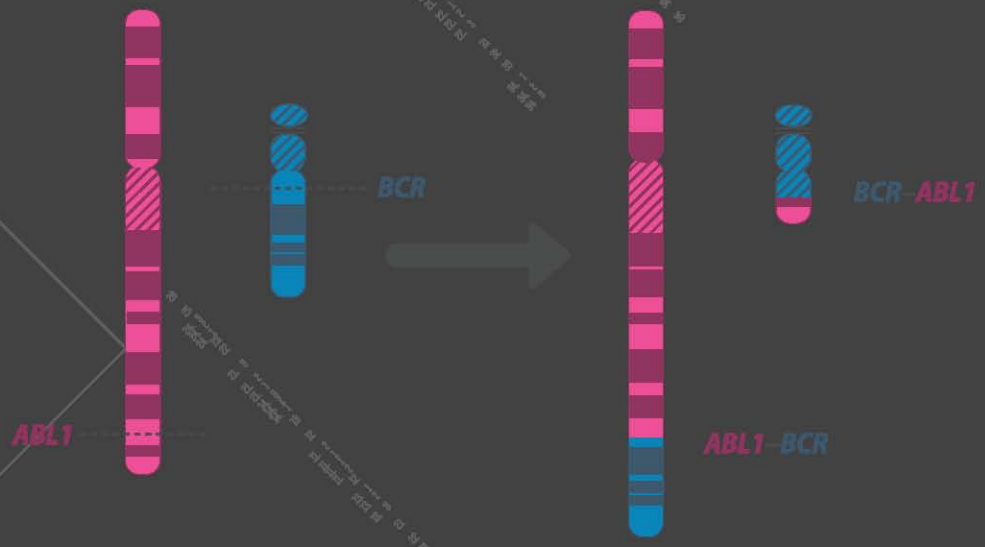
Garland Science

GENETICS OF COMPLEX DISEASE



Peter Donaldson // Ann Daly
Luca Ermini // Debra Bevitt

GENETICS OF COMPLEX DISEASE

GENETICS OF COMPLEX DISEASE

Peter Donaldson // Ann Daly
Luca Ermini // Debra Bevitt

Garland Science
Taylor & Francis Group
LONDON AND NEW YORK

Vice President: Denise Schanck
Senior Editor: Elizabeth Owen
Assitant Editor: David Borrowdale
Production Assistant: Deepa Divakaran
Illustrator: Oxford Designers & Publishers
Layout: Techset Composition Ltd
Cover Designer: Andrew McGee
Copyeditor: Ray Loughlin
Proofreader: Susan Wood

ISBN 9780815344919

Library of Congress Cataloging-in-Publication Data

Donaldson, Peter, 1959-, author.
Genetics of complex disease/Peter Donaldson, Ann Daly, Luca Ermini, Debra Bevitt.
p. ; cm.
Includes bibliographical references.
ISBN 978-0-8153-4491-9 (pbk.)
I. Daly, Ann K., author. II. Ermini, Luca, 1978-, author. III. Bevitt, Debra, 1966-, author. IV. Title.
[DNLM: 1. Disease–genetics. 2. Genetic Diseases, Inborn–genetics. 3. Genetic Predisposition to Disease. QZ 50]
RB155
616'.042–dc23

2015022195

Published by Garland Science, Taylor & Francis Group, LLC, an informa business,
711 Third Avenue, New York, NY, 10017, USA, and 3 Park Square, Milton Park, Abingdon, OX14 4RN, UK.

Printed in the United States of America

15 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Garland Science
Taylor & Francis Group

Visit our web site at http://www.garlandscience.com

Preface

There is a scientific revolution happening in biomedical genetics. The new genetics does not just apply to the well-known and well-described Mendelian diseases with clear patterns of inheritance, nor is it limited to major chromosomal abnormalities. What makes the revolution so exciting is that it includes all human diseases and all aspects of human disease. Diseases that have been largely, but not entirely, ignored in the past are the main focus of this revolution. The potential arising from this work is astounding. It is already having an impact and the impact will only grow over time. There are many books on genetics, but few concentrate on complex diseases—those that do not fit the simple patterns of Mendelian disease and cannot be described as chromosomal abnormalities.

Over the past 15–20 years interest in these genetically complex diseases has taken full flight. Though earlier studies had identified some important genetic links and associations, many of the early studies had failed to be replicated and studies in this area of genetics had developed a poor reputation. There were some good studies and many bad studies. The difference between good and bad studies is quite well known. However, developments in the last 20 years have restored interest and confidence in studies of complex disease.

A number of important developments were the keys to opening up this area for high-quality research. The two most important developments have been the Human Genome Mapping Project and the development of supercomputers along with the necessary systems capable of handling the data that very-large-scale studies produce. These two developments go cap-in-hand, one is not possible without the other. In 2015, we have the human genome sequence, the SNP Map and the HapMap. Of course array platforms for genotyping and application of this knowledge as well as more sophisticated statistical analysis have also filled an essential gap. Indeed, the genetics of today is as much about statistics as it is about biology and there are Professors of Statistical Genetics in our academic institutions who dedicate their research to extracting important facts from the mountains of data that current studies can generate.

This book addresses the subject of genetics of complex disease and is designed in two parts. The first part (Chapters 1–5) provides a basic background to genetically complex diseases, and why and how we study them. The second part (Chapters 6–12) focuses on specific sub-branches of genetics of complex disease and specific examples to highlight the application of genetic data in complex disease and the extent to which this data is fulfilling the promises of the Human Genome Project.

Chapter 1 covers the necessary background to genetic variation in the human population, i.e. our evolutionary past and how genetic variation arises. Chapter 2 goes on to define complex diseases and compare them with Mendelian and chromosomal diseases. Chapter 3 looks at how we investigate complex diseases, including the different plans and strategies available to us. Do we chose a single gene or region to study, or do we throw the net wider and investigate the whole genome in a genome-wide association or linkage study? Chapter 4 considers why we are interested in complex diseases, focusing on the major

promises of the Human Genome Project in relation to complex disease. These suggested that genetic testing will be used in disease diagnosis, patient treatment and management, and in understanding disease pathology. Chapter 5 looks at how data from the studies described below is handled in a range of different statistical tests.

Sufficient information is given in each of Chapters 1–5 to enable students to understand the major points and, where appropriate, examples are used to illustrate the key concepts (e.g. in Chapter 2, where Crohn's disease and Hirschsprung's disease are discussed as two different models of genetically complex diseases). Chapter 4 uses quite a few disease examples to illustrate how the genetic information may be used to meet the promises of the Human Genome Project.

After Chapter 5, the book goes on to look at three specific areas: immunogenetics (Chapter 6), infectious disease (Chapter 7), and pharmacogenetics (Chapter 8).

Chapter 6 on immunogenetics deals with how common variation in genes that regulate the immune response can increase or reduce susceptibility to common diseases. The chapter concentrates on the major histocompatibility complex on chromosome 6p21.3. The chapter includes a considerable number of recently studied examples and discusses the different interpretations that can be applied to the data. In each case, the extent to which these examples do or do not fulfill the promises of the genome project is considered. There are positive examples of how genetics can be used as an aid to diagnosis (e.g. in ankylosing spondylitis), and also how associations and linkage with certain risk alleles may be helping us to understand disease pathogenesis (e.g. in autoimmune liver disease).

Chapter 7 on infectious disease looks at the past and the present considering how genetic variations may influence the likelihood of infection per se and the outcome following exposure to infectious agents. The discussion provides interesting links with mankind's early history. The chapter concentrates on a few selected examples to illustrate the concept and demonstrate how the studies discussed are helping to fulfill the promises of the Human Genome Project. Once again, there are clear examples of genetic investigations impacting on our understanding of disease pathology.

Chapter 8 on pharmacogenetics discusses past and present developments in a fast-expanding field that is at present providing some of the most promising results in complex disease genetics. Studies have shown that responses to commonly used pharmacological agents can be determined by common genetic variation. The impact of this variation ranges from failure to respond to a drug to life-threatening toxic reactions. The potential to use genetics to tailor therapy and also to develop new therapeutic agents is a real possibility in this sub-branch of complex disease genetics, and one that major pharmaceutical companies and academic institutions are aware of.

Chapters 9 and 10 focus on specific disease groups: cancer (Chapter 9) and diabetes (Chapter 10). These two chapters stand alone because one group of diseases (cancer) has a very significant impact in terms of morbidity and mortality in the developed and developing world and the other group (diabetes) is for the most part a perfect example of a complex disease. The potential medical impact of genetic studies in these diseases is vast. More rapid diagnosis, better patient care, personal life planning, and personal treatment planning are all possible. As we gain a greater knowledge of the genetics of these diseases we will start to have a better grasp on the underlying pathology of each disease, which will open up doors for diagnosis, treatment, and management. In some cases, this will mean simple things like changes to a person's diet; in others, selecting the appropriate chemotherapeutic agent to use for a patient. To some extent some of these aims have already been achieved, but as this book indicates, there is still much to be done.

In Chapter 9 (cancer), a selected number of examples are discussed. These include breast cancer, prostate cancer, and lung cancer. The selection is based on the most common cancers, which are also, to some extent, those about which we know the most. Links to useful websites are given for further information and updates. Diabetes (Chapter 10) is discussed in its various forms, especially type 1 and type 2 diabetes and is specifically used to illustrate the difference in the genetic portfolios in type 1 and type 2 diabetes. Here, the question is why are two diseases that have so much in common so different in terms of their genetic profiles?

The last two chapters deal with societal and ethical issues in the new genetic era and the future of genetics in complex disease. This is a fast-moving area of science. The facts being produced today will be marketed as diagnostic or prognostic indicators almost as quickly as they are identified. Genetic testing for risk alleles will soon be normal practice, but this will have ethical and social consequences. The potential for misuse of genetics is discussed in Chapter 11, highlighting the importance of understanding what a genetic test in complex disease is really telling you. You will need to know what a genetic profile is telling you before getting tested. There is considerable commercial interest in genetic profiling and this has ethical and societal impact. Other points discussed include who owns your genome and who can access your genetic data?

Chapter 12 closes the book by looking at the techniques and technologies that have been used and those that will be used in the future. The chapter reminds us that technologies used in the past will also be used in the future, but it also highlights some fascinating new possibilities. Most important will be direct sequencing either at the level of the exome (i.e. protein-coding genes only) or the whole genome.

The structure of the book is designed to provide a basic platform on which students can build their knowledge base. Each of the chapters (including the basic chapters) uses examples of disease to illustrate key specific points and provides a reasonable level of basic current data on each example used. In particular, the book focuses on the promises of the Human Genome Project that suggested genetics will be used to improve disease diagnosis, to develop individual treatment and management plans for patients, and to inform the debate on disease pathogenesis. At each stage and after each example, the text reflects on the extent to which these promises have been or will be met, looking at both the present and, if possible, the future. Links to the web are also provided for access to updates and further information throughout the book. There is an extensive Glossary at the end of the book.

These are very exciting times for genetics, especially in complex disease. They are also fast-moving times. The book is written as a starting point (a first block) and for the most part it is written in an historical style to ensure it remains in date whatever develops in the future.

