2ND EDITION

GENETIC THEORY AND ANALYSIS

FINDING MEANING IN A GENOME

DANNY E. MILLER | ANGELA L. MILLER
R. SCOTT HAWLEY

WILEY

Genetic Theory and Analysis

Genetic Theory and Analysis

Finding Meaning in a Genome

Second Edition

Danny E. Miller
Angela L. Miller
R. Scott Hawley

WILEY

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Library of Congress Cataloging-in-Publication Data applied for
Paperback ISBN: 9781118086926

Cover Design: Wiley
Cover Image: © Romy Honda Oka Romyhonda/EyeEm/Getty Images

Set in 9.5/12.5pt STIXTwoText by Straive, Pondicherry, India

Contents

Preface

For the geneticist there are accordingly three ways of examining anything. Through characters [s]he can examine function; through their changes, [s]he can examine mutation; through their reassortment, [s]he can examine recombination.

– Francois Jacob in The Logic of Life (p. 224)

Although the first edition of this book was intended for an advanced course in genetic analysis, we have realized that motivated undergraduates are as capable of digesting this information as graduate students. This book does assume the reader has a basic familiarity with the genetics of eukaryotes, as well as with the basic biology of prokaryotes and their viruses. We also assume a working knowledge of the three Sirens of molecular biology: transcription, translation, and replication. In cases where specialized techniques are used or concepts required, we have endeavored to provide the essential background material. For ease of use, we have added a glossary and a more comprehensive index.

The focus of this book was, and still is, on the basic principles that underlie genetic analysis: mutation, complementation (and its bridesmaids, suppression and enhancement), recombination, segregation, and regulation. Our goal is to provide insight into the biological and analytical processes that comprise each of these tools and to explain their use. Our basic objective is for you to learn just what each tool or test does and how it can lie to you. Perhaps most importantly, the book is designed to teach you just how much you can learn when nature misunderstands your question. In other words, this is a book about genetic theory.

Although a discussion of genetic analysis invariably requires the presentation of multiple examples, this is not to be considered a textbook of genetic facts. Facts can sometimes change in the blink of an eye; the basic analytical tools change rather more slowly. We have tried to be as comprehensive and catholic as possible in the choice of examples and organisms, drawing on studies from as many genetically tractable model organisms as possible, with even the occasional reference to humans. However, we cannot escape the truth that we are fly biologists by training and so, despite our best efforts, lean a little heavily on Drosophila stories. Moreover, with advancements in genome sequencing and annotation, the model organism landscape is ever growing and evolving. Beautiful and groundbreaking analyses are undertaken daily in organisms whose genomes were not available only a few years ago such as zebrafish, cave fish, planaria, and Nematostella. There are now so many organisms available for study that we must regrettably limit the model organisms described in Appendix A to only those relevant to the examples contained within these pages.

The following nine chapters cover the basic intellectual tools that comprise modern genetics. Some of these techniques, such as mutant hunting, suppressor analysis and complementation

analysis, echo issues derived from even the most current journals, while others, such as the algebraic analysis of recombination data, have fallen into disuse. While we were writing this book, we were often reminded by our colleagues that some things are no longer done in the fashions we have described. This is, after all, the post-genomic era.

Living in an era of sequenced genomes is a heady business, indeed. But we are reminded of a comment by the playwright Noel Coward, when pressured by a friend as to how he was doing on his latest play, he is reported to have answered that he was half done. He had taken all the words out of the dictionary, now he just needed to put them in proper order. The sequencing of many genomes has given us our list of words or genes. Now we, like Mr. Coward, need to put our words in the right order. Unfortunately, unlike Mr. Coward, we have been given many words that we do not understand and must learn what these words (or genes) mean or do. We suspect that the best way to understand what genes do is to mutate them. And the best way to put them in order will be epistasis analysis. One may identify interacting proteins by systems such as the yeast two-hybrid assay or mass spectroscopy, but those interactions will have real meaning only when confirmed by genetic interactions as well. In that sense, then, perhaps the genetics *that was* becomes the genetics *that is*. We suspect that these tools, and the intellectual principles that created them, will have much more than historical value for generations of biologists to come.

Writing a book like this is rarely a solitary process, and we would be remiss if we didn't express our gratitude for those who graciously travelled with us at various points along the journey. We particularly want to thank the Stowers Institute, and specifically members of the Hawley lab, as well as the students in 206H and the patients and their families who keep us motivated. We also enormously appreciate the "community of geneticists"; we are beyond lucky to be part of this community.

Finally, this book contains several hundred references. Despite every attempt to be complete, we know we have missed citing many extremely well done and important examples of genetic analysis. In our defense, we can say only that there came a time to stop reading and to actually write this book. Our apologies to those authors whose work was in the huge stack of "you know, we really ought to discuss this" papers that remain sitting on the sides of our desks.

Introduction

All in due time, my pretties, all in due time.

– The Wicked Witch of the West

One begins a genetics textbook by talking either about Gregor Mendel or about James Watson, Francis Crick, and Rosalind Franklin. That is simply the way things are done. The choice an author makes reflects their basic scientific predilections. Classical geneticists start with Mendel, and thus so will we. We will get to Drs. Watson, Crick, and Franklin in due time.

To Gregor Mendel, a gene was little more than a statistical entity, defined by effects on phenotypic variation and by segregation. Each trait was determined by two copies of a given gene. Differences in the trait (phenotype) were due to differences in the "form" of a given gene. These different forms of a given gene were called alleles. In Mendel's construction, the phenotype was a direct consequence of the alleles present in a given individual. To explain the cases where the effect of one allele predominated over the other, Mendel created the concepts of dominance and recessivity. Thus, to Mendel, the primary definition of a gene was that it was a unit of hereditary information. Genes were not structures or tissues themselves, rather they provided information required to create those things.

Mendel's concept of the gene was also firmly embedded in the idea of a gene as the unit of segregation. This is to say an individual possesses two copies of each gene – one copy that was inherited from the mother and one copy that was inherited from the father. Moreover, an individual passes on only one of those two genes to their own offspring, and they do so at random. The gene pair thus becomes the unit of segregation that ultimately leads to gametes bearing a single hereditary particle (gene) for each trait. Mendel's concept of independent assortment can be thought of as a rather simple extension of this idea: because the individual gene pair is the unit of segregation, the assortment of two gene pairs will occur at random.

As we now understand it, Mendel's concept of the gene was all but Newtonian. He saw genes as small immutable particles whose movement was controlled by natural law in such a way that it could be modeled statistically. We can describe Mendel's concept of the gene in the three laws. The first law, which we will call the Purity and Constancy of the Gene, states that genes themselves are immutable; although genes produce the phenotype, they are not themselves the phenotype, nor are they affected by the phenotype. The second law, the Law of the Gene, states that an individual carries two, and only two, copies of a gene for a given trait. This Law of the Gene also states that each time a gamete (a sperm or an egg) is made, one of those two copies is chosen at random to be included in that gamete. The third law, the Law of Independent Assortment, mandates that for two pairs of gene, the choice of which copy of gene A is included in the gamete does not affect the choice of which copy of gene B was included in the same gamete.

