Review Article

Impact of Next-Generation Sequencing (NGS) technology on cardiovascular disease research

Fengping Xu*, Qin Wang*, Fangfang Zhang*, Yinling Zhu*, Qingquan Gu, Liping Wu, Lin Yang, Xu Yang

BGI-Shenzhen, Shenzhen, China

*These authors contributed equally to this work.

Corresponding to: Xu Yang. BGI-Shenzhen, Shenzhen, China. Email: yangxu@genomics.cn.

Abstract: In recent years, hundreds of gene loci associated with multiple cardiovascular pathologies and traits have been identified through high-throughput Next-Generation Sequencing (NGS) technology. Due to the increasing efficiency and decreasing cost of NGS, rapid progresses anticipated in the field of CVD research. This review summarizes the main strategies of CV research with NGS at the level of genomics, transcriptomics, epigenetics, and proteomics.

Key Words: Cardiovascular disease; Next-Generation Sequencing; genomics; epigenetics; proteomics

Submitted May 19, 2012. Accepted for publication Jun 08, 2012.
DOI: 10.3978/j.issn.2223-3652.2012.06.01
Scan to your mobile device or view this article at: http://www.thecdt.org/article/view/684/753

Introduction

The Human Genome Project was initiated in 1990 in the US with the goal to identify the approximately 20,000-25,000 genes in human DNA, determine the sequences of the 3 billion chemical base pairs that make up human DNA, store this information in databases, and improve tools for data analysis. It was initially anticipated that the project would take 15 years to complete, but rapid technological advances accelerated the completion date to 2003. With the Human Genome Project, and subsequent similar projects including the Hap Map, ENCODE, and the 1,000 Genomes Projects, human genetic disease research has entered into a new era. The basis for the exciting achievements is the availability of high-volume sequencing and analysis of the resulting massive amount of data (1).

Three platforms used widely around the world for massive parallel DNA sequencing are the Roche 454, Illumina Hiseq 2000, and Applied Biosystems SOLiD™ systems, The Illumina Hiseq 2000 has become popular together with the rapid development of bioinformatics methods based on short read nucleotide sequence. Worldwide, genome research centers with extensive sequencing capacity are developing. The Beijing Genomics Institute (BGI) is one of the largest genome sequencing research center in the world, equipped with more than 100 Illumina Hiseq 2000 systems, allowing multilevel high-throughput NGS technologies as shown below (*Figure 1*).

Cardiovascular disease (CVD) is a class of complex pathologies of the heart and blood vessels, including coronary artery disease (heart attack), cerebrovascular disease (stroke), elevated blood pressure (hypertension), peripheral artery disease, rheumatic heart disease, congenital heart disease and heart failure. According to a recent report of world health organization (WHO), CVD is the leading cause of mortality in the world, especially in low and middle-income countries (*Figure 2*) (http://www.who.int/en/). With the development of rapid sequencing technologies, NGS has revolutionized approaches for biomedical cardiovascular research. Considerable progress has been made in the field of genome research related to CVD and hundreds of loci associated with cardiovascular pathologies have been identified (2). However, because of the complexity of CVDs, it is insufficient to focus on the DNA level alone. We suggest that the direction of CVD research will increasingly focus on multilevel NGS analysis, at the level of genomics, transcriptomics, epigenetics and proteomics.

