Post-GWAS Strategies

Sangsoo Kim[1]* and Jong Bhak[2]

[1]Department of Bioinformatics, Soongsil University, Seoul 156-743, Korea, [2]Personal Genomics Institute, Suwon 443-270, Korea

Abstract

Genome-wide association (GWA) studies are the method of choice for discovering loci associated with common diseases. More than a thousand GWA studies have reported successful identification of statistically significant association signals in human genomes for a variety of complex diseases. In this review, I discuss some of the issues related to the future of GWA studies and their biomedical applications.

Keywords: post-GWAS, personal genomics, NGS, SNP

Introduction

Soon after the completion of the Human Genome Project that established a reference human genome sequence, a massive discovery of sequence variations, especially of the type known as single nucleotide polymorphism (SNP), was initiated by the SNP Consortium (TSC) (Thorisson & Stein, 2003). After a few years of efficient international collaborative operations, TSC was succeeded by the International HapMap Project, which aims to catalog common variants in our genomes and investigate the linkage disequilibrium (LD) patterns between SNP markers (The International HapMap Consortium, 2003). The knowledge on LD pattern and haplotype structures quickly helped to design high density SNP chips for genome-wide scanning of disease associated loci (Wang *et al.*, 2005). Commercial availability of such SNP chips spurred GWA studies of a variety of common diseases involving many genes. Before the advent of GWA technologies, the complex diseases were studied by screening genetic association of the variants within or around a few genes of interest, which were selected based on the researcher's prior knowledge on the disease pathogenesis. Compared to this 'candidate gene approach', the GWA approach does not rely on *a priori* biological assumption and knowledge on the disease and thus has the potential to detect novel disease loci, providing a new insight and clues on the disease biology. This advantage has recently outweighed its high cost and GWA method has become the method of choice for studying common diseases involving multiple genes. Public databases now catalog more than a thousand GWA publications reporting intriguing association of novel genomic loci with complex diseases (Hindorff *et al.*, 2009). With these initial successful applications of the GWA method, it is time to plan the next phase of genetic association studies, as we now have the second wave of Next Generation Sequencing (NGS) technologies that are much more powerful than previous versions (Mardis, 2011). NGS will efficiently facilitate the detection of causal mutations that are hidden within the LD blocks encompassing the surrogate markers from GWAS. The 1000 Genomes Project aims to catalog less frequent SNPs (~1% MAF) from the reference populations (http://www.1000genomes.org). By typing those mutations in the study samples, one can save the resequencing efforts in the follow-up studies. It is certain that those studies involving the ethnic groups included in the 1000 Genomes Project will benefit from the project. However, what about the populations not included in the project? In this essay, we will discuss some of the issues concerning the nature and future of genome-wide association study (GWAS), followed by a suggestion for the next phase research strategies.

Nature of GWAS Problems

GWAS utilizes high density SNP chips to scan the genome-wide genotype profiles of sample populations. The present day commercial SNP chips typically include an order of million markers. In a typical analysis setting, the genotype-phenotype association is assessed one by one for each marker. This translates to millions of independent statistical tests. In order to minimize false positives due to multiple testing, one typically imposes a very stringent cutoff for the statistical significance level. However, it is still imaginable that some false positives slip through the threshold. For example, there are a few orders of magnitude more markers than samples, and some peculiar sampling may cause bias in the population structure, causing spurious associations. In order to overcome such problems, the results from a GWAS are typically reexamined by a replication study employing a dataset from independent cohorts. In general strongly associated signals are well replicated in other

*Corresponding author: E-mail sskimb@ssu.ac.kr
Tel +82-2-820-0457, Fax +82-2-824-4383

Accepted 7 March 2011

populations, while moderate ones are not so even within the same ethnic groups. We call the latter issue a 'tip of an iceberg' problem. Identification of the hits of moderate signals tends to be sensitive to the threshold setting. While the association signal of a marker may be above the threshold in one population, it may be just below the threshold in another population. At the same time, we miss bulk of the genuine signals below the threshold. This is similar to the situation observed with the 'tip of an iceberg' . Besides the difficulty in replicating the original GWAS result in independent cohorts, small number of strong signal observed in GWAS hinders both rich biological interpretation of the GWAS results and reasonable risk prediction based on genotypes. This posed a serious threat to GWAS and has been the basis of wide-spread pessimism on the prospect of common disease studies.

