Big Data Research 2 (2015) 12–18

Contents lists available at ScienceDirect

Big Data Research

www.elsevier.com/locate/bdr

The Current Status and Challenges in Computational Analysis of Genomic Big Data

Yiming Qin [a], Hari Krishna Yalamanchili [a], Jing Qin [a,b], Bin Yan [c,d,e], Junwen Wang [a,b,*]

[a] Centre for Genomic Sciences and Department of Biochemistry, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China
[b] Shenzhen Institute of Research and Innovation, The University of Hong Kong, Shenzhen, China
[c] Stem Cell & Regenerative Medicine Consortium, LKS Faculty of Medicine, The University of Hong Kong, Hong Kong, China
[d] Department of Physiology, The University of Hong Kong, Hong Kong, China
[e] Laboratory for Food Safety and Environmental Technology, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

ARTICLE INFO

Article history:
Received 31 October 2014
Received in revised form 18 January 2015
Accepted 11 February 2015
Available online 25 February 2015

Keywords:
Gene regulatory networks
Next generation sequencing
OMICS
Integrative data analysis
Genomic big data

ABSTRACT

DNA, RNA and protein are three major kinds of biological macromolecules with up to billions of basic elements in such biological organisms as human or mouse. They function at molecular, cellular and organismal levels individually and interactively. Traditional assays on such macromolecules are largely experimentally based, which are usually time consuming and laborious. In the past few years, high-throughput technologies, such as microarray and next-generation sequencing (NGS), were developed. Consequently, large genomic datasets are being generated and computational tools to analyzing these data are in urgent demand. This paper reviews several state-of-the-art high-throughput methodologies, representative projects, available databases and bioinformatics tools at different molecular levels. Finally, challenges and perspectives in processing genomic big data are discussed.

1. Introduction

Why do some siblings look alike but have different heights and/or blood types? How do people get old and suffer from diseases? It's so amazing that nature always gets its rules to "design" living organisms. The rules are so stringent that humans are similar to each other, but they are also flexible enough to allow differences between any two individuals.

In biological sciences, an observable trait or characteristic, such as hair color or body height, is called a phenotype. Phenotypes are the results of both genetics and environment, as well as their interactions. Even though we still know very little about how environments affect phenotypes, scientists have a relatively much better understanding in the genetic factors. In 1952, Alfred Hershey and Martha Chase found that DNA is the hereditary material in any organism [1]. DNA, deoxyribonucleic acid, is a double-helix macromolecule. Its two strands are composed of numerous linear-arranged nucleotides. There are four kinds of nucleotides in total, which are distinguished by their nitrogen-containing nucleobases, guanine (G), adenine (A), thymine (T), and cytosine (C). In a DNA molecule, there are fragments called genes that can be coded for proteins, the basic functional elements in an organism. Such protein coding genes only occupy less than 2% of all DNA sequences, but other nearly 98% of the sequences, though not directly coded for proteins, are not junk. Recent studies found that they contain various gene regulatory elements like enhancers and silencers, or code for noncoding RNAs, such as microRNAs, small nuclear RNA and long noncoding RNAs (lncRNA) [2–4]. In Eukaryotes, which have a nucleus in each of their cells, DNAs coil around proteins called histone and are densely organized into chromosomes, and stably exist in the nucleus (see Fig. 1).

Another important biological macromolecule is called ribonucleic acid (RNA), which is the product of DNA transcription. RNAs are transcribed in the nucleus, and most of them are exported to the cytoplasm. They are single-stranded chains of nucleotides, distinguished by guanine (G), adenine (A), cytosine (C) and uracil (U). There are several kinds of RNA, such as messenger RNA (mRNA), transfer RNA (tRNA) and ribosomal RNA (rRNA). Messenger RNA is the most imperative modality of RNA. It can be further translated into proteins with the help of tRNA. Some other forms of RNAs include microRNAs and lncRNAs. Although most of them cannot be translated into functional proteins, they still play their roles in the regulation of gene expression.

Proteins, the end-products of protein-coding genes, are the most essential functional macromolecules. Amino acids are the basic elements of proteins. They are linked to each other with peptide bonds and to form a polypeptide chain. The sequence of

* Corresponding author. Tel.: +852 2831 5075; fax: +852 2855 1254.
E-mail address: junwen@hku.hk (J. Wang).

http://dx.doi.org/10.1016/j.bdr.2015.02.005
2214-5796/

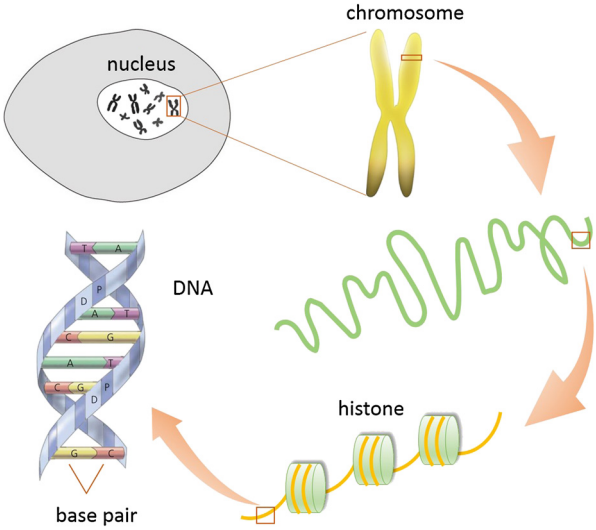

Fig. 1. From nucleotide to nucleus. Note: Base pair is a pair of nucleotides from two chains linked by hydrogen bonds. Such base pairs are formed according to a certain pattern, A–T, C–G. The abbreviation of base pair, bp, refers to the length unit of double-helixed DNA sequence.

