Big Data Research ••• (••••) •••–•••

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 xxxx

*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/

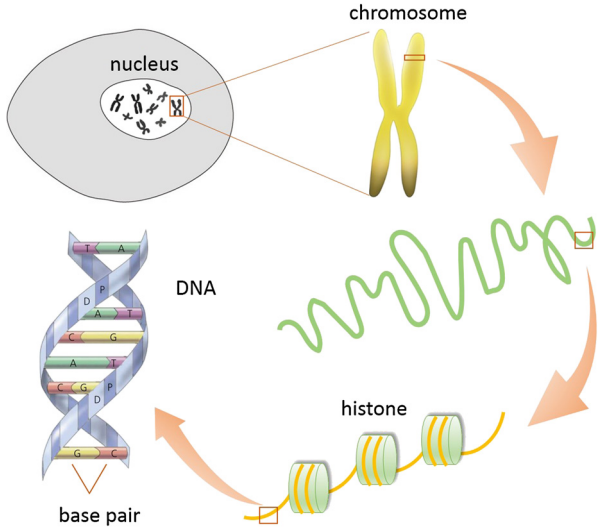


**Fig. 1.** From nucleotide to nucleus. Note: Base pair is a pair of nucleotides from two chains linked by hydrogen bonds. Such base pairs are formed according to a certain pattern, A–T, C–G. The abbreviation of base pair, bp, refers to the length unit of double-helixed DNA sequence.

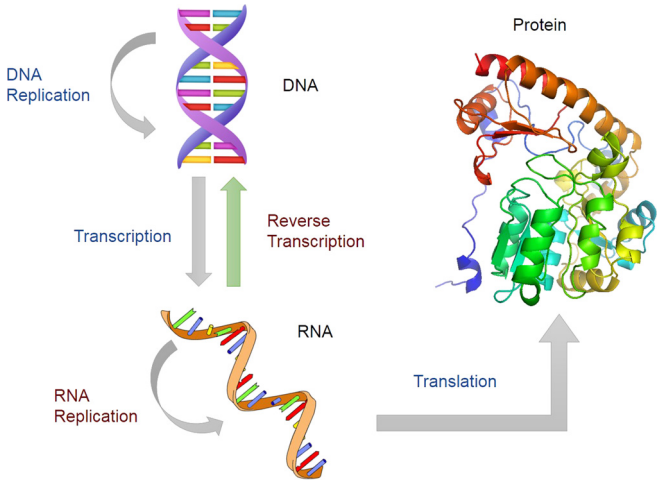


**Fig. 2.** Flow of information among DNA, RNA and protein.

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.

Please cite this article in press as: Y. Qin et al., The Current Status and Challenges in Computational Analysis of Genomic Big Data, Big Data Research (2015), http://dx.doi.org/10.1016/j.bdr.2015.02.005

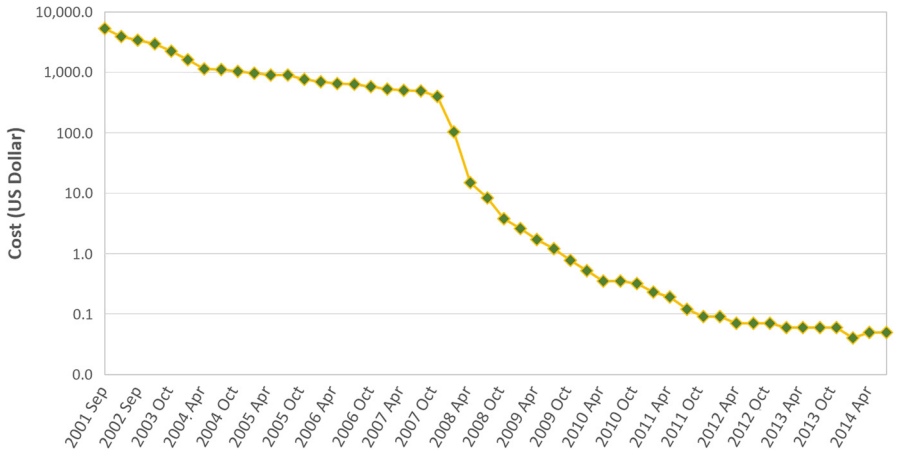


**Fig. 3.** DNA sequencing costs (per raw megabase of DNA) in the past 15 years [10].

can assist understanding disease mechanisms and discovering drug targets [14].

