

# 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

© 2016 by Taylor & Francis Group, LLC  
CRC Press is an imprint of Taylor & Francis Group, an Informa business

No claim to original U.S. Government works  
Version Date: 20160127

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

This book contains information obtained from authentic and highly regarded sources. Reasonable efforts have been made to publish reliable data and information, but the author and publisher cannot assume responsibility for the validity of all materials or the consequences of their use. The authors and publishers have attempted to trace the copyright holders of all material reproduced in this publication and apologize to copyright holders if permission to publish in this form has not been obtained. If any copyright material has not been acknowledged please write and let us know so we may rectify in any future reprint.

Except as permitted under U.S. Copyright Law, no part of this book may be reprinted, reproduced, transmitted, or utilized in any form by any electronic, mechanical, or other means, now known or hereafter invented, including photocopying, microfilming, and recording, or in any information storage or retrieval system, without written permission from the publishers.

For permission to photocopy or use material electronically from this work, please access [www.copyright.com](http://www.copyright.com) (<http://www.copyright.com/>) or contact the Copyright Clearance Center, Inc. (CCC), 222 Rosewood Drive, Danvers, MA 01923, 978-750-8400. CCC is a not-for-profit organization that provides licenses and registration for a variety of users. For organizations that have been granted a photocopy license by the CCC, a separate system of payment has been arranged.

**Trademark Notice:** Product or corporate names may be trademarks or registered trademarks, and are used only for identification and explanation without intent to infringe.

**Visit the Taylor & Francis Web site at**  
**<http://www.taylorandfrancis.com>**

**and the CRC Press Web site at**  
**<http://www.crcpress.com>**

---

# Contents

---

## Section I Introduction to Cellular and Molecular Biology

|  |    |
|--|----|
| <b>1. The Cellular System and the Code of Life</b> .....                         | 3  |
| 1.1 The Cellular Challenge .....   | 3  |
| 1.2 How Cells Meet the Challenge .....   | 4  |
| 1.3 Molecules in Cells .....   | 4  |
| 1.4 Intracellular Structures or Spaces.....                                      | 5  |
| 1.4.1 Nucleus.....   | 5  |
| 1.4.2 Cell Membrane .....  | 6  |
| 1.4.3 Cytoplasm .....  | 7  |
| 1.4.4 Endosome, Lysosome, and Peroxisome .....                                   | 8  |
| 1.4.5 Ribosome.....  | 9  |
| 1.4.6 Endoplasmic Reticulum (ER) .....   | 9  |
| 1.4.7 Golgi Apparatus.....   | 10 |
| 1.4.8 Cytoskeleton .....   | 10 |
| 1.4.9 Mitochondrion.....   | 10 |
| 1.4.10 Chloroplast.....  | 12 |
| 1.5 The Cell as a System .....   | 13 |
| 1.5.1 The Cellular System.....   | 13 |
| 1.5.2 Systems Biology of the Cell .....  | 13 |
| 1.5.3 How to Study the Cellular System .....                                     | 14 |
| <b>2. DNA Sequence: The Genome Base</b> .....                                    | 17 |
| 2.1 The DNA Double Helix and Base Sequence .....                                 | 17 |
| 2.2 How DNA Molecules Replicate and Maintain Fidelity.....                       | 18 |
| 2.3 How the Genetic Information Stored in DNA Is Transferred<br>to Protein ..... | 20 |
| 2.4 The Genomic Landscape.....   | 21 |
| 2.4.1 The Minimal Genome .....   | 21 |
| 2.4.2 Genome Sizes .....   | 21 |
| 2.4.3 Protein-Coding Regions of the Genome .....                                 | 22 |
| 2.4.4 Noncoding Genomic Elements .....   | 23 |
| 2.5 DNA Packaging, Sequence Access, and DNA–Protein<br>Interactions.....         | 25 |
| 2.5.1 DNA Packaging.....   | 25 |
| 2.5.2 Sequence Access.....   | 25 |
| 2.5.3 DNA–Protein Interactions .....   | 26 |
| 2.6 DNA Sequence Mutation and Polymorphism .....                                 | 27 |