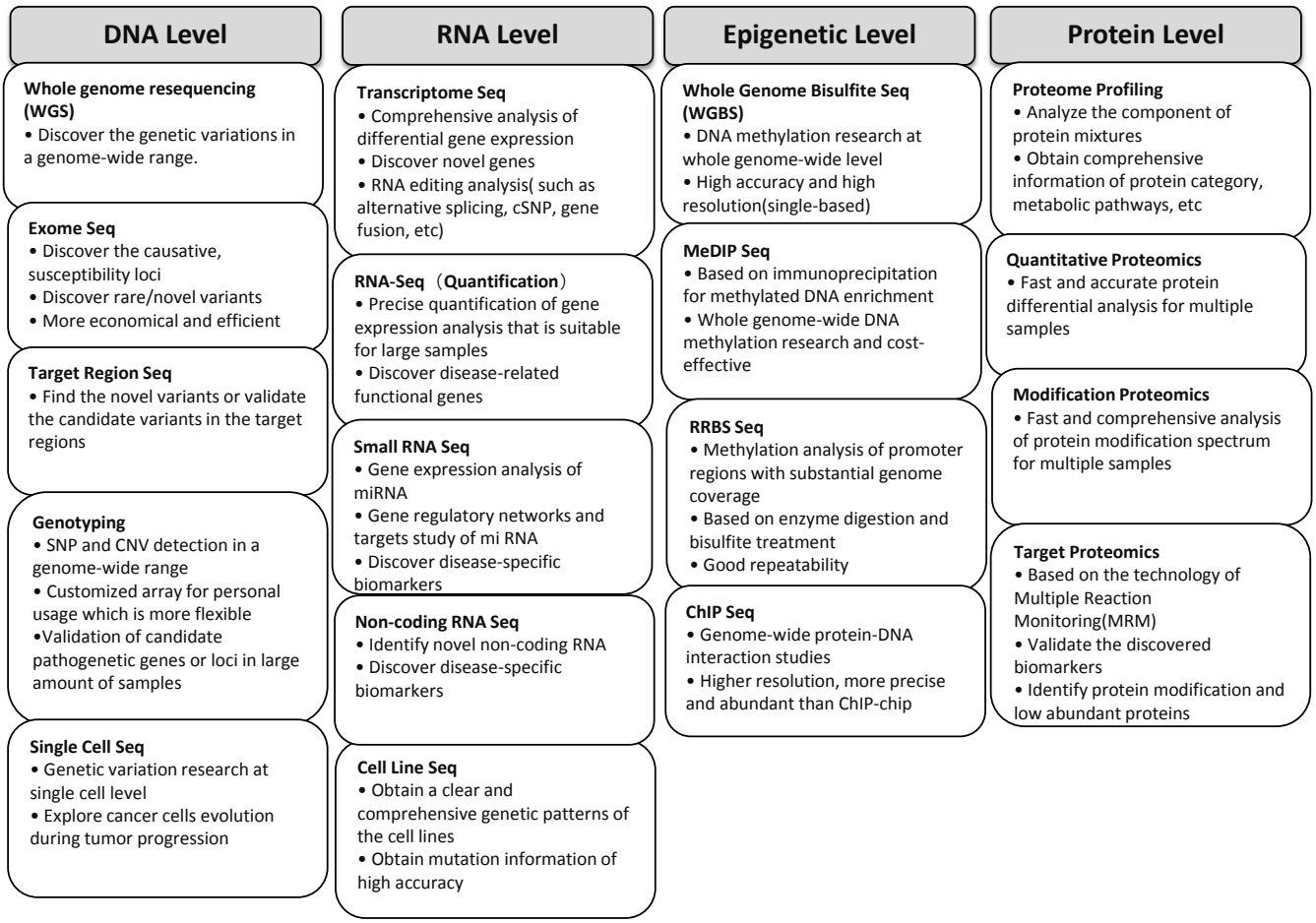

Figure 1 Multilevel high-throughput NGS technologies in BGI

In the era prior to the Human Genome Project, genes associated with Mendelian cardiovascular disease were typically identified through candidate gene studies. Sequencing of the entire human genome (1) has exponentially expanded the understanding of genetic contributions to cardiovascular disease. However, this required sequencing and processing of tremendous amount of data. This has become possible with high-throughput technologies, including NGS and mass spectrometry combined with super-computer technology. For example, exome-capture and whole-genome sequencing could identify rare and novel genetic variants associated with CVDs. However, accumulating research has demonstrated that static variations of DNA sequence can explain only a fraction of the inherited phenotype. These data suggest that additional epigenetic and gene expression mechanisms are necessary to explain the expression of CV disease in both experimental and clinical settings. Eventually, a comprehensive approach will be needed to integrate the accumulated multilevel data (3). As one of the largest genome research centers in the world, BGI has devoted efforts towards this goal.

In this review, the main strategies of CVD genetic research, which are expected to provide novel insights into CVDs, will be summarized with a focus to NGS application at the level of genomics, transcriptomics, epigenetics, and proteomics.

NGS-GWAS strategy to identify susceptibility/causative genes

The strategy of genome-wide association study (GWAS) is widely used, specifically following the completion of the HapMap project, and has largely improved the efficiency of disease-related gene discovery. It is designed to identify candidate variants that contribute to complex diseases.