This book provides a good starting point for anyone studying the genetics of so-called complex diseases. It is written for the undergraduate student and early postgraduate student alike. It is written for the medical and non-medically minded individual. This era is one of the most exciting eras in modern genetics, perhaps as exciting as when the structure of DNA was first revealed to the scientific community.

We would like to thank the staff at Garland Science, Liz Owen, David Borrowdale and Deepa Divakaran, for their support and encouragement in producing this book.

Peter Donaldson, Ann Daly, Luca Ermini, and Debra Bevitt

Acknowledgments

As senior author I would like to give specific thanks to: Robert Taylor (Newcastle University) who provided advice on the mitochondrial genome, John Mansfield (Newcastle University) who provided necessary background on inflammatory bowel disease, Roger Williams (King's College, London) and Oliver James (Newcastle University) both of whom provided a supporting environment within which to learn and develop a background in liver disease and genetics as well as the necessary skills to produce this book, Derek Doherty (Trinity College, Dublin) who worked with me on the molecular genetics of the MHC in liver disease at King's College Hospital, London and the many members of different research teams who have contributed to my research between 1982 and 2015. In addition I would like to give special thanks to the hundreds of students who, through their positive interaction and feedback, have encouraged the writing of this book. Finally, I would like to give very special thanks to Carolyn Donaldson who encouraged and supported production of this book from start to finish, especially during difficult times.

Peter Donaldson

The authors and publisher would like to thank external advisers and reviewers for their suggestions and advice in preparing the text and figures.

Geoffrey Bosson (Newcastle University, UK); Margit Burmeister (University of Michigan, USA); Angela Cox (University of Sheffield, UK); Rachelle Donn (University of Manchester, UK); Yalda Jamshidi (St George's, University of London, UK); Martin Kennedy (University of Otago, New Zealand); Andrew Knight (Newcastle University, UK); Hao Mei (Tulane University, USA); John Pearson (University of Otago, New Zealand); Logan Walker (University of Otago, New Zealand); Kai Wang (University of Iowa, USA); Yun Zhang (Oxford Brookes University, UK).

Contents

CHAPTER 1

Genetic Diversity

Human evolution is driven by a number of different factors, including migration and settlement in different environments, genetic **mutation**, **natural selection**, and **genetic drift**. The product of these different forces is genetic diversity within a population, and understanding this genetic diversity and the reasons for it are essential when considering the genetic basis of common human diseases.

Though the origin of modern humans is relatively recent, humans have managed to colonize almost all possible environments and in doing so have been exposed to considerable **selective pressure**. Consequently, there is extensive variation in the human genome and in the phenotypic traits (e.g. skin color) expressed. In this chapter, we will review the basic background information on mutation, natural selection, and evolution, and the way this helps us to understand the importance of genetic variation in the human genome. We will pinpoint the reasons why genetic variation arises in a population and introduce phenomena such as **epigenetics**. We will also consider the **mitochondrial genome**.

The genetic variation described here creates a basis for genetic risk in the majority of human diseases. Understanding this genetic diversity and how it has arisen is a necessary precursor to understanding the genetics of complex disease. Genetic variations between individuals determine individual susceptibility or protection from a variety of common diseases. This is the basic subject of this book, the idea that common genetic variation gives rise to different levels of susceptibility to common disease. The evolutionary forces that created this genetic variation have enabled populations to thrive, throughout human history, because some population members are likely to be less susceptible to a given illness than others and are thereby more likely to survive even the most catastrophic event.

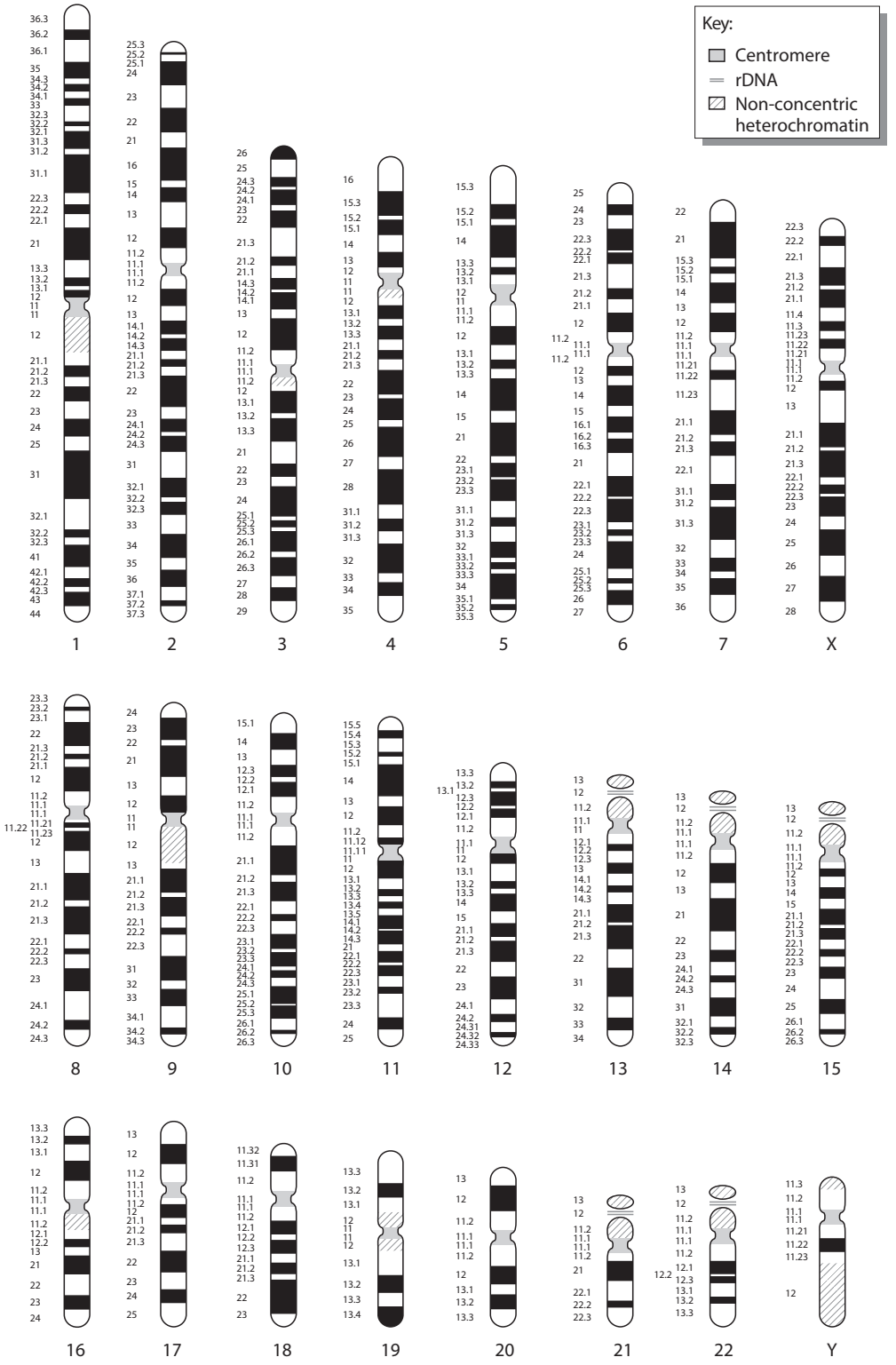
1.1 GENETIC TERMINOLOGY

As with many scientific disciplines, genetics employs a large number of specific terms and this terminology is given in the Glossary at the back of the book. The term **genome** refers to the complete set of genetic information found in a cell and includes 22 pairs of the autosomal chromosomes plus either XX (females) or XY (males) (**Figure 1.1**) and a small amount of DNA found in the **mitochondria (mtDNA)**. Human **chromosomes** are the organized packages of **DNA** found in the nucleus of a cell. DNA is comprised of linear double-stranded molecules that form a helix. The strands of the helix are made up of alternative sugars (deoxyribose) and phosphate groups. Each sugar is attached to one of four bases A, C, G, and T, and the whole molecule is stabilized by cross-linking of the bases A with T and C with G. The DNA structure we are most familiar with looks like a twisted ladder, though when packaged DNA is wound around histone proteins into compressed units. The sequence A–T/C–G provides a code for the production of RNA and RNA production may lead to protein production. The human genome is made up of more than 20,000 **genes**; each gene being a single unit of inheritance that is transmitted from parent to offspring. The location of a gene on a chromosome is referred to as a **locus** (plural: **loci**) and genetic variation at a locus is referred to as **allelic variation**, where the different forms are known as **alleles**. On average, human genes encode approximately 28 kilobases (kb) of DNA with a series of small **exons** (protein-coding sequences) separated by long **introns** (non-coding sequences). Primary transcripts can be differentially spliced into alternative proteins, adding yet another level of complexity to the story of genetics in complex disease. In his book *The Language of Life: DNA and the Revolution in Personalised Medicine* (2010), Francis Collins refers to a single gene in the brain that is capable of making 38,000 different proteins. This is a remarkable and unusual figure. The total impact of intronic genetic variation on common disease is only just beginning to be investigated, but this figure is most likely to be an exception rather than the rule.

The use of the terms genes and alleles varies, though they do have precise definitions

The terms gene and allele are often used as though they are the same, but it is important to note that this is incorrect and that the correct term to use when considering genetic variation is allele. A gene is, as stated above, the basic unit of inheritance. The scientific literature is peppered with examples of incorrect use of this terminology. Writers often refer to the "cystic fibrosis gene" and the "**hemochromatosis** gene" as though only patients with these diseases possess the gene, when actually all members of the population possess these genes. In these two examples, which are both Mendelian autosomal recessive disorders, the difference between affected patients and healthy members of the population is that patients possess two copies of the disease-causing alleles. Unaffected population members may have a single copy of the disease-causing allele or may not carry this allele at all. Instead, they will have one or two copies of the non-disease-causing allele. Thus, it is the possession of the requisite alleles that causes the disease and not the possession of the gene. Finally, the term allele is sometimes used to include any genetic variation within a region,

Figure 1.1: Karyotypes of human chromosomes. The figure illustrates the entire autosome showing banding patterns for each chromosome in size order. Chromosomal banding was (and is) traditionally used to identify chromosomes and chromosomal sites for clinical diagnosis. To obtain these patterns it is necessary to first denature the DNA with enzymes, and then dye the sample to produce light and dark bands. Karyotypes are assigned based on the chromosome length, banding pattern, and position of the centromere. Chromosome 1 is the longest, and chromosome 22 is the shortest among the autosomal chromosomes. (From Strachan T & Read A [2011] Human Molecular Genetics, 4th ed. Garland Science.)


Key:
Centromere
rDNA
Non-concentric heterochromatin
1
2
3
4
5
6
7
X
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Y

whether or not it is part of the exome or intronic sequence. This would not be acceptable to all readers of this book, but those with a focused interest in this area may consider this correct. The use of terminology changes over time.