Missing Heritability

In late 2009, Manolio *et al.* published an article critically reviewing the issue of poor ability to explain phenotypic variation based on genotypes (Manolio *et al.*, 2009). One of the traits given as an example in the review article was human height. The genetic contribution to its variation is known to be around 80%, while the combined alleles identified thus far explained only 5% of the phenotype variance, the rest being claimed as missing heritability. The review also speculated the potential factors contributing to the missing heritability. For example, they posit that rare or structural variants are not well represented in the commercial SNP chips, and that modeling epistatis of gene-gene interaction or incorporation of environmental factors may be necessary. Besides, they also argued that there may be many more weaker signals yet to be found. In order to discover them, much larger sample size will be required. Yang *et al.* reported a simulation result claiming that all the SNPs in the GWAS chip would collectively explain about half of the human height variation (Yang *et al.*, 2010). In fact, a large-scale meta analysis combining 46 data sets comprising 183,727 European individuals identified hundreds more variants related to human height and could explain up to 10% of its variation (GIANT Consortium, 2010). This work confirmed the earlier premise that there are many more weaker signals yet to be found and one way to enhance the statistical power to discover them is by increasing the sample size. Although it is a promising option, not every trait can be studied in this way.

Pathway Analysis

While it is still a long way to go for useful risk prediction based on genotypes, the hefty number of genes whose genetic associations are newly identified by GWAS can provide novel insights regarding trait development or pathogenesis of disease traits. If hundreds of genes are identified by GWAS, gene ontology terms or pathways enriched among them can be identified in a similar manner as well established in gene expression profile analysis (Huang *et al.*, 2009). This method is not applicable if only a handful of genes are identified from GWAS. In such cases, gene-set analysis (GSA) may be considered (Nam & Kim, 2008). GSA is well established in the area of gene expression analysis. Instead of selecting a set of genes passing a certain cutoff and looking for functional terms commonly shared by the gene set, GSA checks a set-wise association score where the contribution by all the member genes in the gene-set are summarized regardless of their significance level. It is important not to filter certain markers or genes based on their significance level. The rationale is as follows: if one wants to calculate a mean property and removes hits with low score values, the mean would be unnecessarily inflated. If a biological pathway or functional module plays a significant role in the trait development, its set-wise association score would be significantly higher than those of the non-associated groups (Nam *et al.*, 2010). Even if individual effect or signal may not be strong enough to pass the threshold, the collective contributions from those genes may add up to surpass the random expectation. This approach offers a benefit of reaching out the 'core of an iceberg' without increasing the sample size. Recently, Wang *et al.*, 2010. reviewed various computational methods that implemented pathway analysis approach for GWAS (Wang *et al.*, 2010). They suggest that pathway analysis may allow rich description of the biological mechanism behind the trait development in an unprecedented manner and has the potential to suggest therapeutic or diagnostic targets without delving into the specific identification of causal variants. Eric Lander, in his review of celebrating tenth anniversary of human genome sequencing, also illustrated several successful examples along this line where GWAS help to advance the biomedical applications (Lander, 2011).

If the biological mechanisms are common for the same traits regardless of populations, GSA would yield highly replicating results, even in the cases where individual markers do not show correlative association across different populations. It would be interesting to see whether this is indeed the case with multi-population meta analysis data sets. It is also informative to know to what extent of ethnic groups we would expect that GSA give overlapping results. For example, may such a high overlap be expected within Asian populations only or even among Asian and Europeans?

