of a tissue or organ to carcinogens, in this case those present in cigarette smoke, and the appearance of the disease.

Like type 2 diabetes or coronary disease, schizophrenia and bipolar disease are genetically complex, as discussed in Chapter 6, "Genes and behavior." There have been many false alarms in identifying susceptibility genes for these and other behavioral conditions—the gay gene controversy comes to mind. But there have also been some notable successes. The chapter begins by recounting the history of the "warrior gene." This odd gene has been implicated in a wide variety of bad or risk-taking behaviors.

Chapter 7, "Genes and IQ: an unfinished story," deals with a subject whose relevance may not seem apparent initially. The reader may rightly ask what on earth this topic has to do with disease. The answer is that not only do quite a number of genetic diseases affect IQ, but in the first half of the last century, the presumption that "feeblemindedness" was inherited was the basis for involuntary sterilizations, particularly of women, in many states in the United States, Scandinavia, and Nazi Germany. To this day, there are those who argue that IQ differences between races and classes are largely genetic in nature and, therefore, explain certain alleged inferiorities.

For better or worse, it seems likely that IQ and related tests will be used to measure intelligence for a long time because they yield numbers and numbers are easier for most people to deal with than descriptions. Take wine, for instance. All that business about tasting like black cherries with a hint of cinnamon loses out to Robert Parker's numbering system. However, his scale is so compressed, between the high 80s and 100, that a Bordeaux wine that rates 96 can command a far greater price than one that Parker grades as 90. IQ scores, in contrast, are not compressed and follow the pleasing shape of the bell curve. Furthermore, IQ does measure something that relates to what we would call intelligence. Most would agree that the cognitive powers of children with Down syndrome are qualitatively different from those of ordinary children. This differ-

ence is captured in IQ distributions for children suffering from Down syndrome and children without this affliction. In both cases, IQs are normally distributed, but the upper end of the Down distribution overlaps with the lower end of the distribution for children who do not have the disease. However, the data on the heritability of IQ rest on shaky underpinnings. They largely depend on comparing the IQs of less than 200 pairs of identical twins reared apart and the assumption that the environments in which these twins were reared are not correlated.

Having dealt at length with genetic diseases, the next question is what to do about them. Chapter 8, "Preventing genetic disease," discusses prevention as the most desirable outcome, particularly for the most severe genetic diseases, but how do we accomplish this? Suppose a man and woman in their late thirties get married and want to have a child while it is still possible. They have a relatively high risk of giving birth to a child with Down syndrome. What should they do? A good place to begin is to initiate a discussion with a genetic counselor. Should amniocentesis or chorionic villus sampling predict the birth of a Down child, the counselor can be helpful in explaining in a nondirective way the options open to the couple. They themselves will have to decide whether the pregnancy should continue or whether to terminate it. Or suppose another couple knows that they may give birth to a Tay-Sachs child. The couple has the choice of initiating the pregnancy and aborting the fetus if it has Tay-Sachs or planning to have a healthy baby following in vitro fertilization and preimplantation genetic diagnosis. This permits the doctor to implant embryos that will not develop into Tay-Sachs babies although some of them may be carriers of the mutant gene. The procedure is not fail-safe, however, and multiple rounds of in vitro fertilization may be required. Moreover, these procedures are costly and the couple may have ethical or religious reasons for not opting either for abortion or in vitro fertilization.

Specific treatments need to be devised for each genetic ailment and many such diseases are not treatable, as explained in Chapter 9, "Treating genetic disease." The first line of defense for diseases like phenylketoneuria is newborn screening. If left untreated, the disease causes a rapid loss of cognition and a precipitous drop in IQ. Fortunately, if a phenylketoneuric infant is given a special diet shortly after birth, these cognitive declines can be avoided. All the states have mandatory newborn screening for this disease and many others where early intervention

can make all the difference. Treatment of some genetic diseases involves administering an enzyme that is missing because of the genetic defect. This sort of therapy is often very expensive and it must be continued for life.

Then there is gene therapy. After 20 years of trying, it is fair to say that, despite all the hype that accompanied gene therapy, particularly in the beginning, gene therapy has delivered very little except in the case of a couple of diseases where the immune system has been rendered nonfunctional. In these cases, insertion of a copy of the normal gene into certain bone marrow stem cells has proven effective. We hope that this heralds the beginning of a new era for gene therapy, possibly in combination with stem cells, a topic that is hardly discussed in this book. The main reason that this book has practically nothing to say about embryonic or adult stem cells is that, despite very encouraging results with mouse models, we have no idea how this technology is going to play out in humans. In fact, the first approved clinical trial got under way late in 2010. We hope that the disappointments that have plagued gene therapy will not also arise in the case of stem cells, but only time will tell.

Today, drugs are being developed to target specific mutational defects for cystic fibrosis and other genetic diseases, as described in Chapter 10, "The dawn of personalized medicine." It has also become clear that certain drugs are effective with people with one genetic background, but not another. Gene testing companies are measuring genetic risk for complex diseases like type 2 diabetes, and genome sequencing will soon cost around $1,000, making it affordable for a lot of people. With regard to their own genomes, the problem for most people will be an overload of information. What are they to do with it? How are they to weigh it? How much do they really want to know? We have entered the era of personalized medicine, an era in which most of us are going to need some guidance. Before proceeding to discuss the array of topics that are the subject of this book, a word about the diverse ways in which human genetic diseases are named is in order.

Genetic diseases are named in various ways. Most commonly, they bear the names of their discoverers. Down syndrome, for instance, is named for its discoverer, John Langdon Down, a nineteenth-century British physician. Sometimes the name is descriptive—sickle cell anemia comes to mind. The red blood cells of people with this disease do sickle. Sometimes the names are misleading or hard to understand. Why would

anyone name a disease that can cause profuse bleeding hemophilia? Only Count Dracula would appreciate that. Or thalassemia. What's that about? It's a disease like sickle cell disease, but its name refers to the sea in Greek. The reason for this odd name is that this disease was once prevalent around the rim of the Mediterranean. Sometimes diseases are named quite specifically for the function they perturb. G6PD refers to a common alteration that results in a deficiency of the enzyme glucose-6-phosphate dehydrogenase.

