Biophysical Reviews (2019) 11:51–54
https://doi.org/10.1007/s12551-018-0493-5

LETTER TO THE EDITOR

Big data challenges in genome informatics

Ka-Chun Wong[1]

Received: 17 September 2018 / Accepted: 13 December 2018 / Published online: 25 January 2019

Abstract
In recent years, we have witnessed a big data explosion in genomics, thanks to the improvement in high-throughput technologies at drastically decreasing costs. We are entering the era of millions of available genomes. Notably, each genome can be composed of billions of nucleotides stored as plain text files in gigabytes (GBs). It is undeniable that those genome data impose unprecedented data challenges for us. In this article, we briefly discuss the big data challenges associated with genomics in recent years.

Introduction

Since 1990s, the whole genomes of different species have been sequenced by different genome sequencing projects. In 1995, the first free-living organism Haemophilus influenzae was sequenced by the Institute for Genomic Research. In 1996, the first eukaryotic genome (Saccharomyces cerevisiase) was completely sequenced. In 2000, the first plant genome, Arabidopsis thaliana, was also sequenced by Arabidopsis Genome Initiative. In 2003, the Human Genome Project (HGP) announced its completion. Following the HGP, the Encyclopedia of DNA Elements (ENCODE) project was started, revealing a massive number of functional elements in the human genome in 2011 (ENCODE Project Consortium et al. 2004). The drastically decreasing cost of sequencing also enables the 1000 Genomes Project and Roadmap Epigenomics Project to be carried out. Their results have been published in 2012 and 2015 respectively (1000 Genomes Project Consortium et al. 2010; Kundaje et al. 2015). Nonetheless, the massive genomic data generated by those projects impose an unforeseen challenge for big data analysis at the scale of gigabytes (GBs) or even terabytes (TBs).

In particular, next-generation sequencing (NGS) technologies have enabled massive data generation for different genomes (Wong and Zhang 2014; Mardis 2008); for instance, DNA sequencing, protein-DNA binding occupancy (Wong et al. 2013) (e.g., ChIP-seq Visel et al. (2009), ChIP-exo Ho and Franklin Pugh (2011), and ChIA-PET Fullwood et al. (2009)), bisulfite sequencing (Bock et al. 2005), transcriptome sequencing (e.g., RNA-seq Mortazavi et al. (2008)), and chromatin interaction sequencing (e.g., Hi-C Lieberman-Aiden et al. (2009)). Thanks to the relatively low costs, those NGS technologies have been readily applied to human genomes nowadays. The international projects aforementioned have been successfully launched, leading to massive NGS data accumulation at an unprecedentedly fast pace. Nonetheless, current integrative analyses are usually limited to traditional machine learning and data mining methods such as pair-wise correlation analysis, statistical tests, classification, and feature extraction (Wong et al. 2015b). Those methods are intentionally designed to generally fit different types of data. However, the data from NGS is unique and different from the traditional data; for instance, the ChIP-seq data is sparse, noisy, and discontinuous. Special care has to be taken to alleviate and transform those challenges to be taken advantages of (Wong et al. 2015a). In addition, the NGS data is huge (in gigabytes per each dataset) which imposes a difficulty in applying some of the existing statistical/computational methods.

Therefore, different genome-scale problems have been defined and framed to harness those genomic data. Figure 1 aims to provide a concise summary of those challenges.

De novo genome assembly

The advancement in DNA sequencing technologies has enabled the assembly of whole genome in an economical and fairly accurate way (Mardis 2011). Nonetheless, a genome cannot be easily identified in one piece from

This article is part of a Special Issue on 'Big Data' edited by Joshua WK Ho and Eleni Giannoulatou.

