Chapter 9

Identifying Genes and Proteins of Interest

After reading this chapter, you should be able to:

- Understand basic concepts in genetics, such as transcription and translation
- Compare and contrast the strengths and limitations of commonly used genetic model organisms
- Describe methods for identifying genes and proteins important to a biological phenotype, including genetic, in silico, and molecular screening techniques

Techniques covered:

- **Genetic screens**: forward and reverse genetic screens
- **In silico screens**: BLAST, Ensembl
- **Molecular screens**: RNA sequencing, CRISPR screens

Virtually every cell in your body contains a copy of your entire genome, the complete set of genetic information embedded in DNA that is necessary for cellular development, metabolism, homeostasis, and plasticity. The functional unit of the genome is the **gene**, a segment of DNA that codes for a functional protein product. **Proteins** are the molecular machines responsible for virtually all of the cell's structural and functional properties. To understand how the brain works, a neuroscientist must understand the anatomy and physiology of neurons—and to understand the anatomy and physiology of neurons, a neuroscientist must understand the genes and proteins that give rise to their properties.

One of the overarching questions in neuroscience is how an animal's **genotype**, the genetic constitution of an animal, determines that animal's **phenotype**, an observable trait or set of traits. In other words, what genes influence a trait, such as proper development of an axon toward its target, or appropriate behavioral response following a stimulus? Each month, scientists publish a variety of studies either identifying a novel gene and demonstrating its important role in an animal's phenotype, or showing that a previously discovered gene also contributes to a separate phenotype. How do scientists identify genes that function as the molecular substrates of behavior?

The purpose of this chapter is to describe methods useful for identifying which genes or proteins are important for a specific biological phenotype.

Guide to Research Techniques in Neuroscience. https://doi.org/10.1016/B978-0-12-818646-6.00009-9

These methods usually involve screening thousands of genes to identify just a few that contribute to a neurobiological phenomenon. In previous chapters, we examined techniques that generally study whole brains, behaviors, or the activity of cells. Starting with this chapter, we begin to study genetic and molecular neuroscience—the contribution of genes and molecules to development, physiology, and behavior. Therefore, we start the chapter with a brief introduction to studying genes and proteins, as well as a survey of commonly used genetic model organisms. The second half of the chapter describes various methods for identifying genes or proteins of interest: using genetic screens, in silico screens, and molecular screens. Once these genes are identified, scientists can perform other experiments to further understand their contribution to a phenotype, as described in later chapters. Scientists can also use knowledge about the genes that specific neurons express to design genetically modified organisms for targeting and manipulating neural populations of interest (Chapter 12).

HOW GENES ENCODE FOR PROTEINS

The number of genetic and molecular techniques used in the literature can often be intimidating to neuroscientists without formal training in genetics or molecular biology. Fortunately, the background information necessary to understand these techniques and why they are used is not complex, and can be summarized in terms of the flow of information within the cell. Highly detailed descriptions of this information flow can be found in other textbooks. Here, we provide the essential information that a neuroscientist needs to know to understand genetic and molecular neuroscience, the subject of the remaining chapters in this book.

The Central Dogma of Molecular Biology

The **central dogma of molecular biology** is a model of the flow of information within a cell. Described by Francis Crick in 1958, this model posits that information in a nucleic acid, encoded by the sequence of specific DNA or RNA bases, can be transferred to another nucleic acid or to a protein, but information in a protein, encoded by the sequence of specific amino acids, is not transferred back to another protein or to a nucleic acid. Many of us learned the simplest version of the central dogma suggesting a unidirectional flow of information—specific sequences of DNA code for molecules of RNA—and these RNA molecules, in turn, code for proteins (Fig. 9.1). This model lays the foundation for all of molecular biology research. While scientists have elaborated on this model over the past 40 years and found more complex situations that contradict the simple version of the central dogma, such as the discovery of microRNAs and other RNA-based methods of DNA regulation, the fundamental concept continues to serve as the framework from which all molecular neuroscientists design and execute experiments.

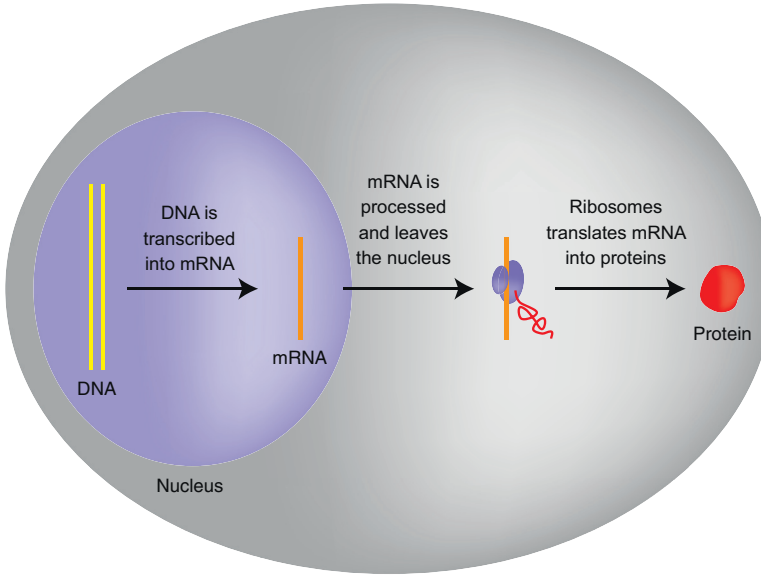

FIGURE 9.1 The flow of information within a cell. DNA resides in the nucleus and codes for relatively short sequences of mRNA. After transcription, the mRNA is processed and leaves the nucleus for the cytoplasm, where it is translated by ribosomes into proteins.

