Genome- Wide Association Studies and Future Aspect

Ritushree Narayan[1] Mohammad Ibrar[2]
[1,2]Department of Computer Science
[1,2]R.L.S.Y. College, Ranchi- 834001, India

Abstract— Genome-wide association study (GWAS) technology has been a primary method of identifying the genes which are responsible for diseases .A genome-wide association study (GWA study, or GWAS), also known as whole genome association study (WGA study, or WGAS), is an examination of a genome-wide set of genetic variants in different individuals to see if any variant is associated with a trait. GWASs typically focus on associations between single nucleotide polymorphism (SNPs) and traits like major human diseases, but can equally be applied to any other organism like plant, animals, and model organisms. In genome-wide association studies (GWAS), researchers analyze the genetic variation across the entire human genome, searching for variations that are associated with observable traits or certain diseases.

Key words: Genome Future Aspect, Genome-Wide Association Studies

I. INTRODUCTION

The basic GWAS approach is to look at approximately a million of positions in the human genome (SNPs) where different people carry different versions of the genetic code. Genome wide association studies (GWAS) are used to rapidly scan a large set of genetic variants and thus to identify associations with a particular trait or disease .The GWAS philosophy indifferent from the conventional candidate gene based approaches ,which directly test the effects of the genetic variants of contributory genes. Genome wide association (GWAS) provides an important way for undertaking evaluation of the association between common genetic variants and risk disease. Understanding of human genetic variation and the technology to measure such variation have made GWAS feasible. GWAS approaches and its benefits helps in overcoming the weaknesses of GWAS.GWAS have convincingly detect hundreds of variants associated with a large number of diseases. Many of these findings are novel, associated SNPs in genes or chromosomal regions were not previously implicated in disease. These results are especially exciting in light of the previous difficulties replicating genetic findings for many diseases. Figure: 1.By the observation of genetic variants detected by most of the studies may not be casual for diseases. In light of the mixed opinions of GWAS, we consider here the following important aspects of these studies: design and analysis, findings and implications, limitations, and future prospects.[1]

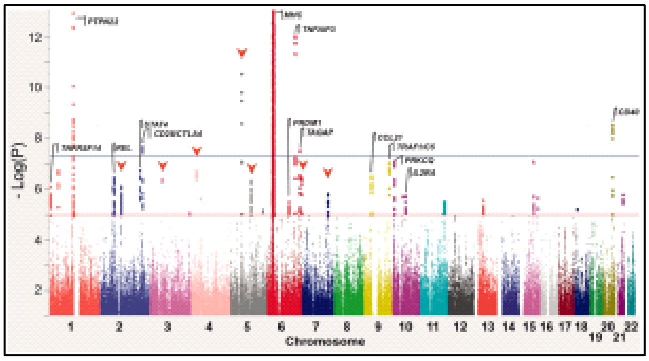

Fig. 1: Manhattan plot for RA GWAS meta analysis. Statistical strength of association (-Log10P) is plotted against genomic position with the 22 autosomal chromosomes in different colors. The blue horizontal line indicates the genom e-wide significance threshold of p= 5*10-8; the red line is a threshold for "suggestive" assoition (P=10-5). SNPs at 29loci known from previous studies (gene symbols shown), and one of the 10 new loci identified in this study (marked by red triangles), achieved genome-wide significance in this meta-analysis(proir to the replication phase of the study). Over 200SNPs representing 35 loci achieved P<10-5, versus roughly 1 expected by chance.

II. DATA COLLECTION AND METHODS

The most common study design GWAS is the case-control design, also known as a retrospective study design, where 'unrelated' affected and healthy individuals are collected for genotyping. When using a family based study design, the samples are collected from families where at least one of the members are affected by the disease.

The case-control design is sensitive to population stratification between case and control samples, which can cause false positives. It is therefore important to consider the optimal selection of samples to minimize or correct for these effects. Family studies are less sensitive to these population substructures, but has a reduced power compared to case-control studies. In case-control studies phenotypic and genetic heterogeneity will often occur in the samples, and family designs are robust against this type of heterogeneity. In addition, case control design has the advantage that it is easier to collect unrelated subjects, compared to families where complete families are not always available.