✉ Ka-Chun Wong
kc.w@cityu.edu.hk

[1] City University of Hong Kong, Kowloon, Hong Kong

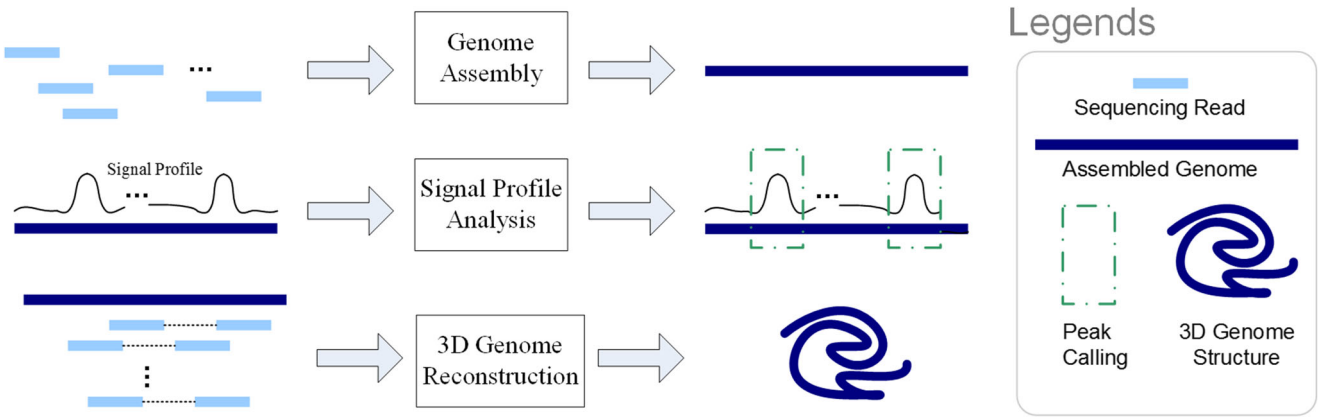

Fig. 1 Big data challenges in genome informatics. The challenges are listed from top to bottom; namely, genome assembly, signal profile analysis, and 3D genome structure reconstruction

wet-labs. Limited by our current DNA sequencing technologies, each genome has to be shattered into many non-overlapping small pieces (short DNA sequence reads) before their DNA nucleotides can get sequenced and identified as shown in Fig. 1. Therefore, we come to the *de novo* genome assembly problem: to sequence and identify a genome, we have to "stitch" those short DNA sequences into a single and consistent DNA genome while allowing for overlaps. There are different benchmark measurements such as N50, total length, and number of missing nucleotides. If we already have a reference genome, then the measurements can be more solid than the previous measures such as NG50 and genome fraction. If a reference genome annotation is available, the number of genes covered can be a good measurement. More details can be found in Gurevich et al. (2013). To solve this kind of genome assembly problems (in GBs or TBs), there are many computational methods proposed in the past. Nonetheless, most of them depend on the construction of *de brujin* graph which is memory-consuming and computationally intensive. According to the recent benchmark study, different genome assembly methods show result disagreement with each other by Bradnam et al. (2013). In addition, the sequencing errors incurred by wet-lab experimental conditions are unavoidable, making the genome assembly problem even more complicated than we have imagined (Mardis 2011). Therefore, the genome assembly problem remains as a big data challenge to be solved.

Genome signal profile analysis

In addition to genome assembly, there are different genome-wide signals such as gene regulation (e.g. protein-DNA binding interactions) and epigenetic interactions (e.g. DNA methylation) as shown in Fig. 1. Therefore, it is essential for us to look into those information. To this end, different genome-wide biotechnologies have been developed such as ChIP-seq, DNase-seq, RNA-seq, CLIP-seq, DNA methylation array assay, bisulphite sequencing, Repli-Seq, and CAGE. To gains insights into those data, tremendous efforts have been made to pre-process the data such as read trimming (Bolger et al. 2014), sequencing error correction (Yang et al. 2013), sequencing replicates (Robasky et al. 2014), and read mapping (David et al. 2011). After the data has been processed, downstream analysis methods can be applied to reveal genome-wide signals from it; for instance, multiple signal profile integrative analysis (Wong et al. 2015a, b) and signal profile peak calling (Zhang et al. 2008). In particular, the multiple signal profile analysis is very important for us to understand the complex behavior of the genome-wide signals (Wong et al. 2015a). Unfortunately, each signal profile is proportional to genome size since it has a genome-wide coverage (usually in GBs). Therefore, if we have multiple signal profiles (e.g., hundreds from the ENCODE consortium), the computational scalability issue has to be taken into serious account. Another issue is that the past wet-lab studies are very limited to fine-scale knowledge (e.g., single gene study). Therefore, the genome-wide result verification is very difficult to be carried out. At the current stage, we heavily rely on null hypothesis testing to ascertain the results' statistical significances. Therefore, we can foresee that the genome signal profile analysis will still be a big data challenge in genome informatics.