Most individuals have two copies of a given gene – one inherited from the mother and one from the father. As a result, they are **diploid**. The **genotype** is the set of alleles that an individual possesses. An individual may have two identical alleles, in which case they are **homozygous**, or two different copies, in which case they are **heterozygous** (**Figure 1.2**). When we consider the expression of a genetic variant we use the term **phenotype**. A phenotype can also be referred to as a **trait** or characteristic and may be either physical, physiological, biochemical, or behavioral. Thus, the condition of having blue eyes or dark hair is a phenotype, but so is having sickle cell anemia. Phenotypes are most often referred to as traits or characteristics when they do not relate to an illness or disease.

In 2001, the first draft of the map of the human genome was published. Even though this was not the complete sequence, it marked the beginning of a new era in genetics frequently referred to as the post-genome era. The great advantage of working in the post-genome era is that we have access to the genome map, and the majority of human genetic variation is known and available through websites such Human Genome Resources (http://www.ncbi.nlm.nih.gov/projects/genome/guide/human/index.shtml) and the **SNP Map** (http://www.ncbi.nlm.nih.gov/SNP/).

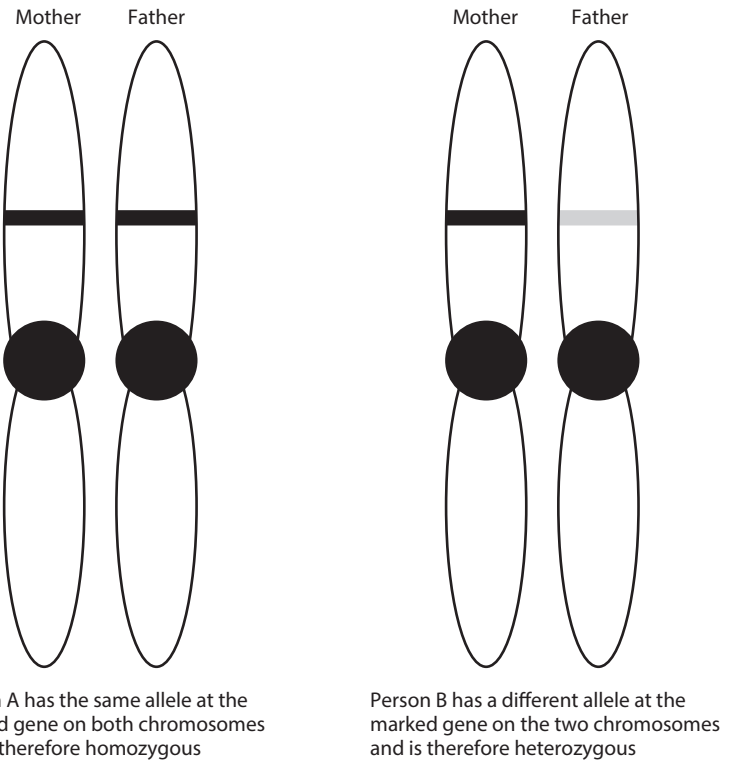


Figure 1.2: Homozygous versus heterozygous. The figure illustrates a single pair of chromosomes in two individuals (A and B). In contrast to the picture in Figure 1.1, the band represents a single gene. Individual A inherits the same allele from both parents (i.e. both black) and is therefore homozygous for this genotype, while individual B inherits different alleles (gray and black) for this gene and is therefore heterozygous.

1.2 GENETIC VARIATION

Genetic variation is by convention discussed in terms of allele frequencies. A frequency is simply a proportion or a percentage usually expressed as a decimal fraction. For example, if 20 out of 100 of the alleles at a particular locus in a population are of the *A* type, we would say that the frequency of the *A* allele in the population is 20% or 0.2.

The term **population** in human genetic studies refers to the group of individuals occupying a defined area such as a country, county, city, or town. Occasionally, a population will be defined by other characteristics, including age, ethnicity, and even in some cases by a particular disease. The complete set of genetic information contained within a population is called the **gene pool**. The gene pool includes all alleles present in the population. Some genes do not encode variation (i.e. they are monomorphic). A monomorphic gene only exists in a single form and therefore has a single allele at a frequency of 100% or 1. However, the majority of our genes are **polymorphic**, existing in two or more (poly) forms in a population. In a population a gene may encode a limited or small number of different alleles or it may encode a much larger number of alleles. One example of the latter is the *HLA-B* gene, which encodes over 3000 **polymorphisms** and mutations. The two terms, mutation and polymorphism, are defined and used differently by different groups. Classical geneticists, particularly those associated with the use of genetics in a clinical setting, who are involved in diagnosis and screening for Mendelian traits, use the term mutation to refer to genetic variations that have a causative effect [i.e. a **disease-causing mutation (DCM)**] and use the term polymorphisms to describe other variations found in the population. Many evolutionary geneticists also prefer to use this definition in this way. However, other geneticists prefer to use the definition provided by Cavalli-Sforza and Bodmer (1971), whereby genetic polymorphism was defined as "the occurrence in the same population of two or more alleles at one locus, each with an **appreciable frequency**." Most authors who apply this definition agree that polymorphic loci are those for which the frequency of the least common allele is greater than 1%. This works well when there are only two alleles. It can become very complicated when considering some of our most complex genes, such as the cystic fibrosis gene ***CFTR*** with over 1910 variations, some of which are common (e.g. Δ508), but the majority of which are rare. In this situation it is difficult to decide which terminology applies; Δ508 is a polymorphism, while most of the other *CFTR* alleles are found at frequencies of less than 1% and are therefore mutations. The dilema is should we call all the *CFTR* variants, including the Δ508 mutations, polymorphisms, or should we apply a mixture of terms as implied in the definition above? There are similar problems with the naming in the **major histocompatibility complex (MHC)** (see Chapter 6).

The problem with the use of this terminology is not simply a matter of choice. Nearly all genetic variation arises through mutation (deletions, **duplication**, insertions, and unrepaired DNA damage). Therefore, most polymorphisms are simply common mutations and, as a consequence, it is not possible to insist on the strict application of this terminology. Though the debate on the correct use of these terms continues, they are used interchangeably in the literature on complex disease. Both terms will be applied throughout this book: polymorphism when describing common variations associated with specific diseases or traits, and mutations when discussing rare variations and evolutionary principals.

A common change in a single base pair (point mutation) is called a **single nucleotide polymorphism (SNP)**. The site at which a SNP is encoded is marked by the "rs" number

(ref-SNP cluster ID number) – a unique ID number based on its position on a chromosome. SNPs are the smallest and most common type of genetic change in humans, and account for an estimated 90% of all variation in the genome. There are currently thought to be more than 38 million SNPs in the genome. Consequently, SNPs are most frequently used as markers to identify genetic variation in human disease. The high frequency of SNPs in the genome enables high-density profiling to be undertaken, which increases the likelihood of accurate identification of disease-promoting alleles. When SNPs were first used for screening for disease alleles on a genome-wide basis, only common SNPs (i.e. those where the least frequent allele was present in 5% or more of the population) were used. The reason for this was that rare SNPs were considered to be less statistically informative. Therefore, it was considered that larger numbers would have to be included in the study to test rare alleles in order to have adequate statistical power. The problem with excluding rare alleles is that potentially important associations with rare SNPs may have been missed. However, as sample collections have become larger the potential to identify statistically significant associations with less common SNPs has grown and the lowest applied limit for SNP frequency has been adjusted downward. For example, instead of a lowest frequency of 5%, a 1% limit can now be applied. The potential for using even lower frequency SNPs will increase as study cohort sizes increase.

Another form of genetic variation that is quite common in the population is **copy number variations (CNVs)**. These occur when there are multiple numbers or copies of a specific gene on a chromosome. CNVs are structural variations that can occur through deletions, duplication, insertions, and translocations. They may represent large or small areas of the chromosome. Good examples can be seen in Chapter 8 on pharmacogenetics.

Genetic variation can be measured by several methods

Though SNPs are the preferred markers for measuring genetic variation, other markers have been used in the past, including **microsatellites**. These are **variable number tandem repeat** (**VNTR**) sequences in the genome. VNTRs can be "short" (involving two to five nucleotide repeats) or "long" (involving more substantial repeat sequences). VNTRs are still used in studies today and are especially useful where the candidate gene is known or a specific region is being scanned. Earlier studies used **restriction enzymes** to identify different VNTRs and SNPs. To determine VNTR genotypes, one or more restriction enzymes that cut the DNA sequence above and below the region encoding the VNTR sequences can be used and DNA fragments of different sizes can be obtained. After digestion with the appropriate enzyme(s), the DNA sample can be run by electrophoresis on either an agarose gel or a polyacrylamide gel to reveal the size(s) of the fragments and thus the number of sequence repeats in each individual sample. Genotypes can be assigned based on the pattern obtained on the gel. This method is known as **restriction fragment length polymorphism (RFLP) analysis**. RFLP analysis was also used to detect SNPs where the differences in the DNA sequence can be detected by use of restriction enzymes that cut the DNA at a particular sequence encoded by one allele, but not the other. Multiple enzymes were often used when genotyping SNPs in order to obtain readable accurate results. Different enzymes are used to detect different polymorphisms. Later studies substituted RFLP genotyping for more reliable **polymerase chain reaction (PCR)** genotyping using primers specific for the gene sequence of interest. This method uses a polymerase enzyme purified from the hot-springs "thermophilic" bacteria *Thermus aquaticus* to amplify multiple copies of the gene sequence. These amplified sequences are then run out on a gel using the same process as that used with RFLP fragments and genotypes can be assigned from the specific banding patterns obtained for each sample (**Figure 1.3**).

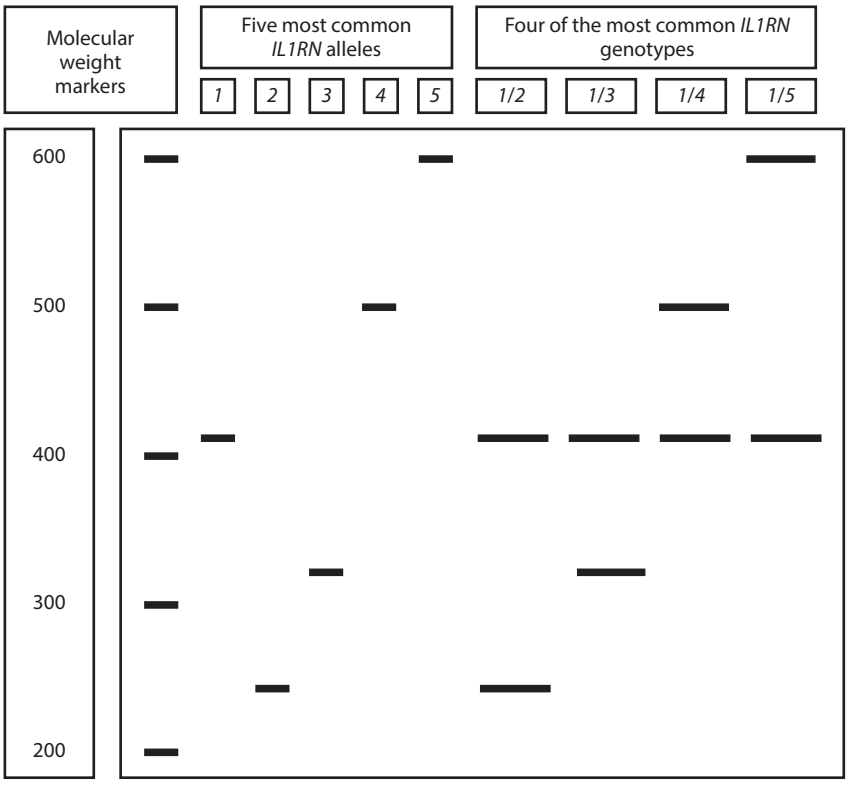


Figure 1.3: VNTRs used to genotype the interleukin (IL)-1 receptor antagonist gene (*IL1RN*). Genotyping the IL-1 receptor antagonist 86 bp VNTR sequence using PCR and agarose electrophoresis. The figure shows the five most common alleles (*1*–*5*) and the four most common genotypes (*1/2*, *1/3*, *1/4*, and *1/5*). The molecular weight markers indicate the approximate band sizes for each allele on the agarose gel as follows: allele *1*, 410 bp; allele *2*, 240 bp; allele *3*, 325 bp; allele *4*, 500 bp; allele *5*, 600 bp. Note the figure does not show the precise position in the gel and the ladder is illustrative only. The genotypes for each sample can be assigned using the band sizes obtained.