The only real error in Mendel's analysis resulted from his lack of understanding of the gene itself. Indeed, the first major advance in genetic thinking would focus on correcting Mendel's view of the gene as an immutable object, his idea of the purity and constancy of a gene. Genes might be very constant, and indeed they are, but they would not turn out to be immutable. Mutations in many, but perhaps not all, genes can produce some phenotypic effects. Moreover, the modern geneticist is savvy enough to know that most biological processes involve the products of multiple genes. For many traits, there may be multiple genes that can mutate to a given phenotype. We are also aware that for some fraction of genes, the connection between genotype and phenotype may be influenced both by other genes and by the environment.

We have come a long way since Gregor Mendel. We have a much clearer view of the gene. Along the way, we have developed some very impressive tools to study gene function. These are the focus of this book. But before we embark on our discussion, perhaps we should ask ourselves, "Why should we bother with doing *genetics* at all?" What can we obtain from the isolation and analysis of mutants (for that is what *genetics* is) that we cannot learn by one of the "-omics" of modern biology?

We will define genetics simply: *Genetics is both the use of mutations and mutational analysis to study a given biological process and the study of the hereditary process itself*. When done right, the two halves of that description are inextricable. If someone is isolating mutants and characterizing those mutants to study flight, they are *doing genetics*. If they are simply isolating genes expressed in bird muscle, they may be doing biochemistry – but they are not doing genetics. The very core of *genetics* is *mutation*. However, the actual, doing of genetics requires more than isolating mutations. Doing genetics well also requires that investigators isolate and characterize those mutants in a fashion that (i) maximizes their chance of answering their initial questions; (ii) provides them with as many novel biological insights as possible; and (iii) facilitates a greater understanding of the structure and function of the genome they are studying? In other words, *doing genetics well* means understanding what types of mutations one can get, how to get them, and how to analyze them. The analysis of suppressor mutants, for example, can be a powerful tool indeed when done correctly.

The proper *doing of genetics* requires that a scientist understand their tools. The basic intellectual tools of genetics are: mutation, complementation, recombination, suppression, and regulation (epistasis). This book is about the proper use of those intellectual tools. Our goal is to give you ideas of what works and cautions as to what doesn't. We will discuss the biology and biochemistry of very processes, but only when we need to do so to describe the mutants. The very essence of our story is: the mutants. So, that is where we begin . . . .

1

Mutation

A **mutation** is a stable and potentially heritable change in a DNA sequence. Mutations may occur in the soma of an organism, affecting only a particular cell or lineage of cells, or they may occur in the germline of an organism and be passed to all of that organism's offspring. Mutations that occur within or near a gene may create a phenotype different from that normally expressed by the wildtype allele of that gene. A number of different types of mutations have been found to cause changes in phenotype. These mutations can be changes in individual base pairs, such as substitutions (e.g. C → T), insertions or deletions of DNA, or they may be chromosomal aberrations such as inversions, translocations, or copy number variants near or within genes. Because this book is fundamentally about mutational analysis, we need to spend some time considering the types of mutations that can occur, both at the molecular level and in terms of the effects they can have. We also need to review the various systems that exist to classify mutations. Such a review is all the more critical because the nomenclature systems that geneticists have developed are keyed to the structure and effects of the mutants they name. Thus, the things themselves – and our names for them – are inextricably intertwined.

1.1 Types of Mutations

Most introductory genetics texts classify mutations simply as recessive or dominant. A mutation (*m1*) is said to be **recessive** if *m1/m1* organisms display a mutant phenotype, but *m1/+* organisms are wildtype. (Note: the symbol "+" denotes the wildtype, or normal, allele of a given gene.) Conversely, a mutation (*M2*) is said to be **dominant** if *M2/+* organisms display a mutant phenotype while +/+ organisms are normal. Some texts use the term **semidominant** to describe cases where a dominant mutation, *M3*, displays a more severe (or extreme) phenotype as a homozygote (*M3/M3*) than it does as a heterozygote (*M3/+*), such that the order of phenotypic severity is *M3/M3* > *M3/+* > +/+. Although such a classification is sufficient for some purposes, it is inadequate to describe the range of mutant types, or phenotypes, that can actually be observed. Accordingly, at least three more-detailed classification systems have been developed and are discussed in this section.[1]

1 Intertwined with the issue of mutant classification is the problem of genetic nomenclature. Unfortunately, each organism uses a different system to symbolize gene names. Appendix A describes the model organisms that we discuss often in this text. Brief summaries of their nomenclature systems and references to more detailed nomenclature guides are provided there. A thorough and comparative summary of the existing nomenclature systems for various model organisms was published in a special issue of *Trends in Genetics* in March 1995.

Genetic Theory and Analysis: Finding Meaning in a Genome, Second Edition.
Danny E. Miller, Angela L. Miller, and R. Scott Hawley.

Muller's Classification of Mutants

The first detailed mutant classification scheme was proposed in 1932 by Herman J. Muller (1932), more than two decades before the first statement of the central dogma of molecular biology – that DNA codes for RNA and RNA then codes for protein (Crick 1958, 1970). Muller classified mutants into five basic groups: nullomorphs, hypomorphs, hypermorphs, antimorphs, and neomorphs. The assignment of a mutant to one of these classes was largely based on Muller's view that mutations can, and should, be described in terms of their effect on activity. A mutation can be assigned to one of these five groups by comparing the phenotypic effects of that mutation in homo-, hetero-, and hemizygotes. (A **hemizygote** is an individual with a single allele at a specific position, instead of two. For example, human males, who are XY, are hemizygous for most X chromosome genes.) Understanding these classifications, and being able to use them, is a critical component of genetic analysis. We will therefore consider each of these types of mutations in some detail. We begin by considering the two classes of loss-of-function mutations: nullomorphs and hypomorphs.

Nullomorphs

Also known as **amorphs**, **nullomorphs** are mutants with no remaining gene function – they produce no functional gene product but may still create part or all of the protein that the gene encodes. They are often, and far more precisely, called **null alleles**, and they are the basic mainstay of genetic analysis. Nullomorphic mutations might correspond to internal deletions, frameshift mutations leading to a premature stop codon, or missense mutations that alter a critical site in the protein in such a way as to fully ablate its activity (see Box 1.1). The most characteristic feature of

Box 1.1 DNA-Level Terminology

While DNA-level terminology is covered in more depth in Section 1.3, many of the terms are useful when discussing both early and modern mutant terminology. Here is a brief overview of this DNA-level vocabulary.