3D genome structure reconstruction

In recent years, Hi-C technology has been developed and applied to reveal the three dimensional organizations of different cell lines by the chromosome conformation capture method (Belton et al. 2012). In particular, there is increasing evidence that long-range chromatin interactions are related to gene co-expression (Babaei et al. 2015; Jin et al. 2013) as well as protein-DNA interactions (Lan et al. 2012; Mifsud et al. 2015). Therefore, it is essential to comprehensively identify and reconstruct the three-dimensional (3D) genome

shape from those long-range chromatin interactions for understanding genomes in the three-dimensional space. Given the GB data size of genome as well as its three-dimensional nature, such a 3D genome reconstruction is doomed to be another big data challenge.

Future perspectives

In this article, we have discussed several big data challenges in genome informatics. Especially, we envision that those challenges will become intense in the near future, given the maturing and cost-effective sequencing technologies. Several future directions are deemed promising: (1) third-generation sequencing technologies (Schadt et al. 2010) have been developed and being refined to be of practical uses. Although its sequencing error rate is still high, we believe that those third-generation sequencing technologies will enable another wave of big data challenges in genome informatics. (2) Single cell sequencing is another promising direction. In the past, we usually studied specific cell types or tissue types using the population-based approaches. However, cell type heterogeneity is often observed in practice. Therefore, our current single-cell sequencing technologies can enable us to look at each of the individual cells; it holds tremendous potential to trigger the next levels of big data challenges. However, the cell-destructive nature of single-cell sequencing may limit its capability such as real-time live tracking, disease prognosis analysis, and stem cell development. To address those limitations, single-cell imaging techniques could be promising; it can even offer insights into the spatial arrangement of individual cells. (3) Given the genome data in GBs or even TBs, high-performance computing frameworks such as MapReduce are definitely needed to handle the exponentially growing genome data in a scalable but still accurate manner. The high-throughput computing technologies such as Hadoop, Spark, and Pig Latin are expected to become more pronounced than now.

Funding information The literature review and writing in this paper were substantially supported by three grants from the Research Grants Council of the Hong Kong Special Administrative Region [CityU 21200816], [CityU 11203217], and [CityU 11200218]. The donation support of the Titan Xp GPU from the NVIDIA Corporation is appreciated.

Compliance with ethical standards

Conflict of interest Ka-Chun Wong declares that he has no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by the author.

References

1000 Genomes Project Consortium et al (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073

Babaei S, Mahfouz A, Hulsman M, Lelieveldt BP, de Ridder J, Reinders M (2015) Hi-C chromatin interaction networks predict co-expression in the mouse cortex. PLoS Comput Biol 11(5):e1004221

Belton JM, McCord RP, Gibcus JH, Naumova N, Zhan Y, Dekker J (2012) Hi-C: a comprehensive technique to capture the conformation of genomes. Methods 58(3):268–276

Bock C, Reither S, Mikeska T, Paulsen M, Walter J, Lengauer T (2005) Biq analyzer: visualization and quality control for dna methylation data from bisulfite sequencing. Bioinformatics 21(21):4067–4068

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30(15):2114–2120

Bradnam KR, Fass JN, Alexandrov A, Baranay P, Bechner M, Birol I, Boisvert S, Chapman JA, Chapuis G, Chikhi R, Chitsaz H, Chou WC, Corbeil J, Del Fabbro C, Docking TR, Durbin R, Earl D, Emrich S, Fedotov P, Fonseca NA, Ganapathy G, Gibbs RA, Gnerre S, Godzaridis E, Goldstein S, Haimel M, Hall G, Haussler D, Hiatt JB, Ho IY, Howard J, Hunt M, Jackman SD, Jaffe DB, Jarvis ED, Jiang H, Kazakov S, Kersey PJ, Kitzman JO, Knight JR, Koren S, Lam TW, Lavenier D, Laviolette F, Li Y, Li Z, Liu B, Liu Y, Luo R, Maccallum I, Macmanes MD, Maillet N, Melnikov S, Naquin D, Ning Z, Otto TD, Paten B, Paulo OS, Phillippy AM, Pina-Martins F, Place M, Przybylski D, Qin X, Qu C, Ribeiro FJ, Richards S, Rokhsar DS, Ruby JG, Scalabrin S, Schatz MC, Schwartz DC, Sergushichev A, Sharpe T, Shaw TI, Shendure J, Shi Y, Simpson JT, Song H, Tsarev F, Vezzi F, Vicedomini R, Vieira BM, Wang J, Worley KC, Yin S, Yiu SM, Yuan J, Zhang G, Zhang H, Zhou S, Korf IF (2013) Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. Gigascience 2(1):10