|           |  |           |
|-----------|--|-----------|
| 2.7       | Genome Evolution.....  | 28        |
| 2.8       | Epigenome and DNA Methylation.....                             | 29        |
| 2.9       | Genome Sequencing and Disease Risk.....                        | 30        |
| 2.9.1     | Mendelian (Single-Gene) Diseases.....                          | 31        |
| 2.9.2     | Complex Diseases That Involve Multiple Genes.....              | 31        |
| 2.9.3     | Diseases Caused by Genome Instability .....                    | 32        |
| 2.9.4     | Epigenomic/Epigenetic Diseases .....                           | 32        |
| <b>3.</b> | <b>RNA: The Transcribed Sequence .....</b>                     | <b>35</b> |
| 3.1       | RNA as the Messenger .....                                     | 35        |
| 3.2       | The Molecular Structure of RNA .....                           | 35        |
| 3.3       | Generation, Processing, and Turnover of RNA as a Messenger ... | 36        |
| 3.3.1     | DNA Template.....  | 37        |
| 3.3.2     | Transcription of Prokaryotic Genes .....                       | 37        |
| 3.3.3     | Initial Transcription of Pre-mRNA from Eukaryotic Genes .....  | 38        |
| 3.3.4     | Maturation of mRNA from Pre-mRNA.....                          | 40        |
| 3.3.5     | Transport and Localization .....                               | 42        |
| 3.3.6     | Stability and Decay.....                                       | 42        |
| 3.3.7     | Major Steps of mRNA Transcript Level Regulation .....          | 43        |
| 3.4       | RNA Is More Than a Messenger.....                              | 44        |
| 3.4.1     | Ribozyme .....   | 45        |
| 3.4.2     | snRNA and snoRNA .....   | 46        |
| 3.4.3     | RNA for Telomere Replication .....                             | 46        |
| 3.4.4     | RNAi and Small Noncoding RNAs .....                            | 46        |
| 3.4.4.1   | miRNA.....   | 47        |
| 3.4.4.2   | siRNA.....   | 49        |
| 3.4.4.3   | piRNA .....  | 49        |
| 3.4.5     | Long Noncoding RNAs .....                                      | 50        |
| 3.4.6     | Other Noncoding RNAs .....                                     | 50        |
| 3.5       | The Cellular Transcriptional Landscape .....                   | 51        |

## Section II Introduction to Next-Generation Sequencing (NGS) and NGS Data Analysis

|           |  |           |
|-----------|--|-----------|
| <b>4.</b> | <b>Next-Generation Sequencing (NGS) Technologies: Ins and Outs .....</b> | <b>55</b> |
| 4.1       | How to Sequence DNA: From First Generation to the Next.....              | 55        |
| 4.2       | A Typical NGS Experimental Workflow.....                                 | 58        |
| 4.3       | Ins and Outs of Different NGS Platforms .....                            | 60        |
| 4.3.1     | Illumina Reversible Dye-Terminator Sequencing .....                      | 61        |
| 4.3.1.1   | Sequencing Principle .....   | 61        |
| 4.3.1.2   | Implementation.....  | 61        |
| 4.3.1.3   | Error Rate, Read Length, Data Output, and Run Time .....                 | 64        |