Now there are even bolder approaches that intend to integrate systems biology to genetics, called systems genetics (Butte *et al*., 2011). Instead of scoring a collective association of a predefined set of genes, this new approach tests the association of protein interaction or gene regulatory networks. While the two approaches may look similar, there is an important distinction. The former mainly relies on evidence collected from literature, and thus of highly reliable, but their coverage is rather low and the interaction context between genes are missing for some gene sets such as those based on gene ontology or molecular signatures. The latter may have the advantage of testing network properties that are likely dependent on the context of the trait, facilitated by the integration of various omics profiling data.

Longitudinal Cohorts

Tom Hudson, President and Scientific Director, Ontario Institute for Cancer Research, Toronto, Ontario, Canada, wrote in an essay celebrating tenth anniversary of genome sequencing that he would invest heavily in developing large clinical resources if he could move the clock back to 2001 when the initial draft of the human "*reference*" genome sequence was published (Hudson, 2011). Once integrated with genomic technologies the cohorts annotated with ten years of detailed clinical history would have profound effect in understanding disease progression.

Personal Genomics

Introduction of NGS technologies has sparked the genome sequencing of individuals (Nature Editorial, 2010). There are now several so-called personal genome projects, such as the 1000 Genomes Project and the Personal Genome Project (http://www.personalgenomes.org). In the era of personal genomics, what would be the future of GWAS? Some think that GWAS will be superseded by variant profiling based on NGS. Certainly high penetrant, monogenic, Mendelian diseases would benefit greatly from NGS, where comparative sequencing of affected and unaffected from the closely related families would reveal the disease mutations. On the other hand, common, complex diseases require a large number of unrelated population samples to locate the statistically confident disease-causing loci. The current NGS technologies are too expensive to replace the traditional SNP chips in population genotyping. In addition, computational cost and time for handling whole genome NGS data for the large number of samples required in GWAS would be prohibitive at the moment and within the foreseeable future. Ideally sequencing technology can detect novel mutations and profile known variants in one shot. However, a two step process would be more practical: (1) discovering novel mutations by sequencing; (2) follow-up genotyping of the variants by chip technology. One of the aims of the 1000 Genomes Project is to provide reference variant profiles for HapMap populations. For a given GWAS, one can look up the database to pick up the potential variants within the LD block of the association signals. The common-disease/common-variant hypothesis tells that if a disease-causing mutation has occurred one time in human history at a certain haplotype, it would be inherited by the descendants who carry the haplotype. On the other hand, those who do not carry this haplotype would be unaffected by the disease. While subsequent recombination would diversify the overall haplotype structure, the core haplotype around the disease-causing variant would be conserved. Therefore, it is important to identify the haplotypes associated with the trait of interest through GWAS and to infer the potential disease-causing variants from the NGS data that match those haplotypes. The success of this approach is then dependent on whether the 1000 Genomes Project covers the samples whose haplotype structures are compatible with those of our GWAS subjects. At the moment, the 1000 Genomes Project is partially complete and we need to wait more to answer this question for non-HapMap population GWAS.

If we focus our attention to Korean GWAS issues, several options can be suggested. Since Korea is geographically between China and Japan and genetically also Koreans are in between Chinese and Japanese populations, we can wait until the 1000 Genomes Project produce sufficiently accurate and useful data for both Chinese and Japanese populations. Alternatively, we can pursue producing Korean reference whole genome sequences, for which the sampling object should be to maximize the genetic diversity among Koreans. The existing GWAS genotype data would facilitate such a sample selection process. In fact these two options are not mutually exclusive, rather complementary. For example, comparison of Korean reference genomes with those from the Chinese and Japanese of the 1000 Genomes Project would tell us how much reference sequencing among Koreans is necessary.

