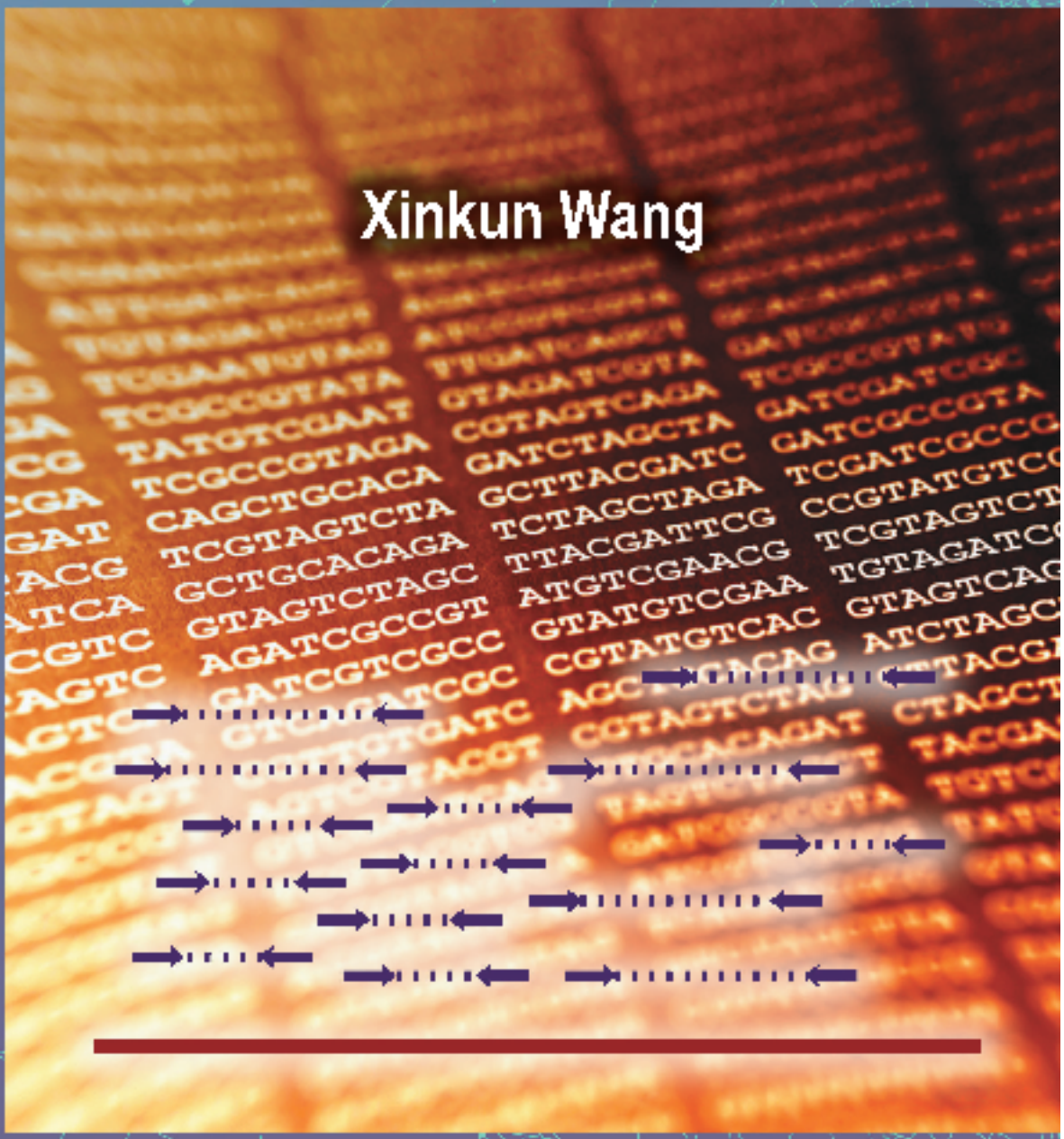
Next-Generation Sequencing Data Analysis