- **Single-nucleotide variant (SNV):** replacement of one nucleotide base with another; also referred to as a substitution mutation or point mutation.
- **Missense mutation:** a type of SNV that changes the amino acid encoded by a codon.
- **Indel:** a DNA insertion or deletion of less than 50 base pairs.
- **Frameshift:** an indel that alters the reading frame.
- **Deficiency:** a large deletion that completely removes an entire gene or region of the genome; this term is used frequently in model organism genetics and less so in human genetics.
- **Transposable element (TE):** a segment of DNA capable of moving around within the genome; also referred to as a transposon.
- **Duplication:** a region of the genome that exists in two or more copies. This could be a tandem duplication where the duplicated segments sit next to each other, or the duplicated segment may reside on another chromosome.[2]
- **Inversion:** a section of a chromosome that has been reversed.
- **Translocation:** the transfer of a section of DNA from one chromosome to another.
- **Structural variant (SV):** a DNA variant greater than 50 base pairs in length; includes insertions, deletions, duplications, inversions, repeat expansions, or translocations.
- **Copy number variant (CNV):** a type of SV that changes the number of copies of a coding or noncoding genomic region; includes large duplications and deletions/deficiencies.

2 Triplications are also possible (Liu et al. 2014).

a nullomorph is that it is the equivalent of a full deletion of the gene in terms of its influence on the final phenotype.

Null alleles lead to the complete absence of a functional protein product via a variety of defects in gene expression. Using the relevant molecular tools (some of which are discussed in Section 2.2), one can discriminate between transcriptional nulls, protein (or translational) nulls, and mutations that produce completely inactive proteins. In the case of a **transcriptional null**, no full-length transcript is produced. Such mutations might reflect, for example, the deletion of crucial elements in the promoter or the insertion of a foreign genetic element (e.g. a transposon) in or near the gene. Transcriptional nulls may be identified by RT-PCR, by examining chromatin structure, or by RNA sequencing (RNA-seq) (see Box 1.2), which would show no transcript of your gene of interest. From experience, we can tell you to be careful when calling a mutant a transcriptional null based on the absence of a transcript. Not every gene is expressed at the same time in the same tissue, so make sure you sample from a tissue and time point where and when you know your gene should be expressed. **Protein nulls** are defined as mutations that fail to produce a protein product at all, as assayed by an antibody specific for that protein. (This classification may include transcriptional nulls in cases where the mutant's effect on transcription has not been assessed.) **Inactivating nulls** produce a protein product, but that product exerts no obvious activity. However, the most obvious type of nullomorphic mutation is a **deficiency** (Df), or a deletion of some or all of the DNA that encompasses the gene in question (Figure 1.1).

Box 1.2 Detecting Gene Expression by RNA-seq

Gene expression can be detected by several different techniques, including in situ hybridization, northern blot analysis, RNA sequencing (RNA-seq), and RT-PCR. **In situ hybridization** is a technique that detects specific RNA fragments in tissue using labeled complementary DNA fragments, while **northern blotting** is a way to detect a specific fragment of RNA using a labeled complementary DNA fragment on a gel. Both RT-PCR and RNA-seq differ from in situ hybridization and northern blotting in that they typically begin by converting messenger RNA (mRNA) into double-stranded DNA via reverse transcription. For the purposes of our discussion here, we can differentiate RT-PCR from RNA-seq by saying that **RT-PCR** uses primers to reverse transcribe the RNA from a specific gene of interest into DNA, while **RNA-seq** may target all mRNA transcripts or only some (in the case of poly-A primers) for reverse transcription to DNA. Note that there are now sequencing techniques that allow direct sequencing of mRNA molecules without a reverse transcription step.

RT-PCR is more sensitive than RNA-seq and in the past has been considered more reliable in measuring changes in gene expression between samples (e.g. control compared to treated sample). The advantage of RNA-seq is that you can amplify a lot more RNA at once, and if you have replicates, the result is likely similar to what you would get with RT-PCR. The caveat is that you must be careful when designing your experiment because there are several decisions you must make that will affect the quality of the data you receive. For example, do you want to perform **poly-A selection**? This allows you to select only those transcripts with poly-A tails, which will include most of the RNA that codes for genes, but you will miss things like small RNAs or ncRNA. What about **ribosomal RNA depletion**? Ribosomal RNA will make up a large percentage of your dataset if you don't select against it, but doing so may deplete other RNAs that you are interested in (whether you know it or not). Additionally, selecting specific RNAs using a capture kit can help you focus on a specific set or class of genes, but again, you may end up missing other important things and artificially influencing your measurement of expression.

If you're only looking for transcripts, as you might when studying a new organism, you might simply prepare RNA from a pool of individuals and sequence them. If you're trying to measure the difference in expression that a particular drug treatment induces, then you need to be a bit more careful. In general, you need at least three biological replicates for each class you're studying. So, in the simplest experiment of drug vs. control, you would need six total samples (three treated, and three control) to determine a difference in the expression of a particular gene. Three feels like a scary number since an error in just one of the three samples can skew your results; five samples per class is perhaps a better goal. Even if you do not care much about expression levels and are instead more interested in things like alternative splicing between organisms under certain conditions, we would still argue that you need multiple biological replicates.

Once you've decided on your experimental design, you need to think about how you're going to analyze the data. We cannot stress enough that the even best analyst cannot turn bad data into good results. So, it is of utmost importance that you design your experiment well before you move to the analysis phase. There are hundreds of software tools available to help you analyze your RNA-seq data. Each one has its own benefits and pitfalls. Our best advice is to consult with a bioinformatician *before* you design your experiment. Several excellent papers offer good information on these topics (Adapted from Jiang and Wong 2009; Katz et al. 2010; Wilhelm et al. 2010; Trapnell et al. 2012; Qin et al. 2015).

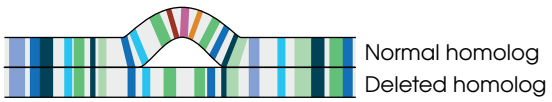

Figure 1.1 Deficiency. A deficiency heterozygote as revealed by polytene chromosome analysis.

According to Muller, most mutations "involve more or less inactivation of the processes governed by the normal gene, and . . . these less active genes should more often act as recessives" (Muller 1932). In the case of genes whose products are enzymes, one can easily understand why most loss-of-function mutations, even nullomorphs, are recessive. For most enzymes in vivo, the reaction rate is limited by substrate concentration, so reducing the concentration of the enzyme by half (as one presumably does in *m*/+ heterozygotes) may be expected to have little effect on the rate of product formation.

