Next-Generation Sequencing in the Genetics of Human Atrial Fibrillation

Chia-Shan Hsieh,[1] Eric Y. Chuang,[1] Jyh-Ming Jimmy Juang,[2] Juey-Jen Hwang,[2] Chuen-Den Tseng,[2] Fu-Tien Chiang,[2] Ling-Ping Lai,[2] Jiunn-Lee Lin[2] and Chia-Ti Tsai[2]

The International Human Genome Sequencing Consortium published the first draft of the human genome in the journal Nature in February 2001, providing the sequence of the entire genome's three billion base pairs. The Human Genome Project involves a concerted effort to better understand the human DNA sequence through identification of all the genes. The knowledge that can be derived from the genome could result in the development of novel diagnostic assays, targeted therapies and the improved ability to predict the onset, severity and progression of diseases. This has been made possible by many parallelized, high-throughput technologies such as next-generation sequencing. In this review, we discuss the possible application of next-generation sequencing in finding the susceptibility gene(s) or disease mechanism of an important human arrhythmia called atrial fibrillation.

Key Words: Arrhythmia • Atrial fibrillation • Genetics • Next-generation sequencing

INTRODUCTION

It was a novel and benchmark accomplishment when the International Human Genome Sequencing Consortium published the first draft of the human genome in the journal Nature in February 2001. The article provided an overview of the sequence of the entire genome's three billion base pairs. The purpose of the Human Genome Project was to provide a known and solid platform of genetic information, to better understand the human DNA sequence and identify all the genes. Ultimately, the full sequence was completed and published in April 2003.[1,2]

Moreover, it was expected that the knowledge obtained from mapping out the genome would result in the development of novel diagnostic assays, targeted therapies and a more efficacious manner of predicting the onset, severity and progression of diseases. As the human genome slowly becomes increasingly unraveled, and better understood by the medical community, it will have a major impact on medical practice. Recently, genetic information was being used to identify mutations in rare and also in undiagnosed genetic disorders, to assist in selecting the therapy best suited for a particular genotype.[3-7] This has been made possible by many parallelized high-throughput technologies such as microarrays and next-generation sequencing. In this review, we discussed the possible application of high-throughput technologies in an effort to find the susceptibility gene(s) or disease mechanism of an important human arrhythmia, atrial fibrillation.

NEXT-GENERATION SEQUENCING TECHNOLOGIES

Over the past several years, there has been a funda-

Received: April 2, 2013 Accepted: June 5, 2013
[1]Genome and Systems Biology Degree Program, Department of Life Science, National Taiwan University; [2]Cardiovascular Center and Division of Cardiology, Department of Internal Medicine, National Taiwan University Hospital and National Taiwan University College of Medicine, Taipei, Taiwan.
Address correspondence and reprint requests to: Dr. Chia-Ti Tsai, Cardiovascular Center and Division of Cardiology, Department of Internal Medicine, National Taiwan University Hospital, No. 7, Chung-Shan South Road, Taipei, Taiwan. Tel: 886-2-2356-2209; Fax: 886-2-2394-1938; E-mail: cttsai1999@gmail.com; cttsai@ntuh.gov.tw

mental shift away from the application of automated Sanger sequencing for genome analysis. The automated Sanger method is considered as a "first-generation" technology, with newer methods collectively referred to as next-generation sequencing (NGS).

There are pros and cons to these evolving major next-generation sequencing technologies, which are listed in overview form in Table 1.

NEXT-GENERATION SEQUENCING APPLICATION

RNA-sequence (RNA-seq)

