Chapter 1
Basics of Molecular Biology for Next-Generation Sequencing

Abstract Organisms can be divided into simple (or unicellular) organisms and complex (or multicellular) organisms. Both simple and complex organisms share major cellular and biological processes that are mediated through proteins and nucleic acids. Proteins are the molecules responsible for every structural or biological process achieved inside living cells or living organisms, while nucleic acids encode the necessary information required for the building and regulation of proteins. In this chapter, we present some basics of molecular biology to provide readers without a biological background with an adequate introduction to the subject. These basics would greatly aid a lay audience in understanding this and other computational biology resources and textbooks. Readers with a firm biological background may choose to skip this chapter.

1.1 Molecular Biology

Molecular Biology can be defined as the study of the molecular principles that govern and regulate biological processes. These biological processes, including the replication, transcription, and translation of genetic material, require the existence, interaction, and regulation of thousands of proteins and their corresponding genes. Thus, the focus of molecular biology starts at divulging and understanding the structure and function of these proteins/genes and continues with the study of the interactions and regulation processes between them. Additionally, the effects of their absence and mutational changes should also be understood [1, 2].

There are major differences that exist between different organisms at the molecular level, which is critical to the diversity observed between simple and complex organisms. In general, organisms are classified into two major classes according to

S. El-Metwally et al., *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*, SpringerBriefs in Systems Biology 7,
DOI 10.1007/978-1-4939-0715-1_1, © The Authors 2014

Table 1.1 Main differences between prokaryotes and eukaryotes

	Prokaryotes	Eukaryotes
Nucleus	Absent	Present
Chromosomes number	One	More than one
DNA	Circular protein-free	Liner, with chromatin
Membrane bound nucleus	Absent	Present
Telomeres	Absent (not needed)	Present
Endoplasmic reticulum	Absent	Present
Mitochondria	Absent or rare	Present
Ribosome	Small	Large
Mitosis	No	Yes
Cell wall	Chemically complex and always present	Simple and only in plants
Cell size	Small (<5 μm)	Large (>10 μm)
Unicellular/multicellular	Always unicellular	Often multicellular
Cytoskeleton	Absent	Present
Reproduction	Always asexual	Asexual or sexual
Metabolic pathways	Varity of pathways	Common set of pathways
Examples	Bacteria and archaea	Plants, animals, fungi

their cellular and genomic structures. Hence, simple organisms with a unicellular structure are called prokaryotes, while more complex organisms that are usually multicellular are called eukaryotes. The differences between prokaryotes and eukaryotes are not simply limited to the number of structure forming cells, but include several other aspects that are of great importance to the topic of this book. A major difference between them is highlighted by the structure of the genome, which is circular and protein-free with the noticeable absence of telomeres in prokaryotes. On the other hand, the eukaryotic genome may possess telomeres and proteins, which is vital for chromatin formation. These differences hold considerable influence in the processes of genome sequencing and the assembly of sequenced genomes, as will be discussed later. The major differences between prokaryotes and eukaryotes are summarized in Table 1.1 [1, 2].

Since the field of molecular biology concerns a comprehensive understanding of the structures of molecules and interactions between them, it overlaps with other fields such as biochemistry. Furthermore, with the advent of modern experimental and analytical tools such as next-generation genome sequencing (NGS) and liquid chromatography mass spectrometry (LC-MS/MS) that generate huge amounts of data due to its high-throughput nature [3], molecular biology developed the need for specialized computational and informatics tools to analyze and process this information. As a result, molecular biology has since considerably overlapped with the field of computational biology and bioinformatics [1, 2].

To develop an understanding of the details of the cellular processes at the molecular level, three particular types of molecules need to be better appreciated: deoxyribonucleic acid (DNA), ribonucleic acid (RNA), and proteins. In the next sections, we will provide a brief introduction to each of these structures.