1

Hunting for disease genes

Leopold George Duncan Albert, Duke of Albany, eighth child and youngest son of Queen Victoria, was buried on Saturday, April 12, 1884, in the Albert Memorial Chapel, Windsor Castle.[1] He was only 31. Leopold's pregnant wife Princess Helene, the daughter of George Victor, reigning Prince of Waldeck-Pyrmont, arrived by carriage to view her husband's remains and to shed some tears over them. Next, the Seaforth Highlanders, in which Leopold was an honorary colonel, arrived. They were wearing their medals and sidearms. The Coldstream Guards followed the Seaforths led by their band. The servants of the late Prince Albert, the servants of the Queen, and then the gentlemen of Leopold's household followed them. The coffin was borne by eight Seaforth Highlanders and followed by the Prince of Wales in the uniform of a field marshal.

Also marching in the funeral procession was a French general who had accompanied Leopold's remains from Cannes, where he had died. On March 27, Leopold had slipped on a tiled floor in the Yacht Club and injured his knee. Although it has been claimed that Prince Leopold died from the effects of the morphine he had been given to ease the pain on top of the claret he had consumed with his dinner, it seems more likely that he died of a cerebral hemorrhage.[2] Leopold was the first victim of what has been called the "Royal Disease" or hemophilia.

Hemophilia A and B, recessive, sex-linked diseases, are normally expressed only in males because a male has a single X chromosome, whereas a female has two, one usually having the normal gene. That is, women are carriers who do not show any symptoms of hemophilia. Hemophilia spread from the British royal line into the Russian,

Prussian, and Spanish royal lines through intermarriage. Its source was Queen Victoria. She had two daughters who were carriers in addition to Leopold, but her other five offspring did not express or transmit the defective gene to their progeny.

Although it is remotely possible that a hemophilia mutation occurred in Queen Victoria very early in egg formation, it is much more likely that Queen Victoria was a carrier of the hemophilia mutation because three of her children had the hemophilia gene. If the Queen was a carrier, the egg from which she arose would either have had to be fertilized by a mutant sperm from her father Edward, Duke of Kent, or else her father was not the duke. After spending many years in Europe in the company of various mistresses, notably Adelaide Dubus and Julie St. Laurent, Edward married Victoire (or Victoria) of Saxe-Coburg-Saalfeld, the widow of the Prince of Leningen, in 1818. Victoria was born the next year and Edward died in 1820.

Perhaps the sperm that fertilized the egg that produced Queen Victoria possessed the hemophilia mutation. If so, the mutation would have arisen during spermatogenesis in the duke as there is no prior evidence of hemophilia in the royal line. In his book *The Victorians,* A. N. Wilson proposes a different theory. Another man may have fathered Victoria. Wilson supposes that man may have been her mother's secretary Sir John Conroy, a man Queen Victoria detested. Conroy and Victoire were widely suspected of being lovers, but there is no evidence he had hemophilia. Even if Conroy was not Queen Victoria's father, Wilson writes, "it seems overwhelmingly probable that Victoire, uncertain of her husband's potency or fertility, took a lover to determine that the Coburg dynasty would eventually take over the throne of England."[3] If so, the presumptive interloper would have needed to work quickly. After all, Victoria of Saxe-Coburg-Saalfeld and Edward, Duke of Kent, were married on May 29, 1818, and Queen Victoria was born just a year later.

It has long been assumed that hemophilia A rather than hemophilia B was the disease transmitted by Queen Victoria because hemophilia A accounts for 85% of all cases and hemophilia B for about 14% with various other clotting defects accounting for the remaining 1%.[4] However, we now know that Queen Victoria carried a hemophilia B mutation. This finding emerges from some remarkable

detective work involving the remains of the murdered family of Nicholas II, the last Russian czar.

On July 16, 1918, the czar, his family, the royal physician, and three servants were herded into the cellar of Ipatiev House in Yekaterinburg where they were held prisoner and shot by a firing squad.[5] The bodies were to be thrown down a mine shaft, but the truck that carried them began to have engine problems so the murderers dug a shallow pit as a grave, poured sulfuric acid on the bodies to impede their identification, covered the bodies, and drove the truck back and forth over the grave site to flatten it. Half a year later, a Russian investigator, Nicholas Sokolov, retrieved some valuable objects from the likely tomb, but reported no evidence of skeletal remains. He concluded that the bodies had been destroyed, but in April 1989, a filmmaker named Geli Ryabov claimed that the bodies had not been destroyed, but that they were located five miles from the site discovered by Sokolov. Ryabov and a geologist colleague had worked out the actual burial place from photographs and the original report written by the head executioner.

DNA analysis confirmed the presence of the skeletal remains of nine people. They included the czar, the czarina, three of their five children, the royal physician, and three servants. However, two of the children were missing. This was in accord with the executioner's report that he had burned two of the bodies, one of which belonged to the czar's only son Alexei, a hemophiliac. Burned bone fragments from two skeletons were found in 2007 in another grave at the site of a bonfire in the same area.[6] The fragments proved to be what was left of Alexei and his sister Alexandra.

The hemophilia A and B genes are called *F8* and *F9* because they encode clotting factors 8 and 9, respectively. DNA analysis of the F8 and F9 genes recovered from the remains revealed that only the latter gene was altered and that Alexandra was a carrier, whereas Alexei's single X chromosome had, of course, the hemophilia mutation.[7]

The pedigree of the "Royal Disease" illustrates how useful a good lineage is in attributing a specific disease to a defective gene. This chapter considers two different approaches to identifying disease and susceptibility genes. The first is to target a specific gene. The example given here is the discovery of the gene whose alteration results in Huntington's chorea. The pedigree that provided the answer was found on

the shores of Lake Maracaibo in Venezuela. Once an approximate chromosomal location had been established for the gene, the investigators, led by James Gusella at Harvard and Nancy Wexler at Columbia, had to inch along the chromosome to the actual gene using various molecular techniques, a method referred to as "positional cloning" (see Glossary).

The second approach is to search for a variety of deleterious genes in a specific sect or group that exhibits characteristics such as originating from a small founding group, inbreeding, or a high incidence of several different disease genes. The Amish, Ashkenazi Jews, and French Canadians are examples. This is one approach favored by many gene-hunting companies. Once again, pedigree analysis and positional cloning play key roles.

