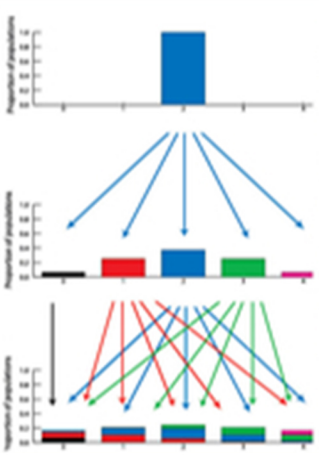
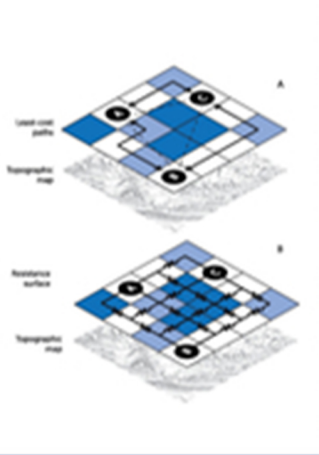
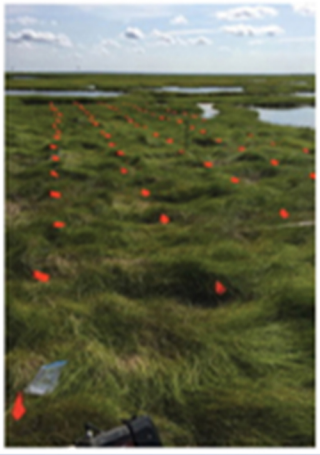
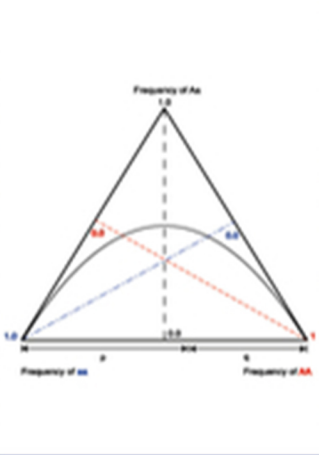
SECOND EDITION

POPULATION GENETICS

MATTHEW B. HAMILTON

with website

WILEY Blackwell

Population Genetics

Population Genetics

Second Edition

Matthew B. Hamilton

WILEY Blackwell

This edition first published 2021

Edition History

Registered Office
John Wiley & Sons, Inc., 111 River Street, Hoboken, NJ 07030, USA

Editorial Office
9600 Garsington Road, Oxford, OX4 2DQ, UK

For details of our global editorial offices, customer services, and more information about Wiley products, visit us at www.wiley.com.

Wiley also publishes its books in a variety of electronic formats and by print-on-demand. Some content that appears in standard print versions of this book may not be available in other formats.

Library of Congress Cataloging-in-Publication Data

Names: Hamilton, Matthew B., author.
Title: Population genetics / Matthew B. Hamilton.
Description: Second edition. | Hoboken, NJ : Wiley-Blackwell, 2021. |
Includes bibliographical references and index.
Identifiers: LCCN 2020025434 (print) | LCCN 2020025435 (ebook) | ISBN 9781118436943 (hardback) | ISBN 9781118436929 (adobe pdf) | ISBN 9781118436899 (epub)
Subjects: LCSH: Population genetics.
Classification: LCC QH455 .H35 2021 (print) | LCC QH455 (ebook) | DDC 576.5/8–dc23
LC record available at https://lccn.loc.gov/2020025434
LC ebook record available at https://lccn.loc.gov/2020025435

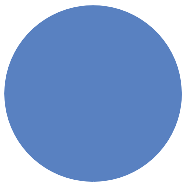
Cover Design: Wiley
Cover Images: Matthew B. Hamilton

Set in 10/12.5pt Photina by SPi Global, Pondicherry, India

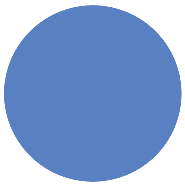
10 9 8 7 6 5 4 3 2 1

Dedication

For my wife and best friend, I-Ling



Contents



Preface and acknowledgements

This book was originally born of two desires, one relatively simple and the other more ambitious, both of which were motivated by my experiences learning and teaching population genetics. My first desire was to create an up-to-date survey text of the field of population genetics. At the same time, I set out with the more ambitious goal of offering an alternative body of materials to change the manner in which population genetics is taught and learned. The first edition of the book made progress toward these goals, and the second edition provides updates and refinements in that same vein.

Much of population genetics during the twentieth century was hypothesis-rich but data-poor. The theory developed between about 1920 and 1980 spawned manifold predictions about basic evolutionary processes. However, many of those predictions could be tested with only very limited power for lack of appropriate or sufficient genetic data. With the advancement of high-throughput DNA sequencing and its still widening employment, population genetics has become much less data-limited. Massive amounts of DNA sequence data are being collected for an expanding set of organisms. Polymorphism and divergence data are now available at a scale of many loci to entire genomes per individual. This has led to a new generation in population genetics that is data-rich. Ironically, this abundance of empirical data has reinforced the central role of models and deductive inference in population genetics. Predictive models have grown to support the genetic data that are now available, fostering innovation at the same time.

Coalescent or genealogical branching models are primary among the models employed in population genetics to make predictions and test hypotheses. During the past few decades, coalescent theory has moved from an esoteric problem pursued for purely mathematical reasons to a central conceptual tool of population genetics. Despite this, the teaching of coalescent theory in undergraduate and graduate population genetics courses has not kept pace with its role in prediction and hypothesis testing. A major impediment has been the lack of teaching materials that make coalescent theory truly accessible to students learning population genetics for the first time. One of my goals was to construct a text that will meet this need with a systematic and thorough introduction to the concepts of coalescent theory and its applications in hypothesis testing. The chapter sections on coalescent theory are presented along with the traditional theory of identity by descent on the same topics to help students see the commonality of the two approaches. However, the coalescence chapter sections could easily be assigned as a group. The second edition retains this focus and adds a section on the ancestral recombination graph.

Another of my goals for this text was to offer a range of explanatory styles. Learning the concepts of population genetics in the language of mathematics is often relatively easy for abstract and mathematical learners. However, my aim was to cater to a wide range of learning styles by building a range of features into the text. A key pedagogical feature of the book is boxes set off from the main text that are designed to engage the various learning styles. Problem boxes placed in the text rather than at the end of chapters are designed to provide practice and to reinforce concepts as they are encountered, appealing to experiential learners. These are now augmented in the second edition with additional end-of-chapter problems. Math boxes that explain mathematical derivations will not only appeal to mathematical and logical learners but also provide insight for all readers into the mathematical reasoning employed in population genetics. In addition, the large number of illustrations in the text were designed to appeal and help cultivate visual learning.