Xinkun Wang

CRC Press
Taylor & Francis Group
A CHAPMAN & HALL BOOK

Next-Generation Sequencing Data Analysis

Next-Generation Sequencing Data Analysis

Xinkun Wang
Northwestern University
Chicago, Illinois, USA

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

Version Date: 20160127

International Standard Book Number-13: 978-1-4822-1789-6 (eBook - PDF)

Contents

Section I Introduction to Cellular and Molecular Biology

Section III Application-Specific NGS Data Analysis

Section I

Introduction to Cellular and Molecular Biology

1

The Cellular System and the Code of Life

1.1 The Cellular Challenge

A cell, although minuscule with a diameter of less than 50 μm, works wonders if you compare it to any human-made system. Moreover, it perpetuates itself using the information coded in its DNA. In case you ever had the thought of designing an artificial system that shows this type of sophistication, you would know the many insurmountable challenges such a system needs to overcome. A cell has a complicated internal system, containing many types of molecules and parts. To sustain the system, a cell needs to perform a wide variety of tasks—the most fundamental of which are to maintain its internal order, prevent its system from malfunctioning or breaking down, and reproduce or even improve the system—in an environment that is constantly changing.

Energy is needed to maintain the internal order of the cellular system. Without constant energy input, the entropy of the system will gradually increase, as dictated by the second law of thermodynamics, and ultimately lead to the destruction of the system. Besides energy, raw "building" material is also constantly needed to renew its internal parts or build new ones, as the internal structure of a cell is dynamic and responds to constant changes in environmental conditions. Therefore, to maintain the equilibrium inside and with the environment, it requires a constant influx of energy and raw material, and excretion of its waste. Guiding the capture of the requisite energy and raw material for its survival and the perpetuation of the system is the information encoded in its DNA sequence.

Because of evolution, a great number of organisms no longer function as a single cell. The human body, for example, contains trillions of cells. In a multicellular system, each cell becomes specialized to perform a specific function, for example, β-cells in our pancreas synthesize and release insulin, and cortical neurons in the brain perform neurobiological functions that underlie learning and memory. Despite this division of labor, the challenges a single-cell organism faces still hold true for each one of these cells. Instead of dealing with the external environment directly, they interact with and respond to changes in their microenvironment.

1.2 How Cells Meet the Challenge

Many cells, like algae and plant cells, directly capture energy from the sun or other energy sources. Other cells (or organisms) obtain energy from the environment as heterotrophs. For raw material, cells can either fix carbon dioxide in the air using the energy captured into simple organic compounds, which are then converted to other requisite molecules, or directly obtain organic molecules from the environment and convert them to requisite materials. In the meantime, existing cellular components can also be broken down when not needed for the reuse of their building material. This process of energy capture and utilization, and synthesis, interconversion, and breaking down for reuse of molecular material, constitutes the cellular metabolism. Metabolism, the most fundamental characteristic of a cell, involves numerous biochemical reactions.

Reception and transduction of various signals in the environment are crucial for cellular survival. Reception of signals relies on specific receptors situated on the cell surface, and for some signals, those inside the cell. Transduction of incoming signals usually involves cascades of events in the cell, through which the original signals are amplified and modulated. In response, the cellular metabolic profile is altered. The cellular signal reception and transduction network is composed of circuits that are organized into various pathways. Malfunctioning of these pathways can have a detrimental effect on the cell's response to the environment and eventually its survival.

Perpetuation and evolution of the cellular system rely on DNA replication and cell division. The replication of DNA (to be detailed in Chapter 2) is a high-fidelity, but not error-free, process. While maintaining the stability of the system, this process also provides the mechanism for the diversification and evolution of the cellular system. The cell division process is also tightly regulated, for the most part to ensure equal transfer of the replicated DNA into daughter cells. For the majority of multicellular organisms that reproduce sexually, during the process of germ cell formation the DNA is replicated once but cell division occurs twice, leading to the reduction of DNA material by half in the gametes. The recombination of DNA from female and male gametes leads to further diversification in the offspring.