Until the last ten years or so, these were the two major approaches to gene identification, but with the discovery that the human genome is riddled with small genetic differences called single nucleotide polymorphisms or SNPs (see Glossary and Chapter 2, "How genetic diseases arise") coupled with the publication of the human genome sequence, two other approaches became popular that do not require information from pedigrees. In the first, called the candidate gene method, the investigator makes an educated guess at a gene or genes mutation of which might lead to a specific genetic disability. The gene and surrounding DNA are compared between people with and without the condition to see whether there are any alterations specific to people having the disease. The second method is completely unbiased and involves comparing entire genomes between the two groups for differences in SNPs. These genome-wide association studies (GWAS) have the potential for discovering differences related to genes that might not normally have been suspected of causing the disease. These methods, especially the latter, are particularly well adapted to finding genetic factors underlying complex genetic diseases like type 2 diabetes (see Chapter 4, "Susceptibility genes and risk factors," for a fuller discussion). However, as the price of whole genome sequencing continues to drop rapidly, whole genome sequencing comparisons will probably replace the candidate gene and GWAS approaches.

Venezuelan adventures: the isolation of the Huntington's gene

One day in 1858, George Huntington, a boy of eight, was riding with his father George Lee Huntington, a physician. His father was making his medical rounds on a wooded road between the towns of Amagansett and Easthampton on the South Fork of Long Island when "we suddenly came upon a mother and a daughter, both bowing, twisting, grimacing. I stared in wonderment, almost in fear. What could it mean?"[8] Thus was George Huntington introduced to the disease that would later bear his name, Huntington's chorea. Huntington's grandfather, a physician like his father, migrated to the eastern end of Long Island from Connecticut in 1797. Both his grandfather and father had observed the "slow onset and gradual development" of this hereditary disease and how some of its victims "worked on their trades long after the choreic movements had developed, but gradually succumbed to the inevitable, becoming more and more helpless as time advanced, and often mind and body failed at an even pace."

Like his father and grandfather before him, George Huntington became a doctor after obtaining his medical degree at Columbia University in 1871. That same year, he moved to Pomeroy, Ohio, to set up a family practice. On February 15, 1872, he traveled five miles across the icy landscape to Middleport, Ohio, to deliver a paper to the Meigs and Mason Academy of Medicine. The academy's membership was made up of physicians from two sparsely populated counties of the same name. In his report titled "On Chorea," Huntington began with a general review pointing out that "chorea" was a disease of the nervous system whose name derived from "the dancing propensities of those who are affected by it." He noted that chorea was principally a disease of childhood. In contrast, "hereditary chorea" as he called it was confined to the few families he had observed in Easthampton as "an heirloom from generations away back in the dim past" and it did not manifest itself until "*adult* or *middle* life."

Huntington's presentation was well received, so he submitted the manuscript to the editors of the *Medical and Surgical Reporter* of Philadelphia, where it was published on April 13, 1872.[9] Huntington's paper describing what he called "hereditary chorea" was short, clear, and concise and was widely discussed, abstracted for international

yearbooks, and published in its entirety in various texts. In 1915, Charles Benedict Davenport, Director of the Eugenics Records Office at Cold Spring Harbor, New York, and a member of the National Academy of Sciences, published a paper on Huntington's chorea in the first volume of its *Proceedings*.[10] Pedigree data from four families suggested strongly that a dominant gene mutation was responsible for the disease, a hypothesis that has proved to be correct.

The discovery of the defective gene that causes Huntington's chorea really begins with the folk singer and songwriter Woody Guthrie.[11] In 1956, he was arrested in New Jersey for "wandering aimlessly," a charge often brought against the mentally ill or confused. He was committed to the Greystone Park Psychiatric Hospital in Morris Plains, New Jersey, a sprawling complex of 43 buildings that opened in 1876. He remained there until 1961 by which time his condition had worsened and he was transferred to the Creedmore facility on Long Island, where he died in 1967. Following Guthrie's death, his widow formed the Committee to Combat Huntington's Disease. Milton Wexler, a doctor, joined Guthrie in her quest. His wife and three brothers-in-law were suffering from the disease.[12] Wexler's daughter Nancy was in graduate school when her mother was diagnosed with Huntington's disease. She realized she had a 50% chance of having the Huntington's mutation herself. In her PhD dissertation at the University of Michigan in clinical psychology, she explored the cognitive and emotional consequences of being at risk for Huntington's disease. She has never revealed publicly whether she has been tested for the gene. However, it seems unlikely that she will become a disease victim because she is now over 60 and has not expressed its symptoms.

Nancy Wexler was determined to try to identify the Huntington's gene. Her big break came in 1972 when Dr. Americo Negrette, a Venezuelan physician, presented a paper at a conference in the United States.[13] Dr. Negrette had set up his practice in 1952 in a remote community near the great saltwater gulf called Lake Maracaibo. He soon noticed that certain individuals were stumbling, weaving, and falling down, and concluded they were probably drunk. He learned from the residents, however, that they were not drunk, but

suffered from a disease that was locally called El Mal. He soon realized that they were expressing the symptoms of Huntington's disease and published a book on the subject in 1955.

Dr. Negrette had already begun to construct a pedigree for Huntington's disease in the Lake Maracaibo population when Nancy Wexler and her team joined him 1979. Members of the relevant families live in three villages on the shores of the lake. The scientists succeeded in tracing the pedigree back to a woman named Maria Concepcion who lived in the early 1800s and had ten children. She may not have had the disease herself. It is likely that the children of hers who suffered from the disease may have inherited it from their father, possibly a sailor from Europe. By 2004, this pedigree numbered 18,149 of whom 15,409 were still living.[14]

The blood samples from the pedigree were shipped to James Gusella's laboratory at Massachusetts General Hospital. In only four years, by mid-1983, Gusella had located a region near the end of the short arm of chromosome 4 that was close to the gene,[15] but, given the technological limitations at the time, it took another ten years to find and sequence the gene itself.[16] Some idea of the enormous amount of work that went into locating and characterizing the Huntington's gene is apparent from the authorship of the paper, which is given simply as "The Huntington's Disease Collaborative Research Group." It turns out this is the collective title for six groups located at different institutions. There are multiple named authors from each institution.