NGS technologies can be used for many applications, including RNA-sequence (RNA-seq). The transcriptome is the set of transcripts, including mRNAs, non-coding RNAs and small RNAs in cells, and their quantity. Understanding the transcriptome is essential for interpreting a specific developmental stage, physiological condition and disease. RNA-seq includes variant discovery by resequencing targeted regions of interest or whole genomes, *de novo* assemblies of bacterial and eukaryotic genomes, and cataloguing the transcriptomes of cells, tissues and organisms.[8] Atrial remodeling is the leading mechanism of atrial fibrillation.[9] RNA expression profiling of atrial remodeling could detect the atrial adaptation to this complex arrhythmia and identify genes involved in the mechanism, or mechanisms that perpetuate this complex arrhythmia. Previously, complementary DNA microarray had been used to study expression profiling in atrial fibrillation.[10] Microarray screens the genetic expression based on a pre-specified set of genes as reference. In RNA sequencing using NGS, no gene is specified a priori, and the new sequencing technology sequences all the RNAs (transcriptome) present in the atrial tissue, and compares them between atrial tissues from patients with atrial fibrillation and those of controls to define which gene(s) are activated in the physiological adaptation process during atrial fibrillation, or the mechanism of atrial fibrillation.

ChIP-sequence (ChIP-seq)

A new method used to analyze protein interactions with DNA is chromatin immunoprecipitation followed by sequencing (ChIP-seq). This ChIP-seq combines chromatin immunoprecipitation (ChIP) with massively parallel DNA sequencing to identify the binding sites of DNA-associated proteins. It can be used to precisely map global binding sites for any protein of interest, and is a technique for genome-wide profiling of DNA-binding proteins, histone modifications or nucleosomes. This ChIP-on-chip technique was the most common

Table 1. Overview of major next-generation sequencing technologies listing pros and cons

Platform	Read length (bases)	Throughput (per run)	Pros	Cons
Roche 454 GS Junior	400	35 Mb	Longer read lengths improve mapping in repetitive regions fast run time low cost instrument	Low throughput Complex data analysis
Life SOLiD	35-75	30 Gb	Low error rate Flexibility and scalability	Complex data analysis High cost of instrument and sequence generation
Illumina MiSeq	35-150	1.5-2 Gb	Low cost instrument Low error rate Short run time Acceptable read length High level of multiplexing	Complex data analysis High cost of data generation
Ion Torrent Personal Genome Machine	100-200	1 Gb	Direct signal detection Short run time Acceptable read length	Low throughput
Pacific Biosciences PACBIO RS	3000-5000	100 Mb	Extraordinarily long reads Extremely high accuracy Shortest run time	High cost of instrument High cost of sequencing

technique utilized to study these protein – DNA relations. Due to the next-generation sequencing technology, ChIP-seq offers higher resolution, less noise and greater coverage than the previous array-based method ChIP-chip. NGS technologies can be used to screen genome-wide profiling of DNA-binding proteins, epigenetic marks and chromatin structure (ChIP-seq, methyl-seq and DNase-seq). ChIP has also become a useful tool for understanding transcriptional cascades and interpreting the information encoded in chromatin. In addition, followed by sequencing, ChIP-seq becomes the dominant profiling approach.[11] To extract the most information from ChIP-seq data, integrative analysis with other data types will be essential. For example, the integration of ChIP-seq data with RNA-seq data may result in the elucidation of gene regulatory networks and the characterization of the interplay between the transcriptome and the epigenome.

Application of ChIP-seq in atrial fibrillation includes genetic profiling for a specific transcriptional factor involved in the mechanism of atrial fibrillation (such as STAT3),[12] and establishes those epigenomic markers by methyl-seq that may help identify new genes involved in the epigenetic mechanism of atrial fibrillation, which has never been addressed before.

Metagenomics

Another method, metagenomics, is based on the genomic analysis of microbial DNA that is extracted directly from microbial communities in environmental samples. Currently, AF is not generally thought to be caused by micro-organisms, though the possibility of such a mechanism does exist. Metagenomics can be used to study species classification or gene discovery. By integrating the information gleaned with information about biological functions within the community, the structure of microbial communities can potentially be probed. Metagenomics could also unlock the massive uncultured microbial diversity present in the environment to provide new molecules for therapeutic and biotechnological applications.[13]