DNA

DNA is the ultimate source of all genetic information within a cell. Residing within the cell nucleus, a DNA molecule is a long polymer made up of repeating units called nucleotides (Fig. 9.2). These nucleotides are composed of a phosphate group, a sugar molecule (2-deoxyribose—what gives DNA its name), and a nitrogenous base. The phosphate and sugar molecules form the backbone of a DNA polymer. There are four types of nitrogenous bases found in DNA that define the properties of the nucleotide: adenine (abbreviated A), thymine (T), guanine (G), and cytosine (C).

In all eukaryotic organisms, DNA exists as a tightly associated pair of two long strands that intertwine, like two handrails of a spiral staircase, forming the shape of a **double helix**. The two strands of DNA are stabilized by hydrogen bonds between the nitrogenous bases attached to the two strands. A base on one strand forms a bond with only one type of base on the opposite strand: A only forms bonds with T, and C only forms bonds with G. The arrangement of two nucleotides binding together across the double helix is referred to as a **base pair**. Strands of DNA that form matches among base pairs are called **complementary strands**. Hydrogen bonds, unlike covalent bonds, are relatively weak and can be broken and rejoined relatively easily. This allows other enzymes to unwind DNA during DNA replication or gene transcription.

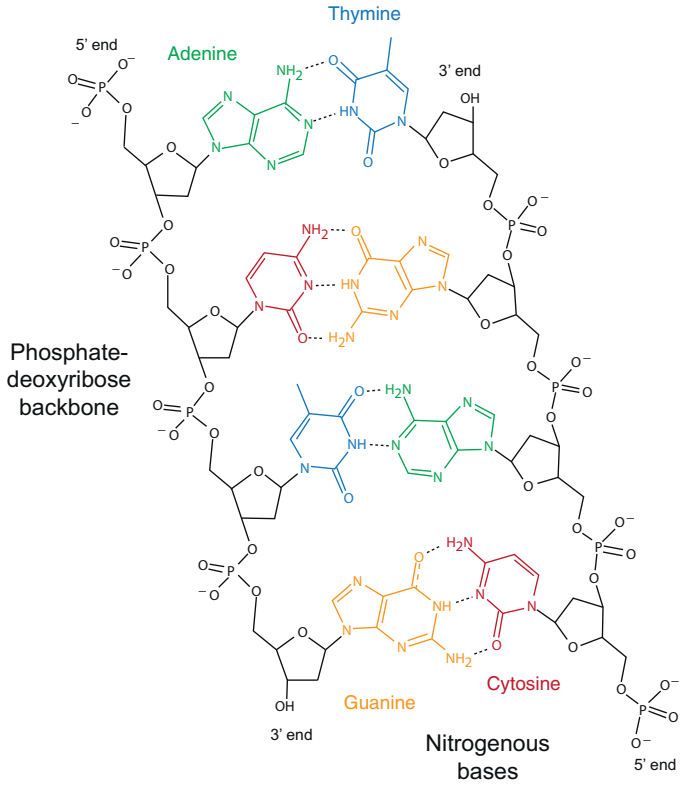

FIGURE 9.2 The molecular structure of DNA. DNA is composed of a sugar–phosphate backbone on the outside and nitrogenous bases on the inside. The four nitrogenous bases form hydrogen bonds in the interior of the DNA molecule and hold the two complementary strands together. The two strands are oriented in opposite directions, with one strand running in the 5′ to 3′ direction and the other running in the 3′ to 5′ direction.

In addition to being complementary, the two strands of DNA are also antiparallel, oriented in opposite directions. The directionality of a DNA strand is defined by the carbon atom in the 2-deoxyribose molecule that attaches to another phosphate molecule in a nucleotide monomer. One strand is said to be oriented in the 5′ (pronounced "five prime") to 3′ ("three prime") direction, while the other strand is oriented in the 3′ to 5′ direction (Fig. 9.2). This orientation has important consequences for DNA synthesis, as DNA can only be synthesized in the 5′ to 3′ direction. This orientation also has consequences for RNA synthesis, as RNA can also only be synthesized in the 5′ to 3′ direction.

Transcription

Transcription is the synthesis of a messenger RNA (mRNA) molecule from a DNA template. There are three major differences between an RNA molecule

and a DNA molecule: (1) RNA molecules use the sugar ribose instead of 2-deoxyribose; (2) RNA molecules are single stranded; and (3) RNA molecules use the nitrogenous base uracil (U) instead of thymine.

In transcription, only one of the two strands of DNA, called the template strand, is transcribed. The other strand, the coding strand, has a sequence that is the same as the newly created RNA transcript (with U substituted for T). An enzyme called **RNA polymerase** reads the template strand in the 3′ to 5′ direction and synthesizes the new mRNA in the 5′ to 3′ direction (Fig. 9.3).