The nature of the defect in Huntington's chorea was unexpected. In the middle of the gene is the sequence CAG. The four bases in DNA are cytosine (C), guanine (G), adenine (A), and thymine (T). In the genetic code, they are read in groups of three. Each base is attached via a sugar molecule (deoxyribose) to a phosphate group that hooks the whole structure into the DNA backbone (see Figure 1–1). This structure is referred to as a nucleotide with a group of three nucleotides being a trinucleotide. The CAG sequence specifies the amino acid glutamine in the middle of a nerve cell protein that was named huntingtin and was encoded by the Huntington's gene. The CAG sequence in the gene and the corresponding glutamine sequence in the protein are repeated a number of times. In Huntington's disease, there are more CAG repeats than normal. The longer

the CAG stretch, the earlier the onset of the disease. This type of trinucleotide repeat mutation is not unique to Huntington's disease, but is characteristic of certain other genetic diseases as well (see Table 1–1).

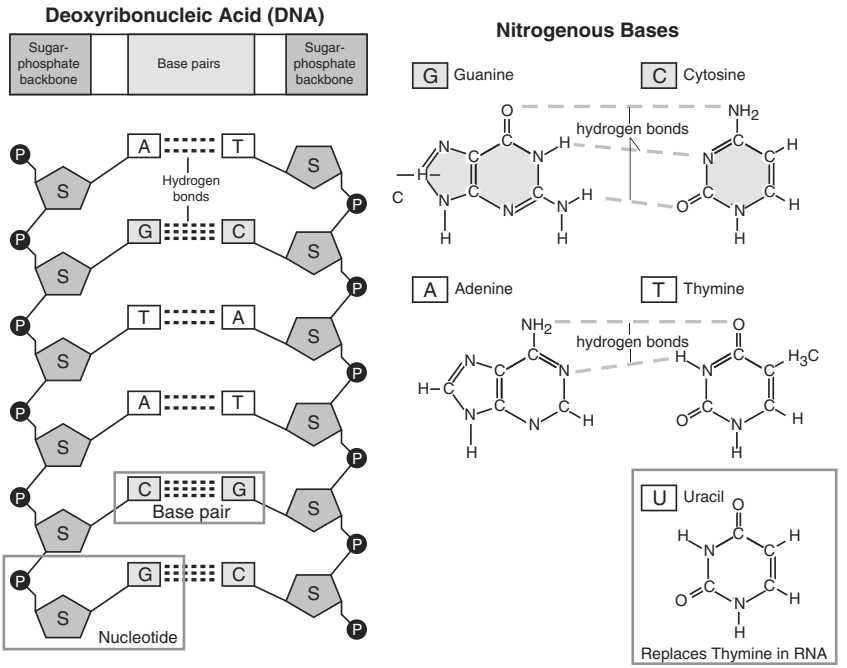

Figure 1–1 Left. A short sequence from the DNA double helix showing the four bases, adenine (A), thymine (T), guanine (G), and cytosine (C). Each base is bonded to a sugar molecule (deoxyribose) which is linked in turn to a phosphate atom to form a nucleotide. The nucleotides are linked to each other via strong, covalent bonds to form the sugar-phosphate backbone of each strand of the helix. The two strands of the helix are held together by hydrogen bonds between the bases with A pairing with T and G with C. Right. G-C and A-T base pairs in more detail. Note that hydrogen bonds are much weaker than covalent bonds and that uracil (U) replaces thymine in RNA.

Courtesy: National Human Genome Research Institute.

Table 1–1 Examples of trinucleotide repeat diseases

		Number of repeats	
Disease	**Trinucleotide**	**Normal**	**Disease**
Huntington's disease	CAG	10–35	40–121
Fragile X syndrome	CGG	5–54	>200–2,000
Friedreich's ataxia	GAA	7–34	200–1,700
Myotonic dystrophy	CTG	5–37	50–11,000

*Huntington's disease is one of eight "polyglutamine diseases" where the sequence CAG is amplified. Each of these sequences occurs in a specific gene encoding a different protein and the disease is neurological in every case. Although the other three diseases shown each involves a specific trinucleotide, these are not within coding regions of the respective genes. Hence, they do not specify different amino acids in the protein products of these genes, although they would do so if they fell within the coding regions. For each disease, there is a numerical gap between the number of repeats found in a normal individual and the number required to cause the disease. This reflects existing uncertainty as to the number of repeats required for the disease to express itself.

Ethnicity, religion, and the gene-hunting companies

As Nancy Wexler's Venezuelan pedigree for Huntington's chorea shows, certain populations are particularly suitable candidates in the search for disease genes. For example, Mormons are a favorable population for the discovery of new disease genes. Mark Skolnick, a University of Utah scientist, realized this many years ago when he became interested in the genetics of breast cancer. In 1991, he was one of the founders of Myriad Genetics of Salt Lake City, Utah, and is currently the chief scientific officer of the company.[17]

Utah's Mormon population was established in the 1840s. Its founders were often polygamous, had large families, and seldom moved. Mormons also marry young so the time span between generations is relatively short. Furthermore, they keep meticulous genealogical records and Myriad Genetics has the rights to these records. As a result, the company has been involved in the identification of the two genes involved in 80% of hereditary breast cancer

cases (*BRCA1*, *BRCA2*) as well as genes important in prostate cancer, colon cancer, melanoma, and also genes disposing individuals to risks other than cancer. The company has developed predictive tests for these genes. Their gene-searching technology is especially useful for such cancer genes because, in addition to having available a computerized genealogy of Mormon pioneers and their descendants, the company can access the Utah Tumor Registry. The registry has required that a record be kept of every cancer occurring in the state since 1973. Using these tools in combination has proved a powerful way of finding tumor genes.