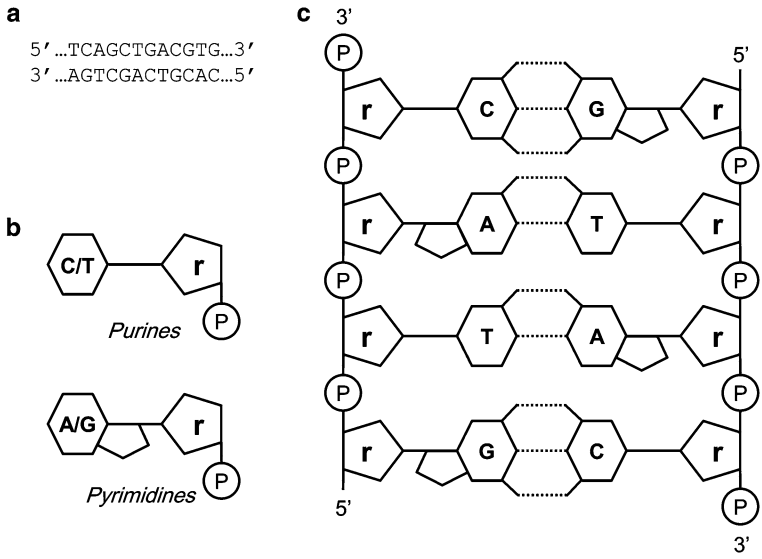

Fig. 1.1 DNA structure. (**a**) Example of complementary bases of 12 base pairs (bp). (**b**) Schematic representation of the DNA nucleotides from purines (*single ring*) and pyrimidines (*double ring*). (**c**) Structure of DNA double-stranded helix. (*r*) Deoxyribose sugar, (*P*) the phosphate group, and the *dotted lines* represent the hydrogen binds between the nucleotides of the two strands

1.1.1 Deoxyribonucleic Acid

DNA is one of the two nucleic acids that exist in living organisms and play a crucial role in cell biology. DNA is a double-stranded chain formed by the repetition of similar basic units called nucleotides. Each nucleotide consists of a sugar molecule called 2'-deoxyribose, phosphate residue, and a nitrogenous base. The sugar molecule contains five carbon atoms (arranged from 1′ to 5′). The phosphate residue is important for creation of the chain through the connecting the 3′ carbon atom of the sugar molecule of one nucleotide with the 5′ carbon atom of the sugar molecule of the next nucleotide. Therefore, the DNA molecule has an orientation that begins at the 5′ end and ends at the 3′ end (Fig. 1.1a). This feature is especially important in DNA sequencing and during sequence assembly as will be discussed later. All DNA sequences available in databases, literature, or books are written from 5′ to 3′ unless otherwise mentioned [2, 4].

The nitrogenous bases are attached to the 1' carbon atom of the nucleotide. There are four types of bases: adenine (A), guanine (G), cytosine (C), and thymine (T). Consequently, there are four types of nucleotides described. Adenine and guanine belong to a group called purines, while cytosine and thymine belong to the pyrimidine group (Fig. 1.1b). When DNA forms the double strand, a nucleotide from the purine group is bound to a nucleotide from the pyrimidine group in the other strand. Adenine is always bound to thymine, while guanine is always bound to cytosine

with weak hydrogen bonds (Fig. 1.1c). This allows the two strands to be tied together and keeps the distance between them the same, and as a result, the double strand forms the familiar DNA double helix shape. These nucleotide pairs are known as complementary bases or Watson-Crick base pairs, and are used as units to measure DNA length. For instance, a DNA sequence of 2,000 nucleotides is referred to as 2,000 base pairs (bp) or 2 kbp [2].

In computational biology and bioinformatics, a DNA sequence is considered as a string (sequence of characters) consisting of a combination of the four letters A, G, C, and T. Therefore, from the complementary bases, the reverse strand of DNA (which starts at 3′ and ends at 5′) can always be predicted for any given DNA sequence by replacing A, T, C, and G with T, A, G, and C in the string, respectively. In fact, this is very similar to what occurs in the living cell, where each strand of the DNA molecule constructs its complementary strand. This process allows the DNA to replicate and make two identical copies of the total DNA during cell division to create two cells, each of which carry an identical copy of the genomic DNA [2, 4].