1.3 Molecules in Cells

Different types of molecules are needed to carry out the various cellular processes. In a typical cell, water is the most abundant, representing 70% of the total cell weight. Besides water, there is a large variety of small and large

molecules. The major categories of small molecules include inorganic ions (e.g., Na^+, K^+, Ca^{2+}, Cl^-, and Mg^{2+}), monosaccharides, fatty acids, amino acids, and nucleotides. Major varieties of large molecules are polysaccharides, lipids, proteins, and nucleic acids (DNA and RNA). Among these components, the inorganic ions are important for signaling (e.g., waves of Ca^{2+} represent important intracellular signal), cell energy storage (e.g., in the form of Na^+/K^+ cross-membrane gradient), or protein structure/function (e.g., Mg^{2+} is an essential cofactor for many metalloproteins). Carbohydrates (including monosaccharides and polysaccharides), fatty acids, and lipids are major energy-providing molecules in the cell. Lipids are also the major component of cell membrane. Proteins, which are assembled from 20 types of amino acids in different order and length, underlie almost all cellular activities, including metabolism, signal transduction, DNA replication, and cell division. They are also the building blocks of many intracellular structures, such as cytoskeleton (see Section 1.4). Nucleic acids carry the code of life in their nearly endless nucleotide permutations, which not only provide instructions on the assembly of all proteins in cells but also exert control on how such assembly is carried out based on environmental conditions.

1.4 Intracellular Structures or Spaces

Cells maintain a well-organized internal structure (Figure 1.1). Based on the complexity of their internal structure, cells are divided into two major categories: prokaryotic and eukaryotic cells. The fundamental difference between them is whether a nucleus is present. Prokaryotic cells, being the more primordial of the two, do not have a nucleus, and as a result their DNA is located in a nucleus-like but nonenclosed area. Prokaryotic cells also lack organelles, which are specialized and compartmentalized intracellular structures that carry out different cellular functions (detailed next). Eukaryotic cells, on the other hand, contain a distinct nucleus dedicated for DNA storage, maintenance, and expression. Furthermore, they contain various organelles including the endoplasmic reticulum (ER), Golgi apparatus, cytoskeleton, mitochondrion, and chloroplast (plant cells). Following is an introduction to the various intracellular structures and spaces, including the nucleus, the organelles, and other subcellular structures and spaces such as the cell membrane and cytoplasm.

1.4.1 Nucleus

Since DNA stores the code of life, it must be protected and properly maintained to avoid possible damage, and ensure accuracy and stability. As proper execution of the genetic information embedded in the DNA is critical