A novel feature of the text is Interact boxes that guide students through semi-structured exercises in computer simulations. These Interact boxes utilize web-based simulations developed specifically for this

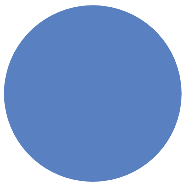
book or public domain software. The simulation problems are an active learning approach and should appeal to experiential or visual learners. Simulations are one of the best ways to demonstrate the outcome of stochastic processes where replication is required before a pattern or generalization can be seen. Because the comprehension of stochastic processes in genetics is a major hurdle for many students, the Interact boxes should aid understanding of central concepts. Additionally, the simulations, spreadsheet models, and scripts provide applications of algorithmic thinking. Algorithmic and computational approaches to problem-solving are now central to prediction and data analysis in population genetics and are useful in most fields of biology and in the sciences more broadly.

The approach to mathematics in the text deserves further explanation. The undergraduate biology curricula employed at most US institutions has students take calculus and applied statistics and usually requires little application of mathematics within biology courses. This leads to students having difficulty in, or avoiding altogether, courses in biological disciplines that require explicit mathematical reasoning. It also leads to courses avoiding explicit mathematical reasoning. Population genetics is built on basic mathematics and probability, and in my experience, students obtain a much deeper understanding of the subject with some comprehension of these mathematical foundations. Therefore, rather than avoid these topics, I have attempted to deconstruct and offer step-by-step explanations of the basic mathematics required for a sound understanding. For those readers with more interest or facility in mathematics, the book presents more detailed derivations in boxes that are separated from the main narrative of the text. There are also some chapter sections containing more mathematically rigorous content. These sections can be assigned or skipped depending on the level and scope of a course supported by this text. This approach will hopefully provide students with the tools to develop their abilities in basic mathematics through application and, at the same time, learn population genetics more fully.

For the second edition, I have tried to incorporate the generous and helpful feedback received from readers of the first edition. John Braverman deserves special mention as a dedicated colleague and friend who has provided sustained suggestions and thoughtful comments. Brent Johnson provided helpful suggestions on statistics topics, and Mak Paranjape helped me understand circuit models. Members of my laboratory and the students who have taken my courses provided feedback on chapter drafts, figures, and effective means to explain the concepts herein. This feedback has been invaluable and has helped me shape the text into a more useful and usable resource for students. The web simulations were developed with the help of Marie Kolawole and Steve Moore, aided by an award from the Georgetown University Initiative on Technology Enhanced Learning.

Many people contributed to the first edition, and their suggestions and input still shapes the book. They include Rachel Adams, Genevieve Croft, John Braverman, Paulo Nuin, James Crow, A.W.F. Edwards, Sivan Rottenstreich Leviyang, Judy Miller, John Dudley, Stephen Moose, Michel Veuille, Eric Delwart, John Epifanio, Robert J. Robbins, Peter Armbruster, Ronda Rolfes, and Martha Weiss. I also thank the anonymous reviewers of the first edition from Aberdeen University, Arkansas State University, Cambridge University, Michigan State University, University of North Carolina, and University of Nottingham. Nancy Wilton, Elizabeth Frank, Haze Humbert, Karen Chambers, and Nik Prowse of Wiley-Blackwell helped bring the first edition to fruition.

Matthew B. Hamilton
October 2020



About the companion websites

This book is accompanied by companion websites for Instructors and Students:

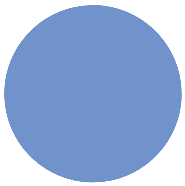
www.wiley.com/go/hamilton/populationgenetics

The Instructor website includes:

- Solutions to the end-of-chapter exercises
- Powerpoints of all figures from the book for downloading, to aid teaching

The Student website includes:

- Chapter resources for Interact Boxes, Problem Boxes, and end-of-chapter exercises



CHAPTER 1

Thinking like a population geneticist

All scientific fields possess a body of concepts as well as a specialized vocabulary used to express these concepts precisely. Population genetics is no different, and the entirety of this book is designed to introduce, explain, and demonstrate these concepts and vocabulary. What may be unique about population genetics among the natural sciences is the way that its practitioners approach questions about the biological world. Population genetics is a dialog between predictions based on the principles of Mendelian inheritance and observations from the empirical measurement of genotype and allele frequencies. Idealized predictions stemming from general principles form the basis of hypotheses that can be tested through observation, experiment, and comparison. At the same time, empirical patterns observed within and among populations are evaluated for evidence of their causes via predictive models. This first chapter will explore some of the ways that population genetics approaches and defines problems that are relevant to the topics in all chapters. The chapter is also intended to give some insight into how to approach the study of population genetics.

1.1 Expectations

- What Do We Expect to Happen?
- Expectations Are the Basis of Understanding Cause and Effect

In our everyday lives, there are many things that we expect to occur or not to occur based on the knowledge of our surroundings and past experience. For example, you probably do not expect to get hit by a meteorite while walking to your next population genetics class. Why not? Meteorites *do* impact the surface of the Earth and, on occasion, strike something noticeable to people nearby. A few times in the distant past, in fact, large meteors have hit the Earth and left evidence like the Chicxulub impact crater on the Yucatán Peninsula in Mexico. What influences your lack of concern? It is probably a combination of basic knowledge of the principles of physics that apply to meteors as well as your empirical observations of the frequency and location of meteor strikes. Basic physics tells us that a small meteor on a collision course with the Earth is unlikely to hit the surface since most objects burn up from the friction they experience traveling through the Earth's atmosphere. You might also reason that even if the object is big enough to pass through the atmosphere intact, and there are far fewer of these, then the Earth is a large place and, just by chance, the impact is unlikely to be even remotely near you. Finally, you have most probably never witnessed a large meteorite impact or even heard of one occurring during your lifetime. You have combined your knowledge of the physical world and your experience to arrive (perhaps unconsciously) at a prediction or an expectation: meteorite strikes are possible but are so infrequent that the risk of being struck while on the way to class is miniscule. In this very same way, you have constructed models of many events and processes in your physical and social world and used the resulting predictions to make comparisons and decisions.

Expectation: The expected value of a random variable, especially the average; a prediction or forecast.

The study of population genetics similarly revolves around constructing and testing expectations for genetic variation in populations of individual

Population Genetics, Second Edition. Matthew B. Hamilton.