Missing heritability of common diseases

Missing heritability of common diseases is probably the most important application of NGS technologies in the genetic research of current common diseases or complex trait diseases, which include atrial fibrillation, hypertension or type 2 diabetes. There are two hypotheses on the inherited basis of complex genetic traits. "Common disease-common variants" consist of many common alleles which have a small effect, and "common disease-rare variants" consisting of a few rare alleles which create a large effect. Both types of genetic loci likely exist, however, and the "common disease-common variants" is the theoretical framework for genome-wide association study (GWAS).[14,15] After years of continuing effort, much of the data generated from GWAS has been published. Among successful GWAS studies, most variants identified confer only a small proportion of heritability, indicating that GWAS based on the "common disease-common variants" is not very effective in identifying genetic variants for complex traits, and common genetic variability is unlikely to explain the entire genetic susceptibility to disease.[16] Results also suggest that rare variants missed by GWAS may account for the "missing" heritability. Such rare variants may have as large an effect as genetic risk factors for complex genetic diseases.[17]

Usually these rare variants could be identified from the familial form of a common disease. For example, atrial fibrillation is a common disease; however, there are several families in which atrial fibrillation segregates in multiple familial members.[18] Several mutations or variants with large effect (significant perturbation of ionic current density) in the potassium channels were found.[19-22] However, screening of these variants in large samples with common atrial fibrillation failed to establish the roles of these potassium channel mutation in the mechanism of common atrial fibrillation.[23,24] Identification of other rare variants using the novel NGS technologies involving DNA samples from patients with extreme phenotypes (such as drug refractory atrial fibrillation) may be possible. These rare variants are supposed to be within the coding region of the susceptibility gene to cause a large pathophysiological effect. Therefore, exon sequencing becomes another popular tool to identify genetic variant(s) within the disease susceptibility gene(s), particularly when explored by a genome-wide approach, the so-called whole exome sequencing.

Whole exome sequencing

One of the remaining challenges for genetic studies

of complex trait diseases will be to define the genetic basis of "missing" heritability. NGS technologies will certainly enable us to identify all the causative variants including "rare variants" within individual human subjects. For example, "whole-genome sequencing (WGS)" will make us understand the genetic contributions to complex diseases, as well as the genetic basis of genomics. The analysis of whole-genome sequence data can be challenging. Because of the current limited ability to make sense of non-coding variation, the analytical components of most "whole genome" studies have focused on variation within the "whole exome". Whole exome sequencing (WES) is a technique to selectively capture and sequence the coding regions of all annotated protein-coding genes. Coupled with next-generation sequencing platforms, it enables the analysis of functional regions of the human genome with unprecedented efficiency. Since its first reported application,[6,25] WES has emerged as a powerful and popular tool for researchers elucidating genetic variants underlying human diseases, especially in the setting of a search for rare or novel variant(s). As mentioned before, rare or novel variant(s) may be responsible for missing heritability of atrial fibrillation, and may be important in the mechanism of atrial fibrillation in certain specific patient populations.

Due to ongoing improvements, sequencing technology has become a method of choice for complex genomic research studies. We are now witnessing its translation into use for clinical diagnostic laboratories involving patient care. Multiple genes for a variety of disorders are now available in several clinical laboratories based on targeted gene enrichment followed by next generation sequencing. The genome-wide study of protein coding regions, or exome sequencing, has been successfully and increasingly applied in the research setting for the elucidation of candidate genes and causal variants in individuals and families with a diversity of rare and complex genetic disorders. Based on this progress, exome sequencing is also beginning a translational process into clinical practice. However, introducing exome sequencing as a diagnostic tool brings new technical and bioinformatical challenges because of the very large quantity of data. Furthermore, most of the time, multiple coding region variants may be identified, leaving clinicians unable to determine which is the true disease-causing variant.

FUTURE DIRECTIONS AND CHALLENGES

Recent research in the field of NGS is very encouraging. However, in order to be widely and routinely used in clinical practice and genetic research, there needs to be additional effort, such as data generation, analysis, management, and measures taken to reduce the cost of sequencing. The major challenges remaining now are making rules and policies around this genetic information easy to interpret, and making sequencing-based genetic tests affordable. This will require sustained collaboration between research labs, clinical physicians and bioinformatic researchers.

Targeted resequencing has already been used to identify genetic mutations underlying disorders such as inflammatory bowel disease,[7] hereditary hearing loss,[26] Miller's syndrome,[27] Kabuki syndrome[28] and hereditary spastic paraparesis.[29] Possible application to other complex trait diseases is encouraging and warrants additional collaborative research.