- 4.3.2 Ion Torrent Semiconductor Sequencing ..... 65
  - 4.3.2.1 Sequencing Principle ..... 65
  - 4.3.2.2 Implementation..... 65
  - 4.3.2.3 Error Rate, Read Length, Data Output,  
and Run Time ..... 66
- 4.3.3 Pacific Biosciences Single Molecule Real-Time  
(SMRT) Sequencing ..... 66
  - 4.3.3.1 Sequencing Principle ..... 66
  - 4.3.3.2 Implementation..... 67
  - 4.3.3.3 Error Rate, Read Length, Data Output,  
and Run Time ..... 67
- 4.4 Biases and Other Adverse Factors That May Affect NGS  
Data Accuracy..... 69
  - 4.4.1 Biases in Library Construction ..... 69
  - 4.4.2 Biases and Other Factors in Sequencing ..... 70
- 4.5 Major Applications of NGS..... 71
  - 4.5.1 Transcriptomic Profiling and Splicing Variant  
Detection (RNA-Seq) ..... 71
  - 4.5.2 Genetic Mutation and Variation Discovery ..... 71
  - 4.5.3 *De novo* Genome Assembly ..... 71
  - 4.5.4 Protein-DNA Interaction Analysis (ChIP-Seq) ..... 71
  - 4.5.5 Epigenomics and DNA Methylation Study (Methyl-Seq)... 72
  - 4.5.6 Metagenomics..... 72
- 5. Early-Stage Next-Generation Sequencing (NGS) Data Analysis:**
  - Common Steps** ..... 73
  - 5.1 Base Calling, FASTQ File Format, and Base Quality Score ..... 73
  - 5.2 NGS Data Quality Control and Preprocessing..... 76
  - 5.3 Reads Mapping..... 78
    - 5.3.1 Mapping Approaches and Algorithms..... 78
    - 5.3.2 Selection of Mapping Algorithms and Reference  
Genome Sequences ..... 80
    - 5.3.3 SAM/BAM as the Standard Mapping File Format ..... 81
    - 5.3.4 Mapping File Examination and Operation ..... 83
  - 5.4 Tertiary Analysis..... 86
- 6. Computing Needs for Next-Generation Sequencing (NGS) Data  
Management and Analysis** ..... 87
  - 6.1 NGS Data Storage, Transfer, and Sharing..... 87
  - 6.2 Computing Power Required for NGS Data Analysis ..... 89
  - 6.3 Software Needs for NGS Data Analysis ..... 90
  - 6.4 Bioinformatics Skills Required for NGS Data Analysis ..... 92

## Section III Application-Specific NGS Data Analysis

|   |     |
|---|-----|
| <b>7. Transcriptomics by RNA-Seq</b> .....  | 97  |
| 7.1 Principle of RNA-Seq.....   | 97  |
| 7.2 Experimental Design.....  | 98  |
| 7.2.1 Factorial Design .....  | 98  |
| 7.2.2 Replication and Randomization .....   | 98  |
| 7.2.3 Sample Preparation .....  | 99  |
| 7.2.4 Sequencing Strategy .....   | 100 |
| 7.3 RNA-Seq Data Analysis .....   | 101 |
| 7.3.1 Data Quality Control and Reads Mapping .....  | 101 |
| 7.3.2 RNA-Seq Data Normalization .....  | 103 |
| 7.3.3 Identification of Differentially Expressed Genes .....                                    | 105 |
| 7.3.4 Differential Splicing Analysis.....   | 107 |
| 7.3.5 Visualization of RNA-Seq Data .....   | 108 |
| 7.3.6 Functional Analysis of Identified Genes .....   | 108 |
| 7.4 RNA-Seq as a Discovery Tool.....  | 109 |
| <b>8. Small RNA Sequencing</b> .....  | 111 |
| 8.1 Small RNA Next-Generation Sequencing (NGS) Data<br>Generation and Upstream Processing ..... | 112 |
| 8.1.1 Data Generation .....   | 112 |
| 8.1.2 Preprocessing .....   | 113 |
| 8.1.3 Mapping .....   | 114 |
| 8.1.4 Identification of Known and Putative Small RNA<br>Species .....                           | 115 |
| 8.1.5 Normalization .....   | 115 |
| 8.2 Identification of Differentially Expressed Small RNAs .....                                 | 116 |
| 8.3 Functional Analysis of Identified Small RNAs .....  | 116 |
| <b>9. Genotyping and Genomic Variation Discovery by Whole<br/>  Genome Resequencing</b> .....   | 119 |
| 9.1 Data Preprocessing, Mapping, Realignment, and Recalibration...                              | 120 |
| 9.2 Single Nucleotide Variant (SNV) and Indel Calling .....                                     | 121 |
| 9.2.1 SNV Calling.....  | 121 |
| 9.2.2 Identification of <i>de novo</i> Mutations.....   | 123 |
| 9.2.3 Indel Calling .....   | 124 |
| 9.2.4 Variant Calling from RNA-Seq Data .....   | 124 |
| 9.2.5 Variant Call Format (VCF) File .....  | 125 |
| 9.2.6 Evaluating VCF Results.....   | 126 |
| 9.3 Structural Variant (SV) Calling.....  | 126 |
| 9.3.1 Read-Pair-Based SV Calling .....  | 126 |
| 9.3.2 Breakpoint Determination.....   | 128 |
| 9.3.3 <i>De novo</i> Assembly-Based SV Detection .....  | 128 |