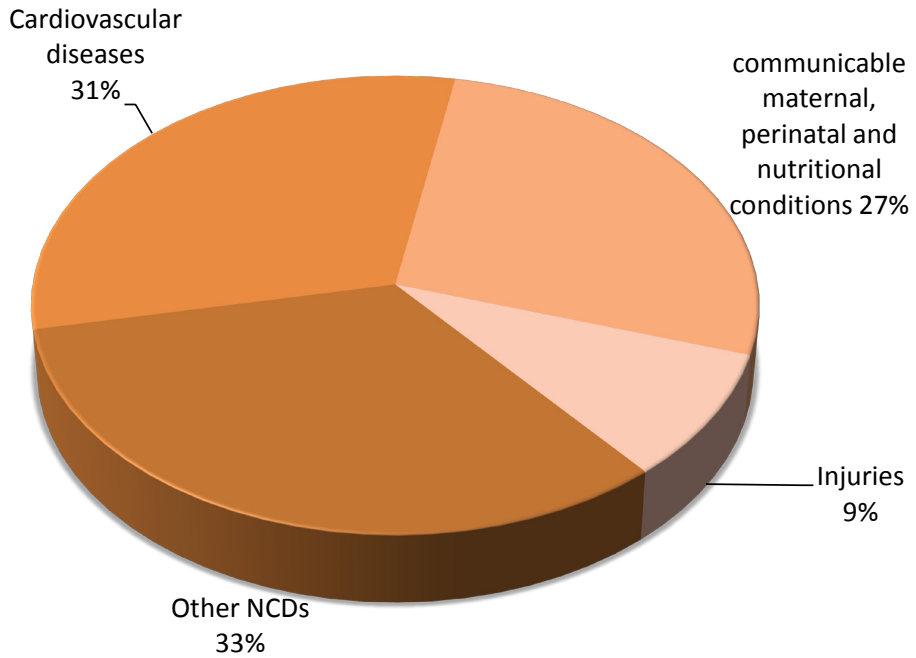

Figure 2 Distribution of major causes of deaths including CVDs

The wide application of GWAS has led to an enormous boost in the discovery of susceptibility genes for CVDs. Multiple novel genetic loci have been identified in common cardiovascular conditions, including myocardial infarction, hypertension, heart failure, stroke and hyperlipidemia. Up to now, 26 risk loci have been identified by GWAS to be associated with coronary artery diseases (CAD) (4-6). However, only a small fraction of the heritable risk for CVDs can be explained by the variants identified by current GWAS.

Genome-wide association studies comprise two or more stages, including the discovery stage, followed by at least one replication stage (*Figure 3*). Only variants of loci for which the association observed at the discovery stage is confirmed at the replication stage(s) are considered 'true hits'. Subsequently, the identified susceptibility loci will be further characterized in molecular and physiological research to determine the mechanisms by which these loci confer susceptibility.

The strategy of NGS-GWAS combines next-generation sequencing and genotyping to uncover novel causative genetic variants of complex diseases. Compared with traditional GWAS, it can provide more detailed information, including not only common SNPs, but also rare variants. Based on the principle of GWAS, we have developed the two-stage NGS-GWAS strategy to find susceptibility or causative loci. It is expected to help understand the pathogenesis of complex diseases at the genome-wide level with whole genome resequencing, whole exome sequencing, target region sequencing or whole genome genotyping. Sporadic or core family samples can be processed. The validation should be followed by characterizing the candidate genes (*Figure 4*). The information obtained from this NGS-GWAS approach will be subjected to bioinformatic analysis to identify the disease-associated genes and better understand the mechanisms of underlying diseases.

Combining chip-based GWAS and RNA sequencing to identify disease-related genes

Traditional chip-based GWAS has the limitation of missing heritability, because it may miss the novel and rare variants related with complex diseases due to the design of chip. However, combined with RNA sequencing, it can complement the disadvantage of chip-based GWAS, and thus identify the disease-associated genes in a cost-effective way.

With this approach, large case/control cohorts are typically genotyped by whole genome genotyping chips, and putative SNPs will be selected. Meanwhile, RNA sequencing will be performed in independent case/control