ENCODE Project Consortium et al (2004) The encode (encyclopedia of DNA elements) project. Science 306(5696):636–640

David M, Dzamba M, Lister D, Ilie L, Brudno M (2011) SHRiMP2: sensitive yet practical Short Read Mapping. Bioinformatics 27(7):1011–1012

Fullwood MJ, Liu MH, Pan YF, Liu J, Xu H, Mohamed YB, Orlov YL, Velkov S, Ho A, Mei PH et al (2009) An oestrogen-receptor-α-bound human chromatin interactome. Nature 462(7269):58–64

Gurevich A, Saveliev V, Vyahhi N, Tesler G (2013) QUAST: quality assessment tool for genome assemblies. Bioinformatics 29(8):1072–1075

Jin F, Li Y, Dixon JR, Selvaraj S, Ye Z, Lee AY, Yen CA, Schmitt AD, Espinoza CA, Ren B (2013) A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503(7475):290–294

Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, Kheradpour P, Zhang Z, Wang J, Ziller MJ et al (2015) Integrative analysis of 111 reference human epigenomes. Nature 518(7539):317–330

Lan X, Witt H, Katsumura K, Ye Z, Wang Q, Bresnick EH, Farnham PJ, Jin VX (2012) Integration of Hi-C and ChIP-seq data reveals distinct types of chromatin linkages. Nucleic Acids Res 40(16):7690–7704

Lieberman-Aiden E, van Berkum NL, Williams L, Imakaev M, Ragoczy T, Telling A, Amit I, Lajoie BR, Sabo PJ, Dorschner MO

et al (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science 326(5950):289–293

Mardis ER (2011) A decade's perspective on DNA sequencing technology. Nature 470(7333):198–203

Mardis ER (2008) The impact of next-generation sequencing technology on genetics. Trends Genet 24(3):133–141

Mifsud B, Tavares-Cadete F, Young AN, Sugar R, Schoenfelder S, Ferreira L, Wingett SW, Andrews S, Grey W, Ewels PA, Herman B, Happe S, Higgs A, LeProust E, Follows GA, Fraser P, Luscombe NM, Osborne CS (2015) Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. Nat Genet 47(6):598–606

Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B (2008) Mapping and quantifying mammalian transcriptomes by rna-seq. Nat Methods 5(7):621–628

Ho SR, Franklin Pugh B (2011) Comprehensive genome-wide protein-dna interactions detected at single-nucleotide resolution. Cell 147(6):1408–1419

Robasky K, Lewis NE, Church GM (2014) The role of replicates for error mitigation in next-generation sequencing. Nat Rev Genet 15(1):56–62

Schadt EE, Turner S, Kasarskis A (2010) A window into third-generation sequencing. Hum Mol Genet 19(R2):R227–R240

Visel A, Blow MJ, Li Z, Zhang T, Akiyama JA, Holt A, Plajzer-Frick I, Shoukry M, Wright C, Chen F et al (2009) Chip-seq accurately predicts tissue-specific activity of enhancers. Nature 457(7231):854–858

Wong KC, Li Y, Peng C, Zhang Z (2015a) SignalSpider: probabilistic pattern discovery on multiple normalized ChIP-Seq signal profiles. Bioinformatics 31(1):17–24

Wong K-C, Peng C, Li Y (2015b) Probabilistic inference on multiple normalized signal profiles from next generation sequencing: Transcription factor binding sites. IEEE/ACM Trans Comput Biol Bioinform 12(6):1416–1428

Wong K-C, Chan T-M, Peng C, Li Y, Zhang Z (2013) Dna motif elucidation using belief propagation. Nucleic Acids Res 41(16):e153–e153

Wong K-C, Zhang Z (2014) Snpdryad: predicting deleterious non-synonymous human snps using only orthologous protein sequences. Bioinformatics page btt769

Yang X, Chockalingam SP, Aluru S (2013) A survey of error-correction methods for next-generation sequencing. Brief Bioinform 14(1):56–66

Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W et al (2008) Model-based analysis of chip-seq (macs). Genome Biol 9(9):R137