1.1.2 Ribonucleic Acid

The second type of nucleic acid that exists in living cells is RNA. RNA has the same general structure and properties of DNA with certain major differences. Unlike DNA, RNA is single stranded with the sugar molecule in its nucleotides being ribose rather than 2'-deoxyribose. Furthermore, the thymine (T) base is absent and another base called uracil (U) exists instead. As a result, uracil (U) binds with adenine (A) in a similar fashion to the thymine (T) binding observed in DNA. Therefore, the RNA sequence can be predicted from the DNA sequence and vice versa through the substitution of A, U, C, and G, by T, A, G, and C, respectively. However, a major difference between DNA and RNA is that the former performs one principle function (the encoding of the genetic information of the organism) while several different types of RNA exist to accomplish a variety of tasks. It is also important to note that RNA can also exist in double strands. In some viruses, it has been observed that the genetic material is double-stranded RNA (ds-RNA) rather than DNA [2, 5]. This viral ds-RNA plays an important role in the detection of viruses by immune systems such as in humans [6].

There are three main types of RNA that play a crucial role in the protein synthesis process: the messenger-RNA (mRNA), ribosomal-RNA (rRNA), and transfer-RNA (tRNA). Furthermore, several additional forms of RNAs exist in the cell to perform critical posttranscriptional modifications and regulatory functions (Table 1.2). Here, we will briefly introduce the major types of RNA that are important to the next-generation sequencing field.

The mRNA results from DNA transcription (Fig. 1.2a), a process that creates a strand of RNA that complements a certain part of the genomic DNA (see below). This RNA is encoded such that it is actually carrying all the information needed to create the protein through the translation of the "encoded" RNA sequence into an

Table 1.2 Examples of RNAs with regulatory and posttranscriptional modification functions

RNA type	Function(s)	Organism	References
Small nuclear RNA (snRNA)[PTM]	Splicing	Eukaryotes and archaea	[20]
Y RNA[PTM]	RNA processing, DNA replication	Eukaryotes (animals)	[21]
Telomerase RNA[PTM]	Telomere synthesis	Most eukaryotes	[22]
Small nucleolar RNA (SnoRNA)[PTM]	Nucleotide modification of RNAs	Eukaryotes and archaea	[23]
Antisense RNA (aRNA)[R]	Transcriptional attenuation, mRNA degradation and stabilization	All organisms	[24]
CRISPR RNAs[R]	Resistance to bacteriophage, prevent plasmid conjugation	Bacteria	[25]
Trans-encoded base pairing sRNAs[R]	Regulation of translation and stability of target mRNAs	Bacteria	[25]
Cis-encoded base pairing sRNAs[R]	Expression regulation	Bacteria	[25]
Small interfering RNA (siRNA)[R]	Gene regulation	Eukaryotes	[26]

[R]Regulatory RNA
[PTM] Posttranscriptional modification RNA

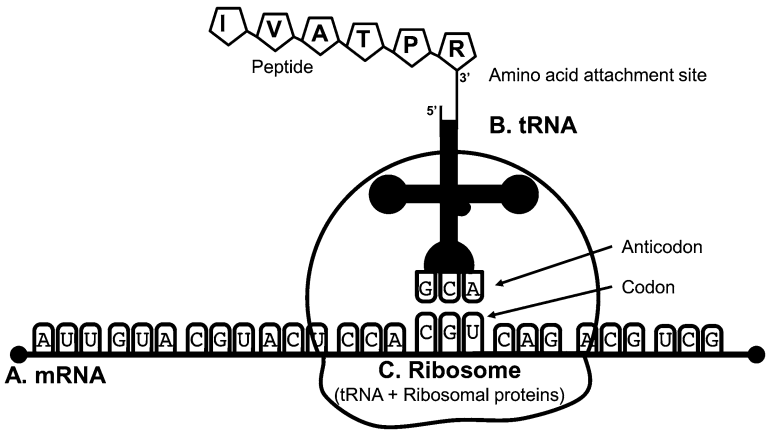

Fig. 1.2 Types of RNA and Translation process. (**a**) messenger-RNA (mRNA). (**b**) transfer-RNA (tRNA). (**c**) ribosomal-RNA (rRNA)

