What have we learned from the past 7 years and the millions of dollars spent on the genome-wide association studies?

Suktitipat B, MD

Bhoom Suktitipat, MD, PhD[1, 2]
Chayanon Peerapittayamongkol, MD, PhD[1]

Keywords: genome-wide association studies, next-generation sequencing

[1] Statistical genetics and genetic epidemiology research laboratory Department of Biochemistry, Faculty of Medicine Siriraj Hospital, Mahidol University, Bangkok, Thailand.
[2] Integrative Computational BioScience Center, Mahidol University, Thailand.

** Address Correspondence to author:*
Bhoom Suktitipat, MD
2 Wanglang Rd., Siriraj, Bangkoknoi
Bangkok 10700, Thailand.
e-mail: bhoom.suk@mahidol.ac.th,
chayanon.pee@mahidol.ac.th

Received: January 17, 2014
Revision received: January 24, 2014
Accepted after revision: February 1, 2014
Bangkok Med J 2014;7:99-101.
E-journal: http://www.bangkokmedjournal.com

The advance in genomic technology has allowed us to genotype millions of single nucleotide polymorphisms (SNPs) across the human genome at the same time. This high-throughput technology, genome-wide association studies (GWAS), enables us to search for genes contributing to many diseases from age-related macular degeneration, type 2 diabetes, coronary heart disease, cancer to common traits such as obesity, height, eye color. To date, there are over 1,779 studies and 12,126 SNPs reported to be associated with more than 300 diseases.[1]

Quite a few of GWASs have been conducted in the Thai population. Searching from HuGE Navigator (an integrated, searchable knowledge base of genetic associations and human genome epidemiology: http://hugenavigator.net/) with a search term **"Thai"** showed 7 articles. Disease phonotypes studied in the Thai population using the GWAS approach so far are: severity in β0-thalassemia/Hb E[2], susceptibility to tuberculosis[3], chronic hepatitis B[4,5], systemic lupus erythematosus (SLE)[6], thyrotoxic hypokalemic periodic paralysis[7] and nevirapine-induced rash in HIV infected patients.[8] The results from these GWAS have been used for personalized medicine ranging from disease risk prediction, treatment selection and medication dosing guide.[9]

The origin of GWAS design came from an attempt to map genomic loci to a disease. Sequencing the whole 3 billion nucleotides in the genome was prohibitively expensive. To study the whole genome, geneticists employed a concept of tagging SNP, using a SNP as a proxy marker for nearby genetic variants. Through correlation between untyped genetic variants and genotyped SNPs, several tagged genetic variants can be represented with a single SNP on a GWAS panel. The SNPs on the GWAS panel are then used to test for association with diseases. Genome-wide significant association signals imply that the genomic location of these variants play a role in disease-causing mechanisms. For example, in a genome-wide association study, rs11591147 was found to have genome-wide significantly associated with hypercholesterolemia.[10] As the rs11591147 is located within *PCSK9* gene, the function of *PCSK9* will be implicated in hypercholesterolemia. Functional implication of these SNPs is quite straightforward if the identified SNPs are functional variants, e.g. non-synonymous mutation, stop codon mutation. However, as more than 80% of the signals from GWAS fall into inter-genic or non-coding regions, establishing the causal relationship between these variants and disease mechanisms is still challenging.

The major success of GWAS is in identification of several new common disease loci. GWAS assume that most complex diseases can be explained by a few common genetic variants. This assumption is known as the common disease/common variant hypothesis

(CDCV). CDCV has been proven in several studies of common complex genetic diseases. For example, the region on chromosome 9p21. 3 near genes have been associated with coronary artery disease in European descent population the studies of three common SNPs (rs1333049, rs10757274, and rs2383207)[11], all have MAF close to 50%.[12]

Although correlations between common variants and some less common variants allow us to identify the variants with MAF < 5% from the GWAS platform, many rare variants that are not tagged by common variants may have been missed. As a result, we are left with an incomplete characterization of the underlying genetic mechanisms of the studied diseases. Correlation between genetic markers (linkage disequilibrium: LD) can also cause a problem in identification of causal loci in a region with high LD among genetic variants.

GWAS also allow us to estimate genetic risks from the associated genetic variants. Using the same CAD associated variants mentioned above as an example, 9p21.3 risk loci and CAD was estimated to contribute to an OR of 1.25 (95% CI, 1.21-1.29).[11] However, for most diseases, the identified variants had small effects on the diseases and explained a small proportion of the estimated heritability. Heritability of height, the proportion of variance in height explained by genetic factors, has been estimated to be as high as 80-90%. But, the results from GWAS revealed genetic loci that explained only 5% of the variation in height in the human population.[13] For risk prediction, the small effects associated with the disease limit the use of genetic markers found from GWAS on disease risk classification, especially for disease with well-established risk factors such as coronary artery disease.[14]

One limitation of GWAS is the ability to characterize structural variation, such as insertion/deletion polymorphisms, duplication, and rearrangement. Although copy number variations (CNV) have been widely studied using GWAS array, the ability to detect CNV vary greatly among different methods for CNV prediction. How well GWAS SNPs correlate with structural variations in that area determines whether GWAS will be able to detect the association between the structural variants and diseases.