|            |  |            |
|------------|--|------------|
| 9.3.4      | CNV Detection .....  | 128        |
| 9.3.5      | Integrated SV Analysis.....  | 129        |
| 9.4        | Annotation of Called Variants .....  | 129        |
| 9.5        | Testing of Variant Association with Diseases or Traits.....                                      | 130        |
| <b>10.</b> | <b><i>De novo</i> Genome Assembly from Next-Generation Sequencing (NGS) Reads.....</b>           | <b>131</b> |
| 10.1       | Genomic Factors and Sequencing Strategies for <i>de novo</i> Assembly .....                      | 132        |
| 10.1.1     | Genomic Factors That Affect <i>de novo</i> Assembly.....   | 132        |
| 10.1.2     | Sequencing Strategies for <i>de novo</i> Assembly.....   | 132        |
| 10.2       | Assembly of Contigs.....   | 134        |
| 10.2.1     | Sequence Data Preprocessing, Error Correction, and Assessment of Genome Characteristics.....     | 134        |
| 10.2.2     | Contig Assembly Algorithms .....   | 136        |
| 10.3       | Scaffolding .....  | 138        |
| 10.4       | Assembly Quality Evaluation .....  | 139        |
| 10.5       | Gap Closure .....  | 140        |
| 10.6       | Limitations and Future Development.....  | 140        |
| <b>11.</b> | <b>Mapping Protein–DNA Interactions with ChIP-Seq.....</b>                                       | <b>143</b> |
| 11.1       | Principle of ChIP-Seq.....   | 143        |
| 11.2       | Experimental Design .....  | 145        |
| 11.2.1     | Experimental Control.....  | 145        |
| 11.2.2     | Sequencing Depth.....  | 145        |
| 11.2.3     | Replication .....  | 146        |
| 11.3       | Read Mapping, Peak Calling, and Peak Visualization.....  | 146        |
| 11.3.1     | Data Quality Control and Read Mapping.....   | 146        |
| 11.3.2     | Peak Calling.....  | 149        |
| 11.3.3     | Peak Visualization .....   | 156        |
| 11.4       | Differential Binding Analysis .....  | 156        |
| 11.5       | Functional Analysis.....   | 159        |
| 11.6       | Motif Analysis .....   | 159        |
| 11.7       | Integrated ChIP-Seq Data Analysis.....   | 160        |
| <b>12.</b> | <b>Epigenomics and DNA Methylation Analysis by Next-Generation Sequencing (NGS).....</b>         | <b>163</b> |
| 12.1       | DNA Methylation Sequencing Strategies.....   | 163        |
| 12.1.1     | Whole-Genome Bisulfite Sequencing (WGBS) .....   | 164        |
| 12.1.2     | Reduced Representation Bisulfite Sequencing (RRBS)...  | 165        |
| 12.1.3     | Methylation Sequencing Based on Methylated DNA Enrichment .....                                  | 165        |
| 12.1.4     | Differentiation of Cytosine Methylation from Demethylation Products in Bisulfite Sequencing..... | 166        |
| 12.2       | DNA Methylation Sequencing Data Analysis .....   | 166        |