Alleles on the same chromosome are physically linked and inherited as haplotypes

As genes are inherited on chromosomes and each chromosome carries a large number of genes, genetic variations on a specific chromosome are inherited en masse as **haplotypes**. Haplotypes do not change from one generation to the next because mutation rates are low, but will change due to recombination during crossover. The potential for change is based on the distance between the genes. One of the very interesting observations to emerge from analysis of haplotypes is that for any small region of a chromosome, most people in a population will carry one of approximately six different haplotypes that can be traced back through history to a shared ancestry in the distant past. However, because **recombination** events exchange pieces of DNA between chromosomes during meiosis, person A may share the same haplotype with person B for a region at one end of a chromosome, yet have a different haplotype compared with person B at a position 1 million base pairs further down the same chromosome. Person B, however, may share the same haplotype in the second region with person C. By studying these haplotypes, it is possible to look back at genetic events that may have happened thousands of years ago.

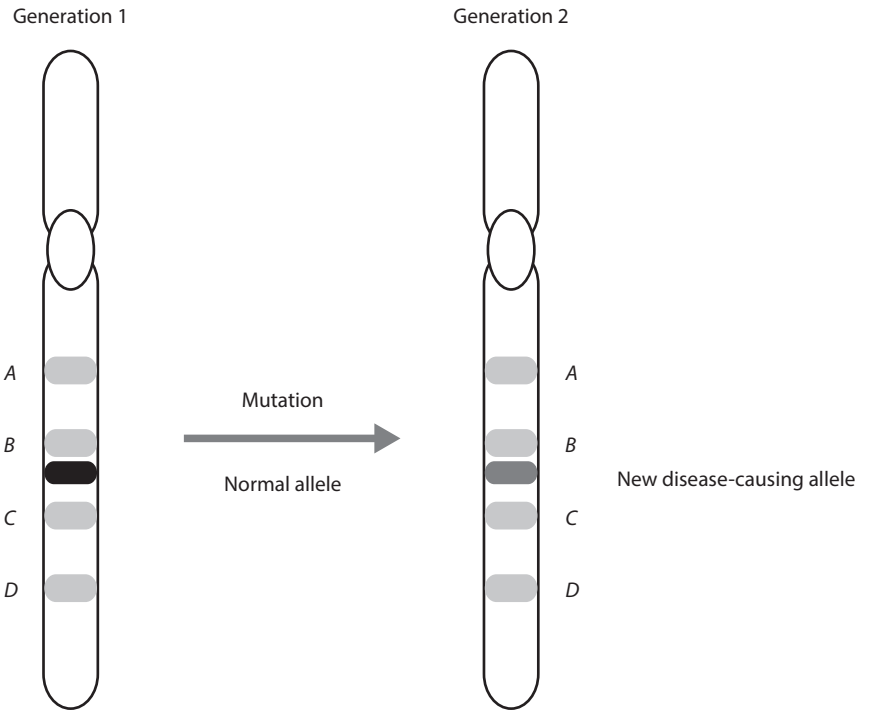


Figure 1.4: The development of a new mutation or allele in the *A–B–C–D* haplotype. The figure illustrates the same chromosome in two individuals in two different generations (generations 1 and 2). On the two chromosomes there are two patterns for the haplotype *A–B–C–D*. One includes a black normal band (representing the normal allele) and one includes a dark gray band (representing a "new" mutated allele). In this illustration the mutation may have arisen through recombination in meiosis. The bands illustrated are not the same as those seen in Figure 1.1, which are the bands based on chromosome staining for karyotyping.

African populations tend to have a greater variety of haplotypes in any given region than other populations. This is expected for a population that is older than all others, and therefore has had more time to diversify and develop more haplotype variations (**Figure 1.4**). In younger populations, such as those in Europe and Asia, fewer haplotypes would be expected because these populations have descended from smaller **founder populations** in which a small subset of the total available haplotypes were present and there has also been less time for new combinations to develop.

Linkage disequilibrium promotes conservation of haplotypes in populations

The term **linkage** refers to the physical association (link) between two alleles that are on the same chromosome. **Linkage disequilibrium** is a population genetics phenomenon whereby alleles on the same chromosome are transmitted together over generations within a population and such pairs or groups of alleles are found together more frequently than expected by chance. In other words, there is **non-random segregation** of the alleles. This is due to the physical proximity of the alleles in question and the low rate of segregation at meiosis. The phenomenon of linkage disequilibrium is common when disease-causing alleles arise in a founder and the alleles are closely linked to other markers along a chromosome. **Crossover**, however, may break up this disequilibrium. When the loci are further apart, linkage disequilibrium breaks down quickly;

Figure 1.5: Extreme linkage disequilibrium. The MHC illustrates extreme linkage disequilibrium whereby alleles at closely linked gene loci are inherited together more often than expected by chance. One example of this is the HLA 8.1 ancestral haplotype (shown above), which is associated with an increased risk of many different autoimmune diseases, but may also convey some survival advantage. The individual alleles are all common in the normal northern European population, but occur together at frequencies far greater than expected by chance. Thus, 60% or more of *HLA-B8*-positives have *HLA-A1*, and 90% or more of *HLA-B8*-positives have *HLA-DRB1*03* and *DQB1*02*. *HLA-B8* is the least common of all of these alleles at around 16%, and if the assortment were random then these pairings would be in equilibrium and the likelihood of finding *HLA-B8* and *HLA-DRB1*03* together would be the sum of their individual frequencies. In this case, the values are approximately 20% for *HLA-DRB1*03* and 16% for *HLA-B8*. This would mean that instead of seeing approximately 14% of the population with the combination *HLA-DRB1*03–HLA-B8*, we would see approximately 3%.

when the loci are close together, crossover is less common and linkage disequilibrium is more likely to persist. Linkage disequilibrium can be used to provide useful information about the distance between genes. Where there is extreme linkage disequilibrium, haplotypes may be conserved and in many cases these conserved haplotypes are common in the population. The human MHC (**Figure 1.5**) illustrates these concepts well (see also Chapter 6). Linkage disequilibrium is a major tool in understanding modern **genome-wide linkage/association studies** (**GWLS/GWAS**).

1.3 GENETICS AND EVOLUTION

Evolution in population genetics refers to changes in the gene pool resulting in the progressive adaptation of populations to their environment. Four main processes account for most of the changes in allele frequency in populations: mutation, migration, natural selection, and random genetic drift. Together these form the basis of cumulative change in the genetic characteristics of populations, leading to the descent with modification that characterizes the process of evolution.

If the population is large enough, allele frequencies remain stable and do not change significantly as a result of random reproduction, and therefore other processes must be responsible for changes in allele frequency. Genetic variation within populations can be increased by migrations and mutations that introduce new alleles into the population. Variation within populations can also be increased by some types of natural selection, such as **over-dominance**, in which both alleles are favored. These evolutionary forces that act to maintain or increase genetic variation are shown in the upper-left quadrant of **Table 1.1**. The lower-left quadrant of Table 1.1 shows evolutionary forces that decrease genetic variation within populations. These forces include genetic drift, which decreases variation through the **fixation** of alleles, and some forms of natural selection, such as **directional selection**, which selectively favors one allele over the other.

Evolutionary forces also affect the **genetic divergence** between populations and are shown on the right quadrants of Table 1.1. Genetic divergence between populations is increased by mutation, genetic drift, and natural selection. Different mutations can arise within each population and therefore mutations almost always increase divergence between populations. **Positive natural selection** can increase or decrease divergence between populations depending on the favored alleles. If different alleles are favored, populations will diverge; however, if natural selection favors the same allele in different populations, genetic divergence between populations will decrease. Migration reduces divergence between populations

Table 1.1 Mutation, migration, genetic drift, and natural selection have different effects on genetic variation within populations and on genetic divergence between populations.

	Within populations	Between populations
Increase genetic variation	Migration Mutation Natural selection	Genetic drift Natural selection Mutation
Decrease genetic variation	Genetic drift Natural (directional) selection	Migration Natural selection

because blending the total gene pool makes populations similar in terms of their genetic composition. Note that migration and genetic drift act in opposite directions: migration increases genetic variation within populations and reduces divergence between populations, whereas genetic drift reduces genetic variation within populations and increases divergence among populations. Mutations mostly increase the genetic variability both within and between populations, though they can occasionally restore the **wild-type**. Natural selection, by contrast, can either increase or reduce genetic variation within a population and increase or reduce genetic divergence between populations.

Finally, before considering each of these processes in turn, it is important to make it clear that populations are simultaneously affected by many evolutionary forces acting at the same time and that evolution results from the complex interplay of these processes.

Mutation is the major cause of genetic variation

Almost all genetic variants arise through some form of mutation. New combinations of these mutations may then arise through recombination in **meiosis**. Meiosis is a process through which cells are able to divide and produce haploid gametes. It is sometimes called a reductive division because there are two stages of cell division, but only one round of DNA replication. Thus, four haploid gametes are created for each diploid spermatocyte (i.e. sperm cell). In oocytes (i.e. egg cells) the situation is different. Division is asymmetric, unlike that for spermatocytes, and the cell division results in a large secondary oocyte and a smaller **polar body** that is discarded. Evolution through natural selection depends on these processes because there has to be genetic variation in the population before evolution can take place. There can be no selection without genetic variation in a population.

A mutation is a heritable change in the DNA sequence. This means that the structure of DNA has been changed permanently and this alteration can be passed from mother to daughter cells during cell division. If a mutation occurs in reproductive cells, it may also be passed from parent to offspring. This kind of mutation is responsible for changing allele frequencies in a population and is an essential process in evolution, as mutations provide the variation that enables humans to change and adapt to their environment when selective pressure is applied. Some mutations may be selectively neutral, which means they do not affect the ability of the organism to survive and reproduce. Only a very few mutations are favorable for the organism and contribute to evolution.