The molecular diagnostic products marketed by Myriad Genetics are "designed to analyze genes and their mutations to assess an individual's risk for developing disease later in life or a patient's likelihood of responding to a particular drug, assess a patient's risk of disease progression and disease recurrence, and measure a patient's exposure to drug therapy to ensure optimal dosing and reduced drug toxicity."[18] Myriad's molecular diagnostic revenues in 2009 were $326.5 million, a 47% increase over the previous year. One reason for this profitability is that Myriad is currently the exclusive provider of tests for the *BRCA1* and *2* genes, which are protected by patents. These patents, which are under legal challenge (Chapter 5, "Genes and cancer"), prevent gene-testing companies like 23andme or deCODE genetics from offering tests for these genes, although they do offer tests for other potential genetic risk factors for breast cancer. This is a potential source of confusion for the uninitiated because these companies assess some, but not all, of the potential breast cancer genetic risk. This is why it is of critical importance for people sending off a saliva or cheek swab sample to a gene-testing company to consult with professionals who can inform them as to exactly what the tests will and will not tell them.

One advantage the Mormon population lacks for gene hunters is that, unlike the Amish, it is not inbred.[19] Although Brigham Young founded Salt Lake City in July 1847 with around 2,000 followers, colonization proceeded rapidly so that by 1890, when most immigration into Utah had ended, there was a total population of 205,889, of whom about 70% were Mormons. They included a great many Mormon converts who frequently came from Great Britain and Scandinavia.

Certain populations benefit the gene hunter by originating from small founding populations. Just by chance, this sometimes means that a deleterious gene may be amplified in the founding population as compared with the population from which it was derived. For example, Tay-Sachs is a common genetic disease among Ashkenazi Jews. Suppose you have a settlement containing 100 couples (excluding children for the purpose of the example) giving a population of 200. Suppose that 10 individuals are carriers of the gene. This yields a carrier frequency of 5%. If 10 couples from the original population decide to form a new settlement and, by chance, they include the 10 carriers, the frequency of carriers rises to 50%. This is the founder effect and accounts for the high frequency of some genetic diseases among certain populations (see Chapter 3, "Ethnicity and genetic disease").

When such a population can be identified and has remained relatively homogenous, it becomes an attractive target for a gene-hunting company. At the Genome 2001 Tri-conference held in San Francisco in March 2001, Phillipe Douville, Vice President and Chief Business Officer of Galileo Genomics Inc. of Montreal, remarked that the "main recognized founder populations in the world are those of Quebec, Finland, Sardinia, Iceland, Costa Rica, the northern Netherlands, Newfoundland, and several discrete ethnic groups including the Ashkenazi Jews."[20]

In fact, it was the population of Quebec that Galileo, now called Genizon BioSciences, planned to focus on. About 15,000 French settlers arrived in eastern Canada in the course of the seventeenth century.[21] Around 2,600 of these hardy souls made their way to Quebec. This population has expanded 800 times over the ten generations since, with intermarriage within the group predominating. Thus, the Quebec founder population is relatively homogenous. Genizon BioSciences claims to have over 47,000 subjects in its biobank, 95% of whom have authorized the company to contact them again.[22] Genizon has research teams investigating eight complex conditions, including Alzheimer's disease, obesity, and schizophrenia by means of genome wide association studies. The company hopes to identify specific patterns of genetic variation that correlate with these conditions and, ultimately, to tie each condition to specific genetic markers (see Chapters 3 and 4 for a fuller discussion).

Gene hunting can be an expensive and unprofitable business. It is usually supported for a while by grants, contracts, venture capital, and deep-pocketed drug companies, but, eventually, it must be profitable. The fate of IDgene Pharmaceuticals Ltd., a Jerusalem genomics start-up founded in 1999, shows what can happen when the path to profitability is not achieved soon enough.

The founder effect, coupled with homogeneity and inbreeding, means that the Ashkenazi Jews are favorable material for the discovery of new disease and susceptibility genes. Hence, IDgene Pharmaceuticals' goal was to search for disease genes among the Ashkenazim.[23] Suitable patients with major chronic ailments that had four Ashkenazi grandparents were asked to donate a single blood sample for genetic testing. Written consent was required and the results were kept anonymous. Israeli Ashkenazi Jews suffering from asthma, type 2 diabetes, schizophrenia, Parkinson's disease, Alzheimer's disease, breast cancer, and colon cancer were studied. Using this method, the company's president, Dr. Ariel Darvasi, and colleagues reported strong genetic evidence supporting the hypothesis that a gene called *COMT* encoding an enzyme involved in the breakdown of certain neurotransmitters is involved in schizophrenia (see Chapter 6, "Genes and behavior").[24] But, subsequently, IDgene failed to raise sufficient capital to continue in operation and closed down in 2004.[25]

The biggest pedigree of all: deCODE genetics and the Icelandic population

A company called deCODE genetics initiated the biggest gene-hunting project of all time. The company proposed to use the entire population of Iceland as a genetic resource because Iceland was founded by a small group of Scandinavian settlers centuries ago. The population is homogenous, and has undergone many population constrictions.

Irish monks, the first inhabitants of Iceland, arrived in the eighth century, but did not become established permanently.[26] A small band of Norsemen who settled Iceland between AD 870 and AD 930 followed them. In the latter year, an annual parliament, the Althing, was established to make laws and solve disputes, making the Althing the oldest parliament in the world. In 1000, Iceland adopted Christianity as its official religion. Iceland's rule over the intervening centuries has

been complex, beginning with its recognition of the King of Norway as its monarch in 1262–1264 and ending with a complete dissolution of Iceland's ties with Denmark in a 1944 referendum.

Viking traders brought the black plague to Iceland. The disease killed as many as 40,000 inhabitants or more than half the population between 1402 and 1404. The plague returned in 1494–95 with a similarly devastating effect. Around 15,000 people, one-third of the population, died during the smallpox epidemic of 1707–09 just as the Icelandic population was recovering from the depredations of the plague and farming was beginning to flourish. In 1783, the Lakigigar eruption resulted in one of the world's worst volcanic disasters. The eruption lasted for eight months. Gases from the eruption reached altitudes of greater than 9,000 feet. The aerosols formed by these gases cooled the Northern Hemisphere by as much as 1 degree centigrade. The haze that formed caused the loss of most of Iceland's livestock from eating fluorine-contaminated grass. Crop failure from acid rain also occurred resulting in the death of 9,000 people, about one-quarter of the population, from the resulting famine.