The 1000 Genomes Project (http://www.1000genomes.org/) is a project aimed at producing a comprehensive resource on human genetic variations. With the sequencing of around 2500 individuals under three projects, the consortium has characterized 15 million SNPs, 1 million indels and 20,000 structural variants with their locations, allele frequencies, and so on [15]. The International HapMap Project (http://hapmap.ncbi.nlm.nih.gov/) provides a map of human haplotype, which is a set of associated SNPs that are likely to be inherited together on a chromosome, to accelerate the progress of disease related gene detection [16]. It describes more than 3.1 million SNPs in human from a population with 270 individuals [17] and further expanded the population scale to comprehensively study SNPs and CNVs [18]. Resources provided by HapMap can aid the implementation of genome-wide association studies (GWAS). GWAS is a powerful method used to identify complex disease associated common genetic variants by statistically analyzing the differences between sequences of normal people and patients at a whole genome level on SNP arrays [17,19]. Moreover, The Cancer Genome Atlas (TCGA, http://cancergenome.nih.gov/) is a project directed against identifying genetic variants responsible for cancer with the utilization of genomic technologies. It aims to make genetic discoveries and characterize over 20 tumor types [20]. Several successful cases indicated the potency of translating cancer genomics into personalized cancer diagnostics and therapeutics [21].

High-throughput sequencing technologies and international collaborations provide opportunities to discover and characterize millions of genetic variants with large populations. To ensure high efficiency, such big data stimulate the development of bioinformatics approaches and tools [22]. As SNPs are thought to be most common form of genetic variation, it is of the most importance to be able to identify them quickly and accurately from NGS data. Tools for SNP calling are developed based on different statistical models. For example, MAQ uses Bayesian-based posterior probabilities to identify SNPs [23]. GATK toolkit also incorporates a Bayesian algorithm [24] to discovery variants for NGS data from multiple sequencing approaches and experiments. Recently, a new developed FaSD program (http://jjwanglab.org/FaSD/) [25] assumed that the number of reads (which are sequenced DNA fragments) mapped and cannot be mapped to the reference follows the binomial distribution. A score is then calculated to measure the polymorphism probability that a certain locus is a SNP. If the score is larger than a given threshold, this position is thought to be a SNP. FaSD exceeds other methods like GATK, SOAPsnp, MAQ, SNVmix2 and Bcftools to a certain degree on both accuracy and processing speed. SNP calling can be done using FaSD in four hours for 30 GB[5] NGS data on a standard desktop computer. Furthermore, FaSD-somatic (http://jjwanglab.org/FaSD-somatic/) has been developed recently to detect somatic mutations in cancer samples [26].

## 3. Transcriptomics

Corresponding to the concept of genome in DNA, the whole set of RNA transcripts in an organism or a certain cell type is called the transcriptome. The study on transcriptome with high-throughput approaches, for example, microarray and RNA-Seq is called transcriptomics [27]. Transcriptomics data measures the expression levels of genes as well as other transcribed DNA elements, and helps discover functional non-coding RNA elements. Expression of a gene varies across different cell types or at different developmental stages. High-throughput expression profiles of over one million samples can be accessed through ArrayExpress (http://www.ebi.ac.uk/arrayexpress/) and Gene Expression Omnibus (GEO, http://www.ncbi.nlm.nih.gov/geo/) [28], and the number of profiles keeps accumulating especially for the data generated by high-throughput sequencing (RNA-Seq), as shown in Fig. 4.

RNA-Seq is an approach for transcriptome profiling with NGS technologies. It was firstly reported in 2008 [29]. Actually it's not the sequencing of RNA directly but its complementary DNA (cDNA) by RNA reverse transcriptase. Such data generated by sequencer is also of large volume. For example, the latest Illumina Hiseq 2500 can produce as many as 300 million to 4 billion reads (approximately 10 GB to 1 TB) in one sequencer run [30]. Bioinformatics analysis is the key to make full use of RNA-Seq data, not only for detecting gene expression levels but also splicing isoforms. Basic RNA-Seq data analysis includes three steps. First, quality control of the raw data, which is to filter reads with low sequencing qualities and trim adapters at the ends of reads. The second step is to align short reads to reference genome or reference transcripts, or *de novo* assemble if the reference genome is not available. TopHat [31] is one of the most commonly used tools for RNA-Seq reads mapping. Then transcripts profiling can be built by Cufflinks [32] and Scripture [33]. These tools can also be used to measure the expression of each transcript and detect genetic variants. Further, differential expressed genes can be identified with Cufflinks by comparison of expression data on different conditions. A more advance usage of RNA-Seq data is construction of gene co-expression networks. SpliceNet [34], a novel method based on Large Dimensional Trace,

[5] GB, gigabyte, is the unit of storage in electronic machines.