Mutation rates are typically low. The mutation rate (μ) is the frequency of such change and it is expressed as the number of mutations per locus per gamete per generation. Estimating the mutation rate is difficult because mutations are rare. In humans, most information on mutation rates comes from studies of rare **Mendelian autosomal dominant** diseases where

it is much easier to estimate mutation rates than it is for **Mendelian autosomal recessive** diseases or for **non-Mendelian complex diseases**. Estimates of mutation rates for a variety of human genes lie between 10^{-6} and 10^{-5} mutations per locus per gamete per generation. However, the estimated mutation rate is higher for some **Mendelian diseases**. For example, in type 1 neurofibromatosis and Duchenne muscular dystrophy the estimated mutation rate is as high as 10^{-4}. This is 10–100 times greater than the general mutation rates.

The **OMIM (Online Mendelian Inheritance in Man)** database (http://www.ncbi.nlm.nih.gov/omim) lists human genes and it is an excellent source for information on specific genetic diseases. For many diseases, a larger number of genetic mutations have been identified than those listed on OMIM, but this is a good starting point to catalog genetic variations that are linked to or associated with a specific disease and it also has a very good bibliography for each disease.

Introducing the Hardy–Weinberg Principle

The **Hardy–Weinberg Principle (HWP)** or **Hardy–Weinberg Equilibrium (HWE)** test is one of the central pillars of statistical analysis in population genetics (**Table 1.2**). The term equilibrium in population genetics refers to something (an allele or gene) that is in a state of balance. Equilibrium arises when alleles remain unchanged over time. The HWE test assesses how allele frequencies have changed from generation to generation. The HWP states that in a large breeding population, provided none of the evolutionary forces described below are operating, allele frequencies will remain the same from generation to generation. In practice, the HWE test can be used to understand the change in allele frequencies over time and indicate whether evolution has taken place. HWE is also used in studies of complex disease to determine whether there is bias in the study sample and in the qualitative assessment of studies. The HWP is a complex principle and the basic concept and its application are discussed in more detail in Section 1.4.

Genetic variation caused by mutation alters allele frequencies in populations

The rate at which a genetic variation increases or decreases is determined by the mutation rate. Consider the example of a single locus with two alleles *A1* and *A2* with frequencies p and q, respectively, in a population of 10 diploid individuals. In this example, the pool of alleles for this gene within the population will consist of 20 allele copies. If there are 15 copies of *A1* and five copies of *A2* in the population, then the frequency of each allele is $p = 0.75$ and $q = 0.25$. If we suppose that a mutation changes one *A1* allele into an *A2* allele, after one mutation there will be 14 copies of *A1* and six copies of *A2*, and the frequency of *A2* will increase from 0.25 to 0.30; a mutation has therefore changed the allele frequency for the population. If copies of *A1* continue to mutate to *A2*, the frequency of *A2* will continue to increase, while the frequency of *A1* will decrease.

Table 1.2 The HWE ($p^2 + 2pq + q^2 = 1$) dictates that the sum of allele genotypes is always 100% and this formula can be used to determine the expected frequency of the different genotypes in a population.

Maternal gamete	Paternal gamete	
	***A* (*p*)**	***a* (*q*)**
A (*p*)	*AA* (p^2)	*Aa* (*pq*)
a (*q*)	*Aa* (*pq*)	*aa* (q^2)

Thus changes in the frequency of the *A2* allele (Δq) depend on:

- μ: the mutation rate *A1* to *A2*
- p: the frequency of the *A1* allele in the population

When p is large, many copies of *A1* are available to mutate to *A2* and the amount of change will be relatively large. However, as more mutations occur and p decreases, fewer copies of *A1* will be available to mutate to *A2*. The change in *A2* frequency as a result of mutation equals the mutation rate multiplied by the allele frequency:

$$\Delta q = \mu p$$

So far, we have considered only the effects of forward mutations (*A1* → *A2*); however, reverse mutations (*A2* → *A1*) can also occur. Reverse mutations will occur at a rate ν, which will probably be different from the forward mutation rate μ. When a reverse mutation occurs, the frequency of the *A2* allele decreases while the frequency of *A1* increases. The overall change in allele frequency (*A1* and *A2*) is a balance between the two opposing forces of forward and reverse mutations:

$$\Delta q = \mu p - \nu q$$

These allele frequencies are determined only by the forward and reverse mutation rates, and they will increase or decrease until the HWE is established. When the equilibrium is established, the HWP indicates that genotype frequencies will remain the same.

The mutation rates of most human genes are low and changes in allele frequencies due to mutation in one generation are very small. Therefore, it may take a long time to reach the HWE. For example, consider a locus where the forward and reverse mutation rates for alleles are $\mu = 1 \times 10^{-5}$ and $\nu = 0.5 \times 10^{-5}$ per generation, respectively, and the allelic frequencies are $p = 0.85$ and $q = 0.15$. The change in allele frequency per generation due to mutation is:

$$\Delta q = \mu p - \nu q = (1 \times 10^{-5})(0.85) - (0.5 \times 10^{-5})(0.15) = 7.75 \times 10^{-5} = 0.0000775$$

This shows that the change due to mutation in a single generation is extremely small and because the frequency of p drops as a result of each mutation, the frequency of change will become even smaller over time, as shown in **Figure 1.6**.

Migration and dispersal cause gene flow

Another process that introduces change in the allele frequencies is the **gene flow**. Gene flow is the result of migration where many individuals of one population move en masse from one geographic location to another. Though migration is the main cause of gene flow, it can also result from population dispersal, i.e. the spreading of individuals away from others. Migration has a similar impact to mutation as new alleles are introduced into a local gene pool by the migrants. In this case, however, the new alleles are new only to the population into which the migrants move and they are not the result of new mutations.

In the absence of migration, the allele frequencies in each local population can change independently through genetic divergence. As a consequence there will be differing frequencies of common alleles among local populations and some local populations will possess certain rare alleles not found in others. This effect of the accumulation of genetic differences among subpopulations can be reduced if subpopulations undergo migration. Human population migration leads to mixing of the gene pool, preventing populations from becoming too different from one another. A relatively small amount of migration among subpopulations, in the order of just a few migrant individuals in each local

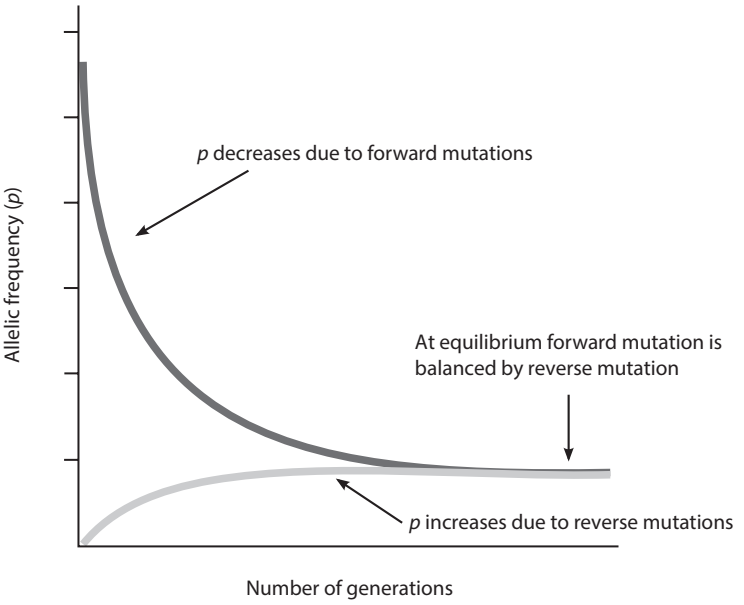


Figure 1.6: Changes due to recurrent mutations slows as the frequency *p* of an allele drops. The figure shows the influence of mutations on the frequency (*p*) of a single allele. The mutation rate in a single generation is exceedingly small and because the frequency of allele *p* drops with each mutation, the rate of change will become even slower over time. Reverse mutations will increase the frequency of allele *p*. Eventually the actions of the opposing forces, i.e. forward and reverse mutations, will establish equilibrium in the frequencies of the alleles *p* and *q*.

population in each generation, can be sufficient to prevent the accumulation of high levels of genetic differentiation between populations. Migration adds genetic variation to populations and increases genetic differences within the recipient population. However, genetic diversification can also occur in spite of migration when other evolutionary forces such as natural selection are sufficiently strong.

Allele frequencies can change randomly via genetic drift

Sewell Wright (1931) introduced the concept of **random genetic drift** into the study of population genetics. Genetic drift refers to changes in allele frequencies in a population due to random fluctuations. These are the frequencies of alleles found in gametes that unite to form **zygotes**. Zygotes are single diploid cells formed by the combination of a single haploid sperm and a single haploid egg, and the alleles found in these gametes vary from generation to generation simply by chance. The zygote referred to is the fertilized egg cell and is the cell from which all other cells in the body are derived. Over time genetic drift usually results in either the loss of an allele or preservation of the allele in the population and fixation at 100%. The rate at which genetic drift occurs depends on the population size and on the initial allele frequencies.

To illustrate the concept of genetic drift we can consider the following hypothetical simulation of changes in allele frequencies for a single gene in five populations of 20 individuals each ($N = 20$) over many generations (**Figure 1.7**). Suppose there are only two alleles, *A* and *B*, and the allele frequencies are identical in all five populations, each with a frequency of 0.5. In the five small populations, the allele frequencies will fluctuate from generation to generation. Eventually, in each of the five populations one of the alleles will be eliminated

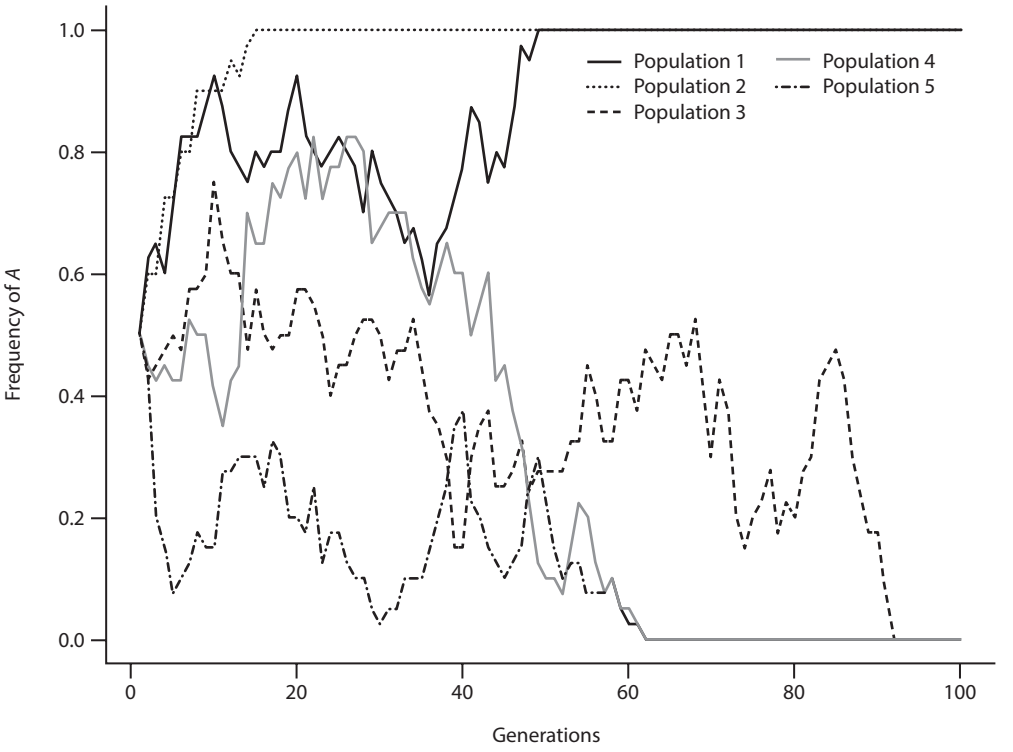


Figure 1.7: Hypothetical model of genetic drift in five different populations. The figure illustrates the potential influence of genetic drift in five populations. The model considers the different outcomes over 100 generations. In all cases the starting allele frequencies are 0.5 for *A* and 0.5 for *a* and each population is assigned 20 individuals ($N = 20$). In all cases the frequency of allele *A* only is considered. The simulations obtained over 100 generations indicate a variety of outcomes with peaks and troughs moving towards frequencies of 1 or 0 for *A* in every case.