amino acid sequence as will be described in more detail later. In prokaryotes, the mRNA is directly translated into a protein under most circumstances, while in eukaryotes, the process is more complex due to the fact that the mRNA consists of coding regions called "exons" and noncoding regions called "introns". The removal of introns from the mRNA is crucial for the creation of mature mRNA that will be translated to a protein, a process named mRNA splicing. Therefore, mRNA splicing

is a major reason for next-generation sequence assembly to be more complicated in eukaryotes in comparison to prokaryotes, where the splicing process is almost absent. Following the process of splicing in eukaryotes, the mRNA is exported to the cytoplasm to be translated. Due to the absence of the nucleus and other cellular compartments, the translation of the mRNA in prokaryotes starts during the transcription process [2, 5].

The tRNA is responsible for the transfer of amino acids to the rRNA and mRNA during the translation process. It consists of a small chain of around 80 nucleotides with a special sequence called anticodon, and has another site for amino acid attachment (Fig. 1.2b). Each tRNA is specially bound with certain amino acids via its amino acid binding site, eventually transferring this amino acid for addition to the protein that is being created. Each three nucleotide sequence in the mRNA represents one codon that corresponds to a particular amino acid. The tRNA anticodon region represents the complementary sequence of these three nucleotides. The position of the amino acid in the protein that is being translated is determined by the anticodon of the tRNA that complements the mRNA codon [2]. The mRNA codon and the tRNA anticodon regions bind with each other through hydrogen bonds, allowing the amino acids to form peptide bonds between each other and therefore, allowing the polypeptide chain (the protein being translated) to grow. tRNA bound to amino acids are termed charged tRNA or aminoacylated tRNA, while amino acid free tRNA are called uncharged tRNA [4, 5].

The rRNA is formed in the nucleus and exported to the cytoplasm where it can bind to the mRNA for translation into protein (Fig. 1.2c). Ribosomes can bind to multiple mRNA at the same time. rRNA is the most abundant type of RNA, numbering up to 80 % of the total RNA isolated from a typical eukaryotic cell [5, 7].

In relation to other forms of RNA, Table 1.2 presents alternate types of these structures that have functions other than protein synthesis e.g., posttranscriptional modifications and regulatory function.

1.1.3 Proteins

Proteins are the result of mRNA translation and form a significant portion of the structures within living cells. Almost all the structural, functional, and regulatory tasks in the cell are performed through the action of proteins. A protein is a chain of amino acids that are joined together with chemical bonds called peptide bonds or amide bonds. Each amino acid consists of a central carbon atom, a hydrogen atom, an amino group (NH_2), a carboxyl group (COOH), and a side chain which distinguishes each of the 20 naturally existing amino acids from each other. The peptide bond is formed between the carboxyl group of one amino acid and the amino group of the other, releasing a water molecule. During the translation process, protein is synthesized through the arrangement of amino acids next to each other as encoded in the genetic information, and then peptide bonds are formed between them sequentially. A short chain of amino acids is called a peptide, where the amino acids are

referred to as residues. Therefore, a protein of 200 amino acids can be described as a polypeptide chain with 200 residues [2, 7].

Similar to DNA and RNA having directions (5′ and 3′), a protein also possesses direction as one of its ends will always end with an amino group while the other ends with a carboxyl group. The end with the amino group is called the N-terminal while the end with the carboxyl group is called the C-terminal. The nitrogen atoms, carbon atoms, and CO- form the protein's backbone, a line that begins from the N-terminal through the C-terminal. Unlike DNA and RNA, proteins are not linear starches of sequences. In fact, a protein's sequence (amino acid order) represents the protein's primary structure. The interactions between the backbone atoms forms a "local structure" termed the protein's secondary structure. An additional layer of folding gives the protein a unique three-dimensional structure called the protein's tertiary structure. In a similar manner, yet another level of packing of the protein or with a group of different proteins is known as the protein's quaternary structure [2, 7].