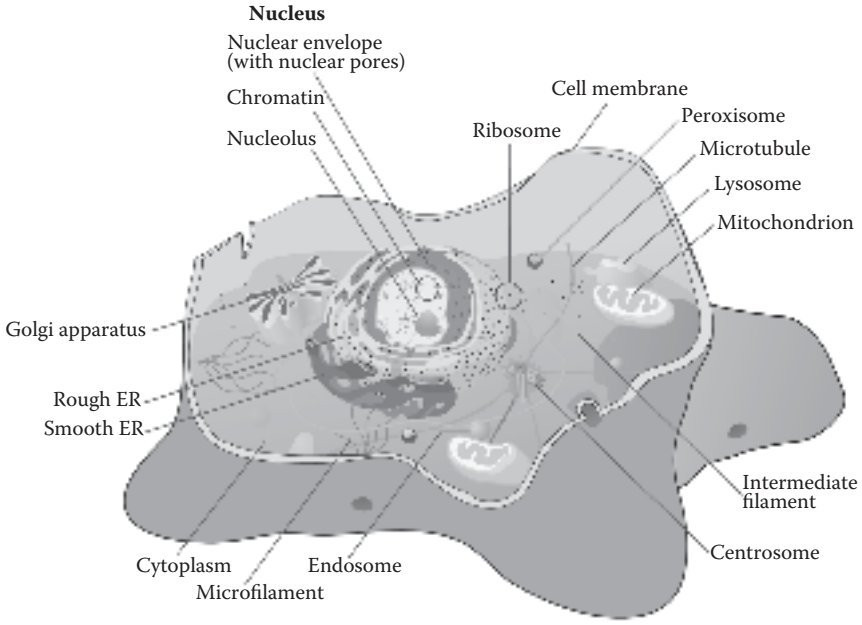


FIGURE 1.1
The general structure of a typical eukaryotic cell. Shown here is an animal cell.

to the normal functioning of a cell, gene expression must also be tightly regulated under all conditions. The nucleus, located in the center of most cells in eukaryotes, offers a well-protected environment for DNA storage, maintenance, and gene expression. The nuclear space is enclosed by a nuclear envelope consisting of two concentric membranes. To allow movement of proteins and RNAs across the nuclear envelope, which is essential for gene expression, there are pores on the nuclear envelope that span the inner and outer membrane. The mechanical support of the nucleus is provided by the nucleoskeleton, a network of structural proteins called lamins. Inside the nucleus, long strings of DNA molecules, through binding to certain proteins called histones, are heavily packed to fit into the limited nuclear space. In prokaryotic cells, a nucleus-like irregularly shaped region that does not have a membrane enclosure called the nucleoid, provides a similar but not as well-protected space for DNA.

1.4.2 Cell Membrane

The cell membrane serves as a barrier to protect the internal structure of a cell from the outside environment. Biochemically, the cell membrane, as well as all other intracellular membranes such as the nuclear envelope, assumes a

lipid bilayer structure. While offering protection to their internal structure, the cell membrane is also where cells exchange materials, and concurrently energy, with the outside environment. Since the membrane is made of lipids, most water-soluble substances, including ions, carbohydrates, amino acids, and nucleotides, cannot directly cross it. To overcome this barrier, there are channels, transporters, and pumps, all of which are specialized proteins, on the cell membrane. Channels and transporters facilitate passive movement, that is, in the direction from high to low concentration, without consumption of cellular energy. Pumps, on the other hand, provide active transportation of the molecules, since they transport the molecules against the concentration gradient and therefore consume energy.

The cell membrane is also where a cell receives most incoming signals from the environment. After signal molecules bind to their specific receptors on the cell membrane, the signal is relayed to the inside, usually eliciting a series of intracellular reactions. The ultimate cellular response that the signal induces is dependent on the nature of the signal, as well as the type and condition of the cell. For example, upon detecting insulin in the blood via the insulin receptor in their membrane, cells in the liver respond by taking up glucose from the blood for storage.

1.4.3 Cytoplasm

Inside the cell membrane, cytoplasm is the thick solution that contains the majority of cellular substances, including all organelles in eukaryotic cells but excluding the nucleus in eukaryotic cells and the DNA in prokaryotic cells. The general fluid component of the cytoplasm that excludes the organelles is called the cytosol. The cytosol makes up more than half of the cellular volume and is where many cellular activities take place, including a large number of metabolic steps such as glycolysis and interconversion of molecules and most signal transduction steps. In prokaryotic cells, due to the lack of a nucleus and other specialized organelles, the cytosol is almost the entire intracellular space and where most cellular activities take place.

