Editorial

Next-Generation Sequencing

Momiao Xiong,¹ Zhongming Zhao,² Jonathan Arnold,³ and Fuli Yu⁴

¹ Human Genetics Center, The University of Texas School of Public Health, Houston, TX 77025, USA

² Departments of Biomedical Informatics and Psychiatry and Vanderbilt-Ingram Cancer Center (VICC), Vanderbilt University Medical Center, Nashville, TN 37232, USA

³ Department of Genetics, The University of Georgia, Life Sciences Building, Athens, GA 30602, USA

⁴ Human Genome Sequencing Center, Baylor College of Medicine, Houston, TX 77030, USA

Correspondence should be addressed to Zhongming Zhao, momiao.xiong@uth.tmc.edu

Received 31 December 2010; Accepted 31 December 2010

Copyright © 2010 Momiao Xiong et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

It has been widely appreciated that the genome sequence is shaping the future biomedical research. The genome sequence provides a general framework for assembling fragmentary DNA information into landscape of biological structure and function [1]. The rapid advances in DNA sequencing technology are revolutionizing biomedical research.

Starting in 2005, a variety of massively parallel sequencing instruments such as the Roche/454, the Life Technologies SOLiD, and the Illumina platforms which were largely different from the Sanger-based capillary sequencing were used to sequence the human and model organism genomes. Although each instrument has its own attributes, all massively parallel sequences machines share some common remarkable features [2]. First, the initial preparatory steps are reduced and simplified. Second, amplification of the library fragments is needed for all platforms. Third, sequencing reactions are performed and detected automatically. In the past decade, the amount of sequence output per run has been dramatically increased, the per-base cost of DNA sequencing has plummeted by ~100,000-fold, and base-calling accuracy has been largely improved. The current second-generation sequencing machines can read ~250 billion bases in a week.

When sequencing becomes simple and inexpensive, it is being routinely applied to biomedical research. To create comprehensive catalogues of genomic variants, the next-generation sequencing technologies have been used to produce sequence data in the 1000 Genomes Project. It plans to sequence more than 2000 individuals to find

essentially all single-nucleotide polymorphisms (SNPs), insertions/deletions (indels), and structural variants with frequency >1% across the genome and >0.1% in protein-coding regions. After the project is completed by 2012, the full spectrum of human genomic variation in large, diverse sample sets will have been identified. The further reduction of cost and improvement of base-calling accuracy will uncover the genetic architectures of complex diseases and make clinical use of genome sequencing a routine practice and create great opportunities for genomic medicine.

Many layers of epigenomic information are being mapped by next-generation sequencing. Chromatic modification and protein binding can be mapped by chromatin immunoprecipitation sequencing (ChIP-Seq). The genome-wide single-base resolution of DNA methylation map has been performed by bisulfate sequencing, in which the methylated cytosines have been chemically modified. Massively parallel sequencing have also been applied to microRNA and mRNA profiling (RNA-Seq) to more accurately measure expressions of microRNA and mRNA, identify variability in microRNA sequence and mRNA sequence, and detect splice form of mRNA expressions.

Massively parallel sequencing platforms have significantly increased our ability to study the human genome and provided powerful new tools for genomic medicine. However, these technologies have also required profound changes to the data analysis. The major obstacle in genomic research is no longer data production. The major challenge in genome sequencing is the methods for data storages,

transfer, and data analysis. The classical statistical methods and computational algorithms are inadequate for analyzing the unprecedented amount of genomic sequence data. Novel analytic strategies for exploring new features of sequencing data, integrating various genomic and epigenomic data, unraveling the structure, organization, and function of the human genome, understanding fundamental principles of genomic biology, and discovering genetic and nongenetic bases of diseases are urgently needed.

This special issue includes six high-quality papers, which were selected after undergoing rigorous peer review. We briefly describe the papers in the following.

The first paper (V. Costa et al., 2010) provides a comprehensive survey of the RNA sequencing methodology. RNA sequencing is a major platform in the next-generation sequencing (NGS), aiming to accurately determine expression levels of specific genes, differential splicing, and allele-specific expression of transcripts at the transcriptome level. So far, RNA sequencing remains the most complex (NGS) application. The authors focus on the challenges that RNA sequencing presents both from a biological and a bioinformatics point of view.

In the second paper (Y. Qi et al., 2010), the authors apply high-throughput sequencing of microRNAs in adenovirus type 3 (AD3) infected human laryngeal epithelial (Hep2) cells. Using the SOLiD sequencing technology, analysis of microRNAs profiles identified 492 precursor microRNAs in the AD3 infected Hep2 cells and 540 precursor microRNAs in the control. Among them, 44 and 36 microRNAs showed high and lower expression in the AD3 infected cells than the control, respectively. The study demonstrates that NGS is efficient and powerful for microRNA profiling in the virus-infected cell lines.

L. Cui et al. (2010) also apply SOLiD sequencing to profile microRNAs involved in the host response to enterovirus 71 (EV71) infection. They found 64 microRNAs whose expression levels changed from more than 2-fold in response to EV71 infection in Hep2 cells. Functional analysis like Gene Ontology enrichment test revealed that many of these microRNAs might be involved in neurological process, immune response, and cell death pathways, which have known to be associated with the extreme virulence of EV71. As authors stated, this is the first paper on host microRNAs expression alteration in response to EV71 infection.

W. Wang et al. (2010) use another NGS technology, ChIP-Seq, to find the targeting microRNA genes of a transcription factor, EGR1, in human erythroleukemia cell line K562. They found EGR1 binding sites near the promoters of 124 distinct microRNA genes, accounting for about 42% of the miRNAs which have high-confidence predicted promoters (294). They also found that EGR1 binds to another 63 pre-miRNAs. This study provides the first global binding profile between the transcription factor EGR1 and its targeting miRNA genes in PMA-treated K562 cells.

S. Hasson et al. (2010) report the cloning of cDNA sequences encoding four groups or isoforms of the haemostasis-disruptive Serine protease proteins (SPs) from the venom glands of *Echis ocellatus*, whose bite is the leading cause of death and morbidity in Africa. Based on

their observation of the extraordinary level of interspecific and intergeneric sequence conservation exhibited by the *Echis ocellatus* EoSPs and analogous serine proteases from other viper species, the authors speculate that antibodies to representative molecules should neutralise the biological function of this important group of venom toxins in vipers that are distributed throughout Africa, the Middle East, and the Indian subcontinent.

The last paper (Z. Zhao and C. Jiang 2010) conducts a comparative genome-wide polymorphism-fixation analysis of human codons, as previously investigators often analyze either interspecies fixed substitutions or intraspecies nucleotide polymorphisms, but not both data types simultaneously. The authors report many features in the recent codon evolution. They conclude that fixation process could effectively and quickly correct the volatile changes introduced by polymorphisms so that codon changes could be gradual and directional and that codon composition could be kept relatively stable during evolution. As numerous mutation data have been identified by sequencing and many more will be identified by NGS in the near future, such analysis may help us understand mutational process in the recent genome evolution.

Acknowledgment

We are especially grateful to the anonymous reviewers who helped improve the quality of the papers in this special issue.

Momiao Xiong
Zhongming Zhao
Jonathan Arnold
Fuli Yu

References

- [1] E. S. Lander, "Initial impact of the sequencing of the human genome," *Nature*, vol. 470, no. 7333, pp. 187–197, 2011.
- [2] E. R. Mardis, "A decade's perspective on DNA sequencing technology," *Nature*, vol. 470, no. 7333, pp. 198–203, 2011.