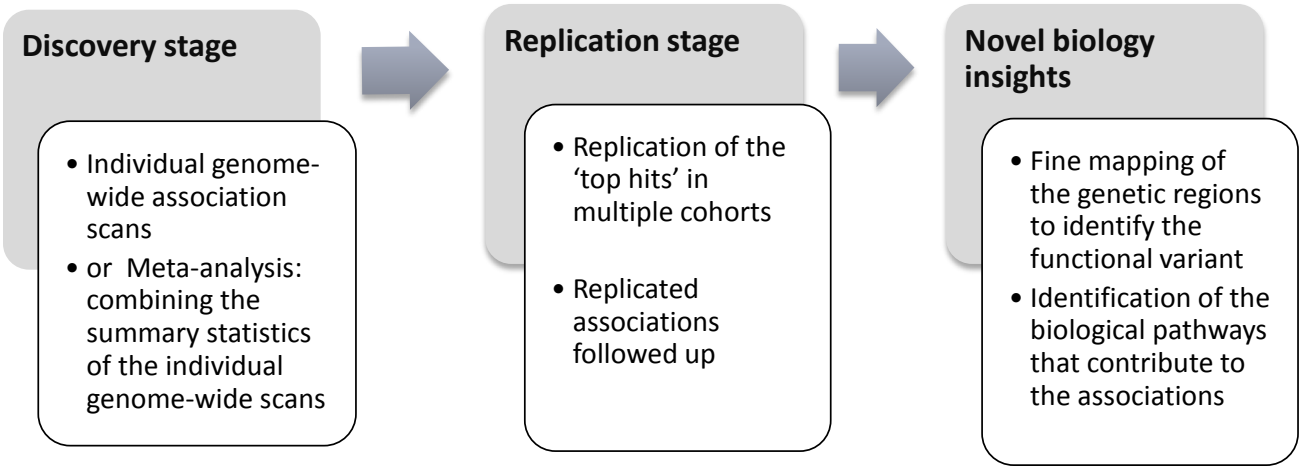

Figure 3 Strategies involved in a genome-wide association study

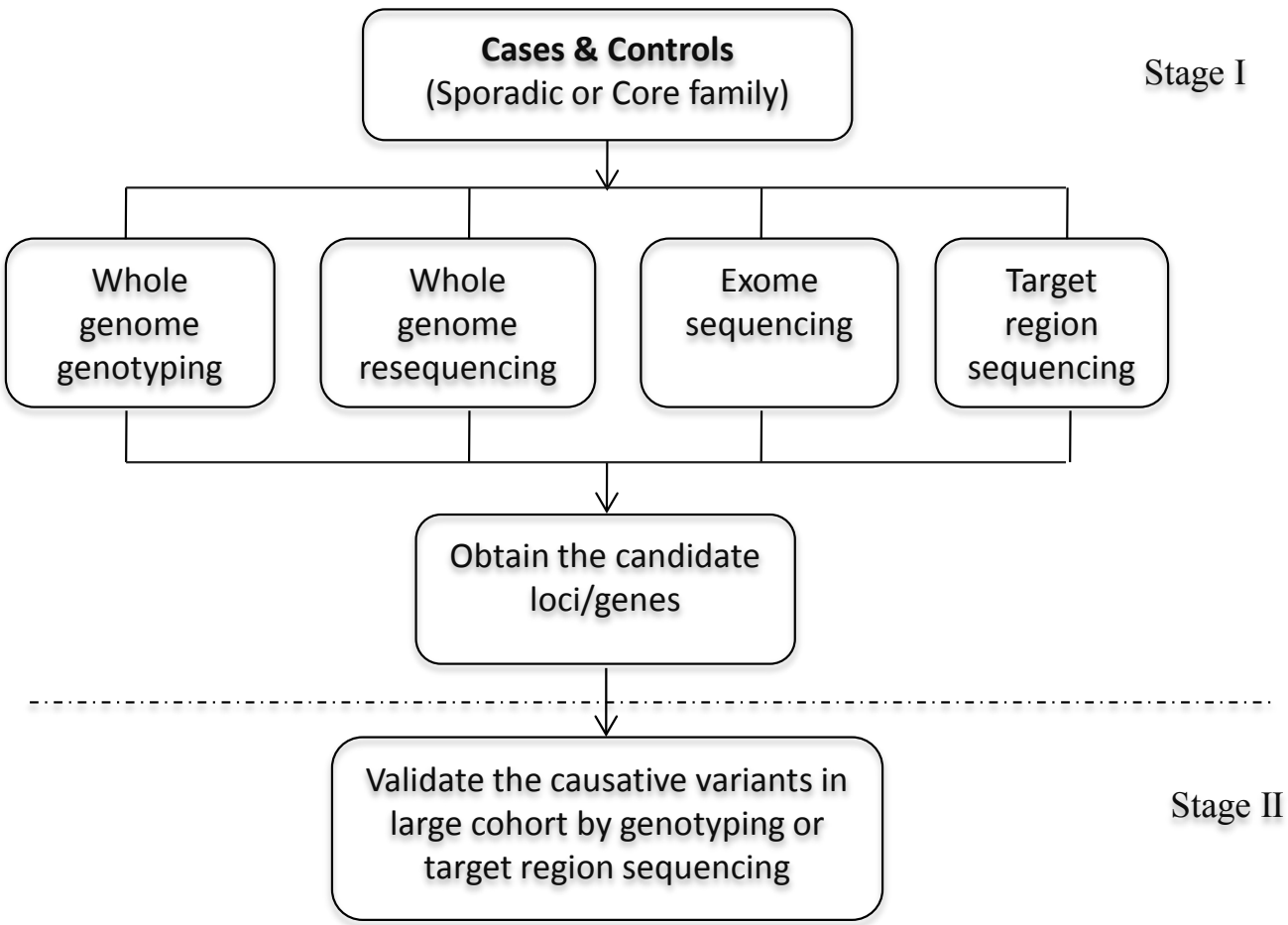

Figure 4 Workflow of NGS-GWAS approach

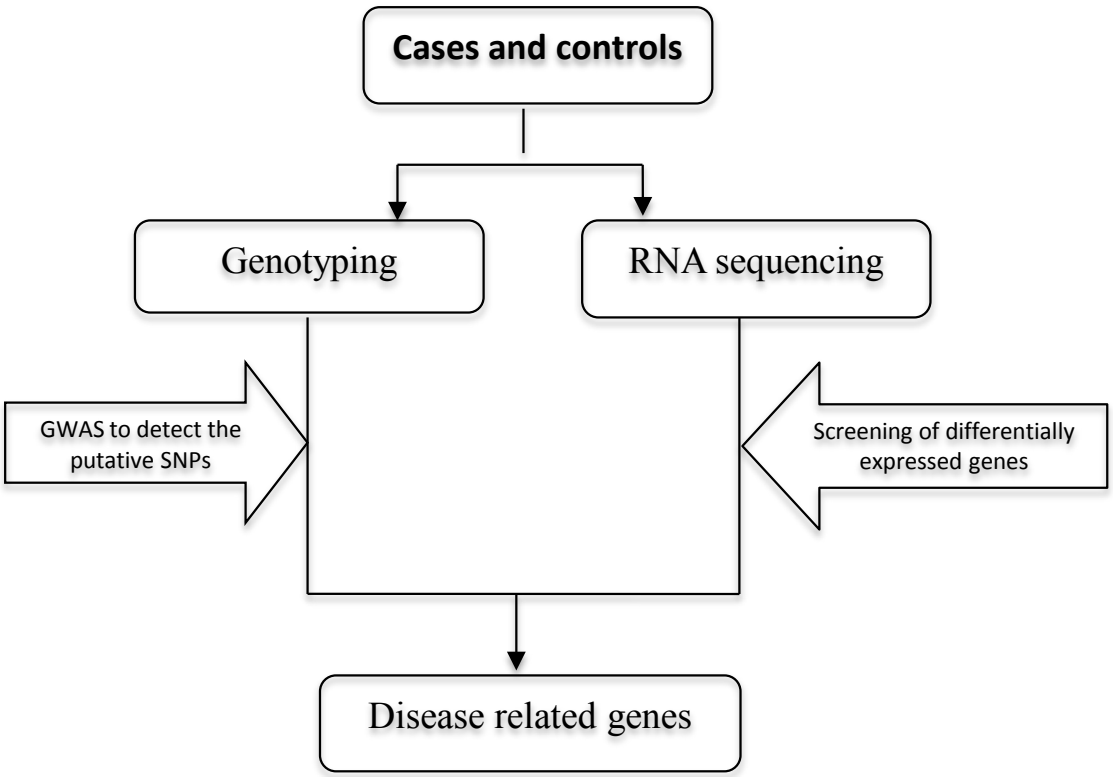

Figure 5 Whole genome genotyping and RNA-seq to find disease related genes

cohorts. The differentially expressed genes will be detected. Finally the genotyping data and expression data will be integrated and analyzed to identify the disease-associated genes (*Figure 5*).

Epigenetic strategies to find epigenetic markers

Epigenetic mark, established almost entirely during embryogenesis and early development, may affect cardiovascular development through differentiation of stem cell, endothelial cells, and cardiac and vascular muscle cells. It may be a uniform principle in the etiology of CVD. However, up to now our knowledge about epigenomic mechanisms has developed on a basic science level, and experience with clinical CVD associations remains limited.

The advances of NGS technology make it possible to assess epigenetic marks at genome-wide scale. For DNA methylation studies, several approaches are used based on NGS, including WGBS (whole genome bisulfite sequencing), MeDIP-Seq (methylated DNA immunoprecipitation sequencing) and RRBS (reduced representation bisulfate sequencing). WGBS is the current standard method, because of its unbiased genome coverage. By combining bisulfite treatment of DNA and sequencing, it allows for a single-base resolution and high accuracy cytosine methylome map. MeDIP is an enrichment-based method for cost-effective DNA methylome study, and RRBS mainly focuses on CpG-rich regions based on enzyme digestion. These methods can be used for different methylated region (DMR) research. For histone modification study, ChIP-Seq (chromatin immunoprecipitation sequencing) is widely used to determine how transcription factors and other chromatin-associated proteins interact with DNA to regulate gene expression. Besides these epigenetic modifications, the increased integration of epigenetic data with genomic and transcriptomic data will enable powerful systematic analyses. Combining epigenetic research with gene expression, it is expected to find candidate epigenetic markers that are related with cardiovascular disease (*Figure 6*).