Following the Welcome Trust Case Control Consortium study [3], it has also become possible to use common controls samples in several studies. One potential problem with such common control samples is unidentified cases among the controls, which might reduce the power if the trait is common. Another possible problem is that some studies use public control data from other countries, not quite matching the case sample. Starting the era of GWAS, 'Population Stratification' was believed to be a major threat to the success of the case-control approach, suggesting family based controls [4]. However, it has turned out not to be a large problem if matching or adjusting for reported ethnicity is applied. It also turns out that the GWA data itself can be used to identify the substructures [5]. In order to have enough power to detect effects with genome wide significance (p-value < 5 · 108) it has been necessary to build consortia for large GWAS. With the possibility of collecting such large samples, it is quite easy to detect and correct for population substructures. However, in many studies it is still either hard to find enough cases to collect, or for financial reasons not enough individuals can be genotyped.

Lately the case-control design has also been extended to population based cohort studies, usually designed to investigate various traits from the same data [2]. These studies are more useful for continuous traits, and still have quite limited power for dichotomous phenotypes. Meta-analysis is another approach to overcome the sample size issue, but unfortunately there are difficulties of standardizing studies performed with varying sampling strategies, genotyping arrays etc.

III. Linkage disequilibrium

Linkage disequilibrium is a property of SNPs on a contiguous stretch of genomic sequence that describes the degree to which an allele of one SNP is inherited or correlated with an allele of another SNP within a population. The term linkage disequilibrium was coined by population geneticists in an attempt to mathematically describe changes in genetic variation within a population over time. It is related to the concept of chromosomal linkage, where two markers on a chromosome remain physically joined on a chromosome through generations of a family. Recombination events within a family from generation to generation break apart chromosomal segments. This effect is amplified through generations, and in a population of fixed size undergoing random mating, repeated random recombination events will break apart segments of contiguous chromosome (containing linked alleles) until eventually all alleles in the population are in linkage equilibrium or are independent. Thus, linkage between markers on a population scale is referred to as linkage disequilibrium in Figure 2.

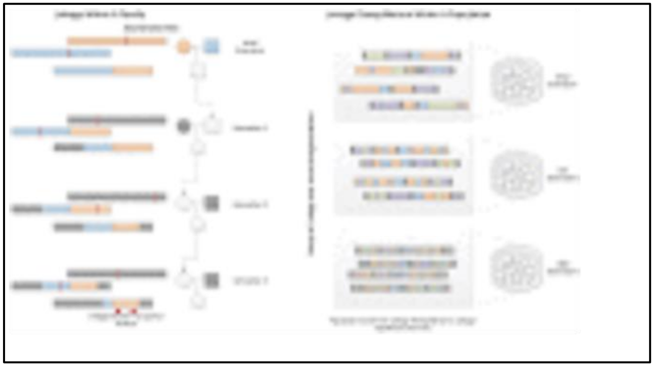

Fig. 2: Linkage and linkage disequilibrium

The rate of LD decay is dependent on multiple factors, including the population size, the number of founding chromosomes in the population, and the number of generations for which the population has existed. As such, different human sub-populations have different degrees and patterns of LD. African-descent populations are the most ancestral and have smaller regions of LD due to the accumulation of more recombination events in that group. European-descent and Asian-descent populations were created by founder events (a sampling of chromosomes from the African population), which altered the number of founding chromosomes, the population size, and the generational age of the population. These populations on average have larger regions of LD than African-descent groups.

Many measures of LD have been proposed , though all are ultimately related to the difference between the observed frequency of co-occurrence for two alleles (i.e. a two-marker haplotype) and the frequency expected if the two markers are independent. The two commonly used measures of linkage disequilibrium are shown in equation 1 and 2. In these equation $\prod$ is the frequency of the haplotype ab, $\prod_1$ is the frequency of the allele a,and $\prod_2$is the frequency of the allele b.

$$D' = \begin{cases} \dfrac{\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}}{\min(\pi_A\pi_b, \pi_a\pi_B)} \; if\, \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB} > 0 \\ \dfrac{\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB}}{\min(\pi_A\pi_B, \pi_a\pi_b)} \; if\, \pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB} < 0 \end{cases} \quad (1)$$

$$r^2 = \frac{(\pi_{AB}\pi_{ab} - \pi_{Ab}\pi_{aB})^2}{\pi_A\pi_B\pi_a\pi_b} \quad (2)$$