However, not all nullomorphic mutations will be recessive. Nullomorphic mutations in genes that code for structural proteins or for enzymes whose function is highly concentration-dependent may well be dominant. The *Minute* mutants of *Drosophila* are excellent examples of nullomorphic mutants that can exert dominant effects. *Minute*/+ flies show delayed development and a variety of morphological anomalies, including short, thin bristles. *Minute*/+ females show reduced fertility or, in extreme cases, are sterile. *Minute/Minute* progeny are inviable. *Minute* mutations are due to loss-of-function mutations in genes that encode ribosomal proteins (Marygold et al. 2007). The reduction of the quantity of even a single ribosomal protein by a factor of two apparently reduces the final number of ribosomes in the cell twofold as well. The consequence of a twofold reduction in the number of ribosomes is a corresponding decrease in protein synthesis capacity, which has dramatic consequences on the phenotype of the organism.

Hypomorphs

Mutants that produce some degree of residual activity, but not enough to show wildtype activity in *m/m* homozygotes, are known as **hypomorphs**. Indeed, the term hypomorph is often synonymous with *weak allele*. Hypomorphic mutations reduce the amount or level of activity of the

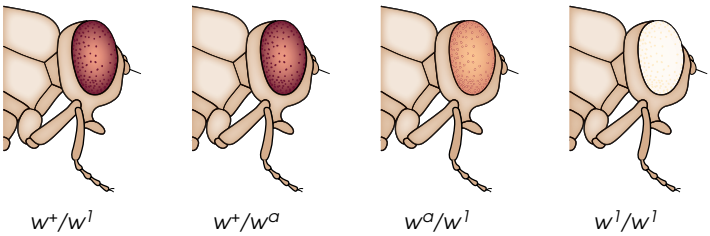

Figure 1.2 Hypomorph. In *Drosophila*, the w^a allele, a hypomorph, produces a diminished quantity of the normal red eye color pigment. Thus, flies heterozygous for w^a and a w^1 allele, which produces no pigment, have orange eyes.

protein product. One can imagine a host of genetic lesions that might produce hypomorphic mutants, ranging from mutations that decrease the level of transcription to mutations that alter messenger stability or the activity or amount of protein product. An excellent example of a hypomorphic mutation is the *white-apricot* (w^a) mutation in *Drosophila* (Figure 1.2). Wildtype (w^+) flies have red eyes, while flies mutant for the *white* gene (w^1) have white eyes. The hypomorphic allele w^a produces some of the protein product needed for red eyes, but not at wildtype levels, so w^a/w^1 flies have orange eyes. This example demonstrates the defining characteristic of a hypomorphic mutant: at least some discernable level of active product is being produced. Most hypomorphic mutants are recessive, but the same caveat about loss-of-function mutations in genes encoding proteins whose dosage is critical applies to partial loss-of-function mutants as well.

One can easily distinguish between hypomorphic and nullomorphic mutants if a deficiency (*Df*) for the gene in question is available. A true nullomorph is the genetic equivalent of a deficiency for a specific gene. Thus, in terms of phenotypic severity, a *m/Df* nullomorph is expected to be equivalent to *m/m*. For a hypomorph, however, *m/Df* is expected to be more severe in phenotype than an *m/m* homozygote. Most of the time the *m/m* homozygote would be expected to produce twice as much active product as the *m/Df* individual. By a similar algebra, an individual with three doses of a hypomorphic mutation (*m/m/m*) is expected to show a less severe phenotype than one with two (*m/m*), whereas adding more doses of a true nullomorphic allele should have no effect.

We've heard many geneticists refer to hypomorphs as the bane of their existence. True enough, the residual level of activity created by such mutations often frustrates the phenotypic or functional analysis of these genes. Nonetheless, hypomorphic mutants are often the first or only mutants to define important genes. It turns out, as we shall discuss in Chapter 2, that many mutant hunts or screens require that homozygotes for the newly induced mutations be viable. Given this restriction, a null allele of a gene required for life would not be recovered in such a screen, even if the protein product of that gene played a critical role in the process under study. In contrast, a hypomorphic mutant can, and often does, produce enough product to allow survival, but not enough to produce a normal phenotype. In such a case, the finding of a hypomorph alerts the investigator to the existence of this gene and heralds its role in the process under study.

Hypermorphs

As the name implies, **hypermorphs** produce either a harmful excess of the normal protein product or a hyperactive one. The defining characteristic of this mutant class is that *m/Df* should be less severe in phenotype than *m/+* because *m/Df* makes less overall protein product than *m/+*. Indeed, in terms of decreasing phenotypic severity, the dosage series should be $m/m > m/+ > m/Df$. Verified examples of hypermorphic mutations are few and far between. The best example of a

hypermorphic mutation in *Drosophila* is a mutant called *Confluens* that affects wing vein morphology. *Confluens* is an allele of *Notch* (*N*). The phenotype of *Confluens* can be mimicked by three doses of *N*+, and *Confluens* over a nullomorphic allele of *Notch* is wildtype. All of this makes perfect sense when one realizes that the *Confluens* mutation is a **tandem duplication** of the *N*+ gene. Clearly, increases in the dose of the *N*+ gene have phenotypic consequences, and however you get to three doses, a phenotype is created. A good example from humans is achondroplasia, a genetic disorder in which individuals have short arms and legs but a normal-sized torso. This is caused by variants in fibroblast growth factor receptor 3 (*FGFR3*) that result in the protein being overactive during development.

Other types of mutations that might upregulate the transcription or translation of a given gene or its mRNA product might also produce observable phenotypes. A technique for creating hypermorphs in *Drosophila* was developed by Pernille Rørth in the late 1990s using **transposable elements** (**transposons**), or DNA segments capable of moving around within the genome. Rørth began by creating a transposon capable of driving the expression of neighboring genes in a fashion that was both tissue-specific and inducible (Rørth et al. 1998). By mobilizing that transposon within the *Drosophila* genome, Rørth and her collaborators created a collection of 2,300 lines of flies, each of which carried an independent transposon insertion. This collection of insertions was screened for the ability to suppress a hypomorphic mutation in the *slow border cells* (*slbo*) gene that confers a cell migration defect. This defect is observed in the *Drosophila* ovary and results in sterility. Because the *slbo* gene encodes a C/EBP transcription factor, Rørth and her collaborators reasoned that the high-expression suppressors "could be genes normally activated by C/EBP in border cells or genes which (in this situation) are rate-limiting for cell migration." They obtained both.

Of 2,082 insertion lines tested, 60 showed clear suppression of the mutant phenotype created by homozygosity for *slbo*. The suppressing insertions resulted in the overexpression of genes that encoded known players in actin cytoskeletal remodeling, a critical process in cell migration. They also recovered insertions in a receptor tyrosine kinase gene (*abl*) that also appears to be involved in the control of actin polymerization. The success of the Rørth suppression screen may have been largely due to the choice of a hypomorphic *slbo* allele that retains some degree of C/EBP activity. Thus, it was only necessary to provide a small increase in border cell migration to suppress the sterility caused by the *slbo* mutation.