The small founding population of Iceland coupled with the population bottlenecks just described, plus the relative isolation of the Icelandic population from immigration, rendered it a natural laboratory for human genetic research. In 1996, Kari Stefansson, a native Icelander and Chief of Neuropathology at Boston's Beth Israel Deaconess Hospital, left his comfortable academic perch to found deCODE genetics, a company whose goal was nothing less than to use the enormous human genetic database of Iceland to identify genetic factors involved in common ailments.[27] His certainty that multiple sclerosis involved such factors and his frustration in trying to identify them was one of the underlying reasons for this move.

Genealogy is a passion in Iceland and local newspaper obituaries give detailed family trees that can extend back a hundred years or more. Furthermore, comprehensive clinical records of Iceland's public health service go back as far as 1915. Stefansson recognized that a computerized database of this information for the entire Icelandic population would be an invaluable tool for tracking down genetic diseases. Even more important, Stefansson knew that an exclusive agreement between his company and the government of Iceland

would be an integral part of any business plan. This would give deCODE a major advantage over potential competitors.

In February 1998, deCODE signed an agreement with Hoffman-La Roche stipulating that Hoffman-La Roche would pay deCODE more than $200 million in "benchmark" payments over five years if the company succeeded in identifying genes associated with common debilitating and often lethal syndromes like stroke, heart disease, Alzheimer's disease, and emphysema.[28] However, these "benchmark" payments required that deCODE achieve specific goals within a given amount of time. In an ominous portent of things to come, deCODE failed to achieve the expected goals and received only around $74.3 million of the original total.

The company initially began its work with DNA donated by small groups of Icelanders.[29] This approach was followed up by a publicity campaign designed to attract donors in larger numbers. But the great coup was the Althing's passage of the Health Sector Database Act in December 1998 by a majority of 37 to 20 with 6 abstentions and with the strong support of the Prime Minister David Oddsson.[30] The database act authorized the development of a Health Sector Database for the collection of genetic and medical information already stored in various places around Iceland as part of the country's national health system.

The government had several altruistic reasons for wanting to form the database.[31] First, the act stated that the comprehensive medical records held by the national health system were a national resource that should be kept intact and utilized in the best way possible. Because government funds were used to support construction of the database, the government rejected the notion that any records submitted to the database could be of a proprietary nature. Neither legal entities nor individuals could be granted ownership of specific medical data. Hence, the database would provide the nation with the opportunity to make use of its information to improve medical services for the people of Iceland.

Second, in 1997, the Ministry of Health and Social Security made public a policy statement regarding its plans for utilizing information technology within the national health system. The idea was to create a number of dispersed personal databases that could be linked. This linked database would include medical records and summarize research in fields of possible relevance to Icelandic health, including

epidemics, demographics, and genetic diseases. The cost of constructing such a database was beyond the capacity of the national government, but deCODE's participation would make the effort possible.

Third, the government hoped the database might reverse the Icelandic brain drain by enticing Icelandic scientists interested in human genetics to return to their country. Fourth, the government expected that the database would provide economic benefits to Iceland.

Further actions favorable to deCODE genetics followed.[32] In January 2000, the minister of health granted a 12-year license to the company to operate the database. In 2002, the Althing passed a bill permitting the government to issue state bonds as security for a $200 million loan to deCODE to show its support for the company and to help in financing construction of the database.

Initially, the idea of establishing such a database met with strong support as the results obtained held the potential of bringing to Iceland enormous sums of money from pharmaceutical companies. Several Icelandic politicians expressed the hope that the deCODE database might be as significant for the country as the discovery of North Sea oil was for Norway.[33] Opposition to the project soon emerged, however, as it became evident that Iceland would be the only country in the world to have passed a law authorizing a private company to collect, store, and analyze the genetic heritage of an entire population for commercial purposes.

Some of the concerns were as follows: First, if an individual's personal health information was accessed from the database by an unauthorized person or company, that individual's privacy would be violated or worse.[34] deCODE countered that a person's information would be encrypted. Second, the database act assumed all Icelanders had given their consent to have their personal statistics entered. Although an individual could opt out of the database at any time, data already recorded on that person remained in the database. Furthermore, Icelanders had only six months from the time that the database was constructed to request that their data not be included in the database. This provision was only added to the act because an earlier version had assumed "presumed consent" rather than informed consent. Additionally, data relating to deceased family members would be included automatically without regard to the possible privacy interests of living relatives.

Third, there was danger of genetic stereotyping. One of the diseases studied for which Hoffman-La Roche provided financing was schizophrenia. If a certain fraction of the population proved to have or be susceptible to this disease, then this might suggest to health insurers that anybody of Icelandic heritage any place on earth might be at risk of becoming schizophrenic. Fourth, as the sole licensee, deCODE had monopoly control of the data, although the database itself was the property of the national health system and was managed by the government. Furthermore, deCODE was to be permitted to use the data for commercial purposes for 12 years and access to the data by others was denied if it threatened the financial interest of the company. Fifth, deCODE would make its data available to pharmaceutical and insurance companies for a price. Furthermore, the arrangement with Hoffman-La Roche, according to which deCODE would exclusively investigate 12 different diseases, prevented others from studying these diseases in Iceland.

Pétur Hauksson, a psychiatrist, founded Mannvernd (an Icelandic word meaning human protection), a nonprofit human rights group. Its goal soon became to overturn the Health Sector Database Act. One of Mannvernd's most important complaints was that the act was based on the presumed consent of Icelandic citizens. In addition, citizens who agreed to give blood for one of deCODE's genetic disease investigations had to consent to have the samples used for other genetic studies without knowledge of what they might be. Because of Mannvernd's efforts, Icelanders were now able to refuse to have their information entered in the database by submitting an appropriate form. By June 2001, 20,000 Icelanders, about 7% of the population, had opted out of the Health Sector Database. The Icelandic Medical Association also voiced its opposition to the database act. Many doctors refused to turn over patients' records without their consent. In April 1999, the Icelandic Medical Association brought the Health Sector Database Act before the World Medical Association. The latter body stated full support for the position taken by its Icelandic member in opposition to the database act. Other international criticism was also on the rise. For example, Harvard's Richard Lewontin, a distinguished population geneticist, published an op-ed piece in the *New York Times* on January 23, 1999, titled "People Are Not Commodities," which argued that the database act had

transformed the "entire population of Iceland into a captive biomedical community."[35]

A major concern of the Icelandic Medical Association was the protection of personal data under the database act. Were the encryption technologies sufficient to prevent some unauthorized individual from linking medical data with a specific individual? The association hired Ross Anderson, a Lecturer in the University of Cambridge Computer Laboratory, in fall of 1998 to look into this question. Anderson concluded that deCODE and the Icelandic Data Protection Commission would have to use coded identifiers that would permit linkage of personal data to specific individuals. Because the encryption system would be broken sooner or later, it seemed to Anderson that informed consent standards would have to apply.