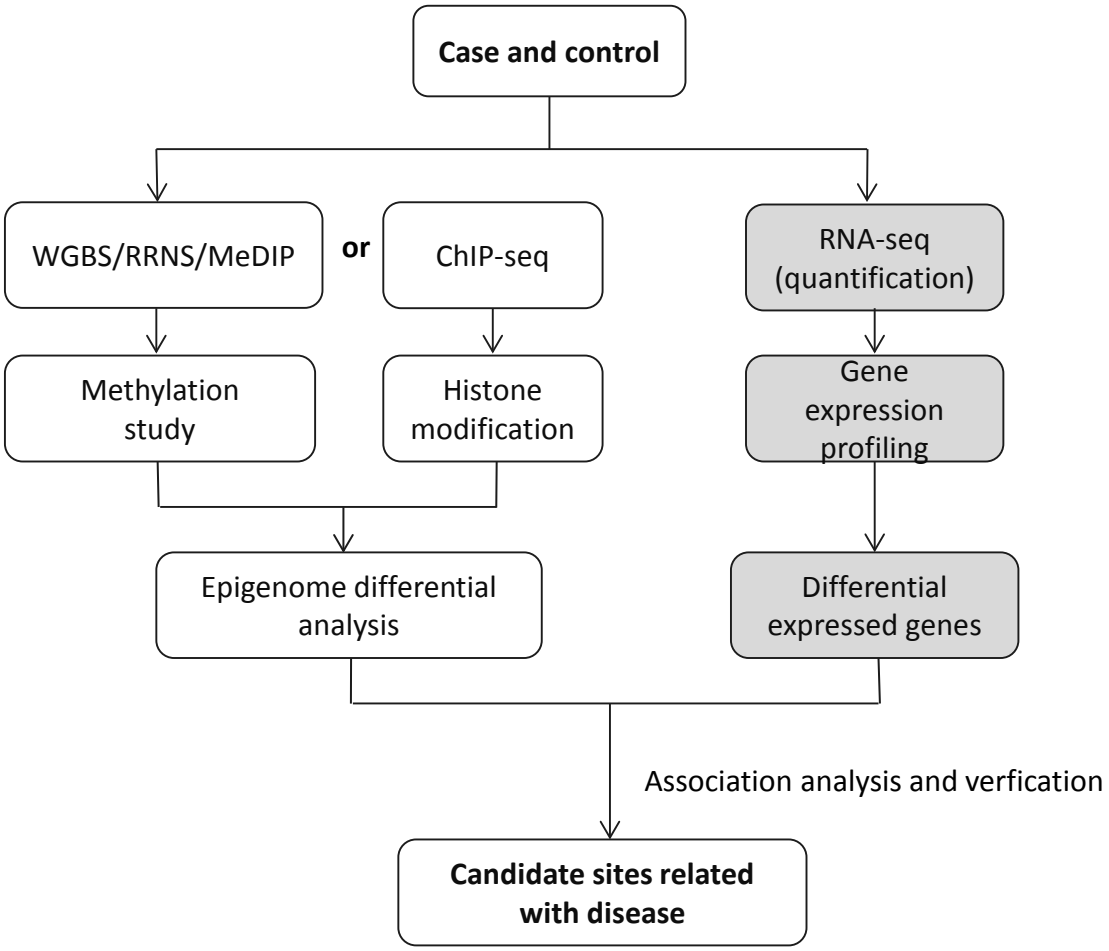

Figure 6 Workflow of identifying the epigenetic marks

Candidate biomarker screening using proteomic strategies

Numerous studies have used proteomic strategies to discover candidate protein biomarkers for a range of diseases, including cardiovascular conditions (7-12). However, up to now, protein biomarker has been identified by proteomics for clinical use. In general, the majority of commonly used clinical biomarkers are derived from blood samples, and it is therefore important to focus during the discovery stage on candidate proteins that are likely to be detectable in plasma. A major challenge is to predict which of the hundreds of differentially abundant proteins detected are truly related with diseases. It is therefore important to design systematic approach integratingproteomic technologies in four stages, including discovery, qualification, verification and validation (*Figure* 7).

Clinical application of multi-omics strategy

Rapid evolution of sequencing technology will enable comprehensive analysis of genomic, epigenomic and transcriptomic data. The integration of multi-omics data will enable clearer understanding of disease-associated loci. This biological knowledge is the first step towards clinical application, including preventive and therapeutic medicine.

The identified susceptibility loci will be verified in molecular and physiological studies to determine the mechanisms through which these loci confer susceptibility. Currently, the predictive value of identified susceptibility/causative loci is too low to be clinically useful. As more variants are identified, certain loci will be predicted to be valuable in clinic. In the era of targeted therapy, molecular biomarkers are becoming increasingly important within both clinical research and clinical practice.

