



Editorial

## Genetic Studies of Complex Diseases in the Sequence Era

Momiao Xiong<sup>1\*</sup>

Fast and cheaper next generation sequencing (NGS) technologies will generate unprecedentedly massive (thousands or even ten thousands of individuals) and highly-dimensional (ten or even dozens of millions) genomic and epigenomic variation data. Due to advances in measurement technologies and communications, a large panel of physiological data including all medical treatments and outcomes accumulated over the patient's lifetime is monitored and collected. Analysis of these extremely big and diverse types of data sets provide invaluable information for holistic discovery of the genetic and epigenetic structure of disease and prediction, prevention, diagnosis and treatment of disease, but also pose great conceptual, analytical and computational challenges.

A deluge of genomic and epigenomic data generated by NGS and enormous amounts of personal phenotype data demand the paradigm shift in genomic and epigenomic data analysis from standard multivariate data analysis to functional data analysis, from low dimensional data analysis to high dimensional data analysis, from independent sampling to dependent sampling, from single type data analysis to integrated multiple types of data analysis, and from individual PC to parallel computing.

The data produced by next-generation sequencing technologies will suffer from three basic problems: high error rates, enrichment of rare variants and large proportion of missing values [1,2]. Since an individual rare variant would have a relatively small impact on the common disease and the rare variants have very low frequencies in the populations, the power of the traditional analytical tools that are mainly designed for the purpose of detecting common variants, for testing association of rare variants will be limited. Developing new analytical tools for the analysis of the massive sequencing data is indispensable.

Population-based sample design is the current major study design for association studies. However, many rare variants are from recent mutations in pedigrees [3]. The inability of common variants to account for most of the supposed heritability and the low power of population-based analysis tests for the association of rare variants have led to a renewed interest in family-based design with enrichment for risk alleles to detect the association of rare variants. It is hypothesized that an individual's disease risk is likely to come from the collected action of common variants segregating in the population and rare variants recently arising in extended pedigrees.

\*Corresponding author: Momiao Xiong, Division of Biostatistics, The University of Texas School of Public Health, Houston, TX 77030, USA, E-mail: momiao.xiong@uth.tmc.edu

Received: June 20, 2012 Accepted: June 21, 2012 Published: June 25, 2012

It is increasingly recognized that analyzing samples from populations and pedigrees separately is highly inefficient. It is natural to unify population and family study designs for association studies.

The NGS technologies will detect millions or even dozens of millions of genetic variants. The extremely large amount of sequence data raises great challenges for data analysis. To simultaneously use all available genomic and epigenomic information will dramatically improve our ability to evaluate the roles of genomic and epigenomic variation in understanding and predicting inter-individual phenotypic variation. The immediate question for jointly using all available genomic and epigenomic information is how to handle millions or even ten millions of genomic and epigenomic variants in the analysis. Graphical representation of genomic and epigenomic variants and high dimensional data reduction techniques will provide a powerful tool to efficiently compress large numbers of genomic and epigenomic variants, improve the prediction of phenotypic variation, and gain new insights into the complex genetic and epigenetic structure of common diseases.

The past decade witnessed rapid development in high-throughput techniques that are capable of discovering millions or even ten millions of genomic and epigenomic variants in the individuals. Massive amount of high-dimensional data will provide holistic and complementary information that will facilitate to unravel mechanism of complex diseases. However, this substantial amount of data will also create a fundamental problem in genomic and epigenomic data analysis, low rate of signal-to-noise. An important strategy for integrating all available genomic and epigenomic information is to conduct analysis of complex genomic and epigenomic networks and study their behaviors under genetic and epigenetic perturbations. The components of complex genomic and epigenomic networks that determine cell function, the phenotype outcomes and response to perturbation of external stimuli include genetic variants such as SNPs and CNVs, mRNAs, miRNAs, and methylation connected by their co-expressions or interactions. It is the whole system and the system dynamics that play an essential role in giving rise to cellular function/dysfunction. Investigating how these networks respond to perturbation of mutations, methylation and environments will greatly facilitate to gain new insights into the complex genetic and epigenetic structure of complex diseases and unravel their mechanism.

Develop new analytic paradigm, novel statistical methods and explore the power of parallel computing for sequence-based genomic and epigenomic data analysis to overcome the serious limitation of the current paradigm and statistical methods for genomic and epigenomic data analysis. Journal of Genetic Disorders & Disease Information provides excellent platforms to present new concepts, methods, and designs for genetic studies of diseases and communicate novel ideas and valuable results among geneticists. We can expect that as new sequence technologies bring individual genome sequencing closer to reality and rich of health care and clinical data are available, publication of the Journal of Genetic Disorders & Disease Information will stimulate the development of novel methods for

disease prediction, prevention, diagnosis and treatment, improve the health care, and hold promise of shifting the focus of health care from the disease to wellness where individual wellness status is monitored.

**References**

1. Bansal V, Libiger O, Torkamani A, Schork NJ (2010) Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11: 773-785.

2. Pool JE, Hellmann I, Jensen JD, Nielsen R (2010) Population genetic inference from genomic sequence variation. *Genome Res* 20: 291-300.  
3. Lupski JR, Belmont JW, Boerwinkle E, Gibbs RA (2011) Clan genomics and the complex architecture of human disease. *Cell* 147: 32-43.

**Author Affiliation**

[Top](#)

*<sup>1</sup>Division of Biostatistics, The University of Texas School of Public Health, Houston, TX 77030, USA*

**Submit your next manuscript and get advantages of SciTechnol submissions**

- ❖ 50 Journals
- ❖ 21 Day rapid review process
- ❖ 1000 Editorial team
- ❖ 2 Million readers
- ❖ More than 5000 
- ❖ Publication immediately after acceptance
- ❖ Quality and quick editorial, review processing

---

Submit your next manuscript at • [www.scitechnol.com/submission](http://www.scitechnol.com/submission)

---