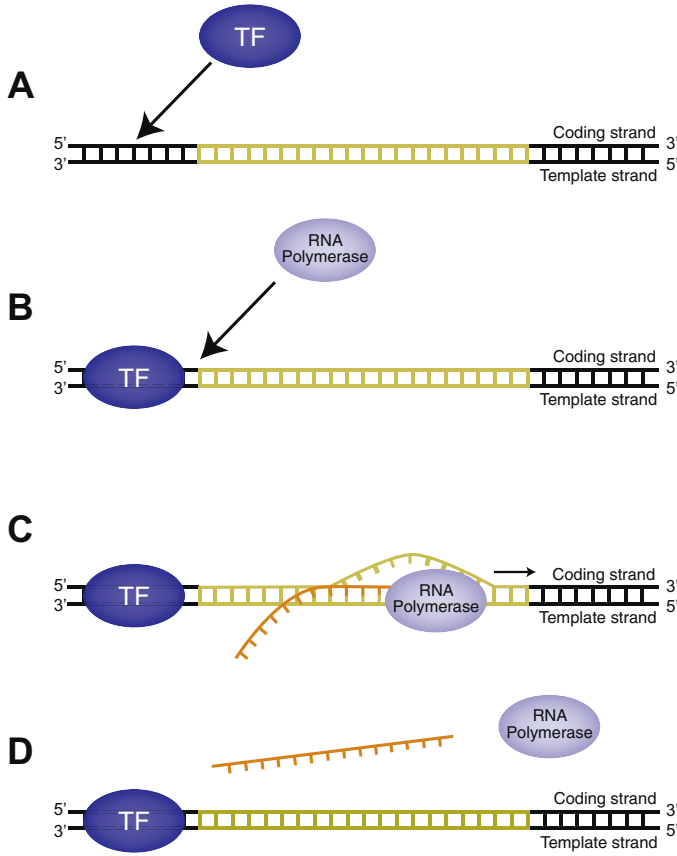

FIGURE 9.3 The transcription of DNA into RNA. (A) The process of transcription begins when specific transcription factors interact with promoter sequences on the genome. These factors ultimately attract RNA polymerase. (B) RNA polymerase binds to the complex between transcription factors and DNA. Once bound, the enzyme begins prying the two strands of DNA apart. (C) RNA polymerase proceeds down the length of the DNA, simultaneously unwinding DNA and adding nucleotides to the 3′ end of the growing RNA molecule. (D) When transcription is complete, RNA polymerase dissociates from the DNA and the two DNA strands reform a double helix conformation. A new mRNA is produced that is complementary to the DNA template strand.

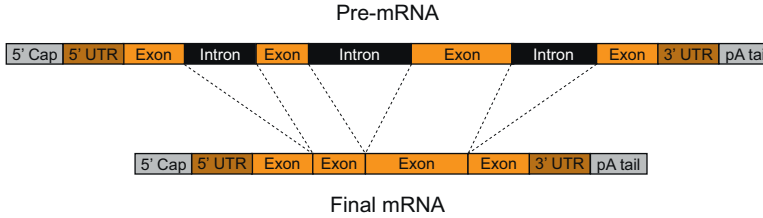

FIGURE 9.4 **RNA splicing.** Before mRNA leaves the nucleus, macromolecular complexes called spliceosomes remove noncoding intron sequences and rejoin exon sequences. Noncoding sequences remain at the 5′ and 3′ ends, as well as a 5′ cap and a 3′ poly-A tail.

Transcription begins when various proteins called **transcription factors** interact with regions of the genome called **promoters**. A promoter is a specific sequence of DNA that facilitates the transcription of a specific gene. These promoters are often located adjacent and upstream of the genes they regulate, toward the 5′ region of the coding strand. When a transcription factor binds to a promoter, multiple proteins are recruited to begin the transcription process.

After transcription terminates, a newly formed RNA molecule receives further processing in the form of **RNA splicing** before leaving the nucleus. At this point, the RNA is composed of exons and introns flanked by untranslated regions (UTRs) at either end (Fig. 9.4). An **exon** is a sequence of an RNA strand that remains after splicing occurs, while an **intron** is a sequence that is lost. The 5′ and 3′ UTRs are sequences that cap the RNA molecule and are not translated into amino acids. RNA splicing is catalyzed by a macromolecular complex known as a **spliceosome**, which cuts exon and intron sequences apart and then joins the exons back together to form a single strand. After RNA splicing, a mature mRNA molecule leaves the nucleus and migrates to the cytoplasm where it can be translated into a protein.

Translation

Translation is the process of using an mRNA molecule as a template to produce a protein. This process is performed by specialized macromolecules in the cytoplasm called **ribosomes**. A ribosome reads an mRNA molecule in the 5′ to 3′ direction and uses this template to guide the synthesis of a chain of amino acids to form a protein. **Amino acids** are the building blocks of all proteins and are assembled as a long strand, much like nucleotides are assembled as a long strand in a nucleic acid molecule. Each sequence of three nucleotides of an mRNA molecule codes for one amino acid. The three nucleotides collectively form a **codon**, a translation of genetic information into an amino acid monomer. There is a precise genetic code for the conversion of codons into amino acids (Fig. 9.5), with each codon precisely coding for one specific amino acid. One codon, “AUG” codes for the amino acid methionine, which is the start of any protein sequence. Other codons (“UAA,” “UAG,” and