If they can be obtained, hypermorphs can be a valuable tool for dissecting a genetic process. This may be especially true in a case where one is dealing with a group of functionally redundant genes. In such a case, a simple loss-of-function mutation in one of these genes may not produce a discernable phenotype but overexpressing one of those genes may create an observable defect. Indeed, in Section 6.7, we will consider the use of high-copy suppression libraries in yeast to mimic the creation of hypomorphic mutations by creating colonies, each of which possesses a high copy number of a plasmid carrying a given gene. Phenotypes created by such methods can also serve as the substrate for enhancer and suppressor screens aimed at identifying other genes in this process.

Antimorphs

An antimorphic mutation results in a protein product that antagonizes, or poisons, the wildtype protein. Thus, the phenotype of a true **antimorph** is expected to mimic the phenotype presented by a strong hypomorph or nullomorph. Antimorphic mutations are dominant by definition. However, increasing the dose of the wildtype allele can sometimes ameliorate the phenotype of an antimorphic mutant. For example, imagine a gene (*muct*) that encodes the protein subunit Muct,

four identical copies of which are required to produce the enzyme muctinase, which is in turn required to synthesize the imaginary substance muctin. Nullomorphic mutations in *muct* should produce a muctin$^-$ phenotype when homozygous, but assuming that half the level of the enzyme is enough, loss-of-function alleles of *muct* will be recessive and *+/muct* heterozygotes should be normal. But what if a mutant allele of *muct* produces an unusual structural variant of Muct, which is incorporated into the polyprotein complex in such a way as to render the entire complex inactive? Assuming the wildtype and mutant Muct subunits are produced with equal abundance, this variant of Muct will inactivate virtually all (15/16) of the muctinase complexes, and the remaining activity may simply not be enough. However, by increasing the dose of the wildtype allele to two, approximately 20% of the muctinase complexes will be composed of normal subunits.

Sickle cell anemia is a type of antimorphic mutation. Hemoglobin A is a tetramer usually formed by two β-globin and two α-globin chains. This is the typical adult configuration. Every adult also has a low level of Hemoglobin A2, formed by two α and two δ chains, as well as Hemoglobin F, with two α chains and two fetal γ chains. A point mutation in the β-globin gene causes the hemoglobin tetramer to become structurally unstable, especially when there is no oxygen bound to the molecule. It is antimorphic in the sense that one mutant β-globin chain disrupts the entire tetramer even if the other β-globin chain is wildtype. Some treatment modalities focus on upregulating the expression of fetal hemoglobin in adults to compensate for the loss of the β-globin chain.

Another example of an antimorph is presented in Figure 1.3. The microtubules that make possible many processes of cellular movement are composed of long arrays of tubulin monomers. Each of these monomers is composed of an α-tubulin and a β-tubulin subunit. Imagine a mutant in the α-tubulin gene that produces a variant subunit that can incorporate into a growing chain but cannot support further growth. Once a mutant subunit is incorporated, chain growth freezes. However, by increasing the dosage of the normal allele, one decreases the probability of incorporating a mutant subunit from 1/2 to 1/3. One *might* see some phenotypic amelioration in such a case. Indeed, a dominant mutant allele ($\beta 2t^D$) of a testis-specific β-tubulin gene has exactly this type of effect on microtubule assembly in the *Drosophila* male germline. Both heterozygotes and homozygotes show dramatic defects in the formation of large microtubule assemblies in the testis and are sterile as males. But heterozygotes carrying an extra dose of the wildtype allele ($\beta 2t^D$/+/+) are weakly fertile (Kemphues et al. 1980, 1983).

As noted here, the vast majority of antimorphs are dominant. If we use the symbol *A* to denote the mutant, the relative phenotypic severity observed in different genotypes can be described as follows:

$$A/A \geq A/Df \geq A/+ >>> +/Df \geq +/+$$

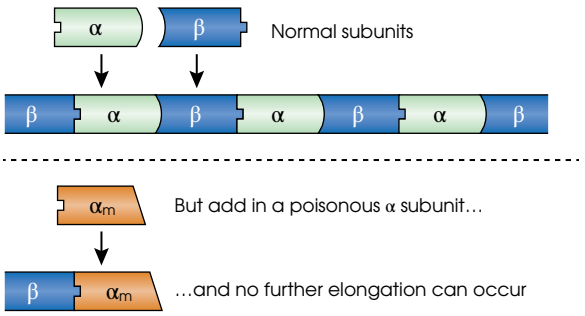

Figure 1.3 Antimorph. Incorporation of the product of an antimorphic allele of tubulin impedes further growth on the microtubule fiber.

The defining characteristic of an antimorph is that one should be able to revert (or, more precisely, pseudorevert) an antimorphic mutation to a nullomorphic mutation of the same gene. In other words, the easiest way to stop this allele from producing a poisonous product is simply to inactivate the gene by a second intragenic mutation (i.e. any mutation that blocks the production of the poisonous protein), thus creating a **pseudorevertant**. One would do this by mutagenizing *A/A* individuals and screening *A*/+* offspring (where *A** indicates a mutagenized chromosome) for revertants. The result should be a new loss-of-function, preferably nullomorphic, allele (denoted r^A). Most critically, the following should be true:

- $+/r^A$ individuals should be phenotypically similar to *+/Df* or +/+ individuals (i.e. the original antimorph should be reverted), and
- r^A/r^A individuals should be phenotypically similar to *A/A*, *A/Df*, or *A/+* individuals

Indeed, the $\beta 2t^D$ allele generated by Kemphues can be reverted to create recessive loss-of-function alleles of the $\beta 2t^D$ gene that follow the rules above (Kemphues et al. 1983).

As a second example of the processes of mutating an antimorph into a nullomorph, consider the *Drosophila* gene *nod*. The *nod* gene is required for female meiosis in *Drosophila*, but it is also expressed in virtually all mitotically dividing cells (Zhang et al. 1990). There are many recessive loss-of-function *nod* alleles that disrupt meiotic chromosome segregation when homozygous. Curiously, none of these mutations has any demonstrable effect on mitotic cell division or mitotic cells. There is, however, one dominant allele called nod^{DTW}. Heterozygous $nod^{DTW}/+$ females show the same defect in chromosome segregation as do females homozygous for complete loss-of-function *nod* mutations (Rasooly et al. 1991).

The normal function of the Nod protein is to stabilize chromosomes along microtubule tracks. The nod^{DTW} mutation, which alters only a single amino acid in a critical region of the protein, poisons that process, and appears to lock the chromosomes in place. Rasooly and colleagues mutagenized males carrying the nod^{DTW} mutation on their X chromosomes and screened for pseudorevertants (mutated X chromosomes that no longer exhibited the dominant meiotic effect). They recovered four such mutants, all of which turned out to be new nullomorphic and fully recessive alleles of *nod*. When sequenced, each of these new mutants carried the original nod^{DTW} mutation as well as a second mutation that inactivated the *nod* gene.