1.2 The Central Dogma of Molecular Biology

The central dogma of molecular biology is a description of the information flow in biological systems. It was first introduced in the middle of the twentieth century by Francis Crick, and then published in 1970 [8]. The central dogma explains a framework of information flow from genetic material to the synthesis of proteins that perform both functional and structural roles in cells. With developments and advancements in biological research methods, analysis instruments, and imaging devices, the details of the original central dogma had been altered (e.g., the addition of the RNA splicing step). Nevertheless, its main description of the basic framework remains valid today. The central dogma states three levels of information flow, from DNA (genes) to RNA (transcripts) to amino acids (proteins) in sequential steps [8]. Here, we will describe two basic steps of the central dogma, transcription and translation, as they are crucial for understanding the subsequent chapters in this book.

1.2.1 Transcription

DNA is the main source of genetic information in organisms, with some notable exceptions where the genetic material may be composed of RNA as in the case of certain viruses [9]. In accordance with the central dogma of molecular biology, genetic information transferred from one cell to another through DNA replication process, which is the first level of information transfer. The next step is transcription, which is the process of creating a piece (sequence or stretch) of mRNA that contains the genetic information stored in corresponding DNA [8]. Transcription is an enzymatic process that is managed by RNA polymerase that sequentially attaches nucleotides to the end of the newly synthesized RNA molecule. Furthermore, the

process is regulated by a group of proteins known as transcription factors that bind to specific DNA sequences and control the transcription process [10].

As mentioned in the RNA section above, in prokaryotes the mRNA is directly translated into protein, while in eukaryotes, the transcription process is more complex. The genetic structure in eukaryotic cells is more complicated in comparison to prokaryotes due to the existence of exons, entrons, and untranslated regions (UTRs). Thus, another process termed mRNA splicing follows transcription. The mRNA splicing process removes introns from the mRNA and joins the exons to create mature mRNA that is ready for translation into protein [11]. Splicing can also result in several mature mRNAs from one mRNA, resulting in several proteins from a single gene accordingly termed alternative splicing variants [12].

1.2.2 Translation

The translation process, also known as the protein synthesis process, involves "translating" the genetic code, which was transferred as nucleotides from the DNA to mRNA into a chain of amino acids (protein). The translation process requires three types of RNA: mRNA, tRNA, and rRNA. The mRNA, as explained above, carries the information required to build the target protein. The tRNA transfers the amino acids sequentially one by one following the encoding of information into the mRNA. Lastly, the rRNA is a complex of two subunits that reads the mRNA code and adds the amino acids in the same order encoded in the DNA (Fig. 1.2) [13].

Genetic information is encoded into the mRNA in triplet codons, where three nucleotides in the mRNA correspond to a specific amino acid. Typically, the reading initiates with an AUG (adenine–uracil–guanine) or imitator methionine codon and ends with a UAA, UGA, or UAG stop codon. Therefore, the rRNA reads the triplet codons and attaches the aminoacylated tRNA (tRNA with added amino acid) to the matching triplet anticodon. Subsequently, a peptide bond joins the newly added amino acid with the preceding one. As the amino acid chain grows, it starts to fold in a specific conformation that confers a three-dimensional shape to the final protein. The translation of mRNA to protein in prokaryotic cells usually occurs in the same vicinity as the transcription process due to the fact that prokaryotic cells do not possess a nucleus. In contrast, transcription takes place in the nucleus of eukaryotic cells after which the mRNA is transferred to the cytoplasm where the translation process can be achieved [1, 13].

1.3 Genetic Information Sources Targeted by Sequencing

The main target of sequencing technologies is to decode the genetic information stored in the molecules of the organisms. With modern sequencing technologies, the genetic information sources became increasingly ubiquitous, involving a

myriad of molecules that lead to the development of an organism. Furthermore, special techniques have been utilized to decode the nucleotide sequences that interact/bind with other non-DNA or RNA molecules such as proteins. However, in the following paragraphs we will introduce four major types of genetic information sources that are the primary targets of available sequencing methods and platforms. In later chapters, we will also elaborate on the details of several other sources and applications.