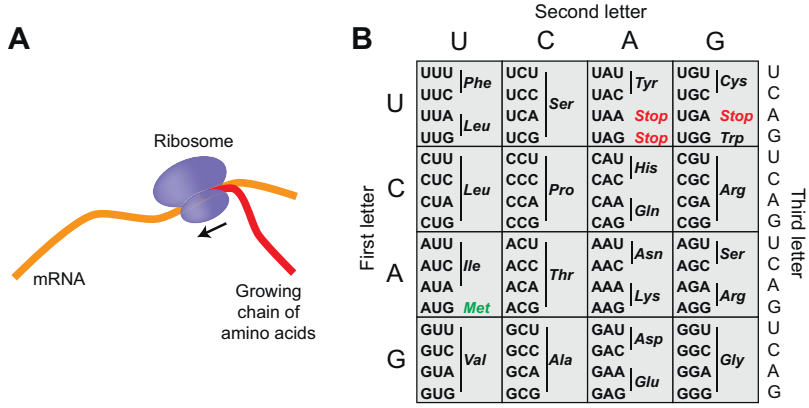

FIGURE 9.5 **The translation of mRNA into proteins.** (A) A ribosome translates the genetic sequence of an mRNA molecule into an amino acid sequence that forms a polypeptide. mRNA sequences are processed in the 5′ to 3′ direction. Each series of three nucleotides, called codons, translates into a single amino acid. (B) The genetic code for the conversion of codons into amino acids. Each codon translates into a specific amino acid (abbreviated here in three letters, such as Phe, which is the abbreviation for phenylalanine). Because there are more possible codons than amino acids, some codons translate into the same amino acids, such as UUU and UUC. The AUG codon translates into methionine and signals the start of translation. The UAA, UAG, and UGA codons do not translate into amino acids, instead serving as a stop signal and terminating translation.

"UGA") serve as "stop" codons and terminate translation. After translation ends, the newly formed polypeptide chain folds to assume its native protein conformation. These proteins can be further processed by other organelles, such as the endoplasmic reticulum, or receive **posttranslational modifications (PTMs)** from other proteins.

This brief survey of DNA, transcription, and translation is no match for much more detailed descriptions found in genetics or molecular biology textbooks. However, this information is the foundation on which most molecular neuroscience techniques are based. Now that we have reviewed these fundamental concepts, we survey various categories of screens used to identify genes of interest.

GENETIC SCREENS

A genetic screen is a process of identifying genes that contribute to a phenotype. Genetic screens fall into two categories: forward or reverse screens. A **forward genetic screen** is a screen used to identify genes important for a biological phenotype. In these screens, an investigator selects a scientifically interesting phenotype and seeks to determine which genes are necessary for the generation of this phenotype. A **reverse genetic screen** is the opposite: an investigator selects an interesting gene and seeks to determine

which phenotypes result from its absence. Forward genetic screens often test thousands of genes at once, while a reverse genetic screen tests only a single gene. Forward genetic screens have been compared to "fishing" because a scientist is never quite sure which genes they will catch; reverse genetic screens have been compared to "gambling" because a scientist places all of their hopes on that gene producing an interesting phenotype. Forward genetic screens can be considered "hypothesis generating," as they typically produce multiple candidate genes that may be important for a phenotype; reverse genetic screens can be considered "hypothesis testing," as a scientist can test specific hypotheses about the role of a gene in a phenotype. Here we briefly describe the process of performing a forward or reverse genetic screen.

Forward Genetic Screen

In a forward genetic screen, a scientist mutagenizes animals to produce thousands of lines that may have a mutation in a single gene, and then screens each individual to identify animals with an abnormal phenotype. These screens are typically performed in flies and worms, although they have occasionally been performed in mice (Box 9.1 and Table 9.1). There are six main steps in performing a forward genetic screen:

BOX 9.1 Genetic Model Organisms

Each week, hundreds of studies are published in peer-reviewed journals that use genetic tools to investigate the nervous system. Interestingly, nearly all of these studies use the same five species as model organisms: worms (*Caenorhabditis elegans*), flies (*Drosophila melanogaster*), zebrafish (*Danio rerio*), mice (*Mus musculus*), and when possible, humans (*Homo sapiens*). There are occasional exceptions, such as the frog (*Xenopus laevis*) or the sea slug (*Aplysia californica*), but these species collectively represent less than 1% of published studies. It is obvious why neuroscientists would want to learn more about the genes of humans, as the ultimate goal of neuroscience is to better understand our own species. But why did the other four species become such tractable organisms for genetic studies? Why use roundworms and fruit flies? Why not flatworms, bees, goldfish, or gerbils? Many of these other organisms have the same short breeding time, high number of offspring, and other factors that might make them attractive model species.

The answer can essentially be summarized in a single word: history. There really is not a good reason why scientists started using mice instead of gerbils, but when multiple laboratories started performing genetic experiments on these animals, it made sense for other labs to follow. It is useful for multiple laboratories to study the same species. When dozens or hundreds of labs each study the same organism, a community develops that allows researchers to better share results, reagents, and genetic tools.

BOX 9.1 Genetic Model Organisms—cont'd

There is no better example of a community forming around a genetic model organism than the community of scientists that study *Drosophila*. Ever since Thomas Hunt Morgan began using the fly as a model organism to study genetics in 1910, a growing number of scientists learned more and more about ideal conditions and practices for using flies in experiments. As the knowledge about fly maintenance increased, so did the number of genetic tools amenable to fly genetic studies. In the late 1960s, Seymour Benzer used *Drosophila* to study the link between genes and behavior. In the 1980 and 1990s, scientists developed dozens of molecular genetic tools for carefully dissecting genetic circuits in flies, and in 2000, the fly became the second eukaryotic organism to have its genome sequenced.

A similar history developed for the worm: the scientist Sydney Brenner began using *C. elegans* as a model organism in the 1960s. After years of study, scientists have given all 302 neurons specific names and mapped out the connections of all 7000 synapses. There is now an energetic *C. elegans* community just as there is a vibrant *Drosophila* community, with many established methods for maintaining and manipulating genetic strains of worms.