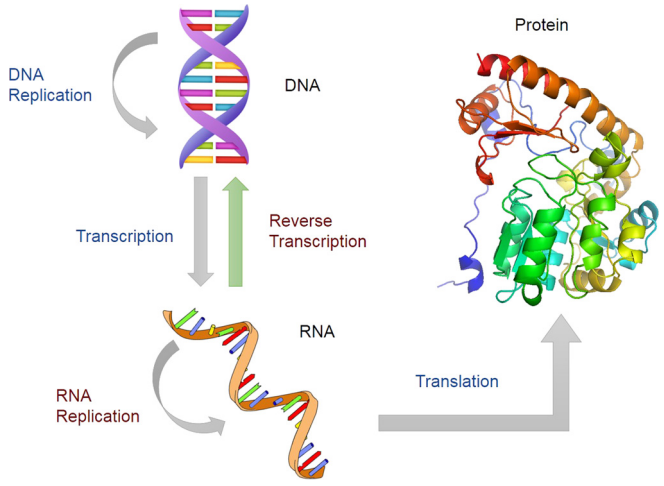

Fig. 2. Flow of information among DNA, RNA and protein.

amino acids on a polypeptide chain is the primary structure of proteins and the chain can be folded to form secondary and tertiary structures. The function of a protein is predominantly associated with its structure [5]. Proteins play vital roles in almost every cellular functions [6], including gene regulation, catalysis, immunity, growth, transport, signaling, and cell differentiation. Thus, comprehending protein functions will help us understand various cellular processes and reveal essential biological pathways.

In 1956, Francis Crick stated that the genetic information flows within biological systems as the central dogma [7], and revised it in 1970 [8]. The common statement of the central dogma is on the flow of sequence information, which draws connections between the sequences of nucleotides in DNA and RNA and of amino acids in proteins. Generally, information in DNA can be replicated with DNA replication, can flow into RNA by transcription and then flow into protein by translation. In some special cases, information in RNA can also flow back into DNA with reverse transcription, or replicate with RNA replication. The process is illustrated in Fig. 2.

The genetic information flow is quite stringent because of the proof reading and repair system in high organisms. However, minor mistakes are inevitable, such as genetic variations at DNA level, gene expression changes at RNA level and amino acid alterations at protein level. Some of them may not affect the function of the end-product protein, but many of them may cause severe diseases. In this regard, it is important to read on these sequence information at different molecular levels and find the real cause of the diseases. However, since there are tens of thousands of different DNA, RNA and protein molecules in a single cell, it is hard to get such large-scale sequence information, let alone to analyze them with traditional approaches. Fortunately, with the advance of high-throughput technologies and consequent OMICS[1] data analysis tools, scientists have the opportunity to retrieve and decode the information for many genes in parallel. Especially with the completion of the Human Genome Project in 2003, scientist obtained nearly 99% of the human genome sequence (3.109 Gbp[2]) and many other organisms' genome sequences [9]. Recently, NGS and even third-generation sequencing (TGS) emerged and have become prevalent, which greatly reduced sequencing prices so that generating multi-dimensional high-throughput data becomes a routine in biomedicine and biological sciences (Fig. 3). One major challenge is how to integrate the various OMICS data to interpret complex biological functions.

In the following sections, we will discuss the available methodologies and databases on NGS data generated from different molecular levels. More importantly, we will present research gaps and challenges we are currently facing, and encourage researchers, particularly these from mathematics, statistics and computer science areas, to mine these genomic big data for biomedical research.

2. Genomics and genetic variants

The whole genetic material of an organism is called a genome. For most living organisms, that is a complete set of DNA. The analysis of genome with bioinformatics approaches is called genomics. Genomics data are usually large in size. For example, the human genome size is 3.2 Gbp and a mouse's is 2.7 Gbp. Such large sequences are not practical to be obtained serially in a single read. Usually, DNAs are "cut" into numerous small pieces and sequenced. Then these small DNA fragments are assembled together and stored as reference genome.

There are several genomic databases available publically for research use. The NCBI Genome database (http://www.ncbi.nlm.nih.gov/genome/), maintained by the US National Institutes of Health (NIH), is the most commonly used one. It contains whole genome sequences or assemblies for over 10,000 organisms of eukaryotes, prokaryotes, viruses, as well as plasmids and organelles. Sequencing data and their annotations can be downloaded freely. UCSC genome browser (https://genome.ucsc.edu/cgi-bin/hgGateway) is another popular database with a visualization function for genomes.

Whole genome/exome sequencing greatly facilitates the detection of genetic variants, including small variations (with range <1 kbp, like Single Nucleotide Polymorphisms (SNPs), microsatellites, small indels[3]) [11] and structural variations (with range around 1 Kbp $\sim$ 3 Mbp, like Copy-Number Variants (CNVs), insertions, deletions, inversions and translocations[4]) [12,13]. Such genetic variants are sources of phenotypic polymorphisms and onsets of diseases. It is no doubt that revealing the function of them

[1] OMICS, a general designation for biological studies with names ending in "-omics", such as genomics, transcriptomics, proteomics, interactomics and epigenomics.

[2] 1 Gbp = 1000 Mbp = 1,000,000 Kbp = 1,000,000,000 bp.

[3] SNP is a single base pair genetic variant with the frequency larger than 1% in a population. Microsatellite is a repeating sequence of 2–5 base pairs of DNA. Indels are the insertions or deletions (of nucleotides) in a small scale in the DNA.

[4] CNVs are the variations in the number of copies of one or more sections of the DNA. Inversion is that a segment of a chromosome is reversed end to end. Translocations are rearrangements of parts between nonhomologous chromosomes.