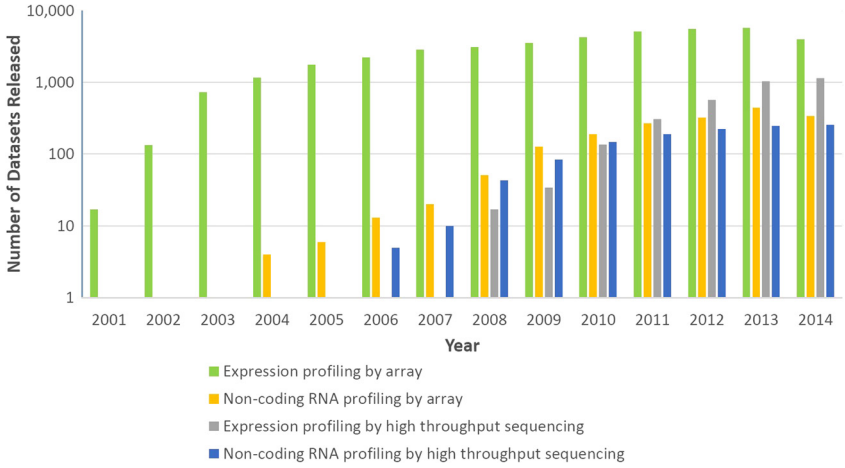


**Fig. 4.** Number of datasets of different types released in GEO each year. Data was extracted by constructing GEO DataSet database queries with 'DataSet type' and 'publication date'.

can infer isoform[6] specific co-expression networks from exon-level RNA-Seq data. It provides a more comprehensive picture to our understanding of complex diseases by inferring network rewiring between normal and diseased samples at isoform resolution. It can be applied to any exon level RNA-Seq or array data.

It is worth mentioning that, together with mRNA, long non-coding RNAs (lncRNAs) can be detected by Total RNA-Seq. The database lncRNAdb (http://www.lncrnadb.org/) provides comprehensive annotations of eukaryotic long non-coding RNAs. It will greatly assist the research of poorly annotated genes [35].

Challenges for the future of transcriptomics include developing fast and accurate transcripts mapping tools which can take all situations of transcription into consideration [27]. Precise measure differential expressed isoforms is also a difficult task nowadays [36]. More recently, "Single-Cell RNA-Seq" [37] was developed to measure the precise differences between transcripts in each single cell. It is able to build transcriptome profile in rare cells, such as stem cells and cancer cells. An unsupervised algorithm called Monocle [38] was applied to temporal transcriptome dynamic analysis with Single-Cell RNA-Seq data. However, a robust and comprehensive analysis methods are still limited [39]. How to extract useful information from a large amount of noise remains a big challenge.

## 4. Proteomics

Similarly, a proteome is the entire set of proteins in a cell and proteomics is the study of proteome on their structures and functions. It is more complicated as compared to genomics and transcriptomics due to complex spatial structure of a protein, the temporality and spatiality of its existence, the post-translational modifications (PTMs) and different isoforms [40]. Furthermore, since transcriptome can only reflect a roughly estimated level of translated proteins [41], quantitatively measuring the presences of proteins in a certain cell type under a certain condition is of great importance to detect key functional and regulatory proteins.

Mass Spectrometry (MS) techniques are the gold standard for both discovery and quantitative proteomics. Reports from MS are in XML format and record the mass-to-charge (m/z) ratio for each peptide precursor [42]. The MS spectra are then used to search the matched peptide sequences with tools such as Mascot (http://www.matrixscience.com/search_form_select.html), MassSearch (http://www.cbrg.ethz.ch/services/MassSearch_new) and MS-Viewer under ProteinProspector (http://prospector.ucsf.edu/prospector/mshome.htm) [43]. Finally, peptide sequences are assembled into proteins and quantified as protein enrichment with statistical significance. Labelling approaches such as SILAC (stable-isotope labelling by amino acids in cell culture) [44] and iTRAQ (isobaric tags for relative and absolute quantitation) [45] are currently prevalent in quantitative and comparative proteomics in conjunction with MS, which can relatively accurately measure the enrichments of protein complexes in different cells or tissues.