Table 9.1 summarizes some of the traits and benefits of these commonly used genetic model organisms. We mention these species not only because they are highly represented in the literature, but also because virtually all novel genes and proteins are discovered in these species. The methods described in this chapter to identify functionally relevant genes depend on the decades-long development of these organisms for biological research.

1. **Design an assay to measure a phenotype.** The first step in determining which genes may be important for a phenotype is to design a specific, quantitative assay to discriminate between wild-type individuals and animals with an aberrant phenotype. This requires fully characterizing the wild-type phenotype and choosing measurable parameters to identify individuals with an abnormal phenotype. For example, a scientist performing a forward genetic screen to determine genes important for axon guidance might choose a specific axon and characterize its normal growth and development over time. When screening individuals, the scientist could look for axons that mistarget to abnormal locations in the brain. Note that this screen requires the scientist to identify the same axon in thousands of individuals. Alternatively, a scientist could choose a behavioral phenotype, characterizing the normal behavioral response to a stimulus and then examining individuals with an abnormal behavior. The abnormal behavior must be statistically different from the normal variation in behavior between individuals.

2. **Mutagenize eggs/larvae.** There are three methods of developing mutant lines of organisms. In **chemical mutagenesis**, a scientist applies a mutagenizing chemical, such as ethyl methane sulfonate (EMS) or N-ethyl-N-nitrosourea (ENU), to thousands of eggs/larvae, which statistically creates lines of animals with mutations in a single gene in the genome. In **irradiation mutagenesis**, a scientist achieves the same goal with high-intensity UV light, which

TABLE 9.1 Relative Advantages and Disadvantages of Model Organisms Used for Genetic Studies in Neuroscience.

Species	Advantages	Disadvantages
Worm (*Caenorhabditis elegans*)	• Simple, multicellular organism • Known genome • Short lifespan • Relatively easy to manipulate genes and screen for genes of interest through mutagenesis • Can be frozen • All neurons and their connections are known	• Invertebrate (brain anatomy not similar to humans) • Genes can have very different functions than mammalian orthologues
Fruit fly (*Drosophila melanogaster*)	• Complex, multicellular organism • Known genome • Short lifespan, rapid reproduction rate • Relatively easy to manipulate genes • Many mutant lines available	• Invertebrate (brain anatomy not similar to humans) • Genes can have very different functions than mammalian orthologues
Zebrafish (*Danio rerio*)	• Vertebrate • Known genome • Can use powerful invertebrate techniques like Gal4/UAS system • Good for imaging studies during development because eggs are clear • Large number of offspring	• Aquatic organism, so needs to be kept in water, making some kinds of experiments impossible
Mouse (*Mus musculus*)	• Complex, higher-order organisms • Known genome • Nervous system is homologous to humans	• Relatively long lifespan and reproductive time • Greater genetic redundancy than invertebrates • More complicated promoter systems than invertebrates
Human (*Homo sapiens*)	• Complex, higher-order organism • Known genome • Can observe naturally occurring genetic mutations and polymorphisms	• Long lifespan • Not tractable for experimental genetic manipulation

damages DNA at a rate such that animals statistically have a mutation in a single gene in the genome. Finally, in **insertional mutagenesis**, a scientist uses various methods to insert mobile genetic elements called **transposons** into the genomes of offspring. These transposons insert at random locations in the genome, occasionally disrupting an endogenous gene. Therefore, they can be used to test thousands of animals, each with a potential loss of function of a single gene.

3. **Screen for abnormal phenotypes.** After hundreds or thousands of mutant lines are produced, a scientist screens each individual in a phenotypic assay, one mutant at a time. This is easily the most time-consuming aspect of a forward genetic screen. Any mutant lines that show an abnormal phenotype are maintained and bred for future experiments. The vast majority of mutant lines that do not show aberrant phenotypes are either used for other genetic screens or discarded.

4. **Perform complementation analysis.** Once a genetic mutant is isolated, investigators often perform a **complementation test** to determine if the mutation is unique or has been previously described. This approach is used when a mutation results in the same phenotype as a previously described mutation. The two strains could have mutations in the same gene or in two different genes. To differentiate between these two possibilities, the two strains are mated and the phenotypes of the offspring are identified. If the offspring exhibits the wild-type phenotype, a scientist concludes that each mutation is in a separate gene. However, if the offspring does not express the wild-type phenotype but rather a mutant phenotype, the scientist concludes that both mutations occur in the same gene.

5. **Map the gene.** Just because a scientist has discovered a mutation in a gene that alters a phenotype does not mean that the scientist knows the molecular identity of that gene. The only way to determine the molecular sequence of the gene and its location in the genome is to map the gene. If mutagenesis was caused by chemical or irradiation methods, investigators can perform a **linkage analysis** to map the gene. This technique is used to identify the relative position of a novel gene in relation to the location of known genes. To map the gene, a scientist mates the mutant animals with animals of a separate mutant phenotype. The two traits are then assessed in the offspring to determine if they are inherited together at relatively high rates. When two traits are closely "linked" in future offspring, they appear together more frequently and there is a greater likelihood that they are located close to each other on a chromosome. Because many genomes have been sequenced, a scientist can identify potential regions on a chromosome where the mutant gene may be found.

   If mutagenesis was caused by insertional mutagenesis methods, it is much easier to identify the molecular identity of the mutated gene. A scientist simply uses the transposon's genetic sequence to design primers for DNA sequencing, and then sequences DNA in either direction of the transposon to identify the region of the genome where it landed (Chapter 10).