Besides water, the cytosol contains large amounts of small and large molecules. Small molecules, such as inorganic ions, provide an overall biochemical environment for cellular activities. In addition, ions such as Na^+, K^+, and Ca^{2+} also have substantial concentration differences between the cytosol and the extracellular space. Cells spend a lot of energy maintaining these concentration differences, and use them for signaling and metabolic purposes. For example, the concentration of Ca^{2+} in the cytosol is normally kept very low at ~10^{-7} M, whereas in the extracellular space it is ~10^{-3} M. The rushing in of Ca^{2+} under certain conditions through ligand- or voltage-gated channels serves as an important messenger, inducing responses in a number of signaling pathways, some of which lead to altered gene expression. Besides small molecules, the cytosol also contains large numbers of macromolecules. Far from being simply randomly diffusing in the cytosol,

these large molecules form molecular machines that collectively function as a "bustling metropolitan city" [1]. These supramacromolecular machines are usually assembled out of multiple proteins, or proteins and RNA. Their emergence and disappearance are dynamic and regulated by external and internal conditions.

1.4.4 Endosome, Lysosome, and Peroxisome

Endocytosis is when cells bring in macromolecules, or other particulate substances such as bacteria or cell debris, into the cytoplasm from the surroundings. Endosome and lysosome are two organelles that are involved in this process. To initiate endocytosis, part of the cell membrane forms a pit, engulfs the external substances, and then an endocytotic vesicle pinches off from the cell membrane into the cytosol. Endosome, normally in the size range of 300 to 400 nm in diameter, forms from the fusion of these endocytotic vesicles. The internalized materials contained in the endosome are sent to other organelles such as lysosome for further digestion.

The lysosome is the principal site for intracellular digestion of internalized materials as well as obsolete components inside the cell. Like the condition in our stomach, the inside of the lysosome is acidic (pH at 4.5–5.0), providing an ideal condition for the many digestive enzymes within. These enzymes can break down proteins, DNA, RNA, lipids, and carbohydrates. Normally the lysosome membrane keeps these digestive enzymes from leaking into the cytosol. Even in the event of these enzymes leaking out of the lysosome, they can do little harm to the cell, since their digestive activities are heavily dependent on the acidic environment inside the lysosome, whereas the pH of the cytosol is slightly alkaline (around 7.2).

Peroxisome is morphologically similar to the lysosome, however it contains a different set of proteins, mostly oxidative enzymes that use molecular oxygen to extract hydrogen from organic compounds to form hydrogen peroxide. The hydrogen peroxide can then be used to oxidize other substrates, such as phenols or alcohols, via peroxidation reaction. As an example, liver and kidney cells use these reactions to detoxify various toxic substances that enter the body. Another function of the peroxisome is to break down long-chain fatty acids into smaller molecules by oxidation. Despite its important functions, the origin of peroxisome is still under debate. One theory proposes that this organelle has an endosymbiotic origin [2]. If this theory holds true, all genes in the genome of the original endosymbiotic organism must have been transferred to the nuclear genome. Another theory proposes that the peroxisome is a remnant of an ancient organelle that served to lower intracellular oxygen levels when the oxygen that we depend on today was still highly toxic to most cells, while exploiting the chemical reactivity of oxygen to carry out useful oxidative reactions for the host cell. Also based on this theory, the mitochondrion (see later) that emerged later releases energy from many of the same oxidative reactions that had previously taken place

in the peroxisome but without generating any energy, thereby rendering the peroxisome largely irrelevant except for carrying out the remnant oxidative functions.

1.4.5 Ribosome

Ribosome is the protein assembly factory in cells, translating genetic information carried in messenger RNAs (mRNAs) into proteins. There are vast numbers of ribosomes, usually from thousands to millions, in a typical cell. Whereas both prokaryotic and eukaryotic ribosomes are composed of two components (or subunits), eukaryotic ribosomes are larger than their prokaryotic counterparts. In eukaryotic cells, the two ribosomal subunits are first assembled inside the nucleus in a region called the nucleolus and then shipped out to the cytoplasm. In the cytoplasm, ribosomes can be either free or get attached to another organelle (the ER). Biochemically, ribosomes contain more than 50 proteins and several ribosomal RNA (rRNA) species. Because ribosomes are highly abundant in cells, rRNAs are the most abundant in total RNA extracts, accounting for 85% to 90% of all RNA species. For profiling cellular RNA populations using next-generation sequencing (NGS), rRNAs are usually not of interest despite their abundance and therefore need to be depleted to avoid generation of overwhelming amounts of sequencing reads from them.