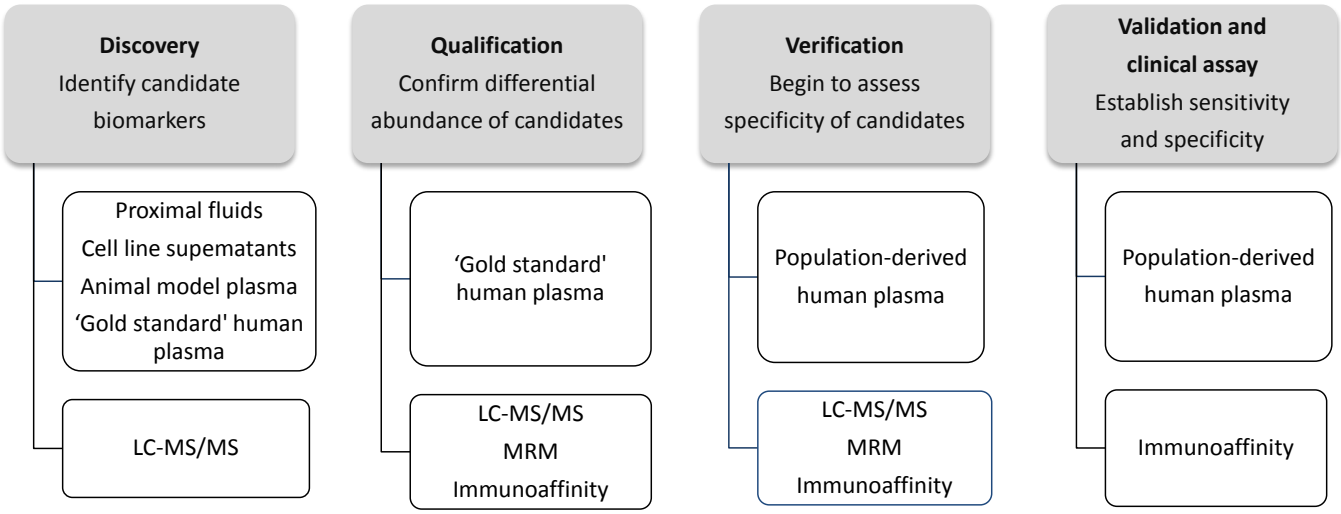

Figure 7 Workflow of identifying protein biomarkers

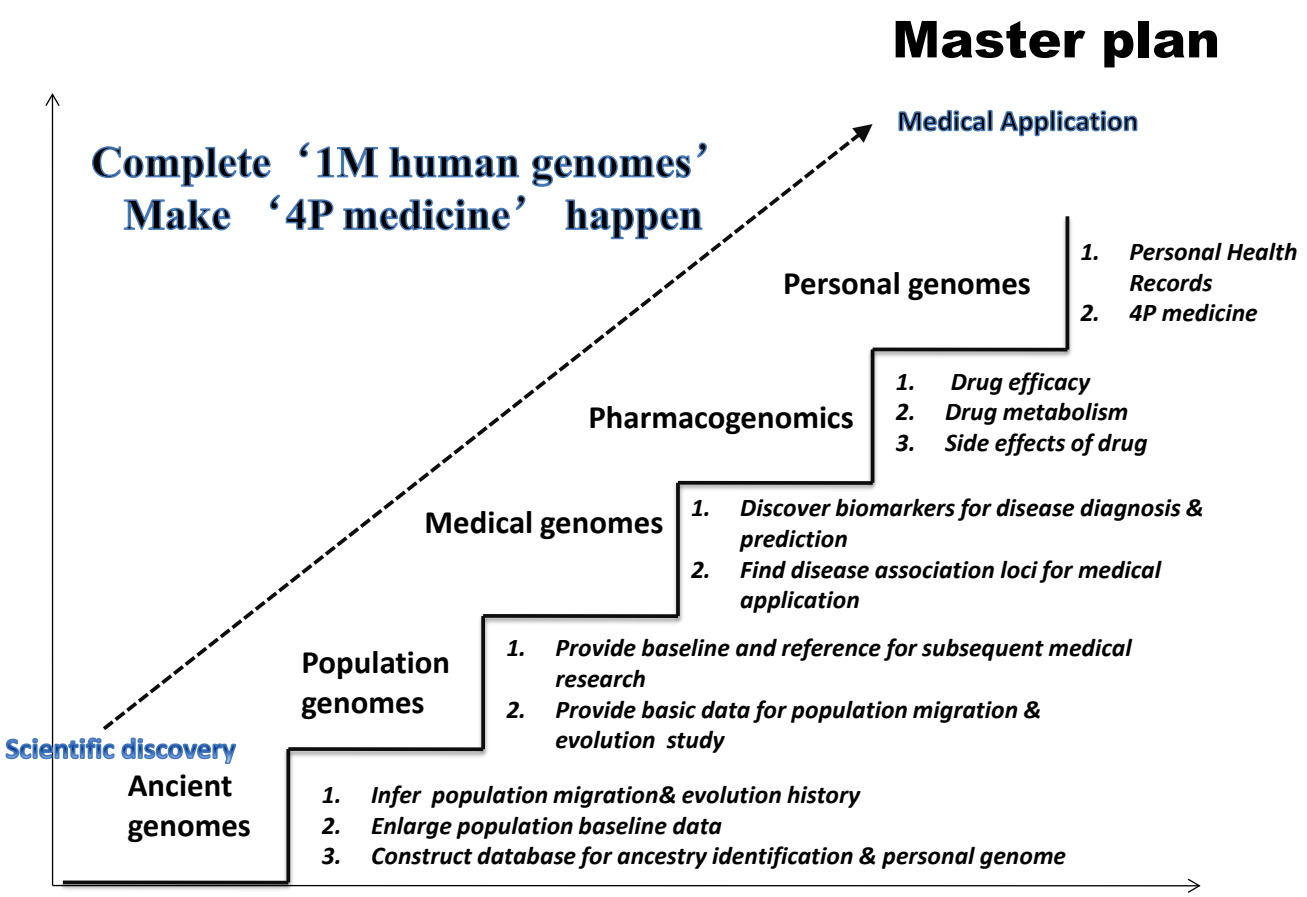

Figure 8 The blueprint of 1 Million Human Genomes Project

1 Million Human Genomes Project launched by BGI

At the ICG-6 conference in 2011, BGI has launched the "1 Million Human Genomes Project". It's a worldwide collaboration project, in which the world's pre-eminent scientists are invited to join together to pave the way for improving human health. The 1 Million Human Genomes Project includes ancient human genome studies to provide an evolutionary perspective, population genomic studies based on large populations or individual specific variation, clinically derived genomes focusing on genetic diseases, and pharmacogenomics. An important goal is to determine drug efficacy in individual persons for optimization of drugs use. The ultimate objective is to develop the "4P model" - predication, prevention, precaution, and personalized healthcare (*Figure 8*).