MS-based methods (e.g. electron transfer dissociation (ETD) [46]) can also greatly facilitate PTMs analysis. Phosphorylation is one of the most widely existed PTMs in biological processes and is well studied (as phosphoproteomics). Specific strategies (e.g. titanium dioxide chromatography [47]) have been developed to quantify phosphopeptide enrichment. With it, over 500,000 sites were predicted as potential phosphorylation sites. Glycoproteomics, studies on another widespread PTM glycosylation, is also well developed along with phosphoproteomics [48]. Sites of two types of glycosylation, N-linked and O-linked, were identified and can assist our comprehension of protein functions in various cell types and biological processes. Biological identification is inseparable from development of computational tools. FindMod [49] and CSS-Palm [50] are frequently used for PTMs prediction. More resources for PTMs can be found here http://www.biocuckoo.org/link.php.

With the maturation of proteomics methodologies and the help of informatics methods, many pilot projects on proteome landscape and comparative proteomics have been implemented. Following human genome project, a draft map of human proteome (http://www.humanproteomemap.org) was drawn recently based on the proteomics profiling of 30 normal human samples [51]. With it, a clear landscape of protein expression was exhibited and 808 novel annotations of the human genome were discovered. Other large amounts of proteomics data generated from high-throughput technologies, like mass spectrometry and microarray, have been collected. ExPASy (http://www.expasy.org/proteomics) lists dozens of databases on proteomics and over 100 tools. Commonly used databases like PROSITE, String, UniProtKB are all included in it. It provides a comprehensive repertory of protein sequences, structures, functions, abundances, protein–protein interactions (PPIs) and associated bioinformatics tools of great help. Aside from that, ProteomeXchange (http://www.proteomexchange.org/mission) would provide a platform for globally coordinated proteomics data submission and dissemination that further promotes data collection, sharing and comprehension [52].

[6] Isoforms are mRNAs transcript from the same gene but may with different splicing so that take different functions.

Please cite this article in press as: Y. Qin et al., The Current Status and Challenges in Computational Analysis of Genomic Big Data, Big Data Research (2015), http://dx.doi.org/10.1016/j.bdr.2015.02.005

Though with booming development, proteomics is still with some bottlenecks, including methods that are used to keep proteins in a certain environment, and to image proteins within living cells.

## 5. Interactomics and epigenomics

Organisms working as complete systems are inseparable from interactions of molecules from genomic, transcriptomic and proteomic layers, for instance, protein-DNA interaction, protein-RNA interaction and PPI as mentioned above. These interactions play key roles in the processing and regulation of transcription, translation and post-translational modifications and can perform cellular functions. The study on the whole set of biological molecular interactions in particular cells is called interactomics. Beside the effects of molecular interactions, DNA methylations and histone modifications can also be critical factors to affect gene expressions, known as epigenetic control [53]. The whole-genome wide studies on epigenetics are called epigenomics. Interactomics and epigenomics will be discussed together not only because they both have connections with gene regulation, but also have common detection methods.

Proteins can communicate with DNA at various levels, such as DNA repair, transcription, and providing structural stability. It is the binding domains of proteins that facilitate their interactions with DNA. Usually, non-coding regions such as enhancers, insulators and promoters are targeted for gene regulation. Proteins binding to such regions are called Transcription Factors (TFs) and the binding sites are called Transcription Factor Binding Sites (TFBS). Interactions amongst TFs and TFBSs regulate the transcription of genes differentially on various developmental stages and tissue types [54]. Chromatin Immunoprecipitation followed by high-throughput DNA Sequencing (ChIP-Seq) has been widely used to detect TF-DNA interactions [55]. It can locate DNA regions bound by a certain TF. Usually, tens of millions of reads are generated from one ChIP-Seq and such raw reads could be as large as several gigabytes. The raw reads are then mapped to the reference genome. Generally tens of thousands of regions enriched with reads are called as potential TFBSs.

Like ChIP-Seq, protein-RNA interactions can also be detected by high-throughput methods. For example, high-throughput sequencing of CLIP[7] cDNA library (HITS-CLIP or CLIP–Seq) provides opportunities to obtain protein-RNA binding landscape comprehensively [56].

In addition to the regulation by proteins or RNAs, distant DNA fragments can also interact with each other spatially and work as regulatory elements. The Chromosome Conformation Capture (3C) methodology was then developed to identify chromosome locations that interact in the nuclear space. Because 3C can be laborious, 4C (circular 3C) [57] and 5C (3C-Carbon Copy) [58] are developed to screen interaction partners for the selected genome sites at the whole genome [59]. Then, Hi-C technology became the first unbiased and real genome-wide adaptation of 3C [60]. If hundreds of millions of reads are generated, the resolution of chromatin interaction map can be detected at 10 Kbp by Hi-C [61]. A recent published study even achieved kilobase resolution with a slightly revised method called *in situ* Hi-C [62].