Companion website: www.wiley.com/go/hamilton/populationgenetics

organisms. Expectations attempt to predict things like how much genetic variation is present in a population, how genetic variation in a population changes over time, and the pattern of genetic variation that might be left behind by a given biological process that acts over time or through space. Building these expectations involves the use of first principles or the set of very basic rules and assumptions that define how natural systems work at their lowest, most basic levels. A first principle in physics is the force of gravity. In population genetics, first principles are the very basic mechanisms of Mendelian particulate inheritance and processes such as mutation, mating patterns, gene flow, and natural selection that increase, decrease, and shape genetic variation. These foundational rules and processes are used and combined in population genetics with the ultimate goal of building a comprehensive set of predictions that can be applied to any species and any genetic system.

Empirical study in population genetics also plays a central role in constructing and evaluating predictions. In population genetics as in all sciences, empirical evidence is drawn from intentional observations, cleverly constructed comparisons, and experiments. Genetic patterns observed in actual populations are compared with expected patterns to test models constructed using general principles and assumptions. For example, we could construct a mathematical or computer simulation model of random genetic drift (change in allele frequency due to sampling from finite populations) based on abstract principles of sampling from a finite population and biological reproduction. We could then compare the predictions of such a model to the observed change in allele frequency through time in a laboratory population of *Drosophila melanogaster* (fruit flies). If the change in allele frequency in the fruit fly population matched the change in allele frequency predicted using the model of genetic drift, then we could conclude that the model effectively summarizes the biological sampling processes that take place in fruit fly populations.

It is also possible to use well-tested and accepted model expectations as a basis to hypothesize what processes caused an observed pattern in a biological population. Again, to use a *D. melanogaster* population as an example, we might ask whether an observed change in allele frequency over some generations in a wild population could be explained by genetic drift. If the observed allele frequency change is within the range of the predicted change in allele frequencies based on a model of genetic drift, then we have identified a possible *cause* of the observed pattern. Comparing observed genetic patterns in populations often requires modifications to existing models or the construction of novel models in order to develop appropriate expectations. For example, a model of genetic drift constructed for *D. melanogaster* might naturally assume that all individuals in the population are diploid (individuals that possess paired sets of homologous chromosomes). If we wanted to use that same model to predict genetic drift in a population of honeybees, we would have to account for the fact that their males are haploid (individuals that possess single copies of each chromosome) while females are diploid. This change in reproductive biology could be taken into account by altering the assumptions of the model of genetic drift to make predictions appropriate for honeybee populations. Note that without some modifications, a single model of genetic drift would not accurately predict allele frequencies over time in both fruit flies and honeybees since their patterns of reproduction and chromosomal inheritance are different.

Parameters and parameter estimates

While developing the expectations of population genetics in this book, we will most often be working with idealized quantities. For example, allele frequency in a population is a fundamental quantity. For a genetic locus with two alleles, A and a, it is common to say that p equals the frequency of the A allele and q equals the frequency of the a allele. In mathematics, **parameter** is another term for an idealized quantity like an allele frequency. It is assumed that parameters have an exact value. Put another way, parameters are idealized quantities where the messy, real-life details of how to measure the quantities they represent are completely ignored.

Empirical population genetics measures quantities such as allele frequencies to give **parameter estimates** by sampling and then measuring the alleles and genotypes present in actual populations. All experiments, observations, and even simulations in population genetics produce parameter estimates of some sort. There is a subtle notational convention used to indicate an estimate, that is, the hat or ^ character above a variable. Estimates wear hats whereas parameters do not. Using allele frequency as an example, we would say $\hat{p}$ (pronounced "p hat") equals the number of A alleles sampled divided by the total number of alleles sampled. Intuitively, we can see from the denominator in the expression for $\hat{p}$ that the allele frequency estimate will depend on the sample we gather to make the estimate.

In actual populations, a parameter has a true value. For the allele frequency p, knowing this true value would require examining the genotype of every individual and counting *all* A and a alleles to determine their frequency in the population. This task is impractical or impossible in most cases. Instead, we rely on an estimate of allele frequency, $\hat{p}$, obtained from a sample of individuals from the population. Sampling leads to some uncertainty in parameter estimates because repeating the sampling and parameter estimate process would likely lead to a somewhat different parameter estimate each time. Quantifying this uncertainty is important to determine whether repeated sampling might change a parameter estimate by just a little or change it by a lot. When dealing with parameters, we might expect that $p + q = 1$ exactly if there are only two alleles with allele frequencies p and q. However, if we are dealing with estimates, we might say the two allele frequency estimates should sum to approximately one ($\hat{p} + \hat{q} \approx 1$) since each allele frequency is estimated with some errors. The more uncertain the estimates of $\hat{p}$ and $\hat{q}$, the less we should be surprised to find that their sum does not equal the expected value of one.

Parameter: A variable or constant appearing in a mathematical expression; a value (usually unknown) used to represent a certain population characteristic; any factor that defines a system and determines or limits its performance.
Estimate: An indication of the value of an unknown quantity based on observed data; an approximation of a true score, parameter, or value; a statistical estimate of the value of a parameter.

It could be said that statistics sits at the intersection of theoretical and empirical population genetics. Parameters and parameter estimates are fundamentally different things. Estimation requires effort to understand sampling variation and quantify sources of error and bias in samples and estimates. The distinction between parameters and estimates is critical when comparing actual populations with expectations to test hypotheses. When large, random samples can be taken, estimates are likely to have minimal errors. However, there are many cases where estimates have a great deal of uncertainty, which limits the ability to evaluate expectations. There are also instances where very different processes may produce very similar expected results. In such cases, it may be difficult or impossible to distinguish the different potential causes of a pattern due to the approximate nature of estimates. While this book focuses mostly on parameters, it is useful to bear in mind that testing or comparing expectations requires the use of parameter estimates and statistics that quantify sampling error. The Appendix provides a review of some basic statistics that are used in the text.