|            |  |            |
|------------|--|------------|
| 12.2.1     | Quality Control and Preprocessing .....  | 166        |
| 12.2.2     | Read Mapping .....   | 167        |
| 12.2.3     | Quantification of DNA Methylation .....  | 169        |
| 12.2.4     | Visualization of DNA Methylation Data .....  | 170        |
| 12.3       | Detection of Differentially Methylated Cytosines or Regions ...                                      | 172        |
| 12.4       | Data Verification, Validation, and Interpretation.....   | 173        |
| <b>13.</b> | <b>Metagenome Analysis by Next-Generation Sequencing (NGS) .....</b>                                 | <b>175</b> |
| 13.1       | Experimental Design and Sample Preparation .....   | 176        |
| 13.1.1     | Metagenome Sample Collection .....   | 177        |
| 13.1.2     | Metagenome Sample Processing .....   | 177        |
| 13.2       | Sequencing Approaches.....   | 178        |
| 13.3       | Overview of Whole-Genome Shotgun (WGS) Metagenome<br>Sequencing Data Analysis .....                  | 179        |
| 13.4       | Sequencing Data Quality Control and Preprocessing.....   | 181        |
| 13.5       | Taxonomic Characterization of a Microbial Community .....  | 181        |
| 13.5.1     | Metagenome Assembly.....   | 181        |
| 13.5.2     | Sequence Binning .....   | 182        |
| 13.5.3     | Calling of Open Reading Frames (ORFs) and Other<br>Genomic Elements from Metagenomic Sequences ..... | 184        |
| 13.5.4     | Phylogenetic Gene Marker Analysis.....   | 184        |
| 13.6       | Functional Characterization of a Microbial Community.....  | 185        |
| 13.6.1     | Gene Function Annotation.....  | 185        |
| 13.6.2     | Metabolic Pathway Reconstruction.....  | 185        |
| 13.7       | Comparative Metagenomic Analysis .....   | 186        |
| 13.7.1     | Metagenome Sequencing Data Normalization .....   | 186        |
| 13.7.2     | Identification of Differentially Abundant Species<br>or Operational Taxonomic Units (OTUs).....      | 187        |
| 13.8       | Integrated Metagenomics Data Analysis Pipelines .....  | 187        |
| 13.9       | Metagenomics Data Repositories.....  | 188        |

## **Section IV The Changing Landscape of Next-Generation Sequencing Technologies and Data Analysis**

|            |  |            |
|------------|--|------------|
| <b>14.</b> | <b>What Is Next for Next-Generation Sequencing (NGS)? .....</b>  | <b>191</b> |
| 14.1       | The Changing Landscape of Next-Generation Sequencing<br>(NGS).....                                       | 191        |
| 14.2       | Rapid Evolution and Growth of Bioinformatics Tools<br>for High-Throughput Sequencing Data Analysis ..... | 193        |
| 14.3       | Standardization and Streamlining of NGS Analytic Pipelines...  | 195        |
| 14.4       | Parallel Computing.....  | 195        |
| 14.5       | Cloud Computing .....  | 196        |

**Appendix A: Common File Types Used in Next-Generation Sequencing (NGS) Data Analysis** ..... 199

**Appendix B: Glossary** ..... 203

**References** ..... 213



---

## **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  $\mu\text{m}$ , 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,  $\beta$ -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.,  $\text{Na}^+$ ,  $\text{K}^+$ ,  $\text{Ca}^{2+}$ ,  $\text{Cl}^-$ , and  $\text{Mg}^{2+}$ ), 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  $\text{Ca}^{2+}$  represent important intracellular signal), cell energy storage (e.g., in the form of  $\text{Na}^+/\text{K}^+$  cross-membrane gradient), or protein structure/function (e.g.,  $\text{Mg}^{2+}$  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.

