Bioinformatics, Proteomics and Immaging Analysis

OMMEGA Publishers
Answering Intelligence
www.ommegaonline.org

Editorial

OPEN ACCESS

The Blooming Era of Genome Informatics: State-of-the-Art and Future Challenges

Ka-Chun Wong*

Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong

Corresponding Author: Ka-Chun Wong, Department of Computer Science, City University of Hong Kong, Kowloon Tong, Hong Kong. Tel: (852)34428618; E-mail: kc.w@cityu.edu.hk

Received Date: October 12, 2015 **Accepted Date: October 23, 2015** **Published Date: October 26, 2015**

Citation: Wong, K.C. The Blooming Era of Genome Informatics: State-of-the-Art and Future Challenges (2016) Bioinfo Proteom Img Anal 1(2): 49- 50.

At the beginning of the current 21st century, we have witnessed the tremendous improvement and growth in parallel sequencing technologies; for instance, the next-generation sequencing (NGS) technologies are well-known for its cost-effectiveness as well as high-throughput capability[1]. Different variants of NGS technologies have been developed; for instance, DNA Sequencing (DNA-seq) is developed to sequence and map genomes; Chromatin ImmunoPrecipitation with Sequencing (ChIP-Seq) is developed to determine the protein-DNA binding locations on genome; RNA Sequencing (RNA-Seq) is developed to quantify the RNAs transcribed within cells in a high throughput manner; RNA ImmunoPrecipitation with Sequencing (RIP-Seq) is similar to ChIP-Seq but for protein-RNA binding location identification; High-throughput Chromosome conformation capture (Hi-C) is developed to capture the three-dimensional structure of genome; Deoxyribonuclease (DNase) hypersensitive sites sequencing (DNase-Seq) is developed to estimate the genome-wide open chromatin regions based on DNA accessibility; Bisulfate sequencing is developed to probe the DNA methylation on genome. Although their objectives are different, their novelty remains the same: taking advantage of NGS to accelerate the existing wet-lab genome technologies to a genome-wide level, resulting in big data challenges.

NGS technology is designed for high-throughput sequencing. Thanks to its parallel nature, a single run can result in millions of sequencing reads in few days or even hours for now. From the computational perspective, the data is massive and should be measured in GigaBytes (GBs), or even TeraBytes (TBs). Such data scales are no longer able to be handled by some of the existing general statistical methods which have been developed in the past[2]. Instead, we have to develop scalable but still accurate bioinformatics methods which scale with the exponentially growing data; for instance, Wong et al. have developed a scalable bioinformatics method called SignalSpider (http://www.cs.toronto.edu/~wkc/SignalSpider/) which can analyse multiple ChIP-Seq signal profiles simultaneously in linear time complexity[3]. SignalSpider has been demonstrated to segment the reference human genome (hg19) into different regulatory regions accurately. In particular, it is very effective in capturing the combinatorial relationships among multiple DNA-binding proteins which are probed by ChIP-Seq. Following SignalSpider, SignalRanker and FullSignalRanker have been developed to harness regression and classification tasks on multiple ChIP-Seq profiles[4]. SignalRanker and FullSignalRanker have been demonstrated more accurate than the traditional machine learning methods such as Gaussian Mixture Regression. It is worth noticing that, although those methods are developed in the context of ChIP-Seq, it can easily be adopted to other NGS technologies such as RNA-Seq.

On the other hand, it cannot be ignored that the maturing sequencing technologies have expanded not only the data amount we can analyse, but also the number of high-impact applications we can build. The growing data can provide new and novel insights into human disease studies; for instance, taking advantage of the available proteomes, Wong and Zhang have developed a deleterious residue change prediction method called SNPdryad[5]. It is significant in the sense that SNPdryad has demonstrated competitive performance edge over the well-established methods such as PolyPhen2 and SIFT on PolyPhen2's own datasets (humdiv and humvar). In addition, SNPdryad has been run on the complete human proteome, generating prediction scores for all the possible residue changes on all the known human proteins. Another interesting direction using the NGS technologies is to develop personalized medicine solutions for long-term healthcare. Before the current 21st century, limited by the available data, medical studies are car-

ried out without taking special attention into individual genetic information; for instance, drugs are designed for human groups without any personal customization. Nonetheless, such a situation will be changed because, for now, we can sequence each human genome at less than USD $1000[6]. The human genome information will enable us to take care of each individual patient separately; personalized medicine solutions can be developed.

In the future, the NGS technologies will be applied further to other areas as illustrated by the on-going GTEx project, leading to high-resolution tissue-specific genotype-phenotype studies on our human bodies. In addition, the third-generation technologies are being developed such as Single-Molecule Sequencing in Real Time (SMRT)[7]. SMART is very cost-effective, fast, and less-sample-required. It has lots of competitive features (e.g. long sequencing reads) which can improve the existing sequencing quality, and thus our understanding on genome. Nonetheless, there is not any free lunch. The evolving sequencing technologies will impose a serious big data challenge because it has been estimated that we will have millions of human genome by 2025[8]). We not only need capable data analysts but also sufficient computational machines as well as efficient methods to harness the exponentially growing genome data in the coming future.

References

1. Mardis, E.R. The impact of next-generation sequencing technology on genetics. (2008) Trends Genet 24(3): 133-141.
2. Wong, K.C. Computational Biology and Bioinformatics: Gene Regulation. (2016) CRC Press (Taylor and Francis Group).
3. Wong, K.C., Li, Y., Peng, C., et al. SignalSpider: Probabilistic pattern discovery on multiple normalized ChIP-Seq signal profiles. (2015) Bioinformatics 31(1): 17-24.
4. Wong, K.C., Peng, C., Li, Y. Probabilistic Inference on Multiple Normalized Signal Profiles from Next Generation Sequencing: Transcription Factor Binding Sites. (2015) ACM/IEEE Transactions on Computational Biology and Bioinformatics (99).
5. Wong, K.C., Zhang, Z. SNPdryad: Predicting deleterious non-synonymous human SNPs using only orthologous protein sequences. (2014) Bioinformatics 30(8): 1112-1119.
6. Dijk Erwin, L.V., Auger, H., Jaszczyszyn, Y., et al. Ten years of next-generation sequencing technology. (2014) Trends Genet 30(9): 418-426.
7. Eid, J., Fehr, A., Gray, J., et al. Real-Time DNA Sequencing from Single Polymerase Molecules. (2009) Science 323(5910): 133-138.
8. Stephens, Z.D., Lee, S.Y., Faghri, F., et al. Big Data: Astronomical or Genomical? (2015) PLoS Biol 13(7): e1002195.

Ommega Online Publisher
Bioinformatics, Proteomics and Immaging Analysis
Short Title : Bioinfo Proteom Img Anal

ISSN: 2381-0793
E-mail : bioinfo@ommegaonline.org
website: www.ommegaonline.org