Neomorphs

It is not clear exactly what Muller intended by this term. But over the years, **neomorph** has come to mean a mutation that causes a gene to be active in an abnormal time or place. One example is a translocation event that occurs in humans and results in a blood cancer known as Burkitt's lymphoma. As a result of a translocation between chromosomes 8 and 14, the coding sequence of the *c-myc* gene on chromosome 8, which acts to promote cell division, now lies downstream from a very powerful set of lymphocyte-specific promoter elements derived from a gene on chromosome 14. After translocation, these promoter elements inappropriately turn on the *c-myc* gene in white blood cells, resulting in uncontrolled cellular proliferation. Many, if not most, human cancers are the result of neomorphic mutations.

Alternatively, one can consider a dominant mutation in *Drosophila*, $Antp^{73b}$, that causes the antennae to be replaced by legs. (Yes, there really are two extra legs sticking out of the head just above the eyes.) The *Antennapedia* (*Antp*) gene is not normally expressed in the head, but rather in the thorax of the developing embryo where it plays a critical role in specifying the development of thoracic structures such as the leg (Struhl 1981). As diagrammed in Figure 1.4, the mutation results from an **inversion** that fuses the 3′ coding sequences of the normal *Antp* gene next to the

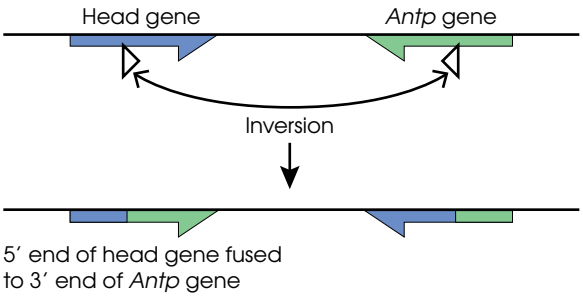

Figure 1.4 Neomorph. *Antp*^{73b} results from an inversion that fuses the 3′ coding sequences of the *Antp* gene with the 5′ coding sequences and regulatory sequences of a gene normally expressed in the head.

5′ coding region of a gene normally expressed in the head (Frischer et al. 1986). As a result of this inversion, a significant portion of the Antp protein – enough to specify leg development – is expressed in the antennae primordia of the developing head. The result is legs where there should be antennae.

Finally, a lovely set of studies by Ganetzky and colleagues has served to elucidate the identity and function of one of the most fascinating of all neomorphic mutants, the *Segregation Distorter* (*SD*) chromosome in *Drosophila* (Kusano et al. 2001). In the *Drosophila* male germline, the *SD* chromosome exhibits a process referred to as **meiotic drive** – when heterozygous with a normal chromosome, *SD* chromosomes have the endearing habit of destroying their wildtype homologs (denoted *SD*$^{+}$). 99% of the sperm produced by *SD/SD*$^{+}$ heterozygotes will carry the *SD* chromosome and less than 1% carry the wildtype *SD*$^{+}$ homolog. The *SD* chromosome does this by causing improper chromosome condensation in *SD*$^{+}$ sperm.

An *SD* chromosome is composed of several genetic units that contribute to its function. The first of these is the *Sd* mutant itself, which acts at a separate site on the chromosome called *Responder* (*Rsp*) to cause spermatid dysfunction (Ganetzky 1977). *Rsp* itself is composed of a repetitive element located in the **centric heterochromatin** (heterochromatin near the centromere) of the second chromosome (Wu et al. 1988). The sensitivity of a given chromosome to destruction by an active *Sd* element increases with the number of copies of the *Rsp* repeat. The *Sd* mutation is the result of a small tandem duplication event involving the *RanGAP* gene (Powers and Ganetzky 1991) that results in a mutant RanGAP protein, truncated by 234 amino acids at the C terminus (Merrill et al. 1999). This truncation creates a novel protein whose expression causes distortion – only the presence of this novel form of the protein gives rise to meiotic drive. Loss-of-function mutations in the wildtype (*SD*$^{+}$) *RanGAP* gene are not expected to create an *SD* phenotype.

There is an important lesson here. Early papers on *SD* often focused their efforts on the idea that understanding the mechanism by which *SD*-induced distortion might provide critical insights into the meiotic mechanism itself. The reality is that the actual function of the *RanGAP* gene is unrelated to the mechanism of maintaining Mendelian fairness. Rather it is a function required to mediate nuclear transport in many, if not most, cell types. But it was this mechanism that lent itself to exploitation in a wonderfully devious way. H.J. Muller was clearly prescient in putting forward the idea of a neomorph, a mutation that creates a novel phenotype unrelated to the usual function of the gene.

One major distinction between antimorphs and neomorphs is that the neomorph is not poisoning the normal function of the unmutated gene, but is instead performing the correct function at the wrong time or place. Operationally, this distinction can be made by attempting to create pseudorevertants of the dominant mutation. Reversion of the *Antp*^{73b} neomorph often results in a complete null mutation in the *Antp* gene. The phenotype of that mutation is early embryonic

lethality due to malformations of the thorax. Compare this to the *nod* antimorph described earlier, where the original antimorph and its loss-of-function revertants had the same phenotype. Neomorphs also differ from antimorphs in that the effects of neomorphic alleles are independent of the presence or dosage of the wildtype allele. Thus, by definition, neomorphic alleles are dominant.

In recent years, the term neomorph has gradually been supplanted by the term gain-of-function mutation. Actually, both terms can be somewhat misleading because in most cases the mutant does not confer a truly new or gained function on the gene, but simply causes the gene to be expressed at the wrong time, in the wrong place, or at a higher rate than wildtype.[3] Nonetheless, it seems odd to us that we cannot cite even a single example that fits the true definition of a neomorph: a mutation that creates a protein with a novel function.

Modern Mutant Terminology

Following the elucidation of the central dogma of molecular biology (DNA → RNA → Protein), some geneticists felt that Muller's classically based system was too awkward to describe the alterations in gene function created by mutants at the molecular level. Thus, we must discuss terminology that is frequently used by geneticists today.

Loss-of-Function Mutants

In modern articles, mutants that reduce the level of gene product are often classified only as **loss-of-function mutants**, but this term often lumps together hypomorphic and nullomorphic mutants. More precisely, loss-of-function mutants can be further characterized as null mutants, partial loss-of-function mutants, or conditional loss-of-function mutants. **Null mutants** are complete loss-of-function mutants. Indeed, in modern parlance, the term *null mutant* is reserved for those cases where the molecular biology of the mutant gene is well enough understood to be confident that there is no functional gene product produced, and often the definition of this term is narrowed further to imply that no product is produced at all. In **partial loss-of-function mutants**, the mutant produces some degree of product activity – or product itself. Partial loss-of-function mutants include both those mutants that impair the level of product formation and those that create a partially functional product. The term **conditional loss-of-function mutant** refers to cases where the loss of activity of the gene protein is observed under one set of conditions (e.g. higher or lower temperature or treatment with a particular drug) and not under another. The canonical case of a **temperature-sensitive mutant** involves the denaturation of the mutant protein at higher temperatures. This is not, however, always the case. In some circumstances, even null mutants can be sensitive to environmental cues and produce a phenotype only in the presence of environmental or genetic stress.