Inductive and deductive reasoning

Population genetics employs both **inductive** and **deductive reasoning** in an effort to understand the biological processes operating in actual populations as well as to elucidate the general processes that cause population genetic phenomena. The inductive approach to population genetics involves assembling measures of genetic variation (parameter estimates) from various populations to build up evidence that can be used to identify the underlying processes that produced the observed patterns. This approach is logically identical to that used by Isaac Newton, who used knowledge of how objects fall to the surface of the Earth as well as knowledge of the movement of planets to arrive at the general principles of gravity. Application of inductive reasoning requires detailed familiarity with the various empirical data types in population genetics, such as DNA sequences, along with the results of studies that report observed patterns of genetic variation. From this accumulated empirical information, it is then possible to draw more general conclusions about the qualities and quantities of genetic variation in populations. Model organisms like *D. melanogaster* and *Arabidopsis thaliana* play a large role in population genetic conclusions reached by inductive reasoning. Because model organisms receive a large amount of scientific effort, for example, to completely sequence and annotate their genomes, a great deal of available genetic data are accumulated for these species. Based on this evidence, many inferences have been made about population genetic processes. Although model organisms are very rich sources of empirical information, the number of species is limited by definition so that any generalizations may not apply universally to all species.

Deductive reasoning: Using general principles to reach conclusions about specific instances.
Inductive reasoning: Utilizing the knowledge of specific instances or cases to arrive at general principles.

The study of population genetics can also be approached using deductive reasoning. The actions of general processes such as genetic drift, mutation, and natural selection are represented by parameters in the mathematical equations that make up population genetic models. These models can then be used to make predictions about the quantity of genetic variation and patterns of genetic variation in space and time. Such population genetic models make general predictions about things like rates of change in allele frequency, the eventual equilibrium of allele or genotype frequencies, and the net outcome of several processes operating at the same time. These predictions are very general in that they apply to any population of any species since the predictions arose from general principles in the first place. At the same time, such general predictions may not be directly applicable to a specific population because the general principles and assumptions used to make the prediction are not specific enough to match an actual population.

Historically, the field of population genetics has developed from an interplay between arguments and evidence developed using both inductive and deductive reasoning approaches. Nonetheless, most of the major ideas in population genetics can be first approached with deductive reasoning by learning and understanding the expectations that arise from the principles of Mendelian heredity. This book stresses on the process of deductive reasoning to arrive at these fundamental predictions. Empirical evidence related to expectations is included to illustrate predictions and to demonstrate hypothesis tests that result from expectations. Because the body of empirical results in population genetics is very large, readers should resist the temptation to generalize too much from the limited number of empirical studies that are presented. Detailed reviews of particular areas of population genetics, many of which are cited, are a better source for comprehensive summaries of empirical studies.

In the next chapter, we will start by building expectations for the frequencies of diploid genotypes based on the foundation of particulate inheritance: that alleles are passed unaltered from parents to offspring. There is ample support for particulate inheritance from both molecular biology, which identifies DNA as the hereditary molecule, and from allele and genotype frequencies that can be observed in actual populations. The general principle of particulate inheritance has been used to formulate a wide array of expectations about allele and genotype frequencies in populations.

1.2 Theory and assumptions

- What Is a Theory and What Are Assumptions?
- How Can Theories Be Useful with So Many Assumptions?

In colloquial usage, the word *theory* refers to something that is known with uncertainty, or a quantity that is approximate. On a day you are running late leaving work, you might say, "In theory, I am supposed to depart at 6:00 pm." In science, theory has a very different meaning. Theory is the accumulation of expectations and observations that have withstood tests and critical scrutiny and are accepted by at least some practitioners of a scientific field. Theory is the collection of all of the expectations developed for specific cases or individual biological processes that together form a more comprehensive set of general principles. The combination of Darwin's hypothesis of natural selection with the laws of Mendelian particulate inheritance is often called the *modern synthesis* of evolutionary biology since it is a comprehensive theory to explain the causes of evolutionary change. The modern synthesis can offer causal explanations for biological phenomena ranging from antibiotic resistance in bacteria to the behavior of elephants to the rate of DNA sequence change, as well as make predictions to guide animal and plant breeders. In all of the modern synthesis, population genetics plays a central role.

It is common for the uninitiated to ask the question "what good is theory if it is based on so many assumptions?" A body of theory is a useful tool to articulate assumptions and generate testable predictions. Theory that generates many testable predictions about the world also offers many opportunities to falsify its predictions and assumptions. Since hypotheses cannot be proven directly, but alternative hypotheses can be disproven, the generation of plausible, testable alternative hypotheses is a requirement for scientific inquiry. Strong theories are able to make accurate

predictions, offer causal explanations for diverse observations, and generate alternative hypotheses based on revised assumptions.

The words *theory* and *assumption* can seem abstract, but you should not be intimidated by them. Theories are just collections of expectations, each with a set of assumptions that place bounds on the prediction being made. If you understand what motivates an expectation, its predictions, and its assumptions, then you understand theory. Most expectations in population genetics will have at least a few, and often many, assumptions used to define and bound the situation. For example, we might assume something about the size of a population or the absence of mutation, or that all genotypes are diploid with two alleles. This is a way of limiting the prediction to appropriate circumstances and a way of defining which quantities and conditions can vary and which are fixed. Each of these assumptions can influence the generality of an expectation. Each assumption can also be relaxed or altered to see how strongly it influences the expectation. To return to the example in the preceding section, if, one day, meteorites start falling around us with regularity, we would be forced to call into question some of the basic assumptions originally used to formulate our expectation that meteorite strikes should be rare events. In this way, assumptions are useful tools to ask "what if…?" as part of the process of developing a prediction. If our initial "what if…?" conditions do not match a situation, then the resulting prediction will probably be inaccurate.

In population genetics, as in much of science where theory and expectations are involved, empirical data and model expectations are routinely compared. Imagine observing a set of genotype frequencies in a biological population. It would then be natural to construct an idealized population by using theory that approximates the biological population. This is an attempt to construct an idealized population that is *equivalent* to the actual population from the perspective of the processes influencing genotype frequencies. For example, a large population may behave exactly like a small, randomly mating ideal population in terms of genotype frequencies. This equivalence allows us to use expectations for ideal populations with one or a few variables specified in order to describe an actual population where there are many more, usually unknown, parameters. What we strive to do is to focus on those variables that strongly influence genotype frequencies in the actual population. In this way, it is often possible to reduce the complexity of a real population and determine the key variables that strongly influence a property like genotype frequencies. The ideal population is not meant to match the actual population in every detail.

Theory: A scheme or system of ideas or statements held as an explanation or account of a group of facts or phenomena; the general laws, principles, or causes of something known or observed.
Infer: To draw a conclusion or make a deduction based on facts or indications; to have as a logical consequence.