Despite the discovery of many novel genes in these complex diseases, researchers worldwide are frustrated that they have not uncovered the complete biological mechanisms of these diseases. The main criticism for GWAS is a small proportion of the phenotypes that can be explained by the genetic variants found on the GWAS panel. Hence, when next-generation sequencing (NGS) came into the picture, many groups of scientists were ready to switch from GWAS to NGS.[15]

NGS technology fills in several gaps that GWAS. Unlike GWAS which genotypes only limited number of variants included on the genotyping platform, NGS enables rapid sequencing of the whole human genome at a much lower cost compared to traditional sequencing methods. All variants in the genome detected including structural variants, small/large insertion/deletion. NGS have been used for several clinical applications such as diagnoses of difficult to diagnose genetic diseases, choosing the appropriate chemotherapy for cancer patients, or investigating novel genetic causes of complex diseases. Although the cost for sequencing may get cheaper, the later steps after sequencing are still a big challenge. The downstream processes to interpret the genome require experts in several fields e.g. bioinformatics, computational biology, statistical genetics, and biostatistics. Hence, it might cost a thousand dollars to sequence the genome, but 100,000 dollars to understand the meaning of.[16]

References

1. Hindorff LA, MacArthur J (European Bioinformatics Institute), Morales J (European Bioinformatics Institute), Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. (Accessed December 22, 2013, at: www.genome.gov/gwastudies).
2. Nuinoon M, Makarasara W, Mushiroda T, et al. A genome-wide association identified the common genetic variants influence disease severity in beta0-thalassemia/hemoglobin E. *Hum Genet* 2010;127:303-14.
3. Mahasirimongkol S, Yanai H, Mushiroda T, et al. Genome-wide association studies of tuberculosis in Asians identify distinct at-risk locus for young tuberculosis *J Hum Genet* 2012;57:363-7.
4. Kamatani Y, Wattanapokayakit S, Ochi H, et al. A genome-wide association study identifies variants in the HLA-DP locus associated with chronic hepatitis B in Asians. *Nat Genet* 2009;41:591-5.
5. Nishida N, Sawai H, Matsuura K, et al. Genome-wide study confirming association of HLA-DP with protection against chronic hepatitis B and viral clearance in Japanese and Korean. *PLoS One* 2012;7:e39175.
6. Yang W, Shen N, Ye DQ, et al. Genome-wide association study in Asian populations identifies variants in ETS1 and WDFY4 associated with systemic lupus erythematosus. *PLoS Genet* 2010;6:e1000841.
7. Jongjaroenprasert W, Phusantisampan T, Mahasirimongkol S, et al. A genome-wide association study identifies novel

ISSN 2287-0237 (online)/ 2287-9674 (print)

susceptibility genetic variation for thyrotoxic hypokalemic periodic paralysis. *J Hum Genet* 2012;57:301-4.
8. Chantarangsu S, Mushiroda T, Mahasirimongkol S, et al. Genome-wide association study identifies variations in 6p21.3 associated with nevirapine-induced rash. *Clin Infect Dis* 2011;53:341-8.
9. Manolio TA. Bringing genome-wide association findings into clinical use. *Nat Rev Genet* 2013;14: 549-58.
10. Sanna S, Li B, Mulas A, et al. Fine mapping of five loci associated with low-density lipoprotein cholesterol detects variants that double the explained heritability. *PLoS Genet* 2011;7: e1002198.
11. Palomaki GE, Melillo S, Bradley LA. Association between 9p21 genomic markers and heart disease: a meta-analysis. *JAMA* 2010; 303:648-56.
12. Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: 139). (Accessed December 25, 2013, at http://www.ncbi.nlm.nih.gov/SNP/).
13. Maher B. Personal genomics: the case of the missing heritability. *Nature* 2008;456:18-21.
14. Van der Net JB, Janssens AC, Sijbrands EJ, et al. Value of genetic profiling for the prediction of coronary heart disease. *Am Heart J* 2009;158:105-10.
15. Cirulli ET and Goldstein DB. Uncovering the roles of rare variants in common disease through whole genome sequencing. *Nature Rev Genet* 2011;11:415-25.
16. Mardis ER. The $1000 genome, the $100,000 analysis? *Genome Medicine* 2010; 2: 84.

ISSN 2287-0237 (online)/ 2287-9674 (print)