Dominant Mutants

Dominant mutants are referred to by several different names, some of which seem more or less useful than their Mullerian counterparts. For example, mutants that produce poisonous products are referred to as **dominant negative** mutants. This is a well-used and often accurate term, but we fail to see its advantage over "antimorph."

---

3 For an interesting example of increased expression of a gene leading to drug resistance in cancer, see Wang et al. (2004).

Gain-of-Function Mutants

Mutants that cause a gene to be inappropriately expressed during development or that cause a gene product to be inappropriately regulated are often lumped together under the term **gain-of-function mutant**. While this term is in general use, it seems to us preferable to use the term **heterochronic mutant** to describe mutants that cause genes to be expressed at the wrong *times*. We can then reserve the term gain-of-function mutant for mutants that, for example, remove the regulatory element from some signal transduction protein and thus lock that protein in the ON state or that, by combining components from more than one gene, allow the creation of a product with novel specificities for binding partners or active sites in the cell.

Separation-of-Function Mutants

Some genes perform two functions – say, converting substrate B into C as well as facilitating the interaction of proteins X and Y. If you can recover a mutant in this gene that only converts substrate B into C but does not facilitate the interaction of proteins X and Y, then you have recovered a **separation-of-function mutant**. Separation-of-function mutants allow you to separate genes into their functional domains. How do you know if your gene is performing multiple functions? One hint may be that your gene seems to be similar to two separate genes in another closely related organism. With this knowledge, you may then decide to delete specific portions of your gene or introduce specific point mutations to gain a deeper understanding of how each region functions. Separation-of-function mutants thus allow you to make claims about how specific regions of a gene's protein product operate in your organism. Keep in mind, however, that what appears to be a separation-of-function mutant might simply be a hypomorph whose level of expression lies between two phenotypic thresholds such that there is enough protein to perform task A but not enough to also perform task B.

DNA-Level Terminology

The ultimate classification scheme for mutations requires the DNA sequence of the wildtype and mutant alleles themselves. The following are the basic classes of mutations at the DNA level.

Base-Pair-Substitution Mutants

Often called **point mutants**, these mutants result from the change of one base in the sequence to another (Figure 1.5). Changes such as A–T to G–C or C–G to T–A that replace a like nucleotide (purine or pyrimidine) on each strand are referred to as **transitions**. Changes such as A–T to C–G, in which a purine on one strand is replaced with a pyrimidine on the same strand (or vice versa), are referred to as **transversions**.

Missense mutants are a class of base-pair-substitution mutants that change the sequence of a given codon, which then directs the incorporation of an amino acid different from the one specified at the same position in the wildtype allele. They can be either conservative or nonconservative. A **nonconservative mutation** changes an amino acid to one with different properties. A **conservative mutation** results in a change that incorporates a chemically or structurally similar amino acid, which makes it less likely to disrupt the function of a protein than a nonconservative missense mutation.

Nonsense mutants are a class of base-pair substitution mutants that alter a given codon to create one of the three stop codons, UAA, UAG, or UGA. Geneticists old enough to remember the moon landing may sometimes refer to these mutants as *ochre* (UAA), *amber* (UAG), and *opal* (UGA).

Silent substitutions (or **silent mutations**) are either mutants in coding sequence that do not change the amino acid directed by that codon, as is often true for third-base substitutions, or

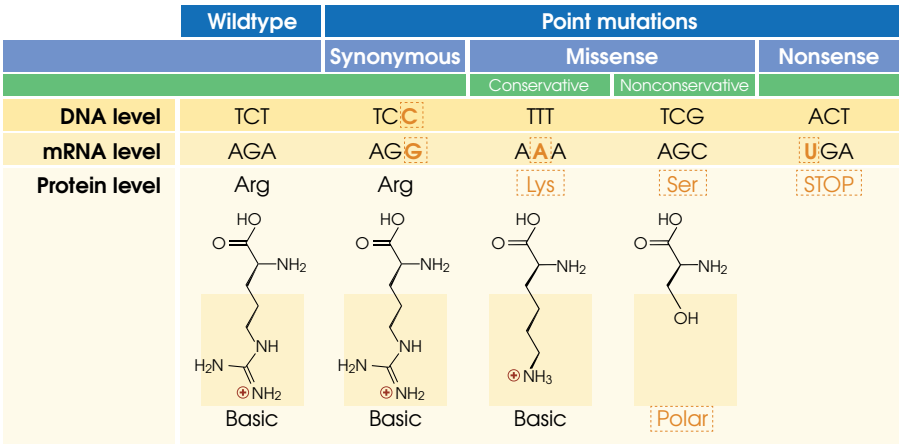

Figure 1.5 Base-pair-substitution mutations. A synonymous mutation is a type of silent substitution that does not change the amino acid expressed by a codon. A conservative missense mutation changes the expressed amino acid to one with a similar structure or functionality as the original. A nonconservative missense mutation results in a different amino acid with different properties than the wildtype amino acid. A nonsense mutation results in a stop codon.

changes in noncoding regions that do not affect gene expression, such as changes in an intron that do not affect splicing or gene function. This class of mutants has taken on real value in providing markers such as **restriction fragment length polymorphisms** (RFLPs) and other single-nucleotide variants for human genetic mapping. Also note that conservative missense mutations might be phenotypically silent if the substitution involves a chemically similar amino acid. (Do not forget that silent mutations can create new splice acceptor or donor sites within a gene, or they might alter an enhancer for your gene of interest or a neighbor – so do not become too comfortable ignoring them.)

Base-Pair Insertions or Deletions

The name says it all. **Frameshift mutants** result from the insertion or deletion of one or more base pairs within the coding sequence that alters the reading frame, typically soon after the insertion or deletion. Because of their tendency to result in premature stop codons, frameshift mutants are often also classified as nullomorphs. Larger deletions are also sometimes referred to as **deficiencies**.

Chromosomal Aberrations

We should also note that a series of terms used to describe molecular events have been borrowed from the terminology used to describe chromosome rearrangement. The meanings of most of these terms, such as **inversion**, **duplication**, and **deficiency** are self-evident (see Box 1.1). The term **translocation** refers to a breakage and rejoining event involving sequences on nonhomologous chromosomes.

The changes we just described can have multiple effects on the function of genes, or they may have no effect at all. For example, a single base change at a splice site might ablate that site, causing the exon to be spliced out as if it were intronic sequence. This kind of change clearly affects the resulting protein structure. Single changes may also ablate start codons or occur close enough to the gene (in the 5′ untranslated region, for example) to change the level of expression of that gene.