1.3.1 The Genome

The genome represented the main target of sequencing efforts as it contained the entire genetic information of an organism. Most genomes are DNA with the exception of certain viral genomes that are RNA-based. In prokaryotes, the genome simply consists of one circular chromosome with most of its sequence represented by coding sequence which can be transcribed to RNA and then translated to proteins. In eukaryotes, the genome is far more complex, existing inside a nucleus and consisting of several pieces each of which is a separate chromosome. In most cases, the eukaryotes carry two copies of each chromosome in each cell except for the gametes in sexually reproductive organisms, which carry only a single copy. Furthermore, the genomes of eukaryotes contain noncoding regions and long intragenic stretches that are not known to encode any genetic information. Such complications bring greater challenges to whole genome sequencing (WGS) technologies and methods as well as the assembly and annotation of the sequenced genomes [1, 4].

1.3.2 The Transcriptome

The transcriptome is the entire set of RNA molecules within a single cell or population of cells. It includes the three main types of RNA (mRNA, tRNA, and rRNA) as well as the short and noncoding RNAs [1, 5]. The transcriptome represents the genes expressed at a given time (such as the time of sample collection). Therefore, it may dynamically change based on age, surroundings, media condition, and treatment of the cell/cell population. Traditionally, gene expression or the transcriptome is measured using DNA microarray techniques that allow for the measurement of a large number of genes simultaneously [14]. However, transcriptome sequencing, also known as RNA sequencing or RNA-seq, became the technology of choice for gene expression studies as its coverage is broader and allows the investigation of known and new transcripts, unlike DNA microarray techniques [15]. Similar to DNA sequence, there are two methods to assemble transcriptome sequence reads in the next-generation environment. These include the utilization of reference sequences or de novo transcriptome assembly, both of which will be discussed later on in this book.

1.3.3 The Exome

WGS identifies the sequences of all genomic DNA in the organism, including coding and noncoding sequences. In many cases, the noncoding regions of the genome are not important to a particular study. For instance, in studies targeting the identification of mutation-based diseases, the investigation of noncoding regions are less critical since 85 % of mutations exist in the coding regions (exons). Therefore, methods for sequencing the total number of exons of the genome were developed to target the whole transcribed exons (or exome) while excluding the entire population of introns. The human exome, for instance, represents 1 % of the human genome [16], which is reflective of the time and costs involved for sequencing as well as the complexity of the analysis required for the assembly and annotation of the associated reads. Therefore, several studies rely on whole exome sequencing (WES) instead of WGS to identify mutations in cancer and inherited human disorders such as Mendelian disorders [17].

1.3.4 The Metagenomes

The Metagenomes are the genomes of several organisms that coexist in a certain environment. They are mainly used during environmental studies such as sequencing and identification of organisms in an environmental sample (e.g., water or soil). Additionally, they may also be utilized in health investigations such as the study of the gut flora of humans and other organisms [18]. Typically, this field is referred to as metagenomics or environmental genomics where the study targets the sequencing and identification of all genes of all member organisms that exist in an environmental or biological sample. Since the standard sequencing procedures of model organisms normally employ cultured clones, metagenomics represents an opportunity to explore the biology and diversity of wild microorganisms in a culture-independent environment [19]. Despite the availability of several types of metagenome sequencing in the first- and next-generation methodology, the sequencing, assembly, and annotation of metagenomes remain a formidable challenge.