From the comparison of expectation and observation, we infer that the first principles used to construct the expectation are sound if they can be used to explain patterns observed in the biological world. However, there is a major distinction between considering an actual and idealized population *equivalent* and considering them *identical*. This is seen in cases where the observed pattern in an actual population is consistent with the expectations from several model populations built around distinct and incompatible assumptions. In such cases, it is not possible to infer the processes that cause a given pattern without additional information. A common example in population genetics are cases of genetic patterns that are potentially consistent with the random process of genetic drift and, at the same time, consistent with some form of the deterministic process of natural selection. In such cases, unambiguous inference of the underlying cause of a pattern is not possible without additional empirical information or more precise expectations.

1.3 Simulation

- A Method of Practice, Trial and Error Learning, and Exploration

Imagine learning to play the piano without ever touching a piano or practicing the hand movements required to play. What if you were expected to play a difficult concerto after extensive exposure (perhaps a semester) to only verbal and written descriptions of how other people play? Such a teaching style would make learning to play the piano very difficult because there would be no opportunity for practice, trial and error, or exploration. You would not have

Interact box 1.1 The textbook website

Throughout this book, you will encounter Interact boxes. These boxes contain opportunities for you to interact directly with the material in the text by using computer simulations designed to demonstrate fundamental concepts of population genetics. Each box will contain step-by-step instructions for you to follow in order to carry out a simulation. By following the instructions, you will get started with the simulation. However, always feel free to use your own imagination and intuition. After following the instructions in the Interact box and understanding the point at hand, enter different values, push more buttons, and even read the documentation. You can also return to Interact boxes at a later time, perhaps after you have read and understood more of the text, to reconsider a simulation or view it in a different light. You can also use the simulations to answer questions that may occur to you or to test hypotheses that you may have. Questions in population genetics that start off "What would happen if…?" can often be answered with simulation.

The book's website gives you the worldwide web address (URL) for each interact box. This prevents problems in case web addresses change because the website can be updated while your copy of the text cannot be updated.

Step 1 Open a web browser and enter http://www.wiley.com/go/hamiltongenetics
Step 2 Click on the Chapter resources link that is associated with Interact boxes.
Step 3 Verify that the page gives links for each of the Interact boxes listed by their number. You could also bookmark this page so you can access it directly in the future.

Congratulations! You have completed the first Interact box.

the opportunity for direct experience nor incremental improvement of your understanding. Unfortunately, this is exactly how science courses are taught to some degree. You are expected to learn and remember concepts with only limited opportunities for directly observing principles in action. In fairness, this is partly due to the difficulty of carrying out some of the experiments or observations that originally lead someone to discover and understand an important principle.

In the field of population genetics, computer simulations can be used to effectively demonstrate many fundamental genetic processes. In fact, computer simulations are an important research tool in population genetics. Therefore, when you conduct simulations, you are both learning by direct experience and learning using the same methods that are used by researchers. Simulations allow us to view how quantities like allele frequencies change over time, observe their dynamics, and determine whether a stable end point is reached: an equilibrium. With simulations, we can view dynamics (change over time) and equilibria over very long periods of time and under a vast array of conditions in an effort to reach general conclusions. Without simulations, it would be impossible for us to directly observe allele frequencies over such long periods of time and in such diverse biological situations.

Simulations are an effective means to understand some of the fundamental predictions of populations genetics. Mathematical expressions are frequently used to express dynamics and equilibria in population genetics, but the equations alone can be opaque at first. Simulations provide a means to explore the relationships among variables that are summarized in the compact language of mathematics. Many people feel that a set of mathematical equations is much more meaningful after having the chance to explore what they describe with some actual numerical values. Simulation provides the means to explore what equations predict and can make learning population genetics an easier, more rewarding experience.

Carrying out simulations has the potential to make the expectations of population genetics much more accessible and understandable. Conducting simulations is not much extra work, especially once you get into the practice of using the text and simulation software in concert. You can approach simulations as if they are games, where each one shows a visual scene that helps to solve a puzzle. In addition, simulations can help you develop a more intuitive understanding of population genetic predictions so

you do not have to approach the expectations of population genetics as disembodied or unanimated "facts."

It is important to approach simulations in a systematic and organized fashion, not as just a collection of buttons to press and text entry boxes to be filled in on a whim. It is absolutely imperative that you understand the meaning behind each variable that you can control as well as the meaning of the results you obtain. To do so successfully, you will need to be aware of both specific details and larger patterns, or both the individual trees and the forest that they compose. For example, in a simulation that presents results as a graph, it is important that you understand the details of what variables are represented on each axis and the range of axis values. Sometimes these details are not always completely obvious in simulation software, requiring you to use both your intuition and knowledge of the population genetic processes being simulated.

Once you are comfortable with the details of a simulation, you will also want to keep track of the "big picture" patterns that emerge as you view simulation results. Seeing these patterns will often require that you examine the results over a range of conditions. Try approaching simulations as experiments by changing only one variable at a time until you understand its effects on the outcome. Changing several things all at once can lead to confusion and an inability to see cause-and-effect relationships, unless you have fully understood the effects of individual variables. Finally, try writing down parameter values you have tried in a simulation and sketching or tabulating results on paper as you work with a simulation. Use all of your skills as a scientist and student when conducting simulations, and they will become a powerful learning tool. Eventually, you may even use scripting and programming to carry out your own simulations specifically designed to explore your own genetic hypotheses.

Chapter 1 review

- Both general principles and direct measurements taken in actual populations combine to form comprehensive expectations about amounts, patterns, and cause-and-effect relationships in population genetics.
- The theory of population genetics is the collection of well-accepted expectations used to articulate a wide array of predictions about the biological processes that shape genetic variation.
- Parameters are idealized quantities that are exact, while parameter estimates wear notational "hats" to remind us that they have statistical uncertainty.
- Population genetics uses both inductive reasoning to generalize from the knowledge of specifics and deductive reasoning to build up predictions from general principles that can be applied to specific situations.
- Population genetics is not a spectator sport! Direct participation through computer simulation provides the opportunity to see population genetic processes in action. You can learn by trial and error and test your own understanding by making predictions and then comparing them with simulation results.

Further reading

For a history of population genetics from Darwin to the 1930s, see:

Provine, W.B. (1971). *The Origins of Theoretical Population Genetics*. Chicago, IL: University of Chicago Press.

For a concise history of population genetics since the mid-1960s that highlights major conceptual advances as well as technical innovations to measure genetic variation, see:

Charlesworth, B. and Charlesworth, D. (2017). Population genetics from 1966 to 2016. *Heredity* 118: 2–9.