1.2 Dominance and Recessivity

We use the term dominant to mean that if *A/a* individuals or cells are phenotypically similar or identical to *A/A* cells or individuals, while the *a/a* genotype confers a different phenotype, then *A* is the dominant allele and *a* is the recessive allele. If *A/a* individuals or cells are phenotypically intermediate between their *A/A* and *a/a* counterparts, then *A* and *a* are said to be **semidominant**. If, on the other hand, *A/a* individuals exhibit the phenotypes of both *A/a* and *a/a* individuals, then *A* and *a* are **codominant**. A simple example for understanding the difference between codominance and semidominance is flower color. If a flower has alleles for both red and white color but produces pink flowers, the alleles are semidominant; if both red and white stripes are produced, the alleles are codominant. These are seemingly straightforward terms that all of us learned in high school biology. Unfortunately, their usage is often rather careless. You need to think about the situation in which the term is being applied.

To a certain extent, the terms dominant and recessive are simply matters of perspective. Suppose we look at a mutant in a human that encodes an essential metabolic enzyme. Heterozygosity for a simple loss-of-function mutant in that enzyme is likely to have little effect on the metabolism in most cell types. Thus, heterozygotes for this mutation are likely to be normal and we would classify this mutation as fully recessive. But, if we refocused our interest only on the amount of active enzyme produced by a given cell type, then our loss-of-function mutant might be codominant. Indeed, in the absence of cellular controls that limit the level of enzyme production, one might expect that virtually all mutants could be shown to be codominant at the molecular level. Thus, our terminology only has meaning when put into proper perspective.

A dramatic example of the importance of perspective can be seen in the human hereditary cancer disorder retinoblastoma. Children who inherit one defective copy of the *Retinoblastoma1* (*RB1*) gene classically develop retinal tumors very early in life and have a high rate of other childhood tumors, such as osteosarcoma. The inherited form of retinoblastoma is a simple autosomal dominant ($RB1^-$) and the disorder behaves in a pedigree as a dominant mutation should. But in $RB1^+/RB1^-$ individuals, only 10 or so cells in the retina of each eye form tumors. These cells only become tumorous because they have, by subsequent somatic loss or mutation, lost the normal $RB1^+$ allele and unmasked the inherited $RB1^-$ allele. The key to this example is that at the cellular level, the defective gene, $RB1^-$, is fully recessive – one wildtype copy of *RB1* is enough for normal function. The $RB1^-$ mutation only induces tumor formation when the wildtype allele is removed. The requisite somatic mutation/loss events are rare, but because there are hundreds of millions of retinal cells in each eye, the somatic mutation events required to unmask the $RB1^-$ allele become virtually certain to occur somewhere in the eye (see Box 1.3 and Chapter 8). So, are $RB1^-$ mutants dominant or recessive? The answer depends on your perspective – in pedigrees they are dominant; in cells they are recessive.

Because it is exactly one's perspective that matters here, we will begin our discussion of dominance and recessivity at the level of the individual cell.

The Cellular Meaning of Dominance

The critical point is that dominance will result when one of three things occur:

1) a single copy of the wildtype gene is insufficient, or
2) the product of the mutant gene is poisonous to the process, or
3) the mutation causes the gene to be incorrectly expressed in a way that creates a phenotype.

Box 1.3 De Novo Mutation

Every cell division is an opportunity for a new error or a **de novo mutation**. Because of this, every multicellular organism is a **mosaic**, where the genome of every cell differs slightly from the original single-celled zygote that the organism arose from (Biesecker and Spinner 2013; Lupski 2013). Examples of visible mosaics in humans can be seen in Figure B1.1, and a detailed discussion of mosaicism is found in Chapter 8: Mosaic Analysis.

The simplest error encountered during cell division is a **single nucleotide variant (SNV)** or point mutation. Each human cell contains about 6.4×10^9 nucleotides (remember, most of your cells are diploid), and the rate of de novo single-nucleotide changes is approximately 5×10^{-11} mistakes per base copied (Drake et al. 1998).[4] This tells you that every time a cell divides, 0.32 new mutations are created, or about one new mutation every three divisions. Fascinating work by Gilissen et al. (2014) shows that de novo SNVs cause many more genetic diseases than previously thought. Studies that carefully compare individuals with a particular disease to their parents are showing us that de novo mutation should be considered in any individual who presents with a genetic disorder and a family history that doesn't seem to suggest a cause (Frank 2014; Acuna-Hidalgo et al. 2015).

The second type of de novo change you may encounter is an insertion or deletion (or **indel**) error. It was previously somewhat difficult to estimate the rate of de novo indel formation, but the falling cost of whole-genome sequencing and the advent of single-cell sequencing exposed a picture of more frequent indel formation than expected (McConnell et al. 2013; Glessner et al. 2014; Hannibal et al. 2014). A third type of error, chromosome segregation errors, are less common than either SNVs or indels simply because they are extreme events that are often not tolerated by a cell.

Individuals can also be germline mosaics due to mutations that occur during **gametogenesis**, or the creation of sperm and eggs. Germline mosaic individuals may have offspring that appear to "inherit" a mutation that the parent doesn't seem to carry. Consider human

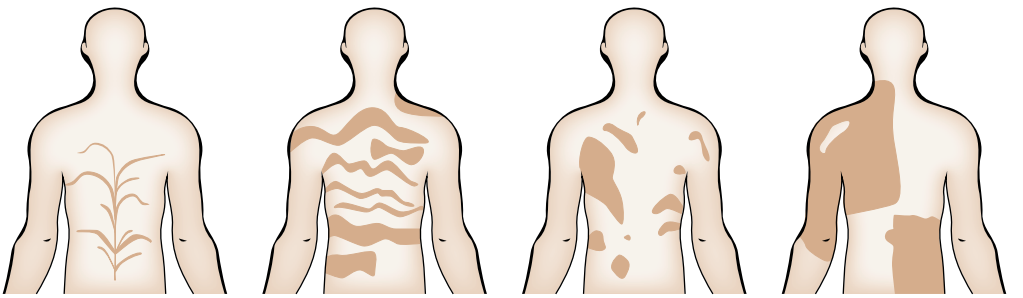

Figure B1.1 Mosaicism. The error inherent in each cell division means that every multicellular organism is a mosaic at some level, with closely related cells being different at a few nucleotides. These minor genetic differences can sometimes be visualized. For example, a variety of genetic disorders can cause patterns of skin discoloration in humans, with cells carrying the mutated gene expressing more or less pigment than normal cells. This visual presentation occurs because of the migratory nature of precursor skin epithelial cells.

4 A fantastic way to find numbers such as this and the references for them is the BioNumbers website at http://bionumbers.hms.harvard.edu.