References

1. Alberts B, Johnson A, Lewis J, Raff M, Roberts K et al. (2007) Molecular Biology of the Cell 5th Edition. Garland Science, New York, USA
2. Setubal C, Meidanis J (1997) Introduction to Computational Molecular Biology. PWS Publishing, Pacific Grove, CA, USA
3. Helmy M, Sugiyama N, Tomita M, Ishihama Y (2010) Onco-proteogenomics: a novel approach to identify cancer-specific mutations combining proteomics and transcriptome deep sequencing. Genome Biol 11. Doi 10.1186/Gb-2010-11-S1-P17

4. Ridley M (2013) Genome: The Autobiography of a Species in 23 Chapters Harper Perennial New York, USA
5. Yarus M (2012) Life from an RNA World: The Ancestor Within. Harvard University Press Cambridge, MA, USA
6. Helmy M, Gohda J, Inoue J, Tomita M, Tsuchiya M et al. (2009) Predicting novel features of toll-like receptor 3 signaling in macrophages. PLoS One 4 (3):e4661. doi:10.1371/journal.pone.0004661
7. Meyers RA (ed) (2006) Proteins Wiley-Blackwell, Hoboken, NJ, USA
8. Crick F (1970) Central dogma of molecular biology. Nature 227 (5258):561-563
9. Patton JT (ed) (2008) Segmented Double-stranded RNA Viruses: Structure and Molecular Biology. Caister Academic Press, Poole, UK
10. Lee TI, Young RA (2000) Transcription of eukaryotic protein-coding genes. Annu Rev Genet 34:77-137. doi:10.1146/annurev.genet.34.1.77
11. Roy SW, Gilbert W (2006) The evolution of spliceosomal introns: patterns, puzzles and progress. Nat Rev Genet 7 (3):211-221. doi:nrg1807
12. Tilgner H, Knowles DG, Johnson R, Davis CA, Chakrabortty S et al. (2012) Deep sequencing of subcellular RNA fractions shows splicing to be predominantly co-transcriptional in the human genome but inefficient for lncRNAs. Genome Res 22 (9):1616-1625. doi:10.1101/gr.134445.111
13. Griffiths JFA, Wessler SR, Lewontin RC, Carroll SB (2008) Introduction to Genetic Analysis (Ninth Edition). W. H. Freeman and Company, New York, USA
14. Bowtell D, Sambrook J (2002) DNA Microarrays: A Molecular Cloning Manual. Cold Spring Harbor Lab Press, New York, USA
15. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10 (1):57-63. doi:10.1038/nrg2484
16. Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW et al. (2009) Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461 (7261):272-276. doi:10.1038/nature08250
17. Kuhlenbaumer G, Hullmann J, Appenzeller S (2011) Novel genomic techniques open new avenues in the analysis of monogenic disorders. Hum Mutat 32 (2):144-151. doi:10.1002/humu.21400
18. Eisen JA (2007) Environmental shotgun sequencing: its potential and challenges for studying the hidden world of microbes. PLoS Biol 5 (3):e82. doi:1544-9173-5-3-e82 [pii]
19. Hugenholtz P, Goebel BM, Pace NR (1998) Impact of culture-independent studies on the emerging phylogenetic view of bacterial diversity. J Bacteriol 180 (18):4765-4774
20. Lui L, Lowe T (2013) Small nucleolar RNAs and RNA-guided post-transcriptional modification. Essays Biochem 54:53-77. doi:10.1042/bse0540053
21. Hall AE, Turnbull C, Dalmay T (2013) Y RNAs: recent developments. Biomolecular Concepts 4 (2):103–110. doi:10.1515/bmc-2012-0050
22. Zhou J, Ding D, Wang M, Cong YS (2014) Telomerase reverse transcriptase in the regulation of gene expression. BMB Rep 47 (2):8-14. doi:2638
23. Bachellerie JP, Cavaille J, Huttenhofer A (2002) The expanding snoRNA world. Biochimie 84 (8):775-790. doi:S0300908402014025
24. Brantl S (2002) Antisense-RNA regulation and RNA interference. Biochim Biophys Acta 1575 (1-3):15-25. doi:S0167478102002804
25. Waters LS, Storz G (2009) Regulatory RNAs in bacteria. Cell 136 (4):615-628. doi:10.1016/j.cell.2009.01.043
26. Ahmad K, Henikoff S (2002) Epigenetic consequences of nucleosome dynamics. Cell 111 (3):281-284. doi:S0092867402010814