NGS data analysis is an evolving field and possibly the biggest bottleneck for routine adoption. A NGS data analysis solution for clinics needs to be simple, fast and accurate and should create output that is easily interpreted by medical staff. Unlike microarrays, which have a number of robust and proven analysis solutions providing almost medical report-like output, NGS data analysis is still largely carried out using open source tools.[30-36] These open source tools have been abundantly useful in analyzing NGS data in research labs, but they cannot be used in clinics. Moreover, commercially available NGS data analysis solutions are geared toward handling the abundant data generated in research settings, but have not been designed from a clinical perspective.

Because of the existing limited ability to make sense of non-coding variation, the analytical components of most "whole genome" studies have focused on variation within the "whole exome". As the cost of sequencing continues to fall, the field will probably gradually move from exome to whole-genome sequencing.[37] However, using these more comprehensive data for disease gene discovery and molecular diagnostics in patients crucially

depends on the development of analytical strategies for comprehending non-coding variation.

There are thousands of poorly defined familial phenotypes that are rare or unique. We have to choose the proper phenotypes, particularly in the context of Mendelian disorders. Development of repositories in which descriptive information about such phenotypes and an accompanying DNA sample could be collected by clinicians would facilitate discovery of the underlying genes. Moreover, some technical, statistical and bioinformatic methods need to be improved, including reducing the rate of false-positive and false-negative variant calls, calling insertion/deletion variants (indels), the ranking of candidate causal variants, and predicting and annotating the potential functional impact for disease gene discovery or molecular diagnostics. Translating exome and even genome sequencing data from bench to clinic still requires further sustained effort. When these challenges are more fully resolved, this would then facilitate applications involving diagnosis and therapy.