and the other will be fixed at 100%. As in this case the gene is now monomorphic (i.e. there is only one allele), all individuals are homozygous for the predominant allele and there can be no further fluctuation in that population. Genetic drift can lead to homozygosity even in large populations, but this will take many more generations to occur.

Figure 1.7 also illustrates another effect of genetic drift. In the example all five populations begin with the same allele frequencies (50% or 0.5 for both alleles), but because genetic drift is random, the frequencies in different populations do not change in the same way and so populations gradually acquire genetic differences. Consequently genetic drift will increase the genetic variation between different populations and there will be genetic divergence over time. In contrast, the opposite effect may also be seen whereby there is reduced genetic variation within populations. Through random change, an allele may eventually reach a frequency of either 100% or 0, at which point all individuals in the population are homozygous for one allele. When an allele has reached a frequency of 1, we say that it has reached fixation. The other allele is lost (reaching a frequency of 0) and can be restored only by migration from another population or by mutation. Fixation leads to a loss of genetic variation within a population. Given enough time, all small populations will become fixed for one allele or the other. Which allele becomes fixed is random in the absence of other forms of selection pressure, though it may be determined by the initial frequency of the allele. If the initial frequency of two alleles is 0.5, both alleles have

an equal probability of fixation; however, if one allele is initially more common, it is more likely to become fixed.

Genetic drift can lead to the fixation of deleterious, neutral, or beneficial alleles, but the effect is greatly influenced by the population size. Allele loss and fixation due to genetic drift occur more rapidly in small populations. Therefore, in nature, both population size and geography can influence genetic drift, and consequently the genetic composition of a population. Some human populations have settled on small islands or in geographically isolated areas and the allele frequencies within these small isolated populations are more susceptible to genetic drift. A population may be reduced in size for a number of generations because of epidemic disease, famine, or other natural or even man-made disasters. As genetic drift is a random process, small isolated populations tend to be more genetically dissimilar to other populations. Geography and population size can influence the effect of genetic drift by creating either a bottleneck or a founder effect.

Bottleneck effect

Changes in population size may influence genetic drift via the **bottleneck effect**. Natural and man-made disasters such as famines or war may reduce the size of the founder population. Depending on the size of the effect and the original population, this can change the degree of genetic variability within the population. Such events may randomly eliminate most of the members of the population with or without regard to the genetic composition or through selection of a group, for example favoring those with specific alleles following infectious epidemics. This can create a bottleneck effect within a population whereby the level of genetic variation is extremely limited (**Figure 1.8**).

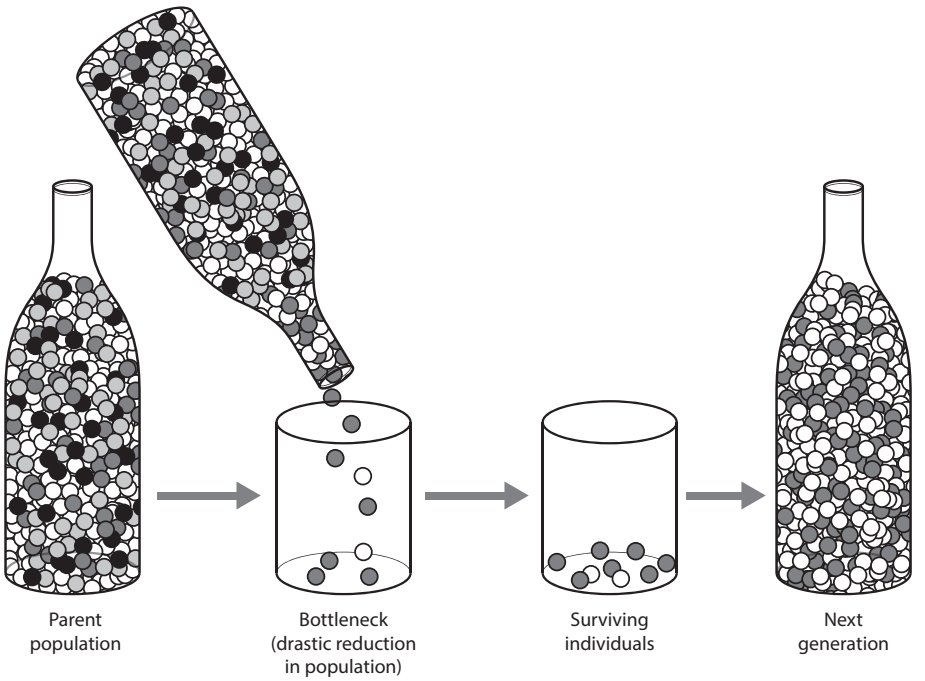


Figure 1.8: The bottleneck effect. The bottleneck effect can occur as a result of major environmental events such as famine or plague whereby the founder or parent population is drastically reduced. This may affect the degree of genetic variability within a population. Natural selection may also operate under these circumstances, favoring those with specific alleles, especially when the disaster involves infectious disease.

The thrifty gene hypothesis

The **thrifty gene hypothesis** was proposed by J. V. Neel in 1962 to explain the growing incidence of diabetes in the Western world. Neel suggested that a thrifty genotype that was more capable of modifying insulin release and glucose storage may have a survival advantage. Though this worked well for our ancestors who had to survive periods of famine, possession of the thrifty genotype in a modern Western society with a plentiful food supply may be a disadvantage as it may cause elevated insulin levels and excessive energy stores. This is seen in clinical cases of type 2 diabetes (T2D).

This hypothesis has been supported by a number of research groups. Work on late-Paleolithic human ancestors indicates alternating periods of abundance and famine and recently two genes or gene sequences have been said to be associated with thrifty characteristics. These are the insulin (*INS*) VNTRs and the apolipoprotein E (*APOE*) gene. More variation has been recorded in the *INS* VNTR genes in African versus non-African populations (27 versus three variants, respectively). *APOE* has been linked with Alzheimer's disease and cardiovascular disease. *APOE2,* which is common in Mediterranean populations, is less common in African and Native American populations. Interestingly, studies have shown that women who possess *APOE4* tend to have more children than those with *APOE2*.

Despite this emerging support for the thrifty gene hypothesis it is still controversial. One reason for this may be that the effects seen are not genetic, but are determined by other factors. Some authors have gone as far as to suggest this may be a thrifty phenotype as opposed to a thrifty genotype effect. In this latter hypothesis the authors suggest that the environment is responsible for the phenotypic variation seen and not the genes, with nutrition in newborn and infant children being particularly important.

Founder effect

Geography and population size may also influence genetic drift via the founder effect. The founder effect involves migration, where a small group of individuals separate from a larger population and establish a colony in a new location. For example, a few individuals may migrate from a large continental population and become the founders of an island population. The founding population is likely to have less genetic variation than the original population from which it was derived and consequently the allele frequencies in the founding population may differ markedly from those of their original population.

Natural selection acting on different levels of fitness affects the gene pool

The final process that brings about changes in allele frequencies is natural selection. Selection is the differential reproduction of genotypes. Selection represents the action of environmental factors on a particular phenotype and genotype through selective pressure. Natural selection is the consequence of differences in the biological fitness of individual phenotypes. Biological fitness is a measure of fertility and reproductive success of a genotype compared with other genotypes in a population. Genotypes with a greater level of biological fitness contribute more to the gene pool of succeeding generations. Differential fitness among genotypes leads to changes in the frequencies of the genotypes over time, which in turn leads to changes in the frequencies of the alleles that make up the gene pool. The effect of natural selection on the gene pool of a population depends on the fitness values of the genotypes in the population. Thus, selection may operate at any time from conception to the end of the reproductive period. There are three major forms of natural selection: purifying or negative selection, positive or adaptive Darwinian selection, and balancing selection (**Figure 1.9**)

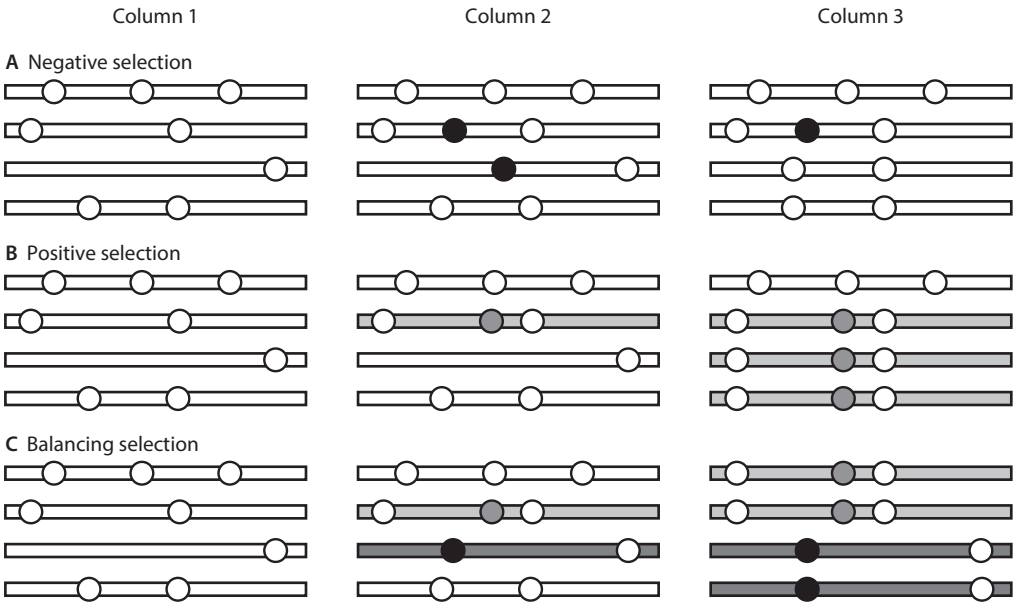


Figure 1.9: Three different models of natural selection. Rows A, B and C illustrate three patterns of natural selection (selection signatures). The columns represent three generations: the first column shows the starting group of four individuals looking at the same chromosome in each, the second column shows the first generation with mutations, and the third column shows the final outcome for the chromosomes in the three different patterns of natural selection. Each circle represents a polymorphism, within a haplotype. White circles represent mutations under neutrality, black circles indicate deleterious mutations, and gray circles indicate advantageous mutations. Pattern A illustrates genetic polymorphisms under negative selection. Deleterious mutations arise (black dot) and they can be removed immediately (if severely deleterious, e.g. line 3 in column 3) or kept at low frequencies (if weakly deleterious, e.g. line 2 in column 3). Linked neutral polymorphism will also disappear (or be kept at low frequencies, e.g. line 3 in column 3). Pattern B illustrates genetic polymorphism under positive selection. When a new advantageous mutation arises (shaded circle in line 2, column 2), the allele increases in frequency (in the population) along with linked neutral polymorphisms (lines 3 and 4 in column 3, which now resemble line 2, column 2). Pattern C illustrates balancing selection. Two new alleles are shown (shaded and black circles) and, if they confer advantage in the heterozygous state, they will increase to intermediate frequencies. Linked neutral polymorphisms will also increase to intermediate frequencies. (From Ermini L, Wilson IJ, Goodship TH & Sheerin NS [2012] *Immunobiology* 217:265–271. With permission from Elsevier.)