6. **Clone the gene.** The final step in the process of identifying novel genes for a phenotype is to clone the DNA encoding the gene. This allows a scientist to perform future genetic experiments that manipulate the gene in vitro or in vivo. Chapter 10 describes methods of cloning genes and producing genetic constructs using recombinant DNA technology.

Reverse Genetic Screen

In a reverse genetic screen, a scientist identifies a gene of interest and then perturbs the gene to determine its role in various phenotypes. For example, if a scientist hypothesizes that a certain gene is necessary for proper formation of neuromuscular junctions in the mouse, the scientist could make a knockout mouse, removing the gene from the genome and examining whether proper synapses form between peripheral nerves and motor units. We describe the process of performing these kinds of experiments in Chapter 12.

IN SILICO SCREENS

With modern online bioinformatics and genomics databases, it is possible to identify genes and proteins that may be relevant for a phenotype of interest. An **in silico screen** is a computer-based method that identifies and compares similar DNA and/or protein sequences from multiple species. For example, let's say that a scientist identifies a gene encoding a novel transmembrane receptor. The scientist may be able to identify and discover other receptors by looking for similar sequences in the genome. Alternatively, if a gene is discovered in one species, a scientist can use bioinformatics to identify orthologous sequences in other species. These genes may have the same function or very different functions. Two examples of publicly available genetic databases are BLAST and Ensembl.

BASIC LOCAL ALIGNMENT SEARCH TOOL (BLAST)

BLAST is a tool used to compare the nucleotide sequences of DNA or the amino acid sequences of proteins. The power of BLAST is the ability to compare a DNA or protein sequence with the entire library of known sequences from all species. The user can set a threshold for the degree to which two sequences form a match. Therefore, if a scientist wants to identify genes with high sequence similarity to a gene of interest, the BLAST program can identify similar genes with high sequence similarity, and potentially, similar functions. Many genes in the mouse genome have been identified due to their similarity with other genes, especially those that encode neuropeptides and ion channels. The best way to learn about BLAST is to try it: http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Ensembl

Ensembl is a bioinformatics browser that functions much like BLAST but with a greater emphasis on genomic analysis. This tool provides the actual genomic sequences for a huge diversity of animal models, allowing comparative genomic searches to be performed. Additionally, this program can be used for analyzing sequence alignment, to investigate gene homology, and to obtain specific gene sequences to be used for subsequent experiments. Ensembl has the complete

genomes and rough drafts of genomes for dozens of animals, including standard model organisms, as well as relatively esoteric organisms, such as the bush baby, the bat, and the platypus. Again, the best way to learn about this genetic tool is to try it: http://www.ensembl.org/index.html.

MOLECULAR SCREENS

A molecular screen uses molecular biology techniques to identify genes important for a specific biological phenotype. Unlike standard genetic screens, molecular screens use high-throughput strategies and thousands of nucleic acid probes to identify the molecular components of a phenotype.

RNA Sequencing

RNA sequencing, commonly referred to as **RNA-seq**, is a high throughput technique that allows researchers to determine the sequences and quantity of all RNA molecules present is a sample—the "transcriptome." RNA-seq can be used as a screen by comparing two or more samples to identify differences in gene expression profiles. For example, a scientist could compare the gene expression profiles between healthy, wild-type cells from healthy animals versus abnormal cells from animal models of disease. Alternatively, a scientist could compare gene expression profiles between cells in culture treated with different chemical compounds.

RNA-seq methods have slight variations depending on the exact cells under investigation and the specific equipment in each lab; however, there is a standard overall procedure (Fig. 9.6). First, the scientist obtains cell samples, either from cultured cells or tissues, or tissue harvested directly from living animals. Next, RNA is isolated from the samples using commercially available kits or classical phenol/chloroform RNA extraction procedures. The RNA molecules are broken up into 200−300 bp fragments and converted into **complementary DNA (cDNA)** fragments using a process call **reverse transcription** (Chapter 10). Converting RNA molecules into DNA molecules is necessary because cDNA is much more stable than RNA. Small DNA adaptor sequences are added to each end of the cDNA fragments—one sequence on the 5′ end and another on the 3′ end. This collection of cDNA fragments, capped with adaptors, is referred to as a cDNA library.

The cDNA library is sequenced using **high-throughput sequencing** (also called **next generation sequencing**). These terms refer to modern DNA sequencing technologies that can rapidly read the sequence of cDNA libraries. Different sequencer machines, manufactured by different companies, use slightly different strategies to sequence cDNA fragments. In general, the adaptors added to cDNA fragments allow the fragments to bind to a solid base such that the cDNA sequences extend from the base. The sequencers then add fluorescent probes that are color coded according to the type of nucleotide they