1.4.6 Endoplasmic Reticulum (ER)

As indicated by the name, the endoplasmic reticulum (ER) is a network of membrane-enclosed spaces throughout the cytosol. These spaces interconnect and form a single internal environment called the ER lumen. There are two types of ERs in cells: rough ER and smooth ER. The rough ER is where all cell membrane proteins, such as ion channels, transporters, pumps, and signal molecule receptors, as well as secretory proteins, such as insulin, are produced and sorted. The characteristic surface roughness of this type of ER comes from the ribosomes that bind to them on the outside. Proteins destined for cell membrane or secretion, once emerging from these ribosomes, are threaded into the ER lumen. This ER-targeting process is mediated by a signal sequence, or "address tag," located at the beginning part of these proteins. This signal sequence is subsequently cleaved off inside the ER before the protein synthesis process is complete. Functionally different from the rough ER, the smooth ER plays an important role in lipid synthesis for the replenishment of cellular membranes. Besides membrane and secretory protein preparation and lipid synthesis, one other important function of the ER is to sequester Ca^{2+} from the cytosol. In Ca^{2+}-mediated cell signaling, shortly after entry of the calcium wave into the cytosol, most of the incoming Ca^{2+} needs to be pumped out of the cell and/or sequestered into specific organelles such as the ER and mitochondria.

1.4.7 Golgi Apparatus

Besides the ER, the Golgi apparatus also plays an indispensable role in sorting as well as dispatching proteins to the cell membrane, extracellular space, or other subcellular destinations. Many proteins synthesized in the ER are sent to the Golgi apparatus via small vesicles for further processing before being sent to their final destinations. Therefore, the Golgi apparatus is sometimes metaphorically described as the "post office" of the cell. The processing carried out in this organelle includes chemical modification of some of the proteins, such as adding oligosaccharide side chains, which serve as "address labels." Other important functions of the Golgi apparatus include synthesizing carbohydrates and extracellular matrix materials, such as the polysaccharide for the building of the plant cell wall.

1.4.8 Cytoskeleton

Cellular processes like the trafficking of proteins in vesicles from the ER to the Golgi apparatus or the movement of a mitochondrion from one intracellular location to another are not simply based on diffusion. Rather, they follow a certain protein-made skeletal structure inside the cytosol, that is, the cytoskeleton, as tracks. Besides providing tracks for intracellular transport, the cytoskeleton, like the skeleton in the human body, plays an equally important role in maintaining cell shape and protecting the cell framework from physical stresses, as the lipid bilayer cell membrane is fragile and vulnerable to such stresses. In eukaryotic cells, there are three major types of cytoskeletal structures: microfilament, microtubule, and intermediate filament. Each type is made of distinct proteins and has its own unique characteristics and functions. For example, microfilament and microtubule are assembled from actins and tubulins, respectively, and have different thicknesses (the diameter is about 6 nm for microfilament and 23 nm for microtubule). Although biochemically and structurally different, both the microfilament and the microtubule have been known to provide tracks for mRNA transport in the form of large ribonucleoprotein complexes to specific intracellular sites, such as the distal end of a neuronal dendrite, for targeted protein translation [3,4]. Besides its role in intracellular transportation, the microtubule also plays a key role in cell division through attaching to the duplicated chromosomes and moving them equally into two daughter cells. In this process, all microtubules involved are organized around a small organelle called centrosome. Previously thought to be only present in eukaryotic cells, cytoskeletal structures have also been discovered in prokaryotic cells [5].

1.4.9 Mitochondrion

The mitochondrion is the "powerhouse" in eukaryotic cells. While some energy is produced from the glycolytic pathway in the cytosol, most energy