Meanwhile, deCODE had begun to achieve scientific success with the more traditional approach by making use of family pedigrees with their informed consent. Hence, the company's obligations under the database act became more of a burden than an opportunity, especially because deCODE was unable to bring the Icelandic Medical Association and the Data Protection Commission on board. The final blow to construction of the database came on November 27, 2003, the day that the Icelandic Supreme Court rendered its verdict in the case of *Gudmundsdóttir v. Iceland*.

The case was prompted by a young woman who wrote to the Icelandic Ministry of Health in February 2000 requesting that any information in her father's medical records and any genealogical or genetic data concerning him not be transferred to the database. The medical director of health denied her request after he had obtained a legal opinion. The Icelandic District Court upheld the director's decision arguing that the medical information available in the database could not be connected to a specific person. But the Supreme Court reversed the lower court decision stating that Gudmundsdóttir had a personal privacy interest in her father's medical data. However, the Court broadened its ruling pointing out that, because by Icelandic law individual medical records were required to contain detailed information on people's health, employment, lifestyles, social circumstances, and so on, a guarantee had to be applied to ensure the individual's freedom from interference with privacy, home, and family life.

Although the database act was dead, deCODE was making good scientific progress in gene discovery. On its Web site, the company claimed to have "discovered risk factors for dozens of common diseases ranging from cardiovascular disease to cancer."[36] deCODE also introduced a new program called deCODEme, which offered customers complete scans that would allow them to discover their "genetic risk for 46 diseases and traits ranging from heart attack and diabetes to alcohol flush reaction and testicular cancer."[37] The company also offered a cardiovascular risk scan, a similar scan for seven common cancers, and a scan of a person's DNA to discover their genetic roots. The problem was that deCODE had never made a profit, was losing money, and was becoming increasingly indebted to its creditors. On November 17, 2009, deCODE filed for bankruptcy under Chapter 11 of the United States Bankruptcy Code.[38] At the same time, it entered into an agreement with Saga Investments LLC to purchase its Iceland-based subsidiary Islensk Erfdagreining and its drug discovery and development programs. Following the sale of these assets, deCODE genetics would be liquidated.

In reporting the bankruptcy of deCODE genetics, the *Times* of London said that it had been assured by Kari Stefansson "that ownership of genetic data remained with the company's customers and that Saga would be bound by a privacy policy that prevents disclosure of data to third parties such as insurers, employers or doctors."[39] But Dan Vorhaus, a lawyer with the American firm of Robinson, Bradshaw, and Hinson, which specializes in genomics, was not convinced. He noted the agreements that deCODE had made with its customers were "often unclear and contradictory."[40]

"The ownership is going to change, and the people making decisions about how to run the company are going to change," Vorhaus said. "This information was held by deCODE, a scientific research organisation. What you have now is Saga, an investment company with a different agenda, very much focused on the bottom line.

Within the range of allowable uses, deCODE's new ownership may choose to use that information in a different way, and possibly to a greater extent, than was previously the case."[41]

So the question of genetic privacy, that became such an issue after the passage of the database act, arises once more with the

bankruptcy of deCODE genetics. It will become an issue again should other gene-hunting companies declare bankruptcy or enter into mergers or takeovers such as the one between deCODE and Saga. In January 2010, deCODE emerged from bankruptcy under the ownership of Saga Investments.[42] Its new CEO was a lawyer named Earl Collier with its founder and former CEO Kari Stefansson now head of research.

How many disease genes are there?

In 1957, Victor McKusick was appointed director of the new Moore Clinic for Chronic Diseases at Johns Hopkins University and head of the newly established Division of Medical Genetics at its medical school.[43] He had come into human genetics via his research on disorders affecting connective tissue, including Marfan's syndrome. Marfan's sufferers typically have long slender limbs and are often taller than normal. The most serious conditions associated with the disease primarily involve the cardiovascular system, as there may be leakage through the mitral or aortic valves that control blood flow through the heart. McKusick noticed that Marfan's syndrome exhibited a familial pattern of occurrence and, indeed, we know today that a dominant genetic mutation is involved. The Marfan's pedigree sparked McKusick's interest and he began to specialize in human clinical genetics.

In 1966, he published his first catalog of all known genes and genetic disorders, *Mendelian Inheritance in Man* (MIM). The 12th edition of his catalog was published in 1998. Meanwhile, a free online version (OMIM) first became available in 1987. It is continuously updated. The database is linked with the National Center for Biotechnology Information and the National Library of Medicine for distribution. In the 1980s, only a few genes were being found each year. By 2000, the number of genes discovered each year was approaching 175. More than 6,000 single gene disorders are currently known,[44] meaning that mutations in somewhere around 24% of the approximately 25,000 human genes found so far can cause genetic disease. Because of the broad interest in disease genes as well as the availability of increasingly sophisticated technical and statistical tools, the rate of disease gene discovery has expanded rapidly. Whether or not it plateaus at some point remains to be seen.

It was originally thought that the human genome might contain as many as 100,000 genes. Once the Human Genome Project was completed in 2003 and a few further revisions were made, this number dropped to around 25,000, roughly the same range as the mouse (see Table 1–2). But the surprising thing is that these protein-encoding genes represent less than 2% of the 3.2 billion base pairs in the human genome.[45] Unlike the even spacing of a string of pearls, our genes often cluster in gene-rich regions separated by gene-poor deserts.