Purifying selection

Purifying natural selection (also called negative selection) reduces the frequency of detrimental alleles in a population. New mutants often have detrimental effects on biological fitness and purifying selection reduces the number of new mutations in the gene pool. In humans, 38–75% of all new non-synonymous mutations are thought to be affected by moderate to strong negative selection. Deleterious mutations are generally found at low frequencies because of the adverse effect they may have on biological fitness. Negative selection is responsible for the removal (or maintenance at low frequencies) of mutations associated with severe Mendelian disorders. Mendelian disease genes come under widespread purifying selection, especially when the disease mutations are dominant.

Positive Darwinian selection

Some mutant alleles introduced to a population by gene flow may be advantageous. In this case a directional genetic change may allow a population to adapt to its environment and new, better adapted alleles may replace old, less well adapted alleles. Such selection of

alleles that are advantageous is called adaptive Darwinian selection or **positive Darwinian selection**. Under the action of positive selection advantageous alleles rapidly achieve high frequencies within the population. This occurs at a rate much faster than that of a neutrally selected allele. As a consequence of this rapid increase few recombination events will take place and any neutral variation linked to selected variants will also increase in frequency within the population. This process often results in a transitory increase in the strength of linkage disequilibrium between alleles on the same haplotype.

Balancing selection

A third form of natural selection is **balancing selection**, whereby **heterozygotes** show a higher level of biological fitness than **homozygotes**. This leads to the maintenance of two or multiple alleles in a population at a given locus. Polymorphisms are maintained in the population for a longer period of time than expected. Balancing selection is often referred to as **heterozygote advantage**, especially in cases where a mutant allele known to cause a disease in homozygotes is found at a high frequency in heterozygous healthy members of the population. Genome scans suggest that balancing selection is less extensive than positive selection. However, balancing selection does occur. The two examples below illustrate heterozygous advantage in two **autosomal recessive Mendelian** diseases.

Cystic fibrosis is one of the most common autosomal recessive diseases in Northern European populations, affecting approximately 1:2500 new born children. The causative gene in cystic fibrosis is the cystic fibrosis transmembrane conductive regulator gene (*CFTR*) and there are currently 1910 mutations on the *CFTR* mutation database (http://www.genet.sickkids.on.ca/StatisticsPage.html) (**Figure 1.10**). Carriers of the *CFTR* mutations (heterozygotes) appear to have, or have had in the past, some reproductive advantage over wild-type normal homozygotes. There has been debate over what this advantage might be. The *CFTR* gene encodes a membrane chloride channel protein that is required by some bacteria such as those belonging to the genus *Salmonella*

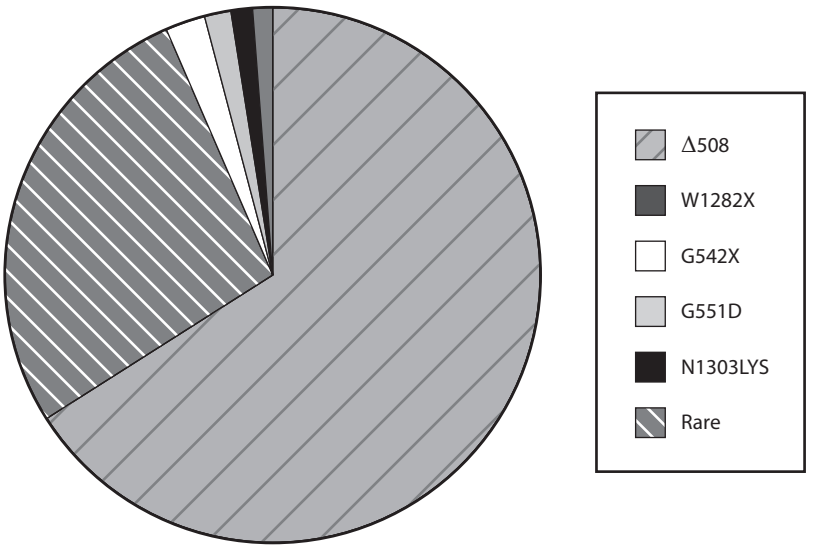


Figure 1.10: The five most common *CFTR* gene mutations. The five mutations listed account for over 70% of the overall mutations and Δ508 is the most common of all, accounting for approximately two-thirds of all cases. All of the other mutations, of which there are at least 1905, are found at frequencies of less than 1% and together these account for approximately 30% of all *CFTR* mutations.

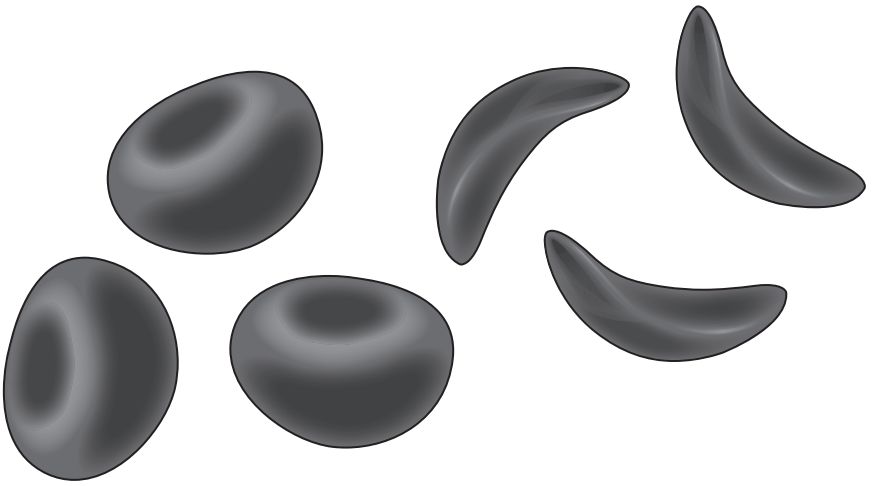


Figure 1.11: Red blood cells in sickle cell disease. Sickle cells are shaped like a harvesting sickle and, unlike the normal doughnut-shaped red blood cells, these cells can be hard with sharp edges that can damage the wall of small blood vessels as they passage through the body. They will often clog the flow of blood and break up as they pass through the small blood vessels.

(e.g. *Salmonella typhi*) to enter into epithelial cells. One explanation is that carriers of a mutant *CFTR* allele may be more resistant to infection by such bacteria than those with two copies of the wild-type gene.

Sickle cell anemia is another example. This is a genetic autosomal recessive blood disorder that is characterized by red blood cells that occasionally assume an abnormal, rigid, sickle shape (**Figure 1.11**). The β-globin allele variant, called *HbS*, is responsible for the sickling of red blood cells seen in the disease. Despite the high mortality associated with homozygosity the sickling allele *HbS* is found at high frequencies in Africa (up to 30%). One possible explanation for the abundance of the *HbS* allele in Africa is that heterozygosity confers some resistance to malaria.

1.4 CALCULATING GENETIC DIVERSITY: DETERMINING POPULATION VARIABILITY

Genotype and allele frequencies illustrate genetic diversity

The genetic diversity of a population can be described using genotype or allele frequencies. A large number of samples from a population are usually collected, and the genotype and allele frequencies are calculated. The genotype and allele frequencies of the sample population are then used to estimate the diversity of the population. To calculate a genotype frequency, the number of individuals having the same genotype is divided by the total number of individuals in the sample (N). For a locus with three genotypes, *AA*, *Aa*, and *aa*, the frequency (f) of each genotype is:

$$f(AA) = \frac{\text{number of } AA \text{ individuals}}{N}$$

$$f(Aa) = \frac{\text{number of } Aa \text{ individuals}}{N}$$

$$f(aa) = \frac{\text{number of } aa \text{ individuals}}{N}$$

The sum of all the genotype frequencies always equals 1 (or 100%).

Genotypes are not permanent. They are disrupted in the processes of segregation and recombination that take place when individual alleles are passed to the next generation through the gametes. Alleles, in contrast, are not broken down and the same allele may be passed from one generation to the next. For this reason the calculation of allele frequencies is often the preferred choice when determining the genetic variability of a population. In addition, there are always fewer alleles than genotypes, e.g. for the gene with two alleles *A* and *a* above, there are two alleles, but there are three genotypes. By using alleles, population diversity can be described in fewer terms than by using genotypes. Finally, by using allele frequencies in case control population studies rather than genotype frequencies, no assumptions about the impact of homozygosity or of heterozygote advantage are being made. This is especially important in the context of complex disease where in the absence of a clear pattern of inheritance it would not be appropriate to make any such assumption, at least in the initial stages of analysis.

Allele frequency refers to the numbers of alleles present in a population

The number of copies of an allele at a locus is divided by the total number of all alleles in the sample:

$$\text{Frequency of an allele} = \frac{\text{number of copies of the allele}}{\text{number of copies of all alleles at the locus}}$$

If we consider a gene with only two alleles *A* and *a* and we suppose the frequencies are *p* for allele *A* and *q* for allele *a*; then *p* and *q* can be calculated as:

$$p = f(A) = \frac{2n_{AA} + n_{Aa}}{2N}$$

$$q = f(a) = \frac{2n_{aa} + n_{Aa}}{2N}$$

In this equation n_{AA}, n_{Aa}, and n_{aa} represent the numbers of *AA*, *Aa*, and *aa* individuals, and *N* represents the total number of individuals in the sample it is necessary to divide by *2N* because being diploid means each individual has two alleles for each gene (one from the maternal locus and one from the paternal locus).