For two personal and historical essays on the past, present, and assumptions of theoretical population genetics, see:

Lewontin, R.C. (1985). Population genetics. In: *Evolution: Essays in Honour of John Maynard Smith* (eds. P.J. Greenwood, P.H. Harvey and M. Slatkin), 3–18. Cambridge: Cambridge University Press.
Wakeley, J. (2005). The limits of theoretical population genetics. *Genetics* 169: 1–7.

CHAPTER 2

Genotype frequencies

2.1 Mendel's model of particulate genetics

- Mendel's breeding experiments.
- Independent assortment of alleles.
- Independent segregation of loci.
- Some common genetic terminology.

In the nineteenth century, there were several theories of heredity, including inheritance of acquired characteristics and blending inheritance. Jean-Baptiste Lamarck is most commonly associated with the discredited hypothesis of inheritance of acquired characteristics (although it is important to recognize his efforts in seeking general causal explanations of evolutionary change). He argued that individuals contain "nervous fluid" and that organs or features (phenotypes) employed or exercised more frequently attract more nervous fluid, causing the trait to become more developed in their offspring. His widely known example is the long neck of the giraffe, which he said developed because individuals continually stretched to reach leaves at the tops of trees. Later, Charles Darwin and many of his contemporaries subscribed to the idea of blending inheritance. Under blending inheritance, offspring display phenotypes that are an intermediate combination of parental phenotypes (Figure 2.1).

From 1856 to 1863, the Augustinian monk Gregor Mendel carried out experiments with pea plants that demonstrated the concept of particulate inheritance. Mendel showed that phenotypes are determined by discrete units that are inherited intact and unchanged through generations. His hypothesis was sufficient to explain three common observations: (i) phenotype is sometimes identical between parents and offspring; (ii) offspring phenotype can differ from that of the parents; and (iii) "pure" phenotypes of earlier generations could skip generations and reappear in later generations. Neither blending inheritance nor inheritance of acquired characteristics are satisfactory explanations for all of these observations. It is hard for us to fully appreciate now, but Mendel's results were truly revolutionary and served as the very foundation of population genetics. The lack of an accurate mechanistic model of heredity severely constrained biological explanations of cause and effect up to the point that Mendel's results were "rediscovered" in the year 1900.

It is worthwhile to briefly review the experiments with pea plants that Mendel used to demonstrate independent assortment of both alleles within a locus and of multiple loci, sometimes dubbed Mendel's first and second laws. We need to remember that this was well before the Punnett square, which originated in about 1905. Therefore, the conceptual tool we would use now to predict progeny genotypes from parental genotypes was a thing of the future. So, in revisiting Mendel's experiments, we will not use the Punnett square in an attempt to follow his logic. Mendel only observed the phenotypes of generations of pea plants that he had hand-pollinated. From these phenotypes and their patterns of inheritance, he inferred the existence of heritable factors. His experiments were actually both logical and clever, but are now taken for granted since the basic mechanism of particulate inheritance has long since ceased to be an open question. It was Mendel who established the first and most fundamental prediction of population genetics: expected genotype frequencies.

Mendel used pea seed coat color as a phenotype he could track across generations. His goal was to determine, if possible, the general rules governing the

Population Genetics, Second Edition. Matthew B. Hamilton.

Companion website: www.wiley.com/go/hamilton/populationgenetics

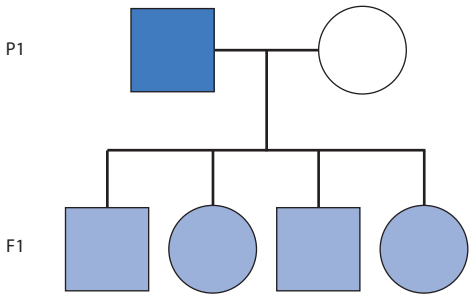


Figure 2.1 The model of blending inheritance predicts that progeny have phenotypes that are the intermediate of their parents. Here, "pure" blue and white parents yield light blue progeny, but these intermediate progeny could never themselves be parents of progeny with pure blue or white phenotypes identical to those in the P1 generation. Crossing any shade of blue with a pure white or blue phenotype would always lead to some intermediate shade of blue. By convention, in pedigrees, females are indicated by circles and males by squares while "P" refers to parental and "F" to filial.

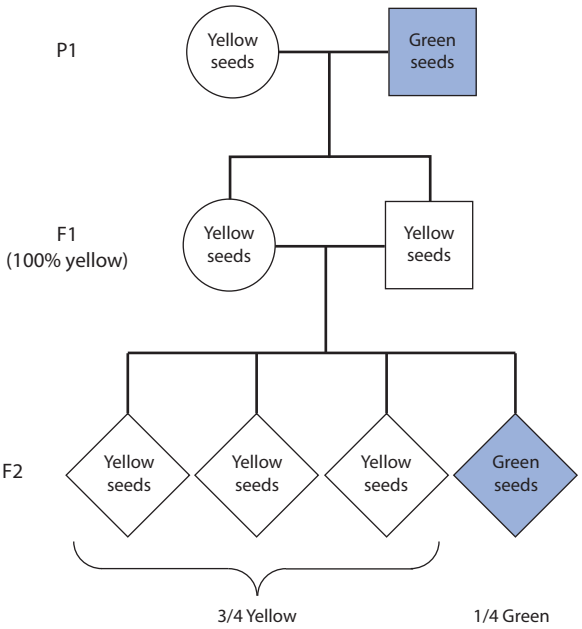


Figure 2.2 Mendel's crosses to examine the segregation ratio in the seed coat color of pea plants. The parental plants (P1 generation) were pure breeding, meaning that if self-fertilized all resulting progeny had a phenotype identical to the parent. Some individuals are represented by diamonds since pea plants are hermaphrodites and can act as a mother, a father, or can self-fertilize.

inheritance of pea phenotypes. He established "pure"-breeding lines (meaning plants that always produced progeny with phenotypes like themselves) of peas with both yellow and green seeds. Using these pure-breeding lines as parents, he crossed a yellow-