Methylation of cytosine residues at special regions (generally enriched with nucleotide C and G) of the DNA molecule typically marks the repression of nearby genes. Bisulfite sequencing (BS-Seq) is one of the most commonly used high-throughput methods to get the whole-genome methylation profile. Special tools were devised for aligning reads generated from BS-Seq, like Bismark [63], BSMAP [64] and RMAPBS [65]. Computational prediction of DNA methylation sites and status with methods like CpGIMethPred [66] is widely adopted to speed up methylation analysis. The result reveals that methylation patterns are related with disease processes [67]. For example, hypermethylation that silencing tumor suppressor genes is often taken as hallmark of cancer cells [68].

Like DNA methylation, histone modification is another well studied type of epigenetics. Many histone modifications have been described and characterized with relations to gene activation. For example, H3K4me3 (trimethylation of histone H3 lysine 4) is usually an active marker on the gene promoters. ChIP-Seq is also used to locate histone modification sites at genome scale. Experimental procedures and computational analysis approaches are quite similar to that used for TFBSs detection.

With the constant generation of interactomics and epigenomics data, several databases and web-based servers have been built up to store and interpret them. For example, the International Molecular Exchange (IMEx) consortium (http://www.imexconsortium.org/home), with 10 active partners, created a curation of non-redundant of PPI data [69]. There are over 300,000 binary interactions contained in IMEx consortium and the number keeps increasing. Constructing and comprehending PPI networks is also a respect of proteomics studies. In contrast, ChIP-Array [70] and CMGRN [71] provided web-based frameworks to integrate gene expression profile and TF binding/epigenetic modification signals of ChIP-Seq data for reconstruction of gene regulatory networks. Web server PTHGRN [72] links PTMs of proteins and transcriptional gene regulation to explore multilayered networks underlying biological complex processes. DDGni [73] can account for expression delays in long time series data, which traditional methods overlooked, by adopting gapped Smith–Waterman algorithm. For epigenomics, Roadmap Epigenomics Project (http://www.roadmapepigenomics.org/) is a pilot project and provides a wide resource of human epigenomics data which is valuable for the research on gene regulation and disease development.

## 6. Conclusions and perspectives

Biological and biomedicine sciences are now coming into multi-dimensional OMICS era with high revolutions. The big data are generated on different biological components and are greatly speeding up clinical translational use. In this paper, we discuss several state-of-the-art high-throughput methodologies and data integrative approaches to solve biomedical questions or reveal biological mechanisms. We also demonstrate that NGS data facilitates the discovery of genetic variants associated with diseases; transcriptomics data creates a landscape of all transcripts in different cell types; proteomics data help quantitatively measure the presence of proteins and monitor the PTMs of proteins. Big data is also generated for multilevel molecular interactions (interactomics) and used to help us in understanding how organisms work as biological systems. Epigenomics data could further open another view and assist us to interpret how epigenetic modifications affect gene expression. At the same time, several major projects, public databases and consortiums regarding big data production and usefulness are introduced.

Just like problems of big data shown in other areas, issues on data generation, transferring, storage, security, visualization and processing exist and challenge scientists in biomedicalfields nowadays. Sequencing data are generally stored and handled in high-performance computing clusters at present. Data analysis needs large storage space and has a high requirement on computational speed. Taking ChIP-Seq data analysis as an example, if there are three mouse cell types to be studied, each cell type would have two ChIPed replicates with two control replicates. Hence, there are

[7] CLIP is known as ultraviolet (UV) crosslinking and immunoprecipitation.

Please cite this article in press as: Y. Qin et al., The Current Status and Challenges in Computational Analysis of Genomic Big Data, Big Data Research (2015), http://dx.doi.org/10.1016/j.bdr.2015.02.005

totally 12 raw data sets and each with around 60 million reads in 3 GB (after compressed). The file generated by BWA alignment is around 10 GB in SAM format and 3 GB in BAM format. The time for alignment on a remote server (GNU/Linux 3.2.0-74-generic x86_64) with a maximum of 12 processors in a node with 20 GB memory is around 6 hours. In total, around 500 GB are needed to store the raw, intermediate and final files and nearly one day to process the data. We believe that these data can be managed better with efforts through interdisciplinary collaborations from different fields. Some ethical issues may need more attention and may be able to reach an agreement with negotiation. Thus, special data processing is the most challenging. Although various tools have been developed for different data types, data integration and interpretation is far from perfect due to large biological noise and technical defects. The mission of bioinformaticians is to reduce noises and improve the accuracies and efficiencies in computational prediction with the comprehension of biological processes, statistics, mathematics and IT sciences.