Conclusions

There has been widely known and somewhat accepted skepticism on GWAS. However, the recent progresses re-shed light on its potential. Recently NIH and NCI, USA funded several projects called the Post-GWA

Initiative (http://grants.nih.gov/grants/guide/rfa-files/RFA-CA-09-002.html). These projects (Monteiro *et al.*, 2010), focusing on GWA studies of cancers, follow up the GWAS outcomes to pin-point the causal variants and understand how they influence disease development and progression. Bioinformatic computational tools would play important roles in prioritizing the potential causal variants through the prediction of the functional consequence of each variant and the analysis of the affected pathways. Experimental validations employing both omics profiling and animal models are also planned. Much of the idea laid out in the projects share the same spirit with this essay.

Acknowledgements

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (NRF-2010-0021811).

References

Butte, A., Califano, A., Friend, S., Ideker, T., and Schadt, E. (2011). Integrative Network-based Association Studies: Leveraging cell regulatory models in the post-GWAS era. *Nature Precedings* doi:10.1038/npre.2011.5732.1

GIANT Consortium. (2010). Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467, 832-838.

Hindorff, L.A., Sethupathy, P., Junkins, H.A., Ramos, E.M., Mehta, J.P., Collins, F.S., and Manolio, T.A. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl. Acad Sci. USA* 106, 9362-9367.

Huang, D.W., Sherman, B.T., and Lempicki, R.A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat. Protoc.* 4, 44-57.

Hudson, T. (2011). Genomics and Clinical Relevance. *Science* 331, 547.

Lander, E.S. (2011). Initial impact of the sequencing of the human genome. *Nature* 470, 187-197.

Manolio, T.A., Collins, F.S., Cox, N.J., Goldstein, D.B., Hindorff, L.A., Hunter, D.J., McCarthy, M.I., Ramos, E.M., Cardon, L.R., Chakravarti, A., Cho, J.H., Guttmacher, A.E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C.N., Slatkin, M., Valle, D., Whittemore, A.S., Boehnke, M., Clark, A.G., Eichler, E.E., Gibson, G., Haines, J.L., Mackay, T.F., McCarroll, S.A., and Visscher, P.M. (2009). Finding the missing heritability of complex diseases. *Nature* 461, 747-753.

Mardis, E.R. (2011). A decade's perspective on DNA sequencing technology. *Nature* 470, 198-203.

Monteiro, A.N.A., Coetzee, G.A., Freedman, M.L., Biasi, M.D., Casey, G., Duggan, D., Risch, A., Plass, C., Liu, P., James, M., Vikis, H.G., Tichelaar, J.W., You, M., Gayther, S.A., and Mills, I.G. (2010). Principles for the post-GWAS functional characterisation of risk loci. *Nature Precedings.* doi:10.1038/npre.2010. 5162.

Nam, D., Kim, J., Kim, S.Y., and Kim, S. (2010). GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucl. Acids Res.* 38, 749-754.

Nam, D., Kim, S.Y., Gene-set approach for expression pattern analysis. (2008). *Brief Bioinform.* 9, 189-197.

Nature Editorial. (2010). The human genome at ten. *Nature* 464, 649-650.

The International HapMap Consortium. (2003). The International HapMap Project. *Nature* 426, 789-796.

Thorisson, G.A., and Stein, L.D. (2003). The SNP Consortium website: past, present and future. *Nucl. Acids Res.* 31, 124-127.

Wang, K., Li, M., and Hakonarson, H. (2010). Analysing biological pathways in Genome-wide association studies. *Nat. Rev. Genet.* 11, 843-854.

Wang, W.Y., Barratt, B.J., Clayton, D.G., and Todd, J.A. (2005). Genome-wide association studies: theoretical and practical concerns. *Nat. Rev. Genet.* 6, 109-118.

Yang, J., Benyamin, B., McEvoy, B.P., Gordon, S., Henders, A.K., Nyholt, D.R., Madden, P.A., Heath, A.C., Martin, N.G., Montgomery, G.W., Goddard, M.E., and Visscher, P.M. (2010). Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.* 42, 565-569.