REFERENCES

1. Lander ES, Linton LM, Birren B, et al. International Human Genome Sequencing Consortium - Initial sequencing and analysis of the human genome. *Nature* 2001;409:860-921.
2. Venter JC, Adams MD, Myers EW, et al. The sequence of the human genome. *Science* 2001;291:1304-51.
3. Bell CJ, Dinwiddie DL, Miller NA, et al. Carrier testing for severe childhood recessive diseases by next-generation sequencing. *Sci Transl Med* 2011;3:65ra4.
4. Sobreira NL, Cirulli ET, Avramopoulos D, et al. Whole-genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet* 2010;6: e1000991.
5. Lam K, Guo H, Wilson GA, et al. Identification of variants in CNGA3 as cause for achromatopsia by exome sequencing of a single patient. *Arch Opthalmol* 2011;129:1212-7.
6. Choi M, Scholl UI, Ji W, et al. Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *PNAS* 2009; 106:19096-101.
7. Worthey EA, Mayer AN, Syverson GD, et al. Making a definitive diagnosis: successful clinical application of whole exome sequencing in a child with intractable inflammatory bowel disease. *Gen Med* 2011;13:255-62.
8. Wang Z, Gerstein M, Snyder M. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Rev Genet* 2009;10:57-63.
9. Nattel S, Burstein B, Dobrev D. Atrial remodeling and atrial fibrillation: mechanisms and implications circulation. *Circ Arrhythm Electrophysiol* 2008;1:62-73.
10. Lai LP, Lin JL, Lin CS, et al. Functional genomic study on atrial fibrillation using cDNA microarray and two-dimensional protein electrophoresis techniques and identification of the myosin regulatory light chain isoform reprogramming in atrial fibrillation. *J Cardiovasc Electrophysiol* 2004;15:214-23.
11. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. *Nature Rev Genet* 2009;10:669-80.
12. Tsai CT, Lai LP, Kuo KT, et al. Angiotensin II activates signal transducer and activators of transcription 3 via Rac1 in atrial myocytes and fibroblasts: implication for the therapeutic effect of statin in atrial structural remodeling. *Circulation* 2008;117:344-55.
13. Petrosino JF, Highlander S, Luna RA, et al. Metagenomic pyrosequencing and microbial identification. *Clin Chem* 2009;55: 856-66.
14. Manolio TA, Collins FS, Cox NJ, et al. Finding the missing heritability of complex diseases. *Nature* 2009;461:747-53.
15. Schork NJ, Murray SS, Frazer KA, Topol EJ. Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 2009; 19:212-9.
16. Singleton AB, Hardy J, Traynor BJ, Houlden H. Towards a complete resolution of the genetic architecture of disease. *Trends Genet* 2010;26:438-42.
17. Frazer KA, Murray SS, Schork NJ, Topol EJ. Human genetic variation and its contribution to complex traits. *Nat Rev Genet* 2009;10:241-51.
18. Brugada R, Tapscott T, Czernuszewicz GZ, et al. Identification of a genetic locus for familial atrial fibrillation. *N Engl J Med* 1997; 336:905-11.
19. Chen YH, Xu SJ, Bendahhou S, et al. KCNQ1 gain-of-function mutation in familial atrial fibrillation. *Science* 2003;299:251-4.
20. Yang Y, Xia M, Jin Q, et al. Identification of a KCNE2 gain-of-function mutation in patients with familial atrial fibrillation. *Am J Hum Genet* 2004;75:899-905.
21. Hong K, Bjerregaard P, Gussak I, Brugada R. Short QT syndrome and atrial fibrillation caused by mutation in KCNH2. *J Cardiovasc Electrophysiol* 2005;16:394-6.
22. Xia M, Jin Q, Bendahhou S, et al. A Kir2.1 gain-of-function mutation underlies familial atrial fibrillation. *Biochem Biophys Res Commun* 2005;332: 1012-9.
23. Ellinor PT, Moore RK, Patton KK, et al. Mutations in the long QT gene, KCNQ1, are an uncommon cause of atrial fibrillation. *Heart* 2004;90:1487-8.
24. Ellinor PT, Petrov-Kondratov VI, Zakharova E, et al. Potassium channel gene mutations rarely cause atrial fibrillation. *BMC Med Genet* 2006;7:70.
25. Ng SB, Turner EH, Robertson PD, et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* 2009;461:272-6.
26. Shearer AE, DeLuca AP, Hildebrand MS, et al. Comprehensive genetic testing for hereditary hearing loss using massively parallel sequencing. *Proc Natl Acad Sci USA* 2010;107:21104-9.
27. Ng SB, Buckingham KJ, Lee C, et al. Exome sequencing identifies

the cause of a Mendelian disorder. *Nat Genet* 2010;42:30-5.
28. Ng SB, Bigham AW, Buckingham KJ, et al. Exome sequencing identifies MLL2 mutations as a cause of Kabuki syndrome. *Nat Genet* 2010;42:790-3.
29. Erlich Y, Edvardson S, Hodges E, et al. Exome sequencing and disease-network analysis of a single family implicate a mutation in KIF1A in hereditary spastic paraparesis. *Genome Res* 2011;21:658-64.
30. Simpson JT, Wong K, Jackman SD, et al. ABySS: a parallel assembler for short read sequence data. *Genome Res* 2009;1117-23.
31. Gnerre S, Maccallum I, Przybylski D, et al. High quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proc Natl Acad Sci USA* 2011;108:1513-8.
32. Grabherr MG, Haas BJ, Yassour M, et al. Full-length transcriptome assembly from RNA-Seq data without reference genome. *Nat Biotechnol* 2011;29:644-52.
33. Hu J, Ng PC. Predicting the effects of frameshifting indels. *Genome Biol* 2012;13:R9.
34. Katz Y, Wang ET, Airoldi EM, Burge CB. Analysis and design of RNA sequencing experiments for identifying isoform regulation. *Nat Methods* 2010;7:1009-15.
35. Kim D, Salzberg SL. Top-Hat fusion: an algorithm for discovery of novel fusion transcript. *Genome Biol* 2011;12:R72.
36. Korbel JO, Abyzov A, Mu XJ, et al. PEMer: a computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 2009;10:R23.
37. Green ED, Guyer MS. Charting a course for genomic medicine from base pairs to bedside. *Nature* 2011;470:204-13.