Table 1–2 Genome sizes and gene density in humans as compared with other organisms frequently used in genetic research

Organism	Estimated size (base pairs)	Estimated gene number	Average gene density	Chromosome number
Homo sapiens (human)	3.2 billion	~25,000	1 gene per 100,000 bases	46
Mus musculus (mouse)	2.6 billion	~25,000	1 gene per 100,000 bases	40
Drosophila melanogaster (fruit fly)	137 million	13,000	1 gene per 9,000 bases	8
Arabidopsis thaliana (plant)	100 million	25,000	1 gene per 4,000 bases	10
Caenorhabditis elegans (roundworm)	97 million	19,000	1 gene per 5,000 bases	12
Saccharomyces cerevisiae (yeast)	12.1 million	6,000	1 gene per 2,000 bases	32
Escherichia coli (bacteria)	4.6 million	3,200	1 gene per 1,400 bases	1
Hemophilus influenzae (bacteria)	1.8 million	1,700	1 gene per 1,000 bases	1

From Human Genome Project Information: Functional and Comparative Genomics Fact Sheet. www.ornl.gov/sci/techresources/Human_Genome/faq/compgen.shtml

The human genome is distributed between 23 chromosomes. These are found singly in sperm and eggs (haploid), but in pairs in all of the rest of our cells (diploid). This halving in chromosomes number in eggs and sperm is achieved during the two cell divisions of meiosis. During the first division, homologous paternal and maternal chromosomes pair respectively with paternal and maternal chromosomes assorting independently of each other. During the pairing, chromosome segments are exchanged between homologs, a process called genetic recombination (see Glossary for a brief introduction to Mendelian genetics). Although not generating new genetic alterations, the processes of independent assortment and recombination provide the opportunity to assort existing parental genes in a variety of new combinations. Creation of all of this new genetic variability on which natural selection can act is a major reason why sexual reproduction predominates in animals and plants.

Like the genes of other higher organisms, human genes themselves are not single blocks of DNA that encode specific proteins. Instead, they are broken up into coding sequences (exons) and noncoding sequences (introns). Following the process of transcription, when the information in a gene is copied into a messenger RNA molecule, the intron sequences are spliced out of the message so only the coding sequences in the messenger RNA can be translated into protein sequence.

What is all that other DNA doing that has no obvious genetic function? We know that at least 50% of the genome, perhaps more, is made up of repeated sequences that do not encode human proteins and often no proteins at all. These repeats are of several kinds, but the most abundant are "mobile" genetic elements that make up roughly 43% of the mammalian genome.[46] They either are or at one time were capable of movement from one site in the genome to another.

Transposons are the first group of mobile elements. They comprise around 3% of the genome. The name transposon evokes the word transposition and, indeed, these elements are capable of moving from one to another place in the genome. The easiest way to think about transposition is as a "cut-and-paste" process. One cuts out a word, or a group of words, in a text and then pastes those words into a specific place elsewhere in the text. The important difference

between transposition and cutting and pasting is that, although transposition will take place only into its target DNA sequence, the element can be pasted into that sequence anywhere in the genome. An enzyme called a transposase encoded by the element catalyzes the transposition process. Hence, transposons are sometimes called jumping genes.

The second group includes several sets of elements of which three are the most abundant. The first are endogenous retroviruses. These are viruses whose genetic material is RNA. An enzyme called a reverse transcriptase encoded by the virus catalyzes synthesis of DNA copies of the viral RNA. These DNA copies are then inserted into the genome. The AIDS virus is the best-known retrovirus, but unlike AIDS, the retroviral fragments that inhabit our genomes today are, for the most part, the remains of ancient retroviruses that have lost their ability to become independent of the genome.

LINES (long interspersed nuclear elements) comprise the second group (see Glossary for a more complete discussion of LINES and SINES). They are retrotransposons. One way to think about a retrotransposon is as an odd sort of printing press. An RNA copy is transcribed from the retrotransposon DNA. In the case of LINES, translation of the RNA copy results in the production of two proteins. One of these proteins is essential for the transposition process. The second catalyzes synthesis of a DNA copy of the RNA and then makes a cut in a specific DNA sequence (e.g., TTTTAA/AAAATT for L1) in the genome where the newly made retrotransposon can insert. This method of reproduction has the potential for enormously amplifying the number of retrotransposons in the genome that can then home into their target sequences wherever they are in the genome.

There are several different kinds of LINE elements, but L1, which predominates in the human genome, has evolved along with the mammals over the past 160 million years or more. Expansion in the number of L1s in the genome was rapid, but appears to have slowed down about 25 million years ago. The 500,000 or so copies of L1 present today in the human genome amount to around 18% of its content. The intact L1 element is about 6,000 base pairs in length, but truncated versions are common. L1s are the only active transposons in the human genome today.

SINES (short interspersed nuclear DNA elements) are short DNA sequences of less than 500 base pairs. SINES do not encode any proteins and are not autonomous. They can only transpose with the aid of the two proteins made by active LINE elements. The most important SINES are the Alu elements.[47] More than a million copies of these short DNA sequences are found in the human genome. They represent around 13% of the total DNA. Alu elements originated and coordinated their amplification with the radiation of the primates about 65 million years ago.

Because nobody is exactly sure why human and other animal and plant genomes contain so many repeated elements, they have sometimes been treated as irritants with regard to the real genes, gaining them epithets such as "junk DNA" and "selfish DNA." In a recent review, Goodier and Kazazian point out that a more sophisticated name "dark matter" is coming into vogue for these repeated elements, acknowledging the fact that we don't really understand whether they have an as-yet-to-be-discovered function. Goodier and Kazazian prefer to think of mobile elements as "dark energy."[48] They are "a dynamic force that not only accelerates expansion but also helps set the warp and weft of genomes for better and for worse. Transposable elements arose as intracellular parasites that became domesticated."

Well not entirely. Transposition of these elements can disrupt gene function. In a 1998 paper in *Nature,* Kazazian and his colleagues reported two unrelated cases of hemophilia A for which there was no family history, suggesting that the mutations had arisen de novo.[49] Each of them involved the insertion of L1 sequences into the *F8* gene. So we end this chapter where we began it—with hemophilia. Transposon insertions have also been implicated in a wide spectrum of genetic diseases other than hemophilia.[50]

This page intentionally left blank

Index

Numbers

A

B

C

D

E

F

G

H

I

J

K

L

M

N

O

P

Q–R

S

T

U

V

W

X–Y–Z