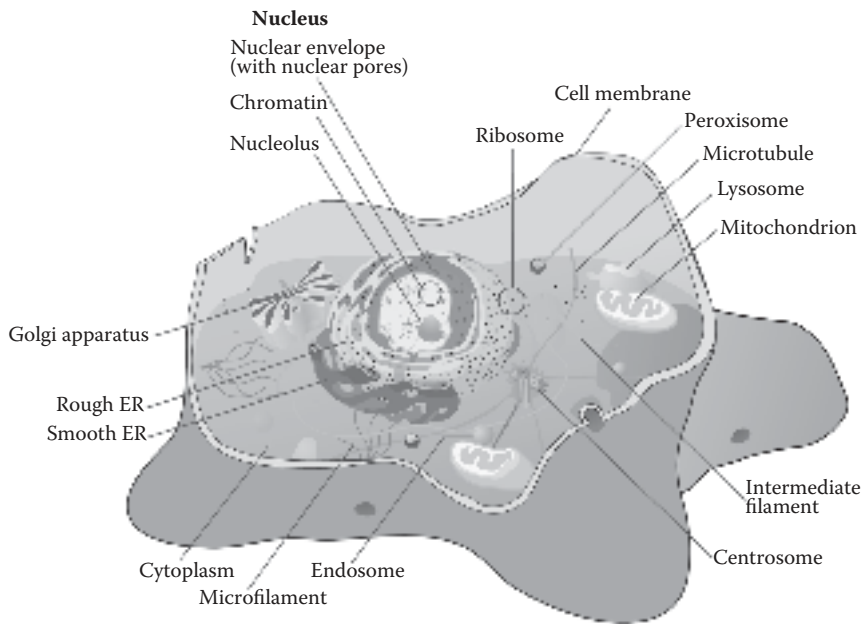
---

## 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  $\text{Na}^+$ ,  $\text{K}^+$ , and  $\text{Ca}^{2+}$  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  $\text{Ca}^{2+}$  in the cytosol is normally kept very low at  $\sim 10^{-7}$  M, whereas in the extracellular space it is  $\sim 10^{-3}$  M. The rushing in of  $\text{Ca}^{2+}$  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  $\text{Ca}^{2+}$  from the cytosol. In  $\text{Ca}^{2+}$ -mediated cell signaling, shortly after entry of the calcium wave into the cytosol, most of the incoming  $\text{Ca}^{2+}$  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



- [\*Routledge Philosophy Guidebook To Derrida on Deconstruction \(Routledge Philosophy Guidebooks\) pdf, azw \(kindle\), epub\*](#)
- **read Plexus (The Rosy Crucifixion, Book 2)**
- [click Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments](#)
- [click The Control Book online](#)
- [Spin Control \(Spin Trilogy, Book 2\) pdf](#)
- [read Inventing Falsehood, Making Truth: Vico and Neapolitan Painting \(Essays in the Arts\) book](#)
  
- <http://toko-gumilar.com/books/Routledge-Philosophy-Guidebook-To-Derrida-on-Deconstruction--Routledge-Philosophy-Guidebooks-.pdf>
- <http://unpluggedtv.com/lib/Captains-of-the-Sands--Penguin-Classics-.pdf>
- <http://interactmg.com/ebooks/Development-of-Professional-Expertise--Toward-Measurement-of-Expert-Performance-and-Design-of-Optimal-Learning-E>
- <http://interactmg.com/ebooks/The-Control-Book.pdf>
- <http://sidenoter.com/?ebooks/Spin-Control--Spin-Trilogy--Book-2-.pdf>
- <http://tuscalaural.com/library/All-the-Light-We-Cannot-See--A-Novel.pdf>