In a fairly short time, collaborations and partnerships have been initiated with many institutions worldwide. These include collaboration between Autism Speaks and BGI, with the goal to generate the world's largest library of sequenced genomes of individuals with Autism Spectrum Disorders. Using the Autism Speaks Autism Genetic Resource Exchange (AGRE), this collaboration will perform whole genome sequencing on more than 2,000 participating families who have two or more children on the autism spectrum. The data from the 10,000 AGRE participants will enable new research in the genomics of ASD, and significantly enhance the science and technology networks of both Autism Speaks and BGI. In addition, Autism Speaks and BGI will collect and sequence genome samples from individuals in China. This collaboration will be conducted over a two-year period. The initial pilot sequencing of 100 genomes will be directly funded by the Autism Speaks science portfolio. Additional funding will be secured from government, donors, and public and private sources. In November 2011, the Children's Hospital of Philadelphia and BGI announce a partnership establishing a new joint genome center to target pediatric diseases. BGI's capabilities and expertise in whole genome sequencing and analysis, combined with Children's Hospital's extensive biobank and expertise in clinical phenotyping, will allow scientists and clinicians to harness the power of large, detailed data sets to improve the lives of patients and families. Together the partners will also host an international conference on Genomics in Philodelphia in 2012 (http://www.icgamericas.org/). The Inner Mongolia University for the Nationalities (IMUN) and BGI recently announced to cooperate in the first complete sequencing of Mongolian genome. This genomic study will help researchers to better understand the evolutionary process and migration of Mongolians and their ancestors from Africa to Asia, which also lays an important genomic foundation for further development of human genetic diseases research.

As a scientific partner, BGI offers flexible collaboration models in the context of the "1 Million Human Genomes Project". BGI intends to work with principal investigators around the world covering established and emerging fields. Both partners will participate in research design and work together to translate scientific findings to clinical application.

About BGI

BGI was founded in Beijing, China on September 9th, 1999 with the mission of being a premier scientific partner to the global research community. The goal of BGI is to make leading-edge genomic science wildly accessible. BGI, and its affiliates, BGI Americas and BGI Europe, have established partnerships and collaborations with leading academic and government research institutions as well as global biotechnology and pharmaceutical companies, supporting a variety of disease, agricultural, environmental, and related applications.

BGI has established a proven track record of excellence, delivering results with high efficiency and accuracy for innovative, high-profile research which has generated over 170 publications in top-tier journals such as Cell, Nature and Science (http://en.genomics.cn/navigation/show_navigation.action?navigation.id=97). These accomplishments include sequencing one percent of the human genome for the International Human Genome Project, contributing 10 percent to the International Human HapMap Project, join the 1,000 genomes project, carrying out research to combat SARS and German deadly E. coli.

Acknowledgements

Disclosure: The authors declare no conflict of interest.

References

1. Torkamani A, Scott-Van Zeeland AA, Topol EJ, et al. Annotating individual human genomes. Genomics 2011;98:233-41.
2. Hellings WE, Moll FL, De Kleijn DP, et al. 10-years experience with the Athero-Express study. Cardiovasc Diagn Ther 2012;2:63-73.
3. Schnabel RB, Baccarelli A, Lin H, et al. Next steps in cardiovascular disease genomic research--sequencing,

epigenetics, and transcriptomics. Clin Chem 2012;58:113-26.
4. O'Donnell CJ, Nabel EG. Genomics of cardiovascular disease. N Engl J Med 2011;365:2098-109.
5. Zeller T, Blankenberg S, Diemert P. Genomewide association studies in cardiovascular disease--an update 2011. Clin Chem 2012;58:92-103.
6. Harismendy O, Notani D, Song X, et al. 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature 2011;470:264-8.
7. Edwards AV, White MY, Cordwell SJ. The role of proteomics in clinical cardiovascular biomarker discovery. Mol Cell Proteomics 2008;7:1824-37.
8. Jacquet S, Yin X, Sicard P, et al. Identification of cardiac myosin-binding protein C as a candidate biomarker of myocardial Infarction by proteomics analysis. Mol Cell Proteomics 2009;8:2687-99.
9. Fu Q, Van Eyk JE. Proteomics and heart disease: identifying biomarkers of clinical utility. Expert Rev Proteomics 2006;3:237-49.
10. Anderson NL. The clinical plasma proteome: a survey of clinical assays for proteins in plasma and serum. Clin Chem 2010;56:177-85.
11. Addona TA, Shi X, Keshishian H, et al. A pipeline that integrates the discovery and verification of plasma protein biomarkers reveals candidate markers for cardiovascular disease. Nat Biotechnol 2011;29:635-43.
12. Rifai N, Gillette MA, Carr SA. et al. Protein biomarker discovery and validation:the long and uncertain path to clinical utility. Nat Biotechnol 2006;24:971-83.

Cite this article as: Xu F, Wang Q, Zhang F, Zhu Y, Gu Q, Wu L, Yang L, Yang X. Impact of Next-Generation Sequencing (NGS) technology on cardiovascular disease research. Cardiovasc Diagn Ther 2012;2(2):138-146. DOI: 10.3978/j.issn.2223-3652.2012.06.01