To make full use of the available data, integrative analytics is becoming more and more popular currently. Data generated from different platforms at various molecular levels can be integrated together and draw more conclusions. For example, GWASdb (http://jjwanglab.org/gwasdb) is a comprehensive database with traits/diseases associated SNPs and their comprehensive functional annotations, as well as disease classifications [74]. Such well-rounded data can offer researchers full insights of recent GWA studies and help them get as much information from a single website. To get comprehensive understanding of the function of a TF, ChIP-Seq data is often integrated with expression data. ChIP-Array (http://wanglab.hku.hk/ChIP-Array), PTHGRN (http://www.byanbioinfo.org/pthgrn) and BETA (http://cistrome.org/BETA/) are well-equipped tools for such integration [70,72,75]. We expect that, with integration of big data generated by several platforms, biomedicine mechanisms can be explored more easily and thoroughly. Furthermore, with sequencing costs going down, and with data generation and analysis speed going up, personalized diagnosis and therapy can become more and more common.

## Acknowledgement

We would like to acknowledge the financial support of the Research Grants Council, University Grants Committee, Hong Kong (Grant No. T12-708/12-N, 781511M, 17121414M) and National Natural Science Foundation of China (Grant No. 91229105). We thank Mr. Ken Yip in the lab for editing.

## References

[1] A.D. Hershey, M. Chase, Independent functions of viral protein and nucleic acid in growth of bacteriophage, J. Gen. Physiol. 36 (1952) 39–56.
[2] P. Kapranov, et al., RNA maps reveal new RNA classes and a possible function for pervasive transcription, Science 316 (2007) 1484–1488.
[3] T.R. Mercer, M.E. Dinger, J.S. Mattick, Long non-coding RNAs: insights into functions, Nat. Rev. Genet. 10 (2009) 155–159.
[4] K. Ranganathan, V. Sivasankar, MicroRNAs – biology and clinical applications, J. Oral Maxillofac. Pathol. 18 (2014) 229–234.
[5] J. Skolnick, J.S. Fetrow, From genes to protein structure and function: novel applications of computational approaches in the genomic era, Trends Biotechnol. 18 (2000) 34–39.
[6] B. Boeckmann, et al., Protein variety and functional diversity: Swiss-Prot annotation in its biological context, C. R. Biol. 328 (2005) 882–899.
[7] F. Crick, On protein synthesis, Symp. Soc. Exp. Biol. 12 (1958) 138–163.
[8] F. Crick, Central dogma of molecular biology, Nature 227 (1970) 561–563.
[9] I. Noble, Human genome finally complete, 2003 [cited 2014]; Available from: http://news.bbc.co.uk/2/hi/science/nature/2940601.stm.
[10] K. Wetterstrand, DNA sequencing costs: data from the NHGRI Genome Sequencing Program (GSP), 2014 [cited August 4, 2014]; Available from: www.genome.gov/sequencingcosts.
[11] A.F. Wright, Nature Encyclopedia of the Human Genome, Nature Publishing Group, London, 2003, pp. 959–968.
[12] J.R. MacDonald, et al., The database of genomic variants: a curated collection of structural variation in the human genome, Nucleic Acids Res. 42 (2014) D986–D992.
[13] L. Feuk, A.R. Carson, S.W. Scherer, Structural variation in the human genome, Nat. Rev. Genet. 7 (2006) 85–97.
[14] J.T.L. Mah, E.S.H. Low, E. Lee, Insilico SNP analysis and bioinformatics tools: a review of the state of the art to aid drug discovery, Drug Discov. Today 16 (2011) 800–809.
[15] Genomes Project Consortium, et al., A map of human genome variation from population-scale sequencing, Nature 467 (2010) 1061–1073.
[16] International HapMap Consortium, A haplotype map of the human genome, Nature 437 (2005) 1299–1320.
[17] Genomes Project Consortium, et al., A second generation human haplotype map of over 3.1 million SNPs, Nature 449 (2007) 851–861.
[18] International HapMap Consortium, et al., Integrating common and rare genetic variation in diverse human populations, Nature 467 (2010) 52–58.