The sum of the allele frequencies is always 1 (100%) ($p + q = 1$); therefore where there are only two alleles, *q* can be determined by simple subtraction after *p* has been calculated:

$$q = 1 - p$$

These calculations apply only where there are two alleles. In cases where there are several different alleles at a locus the calculation used is based on the same principle, but is more complicated. Statistical software will usually be used to perform complex calculations, but it is important to understand the underlying principles in any analysis.

Heterozygosity provides a quantitative estimation of genetic variation

Knowing the frequency of heterozygotes (i.e. those carrying both wild-type and mutant alleles for the same gene) can be a very useful tool for the quantitative estimation of genetic variation in a population. Where mutations are common, heterozygotes are common and homozygotes can be quite rare. Heterozygosity can provide information on the structure and even the history of a population. High levels of heterozygosity reflect high levels of genetic variability, while low levels of heterozygosity indicate low levels of genetic variability. Very low levels of heterozygosity can indicate the effects of small population sizes created by population bottlenecks. Often the observed levels of heterozygosity are compared with what is expected under HWE (see below). If the observed heterozygosity deviates from HWE or is lower than expected, this discrepancy may be attributed to non-random mating. This can occur in small isolated populations when individuals select a closely related mate more often than would be expected by chance in a larger population.

Non-random mating does not change the allele frequencies, but leads to an increase in homozygous offspring over time because the parents are more likely to be genetically similar. Consequently, there will be a decrease in heterozygosity in such populations. This places individuals and the population at a greater risk from Mendelian recessive diseases. The impact of accumulating deleterious homozygous traits is called **inbreeding depression**. This term refers to the loss of **population vigor** due to reduced genetic variability or reduced biological fitness in a given population as a result of the breeding between related individuals. This phenomenon is often the result of a population bottleneck. If heterozygosity is higher than expected, an isolated breakout may have taken place through contact with individuals from another population, which can introduce a temporary excess of heterozygotes.

Expected heterozygosity can be measured using the simple formula:

$$H_{\mathrm{E}} = 1 - \sum_{i=1}^{n} p_i^2$$

In this equation n is the number of alleles and p_i is the frequency of the ith allele at a locus.

The value of this measure ranges from 0 for no heterozygosity to nearly 1 (i.e. 100%) for a system with a large number of equally frequent alleles.

The HWP is a complex but essential concept in population genetics

The HWP, which was introduced earlier in this chapter, is one of the most important statistical principles in population genetics, and because it is an abstract and quantitative principle it is one of the hardest concepts to understand. Therefore, it needs to be discussed in some detail. We may wonder why a recessive trait is not gradually eliminated over the course of time or how the O blood type can be the most common blood type if it is a recessive trait. These questions reflect the assumption that the dominant allele in a population will always be found at the highest frequency and the recessive allele will always be less common. The HWP addresses these questions and enables us to consider the frequency of alleles over the time.

The HWP depends on certain assumptions, of which the most important are:

- Mating in a population is random – there are no subpopulations that differ in allele frequency.
- Allele frequencies are the same in males and females.
- All the genotypes are equal in viability and fertility – selection does not operate.
- Mutation does not occur.
- Migration into the population is absent – gene flow does not occur.
- Genetic drift does not occur.
- The population is sufficiently large that the frequencies of alleles do not change from generation to generation.

The HWP states that after one generation of random mating, in a large breeding population, where the restrictions listed above all apply, single-locus genotype frequencies can be presented as a binomial function (where there are only two alleles) or multinomial function (where there are multiple alleles). Under the above conditions and over time, allele frequencies will reach equilibrium and remain constant from generation to generation.

Calculating expected genotype frequencies using the HWP

The HWE can be calculated using the simple mathematical formula:

$$p^2 + 2pq + q^2 = 1$$

In this equation p and q represent the frequencies of alleles (**Figure 1.12**). It is important to note that the sum of the allele frequencies ($p + q$) *is* always equal to 1. To illustrate HWE, we can consider a single gene with two alleles A and a in a large population with frequencies p and q, respectively (**Box 1.1**). First, we must assume that male and female gametes interact randomly, and all of the major assumptions above remain true (i.e. there is no mutation, selection, random genetic drift, or gene flow). We can then use a simple calculation based on the **Punnett Square** illustrated in Table 1.2. The Punnett square is perhaps the most common of all mathematical representations used in the study of the genetics of complex disease. The data entered in the table can be used to generate odds ratios (ORs) and significance values via the χ^2 test. It is important to note that the exact numbers and not the percentages must be used in the calculations to generate accurate outcomes.

Different populations may have different allele frequencies

Note that heterozygotes are more common when allele frequencies are intermediate; however, when one allele is more common than the other, homozygosity for that allele is increased and heterozygosity reduced. In this illustration heterozygotes have a maximum frequency of 50%, which is achieved when $p = q = 0.5$. When either locus is monomorphic ($p = 1$ or $q = 1$), there are no heterozygotes.

HWE allows us to describe a population only considering the frequencies of n alleles at a particular locus. For the less statistically inclined geneticists, the HWE principle is mostly applied to ensure validation of data from population studies. Departure from the expected distribution of genotypes generally indicates problems in sample recruitment or some other form of population bias. Conversely, there is more confidence in the results when there is equilibrium.

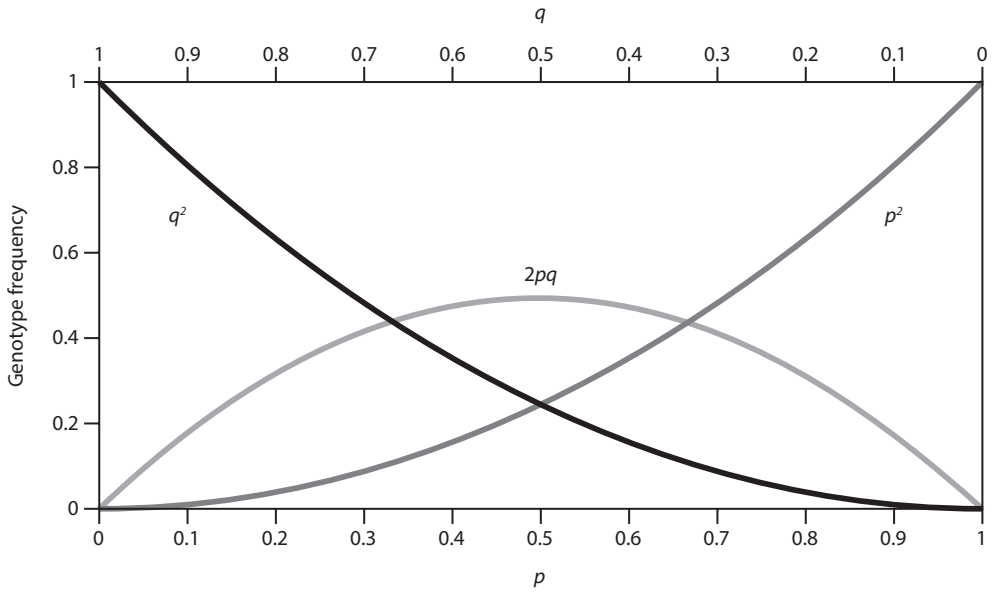


Figure 1.12: A plot of the HWE-based genotype frequencies (p^2, $2pq$, and q^2) as a mathematical function of allele frequencies (p and q). The plot illustrates the influence of allele frequencies on genotype frequencies and shows what we can expect from the HWE test. The plot shows how the two alleles p and q determine genotype frequencies and these change as the allele frequencies change. For example the closer q is to 1, the lower the value of p and the higher the value for the q^2 genotype (homozygous q). When p and q are both set at 0.5 (50%), then the frequency of pq heterozygotes is high.

BOX 1.1 CALCULATING EXPECTED GENOTYPE FREQUENCIES USING THE HWP

The HWE can be calculated using the simple mathematical formula:

$$p^2 + 2pq + q^2 = 1$$

In this equation p and q represent the frequencies of alleles. It is important to note that the sum of the allele frequencies ($p + q$) is always equal to 1.

To illustrate HWE, we can consider a single gene with two alleles A and a in a large population with frequencies p and q, respectively. First, we must assume that male and female gametes interact randomly and all of the following assumptions are true:

- Mating in a population is random – there are no subpopulations that differ in allele frequency.
- Allele frequencies are the same in males and females.
- All the genotypes are equal in viability and fertility – selection does not operate.
- Mutation does not occur.
- Migration into the population is absent – gene flow does not occur.
- Genetic drift does not occur.
- The population is sufficiently large that the frequencies of alleles do not change from generation to generation.

BOX 1.1 CALCULATING EXPECTED GENOTYPE FREQUENCIES USING THE HWP (*Continued*)

Then we can use a simple calculation based on the Punnett Square illustrated in Table 1.2. The top side of the square is divided into proportions p and q representing the frequencies of the male alleles A and a, respectively. The left side represents the same proportions, but for the female alleles. If we assume there is random union of gametes we can apply the product rule of probabilities. Imagine a pool with male and female gametes, p with A alleles and q with a alleles, and where zygote formation occurs by random union. The upper-left square represents the frequency of the homozygous genotype AA. The expected frequency is simply the product of the separate allele frequencies.

$$\text{Frequency of } AA = p \times p = p^2 \text{ (Homozygous for } A\text{)}$$

The frequency of homozygous genotype aa is shown in the lower-right square:

$$\text{Frequency of } aa = q \times q = q^2 \text{ (Homozygous for } a\text{)}$$

The other two squares illustrate the third possibility, i.e. Aa heterozygotes. The total proportion of Aa heterozygotes can be calculated:

$$\text{Frequency of upper-right square } Aa = pq \text{ (Heterozygous)} \quad (1)$$

$$\text{Frequency of lower-left square } Aa = pq \text{ (Heterozygous)} \quad (2)$$

Total frequency of $Aa = Aa\ (1) + Aa\ (2) = 2pq$ (Heterozygous Aa)

The three different genotypes AA, Aa, and aa are formed in proportions p^2, $2pq$, and q^2, respectively. The sum of allelic frequencies is:

$$(p + q)^2 = p^2 + 2pq + q^2$$

This illustrates the HWE. It is important to note that Hardy–Weinberg proportions are binomial in this case (i.e. there are only two alleles). Given any set of genotype frequencies (AA, Aa, and aa), the HWE predicts that after one generation of random mating, provided the assumptions above are all met, the genotypic frequencies will be in the proportions p^2, $2pq$, and q^2. For example, given the initial genotype frequencies of $AA = 0.4$, $Aa = 0.4$, and $aa = 0.2$, where $p = 0.6$ (frequency of allele A) and $q = 0.4$ (frequency of allele a), after one generation the genotype frequencies become:

$$p^2, 2pq, q^2 = (0.6)^2, 2(0.6)(0.4), (0.4)^2 = 0.36, 0.48, 0.16.$$

The genotype frequencies will stay in these proportions generation after generation provided mating is random and the assumptions above are all met. Deviation from any of the above conditions can led to an increase or decrease in allele frequencies from one generation to another and this will impact on the genotype distribution. Finally, it is important to note that this example deviates from the statement earlier in this chapter that states it takes a long time to reach HWE. This is because mutation rates in human genes are low; a full explanation of this is given earlier in this chapter.