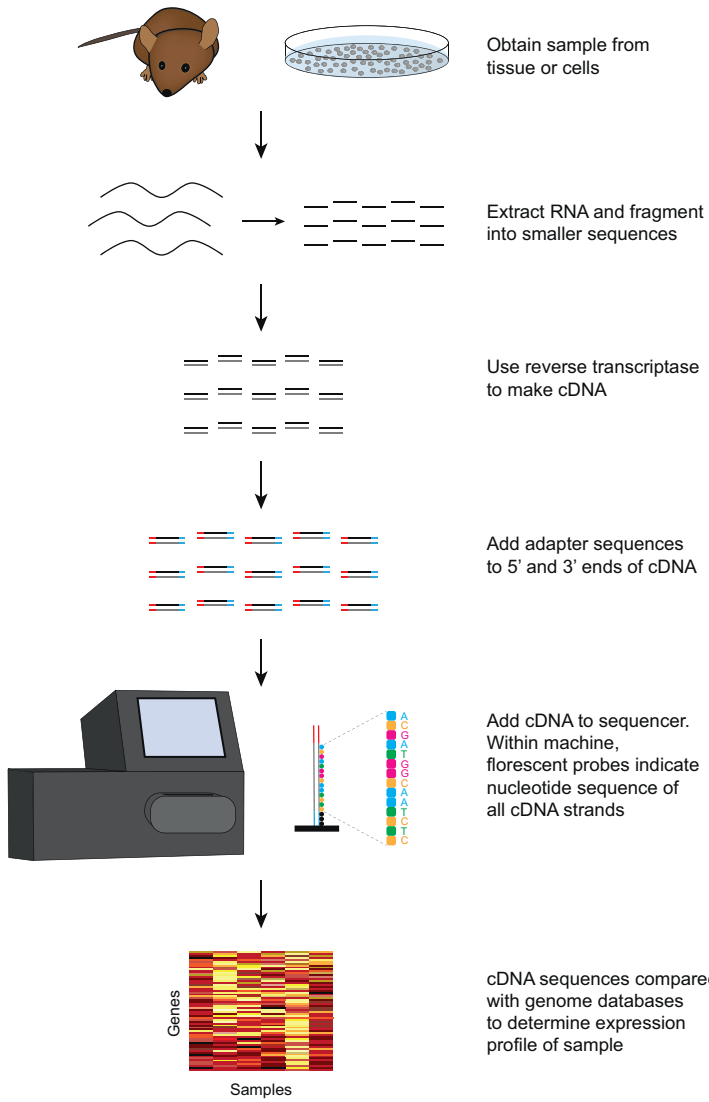

FIGURE 9.6 **RNA sequencing.** Samples are obtained from animal tissue or cultured cells. RNA is extracted from the samples and initially fragmented. Reverse transcriptase is used to produce cDNA, and short adapters are added to the 5′ and 3′ ends of the cDNA sequences. The entire cDNA library is added to a high-throughput sequencing machine, which uses fluorescent probes to bind and identify nucleotides on each cDNA strand. These fluorescent probes are read by a detector. When sequencing is complete, each read is compared with genomic databases to determine the identity and expression level of all mRNA present. Gene expression profiles of each sample are compared and visualized, for example, as a heat map.

can bind. For example, a red fluorescent probe may bind to adenine nucleotides, while a green probe may bind to cytosine nucleotides. The probes bind to the nucleotides closest to the solid base, then a camera takes an image of all the colors and where they are located on the base. The fluorescence is neutralized, and a new round of fluorescent probes are added that bind to the next nucleotides in the cDNA sequence. This process is repeated several times such that a computer keeps track of the nucleotide order of each sequence. In this way, the computer "reads" each cDNA by tracking the addition of fluorescent probes, one by one, throughout the length of the entire cDNA sequence. A single sequencing session can read the sequences of hundreds of millions of cDNA fragments.

Because the computer keeps track of all sequences within the cDNA library, sequences can be compared with genome databases. Therefore, a computer can identify which genes are represented in the cDNA library for any model organism, and quantify the number of reads for each gene. High-throughput sequencing produces an incredible amount of data for analysis. Multivariate statistical analyses are performed to determine transcriptome-wide differences between different samples.

There are many varieties of RNA-seq, allowing for many kinds of molecular screens. For example, **single-cell RNA sequencing** allows a scientist to determine the RNA expression profiles of individual cells. This process could be used to identify and classify unique genetic profiles of cells within a single brain region. Alternatively, **TRAP-seq** is used to determine which RNA species are actively translated by a ribosome at a specific moment in time. **ChIP-Seq** combines chromatin immunoprecipitation techniques (Chapter 14), used to examine protein−DNA interactions, with high-throughput sequencing to determine the identity of all DNA sequences that are bound to a protein of interest. These, and many other variations of RNA-seq technology, allow for unparalleled insights into the expression patterns of genes in a variety of contexts. One disadvantage of RNA-seq experiments is that they only examine gene expression—there is no information about whether the RNA transcripts are actually translated into proteins. Therefore, if there are differences in expression of RNA between two samples, there is no guarantee that there is also a change in expression of the relevant translated proteins. However, once interesting genes are identified, a scientist can pursue many follow-up experiments to validate differences in gene/protein expression using other methods.

CRISPR-Cas9 Screens

CRISPR-Cas9 is a genome editing technique discussed in Chapter 12. Cas9 is a protein that can cut very specific sequences of DNA in the genome, allowing a scientist to essentially knock out a gene of interest. Cas9 is directed to a specific location in the genome by the presence of a single guide RNA (sgRNA) that serves as a template for the DNA sequence to be cut.

The CRISPR-Cas9 system can be used for molecular screens in cultured cells by causing a Cas9-mediated cut in a distinctive gene in different cells and then searching for cells with an aberrant phenotype. To perform a CRISPR-Cas9 screen, a scientist uses a collection of thousands of sgRNAs, known as sgRNA libraries. These sgRNAs are packaged into viruses that can deliver unique sgRNAs, as well as the DNA that encodes the Cas9 protein, into cells. Therefore, the Cas9 and sgRNAs cause a unique loss of genes among the different cells.