[19] S.F. Kingsmore, et al., Genome-wide association studies: progress and potential for drug discovery and development, Nat. Rev. Drug Discov. 7 (2008) 221–230.
[20] NHGRI, N.a.t. TCGA Program Overview, 2014 [cited August 2014]; Available from: http://cancergenome.nih.gov/abouttcga/overview.
[21] L. Chin, J.N. Andersen, P.A. Futreal, Cancer genomics: from discovery science to personalized medicine, Nat. Med. 17 (2011) 297–303.
[22] L. Chin, et al., Making sense of cancer genomic data, Genes Dev. 25 (2011) 534–555.
[23] H. Li, J. Ruan, R. Durbin, Mapping short DNA sequencing reads and calling variants using mapping quality scores, Genome Res. 18 (2008) 1851–1858.
[24] M.A. DePristo, et al., A framework for variation discovery and genotyping using next-generation DNA sequencing data, Nat. Genet. 43 (2011) 491–498.
[25] F. Xu, et al., A fast and accurate SNP detection algorithm for next-generation sequencing data, Nat. Commun. 3 (2012) 1258.
[26] W. Wang, et al., FaSD-somatic: a fast and accurate somatic SNV detection algorithm for cancer genome sequencing data, Bioinformatics 30 (2014) 2498–2500.
[27] Z. Wang, M. Gerstein, M. Snyder, RNA-Seq: a revolutionary tool for transcriptomics, Nat. Rev. Genet. 10 (2009) 57–63.
[28] T. Barrett, et al., NCBI GEO: archive for functional genomics data sets—update, Nucleic Acids Res. 41 (2013) D991–D995.
[29] U. Nagalakshmi, et al., The transcriptional landscape of the yeast genome defined by RNA sequencing, Science 320 (2008) 1344–1349.
[30] HiSeq System Performance Parameters, 2014 [cited September 2014]; Available from: http://systems.illumina.com/systems/hiseq_2500_1500/performance_specifications.ilmn.
[31] C. Trapnell, L. Pachter, S.L. Salzberg, TopHat: discovering splice junctions with RNA-Seq, Bioinformatics 25 (2009) 1105–1111.
[32] C. Trapnell, et al., Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation, Nat. Biotechnol. 28 (2010) 511–515.
[33] M. Guttman, et al., Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs, Nat. Biotechnol. 28 (2010) 503–510.
[34] H.K. Yalamanchili, et al., SpliceNet: recovering splicing isoform-specific differential gene networks from RNA-Seq data of normal and diseased samples, Nucleic Acids Res. (2014).
[35] P.P. Amaral, et al., lncRNAdb: a reference database for long noncoding RNAs, Nucleic Acids Res. 39 (2011) D146–D151.
[36] H.D. Li, et al., The emerging era of genomic data integration for analyzing splice isoform function, Trends Genet. 30 (2014) 340–347.
[37] F. Tang, et al., RNA-Seq analysis to capture the transcriptome landscape of a single cell, Nat. Protoc. 5 (2010) 516–535.
[38] C. Trapnell, et al., The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells, Nat. Biotechnol. 32 (2014) 381–386.
[39] G.K. Marinov, et al., From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing, Genome Res. 24 (2014) 496–510.
[40] H.K. Yalamanchili, Q.W. Xiao, J. Wang, A novel neural response algorithm for protein function prediction, BMC Syst. Biol. 6 (Suppl. 1) (2012) S19.
[41] V. Dhingra, et al., New frontiers in proteomics research: a perspective, Int. J. Pharm. 299 (2005) 1–18.
[42] A.F. Altelaar, J. Munoz, A.J. Heck, Next-generation proteomics: towards an integrative view of proteome dynamics, Nat. Rev. Genet. 14 (2013) 35–48.
[43] P.R. Baker, R.J. Chalkley, MS-viewer: a web-based spectral viewer for proteomics results, Mol. Cell. Proteomics 13 (2014) 1392–1396.
[44] M. Mann, Functional and quantitative proteomics using SILAC, Nat. Rev. Mol. Cell Biol. 7 (2006) 952–958.
[45] S. Wiese, et al., Protein labeling by iTRAQ: a new tool for quantitative mass spectrometry in proteome research, Proteomics 7 (2007) 340–350.
[46] J.E. Syka, et al., Peptide and protein sequence analysis by electron transfer dissociation mass spectrometry, Proc. Natl. Acad. Sci. USA 101 (2004) 9528–9533.
[47] M.W. Pinkse, et al., Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns, Anal. Chem. 76 (2004) 3935–3943.