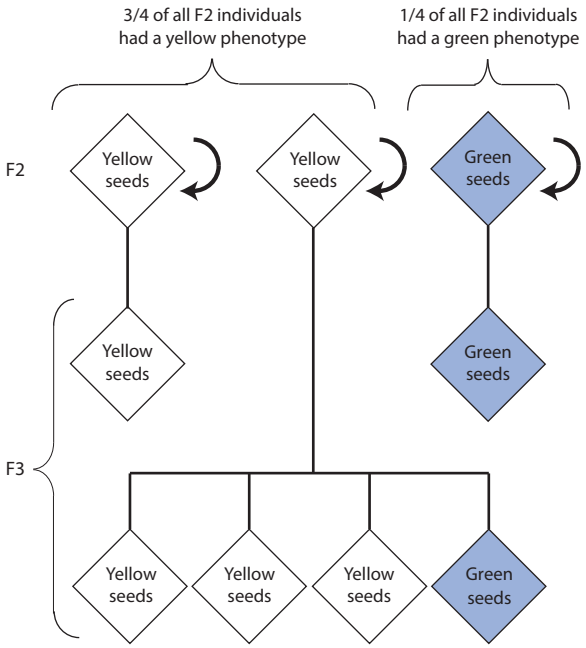


Figure 2.3 Mendel self-pollinated (indicated by curved arrows) the F2 progeny produced by the cross shown in Figure 2.2. Of the F2 progeny that had a yellow phenotype (3/4 of the total), 1/3 produced all progeny with a yellow phenotype and 2/3 produced progeny with a 3 : 1 ratio of yellow and green progeny (or 3/4 yellow progeny). Individuals are represented by diamonds since pea plants are hermaphrodites.

and a green-seeded plant. The parental cross and the next two generations of the progeny are shown in Figure 2.2. Mendel recognized that the F1 plants had an "impure" phenotype because of the F2 generation plants, of which three-quarters had yellow and one-quarter had green seed coats.

His insightful next step was to self-pollinate a sample of the plants from the F2 generation (Figure 2.3). He considered the F2 individuals with yellow and green seed coats separately. All green-seeded F2 plants produced green progeny and thus were "pure" green. However, the yellow-seeded F2 plants were of two kinds. Considering just the yellow F2 seeds, one-third were pure and produced only yellow-seeded progeny, whereas two-thirds were "impure" yellow since they produced both yellow- and green-seeded progeny. Mendel combined the frequencies of the F2 yellow and green phenotypes along with the frequencies of the F3 progeny. He reasoned that three-quarters of all F2 plants had yellow seeds, but these could be divided into plants that produced pure yellow F3 progeny (one-third) and plants that produced both yellow and green F3 progeny

(two-thirds). So, the ratio of pure yellow to impure yellow in the F2 was (1/3 × 3/4 =) 1/4 pure yellow to (2/3 × 3/4 =) 1/2 "impure" yellow. The green-seeded progeny comprised one-quarter of the F2 generation and all produced green-seeded progeny when self-fertilized, so that (1 × 1/4 green =) 1/4 pure green. In total, the ratios of phenotypes in the F2 generation were 1 pure yellow : 2 impure yellow : 1 pure green or 1 : 2 : 1. Mendel reasoned that "the ratio of 3 : 1 in which the distribution of the dominating and recessive traits take place in the first generation therefore resolves itself into the ratio of 1 : 2 : 1 if one differentiates the meaning of the dominating trait as a hybrid and as a parental trait" (quoted in Orel 1996). During his work, Mendel employed the terms "dominating" (which became dominant) and "recessive" to describe the manifestation of traits in impure or heterozygous individuals.

With the benefit of modern symbols of particulate heredity, we could diagram Mendel's monohybrid cross with pea color in the following way.

P1	Phenotype	Yellow × green	
	Genotype	GG	Gg
	Gametes produced	G	G
F1	Phenotype	All "impure"yellow	
	Genotype	Gg	
	Gametes produced	G, g	

A Punnet square could be used to predict the phenotypic ratios of the F2 plants

	G	G
G	GG	Gg
G	Gg	Gg

F2	Phenotype	3 Yellow : 1 green		
	Genotype	GG	Gg	Gg
	Gametes produced	G	G, g	G

and another Punnet square could be used to predict the genotypic ratios of the two-thirds of the yellow F2 plants

	G	G
G	GG	Gg
G	Gg	Gg

Mendel's first "law": Predicts independent segregation of alleles at a single locus: two copies of a diploid locus (a pair of alleles that make a diploid genotype) segregate independently into gametes so that in a large number of gametes half carry one allele and the other half carry the other allele.

Individual pea plants obviously have more than a single phenotype, and Mendel followed the inheritance of other characters in addition to seed coat color. In one example of his crossing experiments, Mendel tracked the simultaneous inheritance of both seed coat color and seed surface condition (either wrinkled ["angular"] or smooth). He constructed an initial cross among pure-breeding lines identical to what he had done when tracking seed color inheritance, except now there were two phenotypes (Figure 2.4). The F2 progeny appeared in the phenotypic ratio of 9 round/yellow : 3 round/green : 3 wrinkled/yellow : 1 wrinkled/green.

How did Mendel go from this F2 phenotypic ratio to the second law? He ignored the wrinkled/smooth phenotype and just considered the yellow/green seed color phenotype in self-pollination crosses of F2 plants just like those for the first law. In the F2 progeny, 12/16 or three-quarters had a yellow seed coat and 4/16 or one-quarter had a green seed coat, or a 3 yellow : 1 green phenotypic ratio. Again using self-pollination of F2 plants like those in Figure 2.3, he showed that the yellow phenotypes were (1/3 × ¾) one-quarter pure and (2/3 × ¾) one-half impure yellow. Thus, the segregation ratio for seed color was 1 : 2 : 1 and the wrinkled/smooth phenotype did not alter this result. Mendel obtained an identical result when considering instead only the wrinkled/smooth phenotype and ignoring the seed color phenotype.

Mendel concluded that a phenotypic segregation ratio of 9 : 3 : 3 : 1 is the same as combining two independent 3 : 1 segregation ratios of two phenotypes since (3 : 1) × (3 : 1) = 9 : 3 : 3 : 1. Similarly, the multiplication of two (1 : 2 : 1) phenotypic ratios will predict the two phenotype ratios (1 : 2 : 1) × (1 : 2 : 1) = 1 : 2 : 1 : 2 : 4 : 2 : 1 : 2 : 1. We now recognize that dominance in the first two phenotype ratios masks the ability to distinguish some of the homozygous and heterozygous genotypes, whereas the ratio in the second case would result if there was no