There are two general categories of CRISPR-Cas9 screens. In an "arrayed screen," each sample contains a known sgRNA that targets a specific gene. Because the sgRNA is known, genes of interest that result from the screen can be readily identified. These screens typically do not test all possible sgRNAs but allow for a time-consuming, in-depth analysis of cellular phenotypes. Alternatively, in a "pooled" screen, cells are all grown together and are randomly infected with viruses such that they express a unique but unknown sgRNA. Once cells with aberrant phenotypes are identified, the scientist must perform sequencing reactions to identify the specific sgRNA that caused the mutation. These screens allow an investigator to test a large sgRNA library, and thus potentially identify a wide variety of genes of interest that play a role in a cellular phenotype.

CONCLUSION

This chapter serves as an introduction to techniques used to identify new genes and their protein products. These techniques are often the first step in discovering the role of a gene in a biological process or phenotype. Other methods are necessary to validate this discovery, such as methods that perturb the expression of the gene to identify its necessity and/or sufficiency for a phenotype. The remaining chapters of this book use molecular and genetic methods to gain insight into newly discovered genes and their role in vitro and in vivo.

SUGGESTED READING AND REFERENCES

Books

Greenspan, R.J., 2004. Fly Pushing: The Theory and Practice of *Drosophila* Genetics, second ed. Cold Spring Harbor Laboratory Press.

Griffiths, A.J.F., Wessler, S.R., Carroll, S.B., 2015. Introduction to Genetic Analysis, eleventh ed. Freeman.

Krebs, J.E., Goldstein, E.S., Kilpatrick, S.T., 2020. Lewin's Essential Genes, fourth ed. Jones and Bartlett.

Neelakanta, P.S., 2020. A Textbook of Bioinformatics: Information-Theoretic Perspectives of Bioengineering and Biological Complexes.

Picardi, E.P., 2021. RNA Bioinformatics. Humana Press.

Review Articles

Goldowitz, D., Frankel, W.N., Takahashi, J.S., et al., 2004. Large-scale mutagenesis of the mouse to understand the genetic bases of nervous system structure and function. Brain Res. Mol. Brain Res. 132, 105–115.

Medland, S.E., Jahanshad, N., Neale, B.M., Thompson, P.M., 2014. Whole-genome analyses of whole-brain data: working within an expanded search space. Nat. Neurosci. 17, 791–800.

Ryder, E., Russell, S., 2003. Transposable elements as tools for genomics and genetics in *Drosophila*. Briefings Funct. Genomics Proteomics 2, 57–71.

Stark, R., Grzelak, M., Hadfield, J., 2019. RNA sequencing: the teenage years. Nat. Rev. Genet. 20, 631–656.

Wu, S., Ying, G., Wu, Q., Capecchi, M.R., 2007. Toward simpler and faster genome-wide mutagenesis in mice. Nat. Genet. 39, 922–930.

Primary Research Articles—Interesting Examples from the Literature

Boisvert, M.M., Erikson, G.A., Shokhirev, M.N., Allen, N., 2018. The aging astrocyte transcriptome from multiple regions of the mouse brain. Cell Rep. 22, 269–285.

Blum, J.A., et al., 2021. Single-cell transcriptomic analysis of the adult mouse spinal cord reveals molecular diversity of autonomic and skeletal motor neurons. Nat. Neurosci. 24, 572–583.

Crick, F.H., 1970. Central dogma of molecular biology. Nature 227, 561–563.

Devineni, A.V., Eddison, M., Heberlein, U., 2013. The novel gene tank, a tumor suppressor homolog, regulates ethanol sensitivity in *Drosophila*. J. Neurosci. 33, 8134–8143.

Haney, M.S., Bohlen, C.J., Morgens, D.W., et al., 2018. Identification of phagocytosis regulators using magnetic genome-wide CRISPR screens. Nat. Genet. 50, 1716–1727.

Husken, U., et al., 2014. Tcf7l2 is required for left-right asymmetric differentiation of habenular neurons. Curr. Biol. 6, 2217–2227.

Marques, S., et al., 2016. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. Science 352, 1326–1329.

Schuldiner, O., Berdnik, D., Levy, J.M., et al., 2008. piggyBac-based mosaic screen identifies a postmitotic function for cohesin in regulating developmental axon pruning. Dev. Cell 14, 227–238.

Tracey Jr., W.D., Wilson, R.I., Laurent, G., Benzer, S., 2003. Painless, a *Drosophila* gene essential for nociception. Cell 113, 261–273.

Ule, J., Ule, A., Spencer, J., et al., 2005. Nova regulates brain-specific splicing to shape the synapse. Nat. Genet. 37, 844–852.

Walsh, T., McClellan, J.M., McCarthy, S.E., et al., 2008. Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. Science 320, 539–543.

Zarbalis, K., May, S.R., Shen, Y., Ekker, M., Rubenstein, J.L., Peterson, A.S., 2004. A focused and efficient genetic screening strategy in the mouse: identification of mutations that disrupt cortical development. PLoS Biol. 2, E219.

Protocols

Bökel, C., 2008. EMS screens : from mutagenesis to screening and mapping. Methods Mol. Biol. 420, 119–138.

Mackenzie, R.J., 2018. RNA-Seq: Basics, Applications, and Protocol.

Mereu, E., Lafzi, A., Moutinho, C., et al., 2020. Benchmarking single-cell RNA-sequencing protocols for cell atlas projects. Nat. Biotechnol. 38, 747–755.

Websites

BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi.

Ensembl: http://www.ensembl.org/index.html.