Please cite this article in press as: Y. Qin et al., The Current Status and Challenges in Computational Analysis of Genomic Big Data, Big Data Research (2015), http://dx.doi.org/10.1016/j.bdr.2015.02.005

[48] B. Tissot, et al., Glycoproteomics: past, present and future, FEBS Lett. 583 (2009) 1728–1735.
[49] M.R. Wilkins, et al., High-throughput mass spectrometric discovery of protein post-translational modifications, J. Mol. Biol. 289 (1999) 645–657.
[50] J. Ren, et al., CSS-Palm 2.0: an updated software for palmitoylation sites prediction, Protein Eng. Des. Sel. 21 (2008) 639–644.
[51] M.S. Kim, et al., A draft map of the human proteome, Nature 509 (2014) 575–581.
[52] J.A. Vizcaino, et al., ProteomeXchange provides globally coordinated proteomics data submission and dissemination, Nat. Biotechnol. 32 (2014) 223–226.
[53] A. Bird, Perceptions of epigenetics, Nature 447 (2007) 396–398.
[54] S. Yang, et al., Correlated evolution of transcription factors and their binding sites, Bioinformatics 27 (2011) 2972–2978.
[55] S.G. Landt, et al., ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia, Genome Res. 22 (2012) 1813–1831.
[56] J. Konig, et al., Protein–RNA interactions: new genomic technologies and perspectives, Nat. Rev. Genet. 13 (2011) 77–83.
[57] Z. Zhao, et al., Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions, Nat. Genet. 38 (2006) 1341–1347.
[58] J. Dostie, et al., Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements, Genome Res. 16 (2006) 1299–1309.
[59] A. Gavrilov, et al., Chromosome conformation capture (from 3C to 5C) and its ChIP-based modification, Methods Mol. Biol. 567 (2009) 171–188.
[60] N.L. van Berkum, et al., Hi-C: a method to study the three-dimensional architecture of genomes, J. Vis. Exp. (2010).
[61] J. Dekker, M.A. Marti-Renom, L.A. Mirny, Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data, Nat. Rev. Genet. 14 (2013) 390–403.
[62] S.S. Rao, et al., A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping, Cell 159 (2014) 1665–1680.
[63] F. Krueger, S.R. Andrews, Bismark: a flexible aligner and methylation caller for Bisulfite-Seq applications, Bioinformatics 27 (2011) 1571–1572.
[64] Y. Xi, W. Li, BSMAP: whole genome bisulfite sequence MAPping program, BMC Bioinform. 10 (2009) 232.
[65] A.D. Smith, et al., Updates to the RMAP short-read mapping software, Bioinformatics 25 (2009) 2841–2842.
[66] H. Zheng, et al., CpGIMethPred: computational model for predicting methylation status of CpG islands in human genome, BMC Med. Genomics 6 (Suppl. 1) (2013) S13.
[67] M. Esteller, Epigenetics in cancer, N. Engl. J. Med. 358 (2008) 1148–1159.
[68] P.A. Callinan, A.P. Feinberg, The emerging science of epigenomics, Hum. Mol. Genet. 15 (Spec. No. 1) (2006) R95–R101.
[69] S. Orchard, et al., Protein interaction data curation: the International Molecular Exchange (IMEx) consortium, Nat. Methods 9 (2012) 345–350.
[70] J. Qin, et al., ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor, Nucleic Acids Res. 39 (2011) W430–W436.
[71] D. Guan, et al., CMGRN: a web server for constructing multilevel gene regulatory networks using ChIP-seq and gene expression data, Bioinformatics (2014).
[72] D. Guan, et al., PTHGRN: unraveling post-translational hierarchical gene regulatory networks using PPI, ChIP-seq and gene expression data, Nucleic Acids Res. 42 (2014) W130–W136.
[73] H.K. Yalamanchili, et al., DDGni: dynamic delay gene-network inference from high-temporal data using gapped local alignment, Bioinformatics 30 (2014) 377–383.
[74] M.J. Li, et al., GWASdb: a database for human genetic variants identified by genome-wide association studies, Nucleic Acids Res. 40 (2012) D1047–D1054.
[75] S. Wang, et al., Target analysis by integration of transcriptome and ChIP-seq data with BETA, Nat. Protoc. 8 (2013) 2502–2515.

Please cite this article in press as: Y. Qin et al., The Current Status and Challenges in Computational Analysis of Genomic Big Data, Big Data Research (2015), http://dx.doi.org/10.1016/j.bdr.2015.02.005