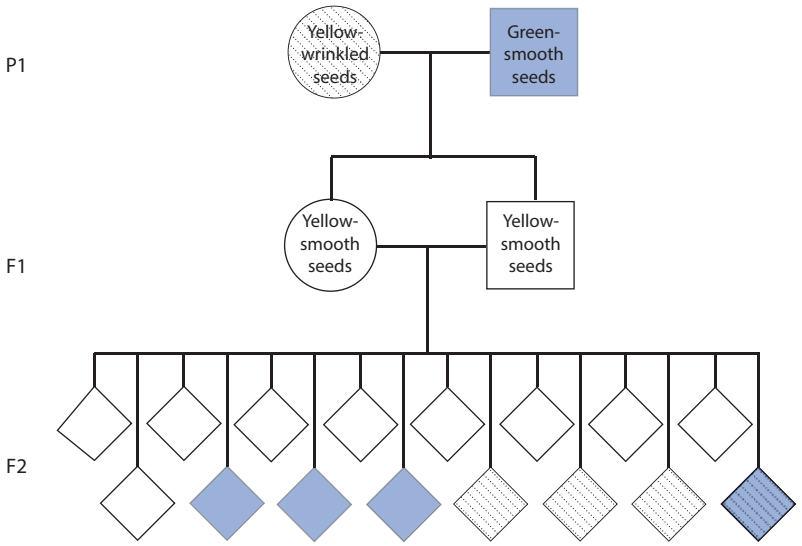


Figure 2.4 Mendel's crosses to examine the segregation ratios of two phenotypes, seed coat color (yellow or green) and seed coat surface (smooth or wrinkled), in pea plants. The stippled pattern indicates wrinkled seeds, while the solid color indicates smooth seeds. The F2 individuals exhibited a phenotypic ratio of 9 round-yellow: 3 round-green: 3 wrinkled-yellow: 1 wrinkled-green.

dominance. You can confirm these conclusions by working out a Punnett square for the F2 progeny in the two-locus case.

Mendel's second "law": Predicts independent assortment of multiple loci: during gamete formation, the segregation of alleles of one locus is independent of the segregation of alleles of another locus.

Mendel performed similar breeding experiments with numerous other pea phenotypes and obtained similar results. Mendel described his work with peas and other plants in lectures and published it in 1866 in the *Proceedings of the Natural Science Society of Brünn* in German where it went unnoticed for nearly 35 years. However, Mendel's results were eventually recognized, and his paper was translated into several languages. Mendel's rediscovered the hypothesis of particulate inheritance was also bolstered by evidence from microscopic observations of chromosomes during cell division that led Walter Sutton to propose in 1902 that chromosomes are the physical basis of heredity, supported by results obtained independently by Theodor Boveri at around the same time (see Crow and Crow 2002).

Much of the currently used terminology was coined as the field of particulate genetics initially developed. Therefore, many of the critical terms in genetics have remained in use for long periods of time. However, the meanings and connotations of these terms have often changed as our understanding of genetics has also changed.

Unfortunately, this has led to a situation where words can sometimes mislead. A common example is equating *gene* and *allele*. For example, it is commonplace for news media to report scientific breakthroughs where a "gene" has been identified as causing a particular phenotype, often a debilitating disease. Very often what is meant in these cases is that a genotype or an *allele* with the phenotypic effect has been identified. Both unaffected and affected individuals all possess the gene, but they differ in their alleles and therefore in their genotype. If individuals of the same species really differed in their gene content (or loci they possessed), that would provide evidence of additions or deletions to genomes. For an interesting discussion of how terminology in genetics has changed – and some of the misunderstandings this can cause, see Judson (2001).

Gene: A unit of particulate inheritance; in contemporary usage, it usually means an exon or series of exons, or a DNA sequence that codes for an RNA or protein.
Locus (plural **loci**, pronounced "low-sigh"): Literally "place" or location in the genome; in contemporary usage, it is the most general reference to *any* sequence or genomic region, including non-coding regions.
Allele: A variant or alternative form of the DNA sequence at a given locus.
Genotype: The set of alleles possessed by an individual at one locus; the genetic composition of an individual at one locus or many loci.
Phenotype: The morphological, biochemical, physiological, and behavioral attributes of an individual; synonymous with character and trait.
Dominant: Where the expressed phenotype of one allele takes precedence over the expressed phenotype of another allele. The allele associated with the expressed phenotype is said to be dominant. Dominance is seen on a continuous scale that includes "complete" dominance (one allele completely masks the phenotype of another allele so that the phenotype of a heterozygote is identical to a homozygote for the dominant allele) and "partial" or "incomplete" dominance (masking effect is incomplete so that the phenotype of a heterozygote is intermediate to both homozygotes) and includes over- and under-dominance (phenotype is outside the range of phenotypes seen in the homozygous genotypes). The lack of dominance (heterozygote is exactly intermediate to the phenotypes of both homozygotes) is when the effects of alleles are additive, a situation sometimes termed "codominance" or "semi-dominance."
Recessive: The expressed phenotype of one allele is masked by the expressed phenotype of another allele. The allele associated with the concealed phenotype is said to be recessive.

2.2 Hardy–Weinberg expected genotype frequencies

- Hardy–Weinberg and its assumptions.
- Each assumption is a population genetic process.
- Hardy–Weinberg is a null model.
- Hardy–Weinberg in haplo-diploid systems.

Mendel's "laws" could be called the original expectations in population genetics. With the concept of particulate genetics established, it was possible to make a wide array of predictions about genotype and allele frequencies as well as the frequency of phenotypes with a one-locus basis. Still, progress and insight into particulate genetics were gradual. Until 1914, it was generally believed that rare (infrequent) alleles would disappear from populations over time. Godfrey H. Hardy (1908) and Wilhelm Weinberg (1908) worked independently to show that the laws of Mendelian heredity did not predict such a phenomenon (see Crow 1988). In 1908, they both formulated the relationship that can be used to predict allele frequencies given genotype frequencies or predict genotype frequencies given allele frequencies. This relationship is the well-known Hardy–Weinberg equation.

$$p^2 + 2pq + q^2 = 1 \qquad (2.1)$$

where p and q are allele frequencies for a genetic locus with two alleles.

Genotype frequencies predicted by the Hardy–Weinberg equation can be summarized graphically. Figure 2.5 shows Hardy–Weinberg expected genotype frequencies on the y axis for each genotype for any given value of the allele frequency on the x axis. Another graphical tool to depict genotype and allele frequencies simultaneously for a single locus with two alleles is the de Finetti diagram (Figure 2.6). As we will see, de Finetti diagrams are helpful when examining how population genetic processes dictate allele and genotype frequencies. In both graphs, it is apparent that heterozygotes are most frequent when the frequency of the two alleles is equal to 0.5. You can also see that when an allele is rare, the corresponding homozygote genotype is even rarer since the genotype frequency is the square of the allele frequency.

A single generation of reproduction where a set of conditions, or assumptions, is met will result in a