D' is a population genetics measure that is related to recombination events between markers and is scaled between 0 and 1. A value of 0 indicates complete linkage equilibrium, which implies frequent recombination between the two markers and statistical independence under principles of Hardy-Weinberg equilibrium. A of 1 indicates complete LD, indicating no recombination between the two markers within the population. For the purposes of genetic analysis, LD is generally reported in terms of, a statistical measure of correlation. High values indicate that two SNPs convey similar information, as one allele of the first SNP is often observed with one allele of te second SNP, so only one of the two SNPs needs to be genotyped to capture the allelic variation. There are dependencies between these two statistics; is sensitive to the allele frequencies of the tow markers, and can only be high in regions of high D'.[9]

One often forgotten issue associated with LD measures is that current technology does not allow direct measurement of haplotype frequencies from a sample because each SNP is genotyped independently and the *phase* or chromosome of origin for each allele is unknown. Many well-developed and documented methods for inferring haplotype phase and estimating the subsequent two-marker haplotype frequencies exist, and generally lead to reasonable results.

SNPs that are selected specifically to capture the variation at nearby sites in the genome are called *tag SNPs* because alleles for these SNPs tag the surrounding stretch of LD. As noted before, patterns of LD are population specific and as such, tag SNPs selected for one population may not work well for a different population. LD is exploited to optimize genetic studies, preventing genotyping SNPs that provide redundant information. Based on analysis of data from the HapMap project, >80% of commonly occurring SNPs in European descent populations can be captured using a subset of 500,000 to one million SNPs scattered across the genome.

IV. META-ANALYSIS

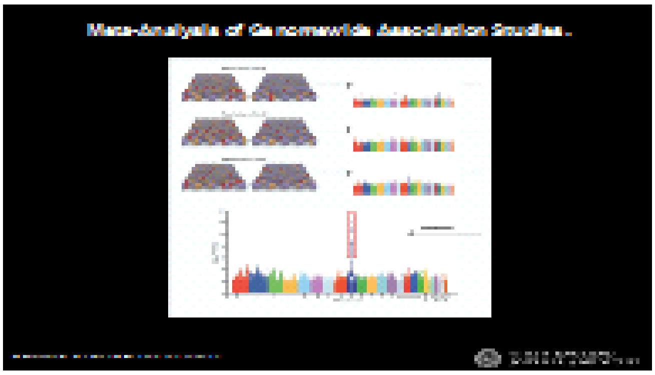

Fig. 3: Meta-Analysis of Genom eewide association studies

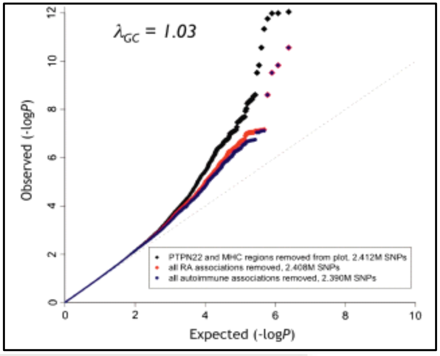

Fig. 4: Q-Q plot of the RA GWAS meta-analysis (Stahlet al. 2010). Results for all SNPs excluding the strongly associated PTPN22 (chr1, 113.5-114.5Mb) and MHC (chr6, 26-34 Mb) regions (which would otherwise dominate the tail of the distribution) are plotted in black. Results excluding SNPs in LD (r2>0.1) with previously known RA risk associations are plotted in red, showing that substantial association signal remains in the data. And results excluding SNPs in LD with validated autoimmune disease associations are plotted in blue, showing a degree of overlap between RA and related complex diseases. Geonomic control λGC (scaled for 1000 cases and 1000 controls) for the data excluding PTPN22 and MHC (black) is shown in the inset. Image adapted from Stahlet al. 2010

Meta-analysis techniques were originally developed examine and refine significance and effect size estimates from multiple studies examining the same hypothesis in the published literature. With the development of large academic consortia, meta-analysis approaches allow the synthesis of results from multiple studies without requiring the transfer of protected genotype or clinical information to parties who were not part of the original study approval – only statistical results from a study need be transferred. For example, a recent publication examining lipid profiles was based on a meta-analysis of 46 studies. Figure 3 and 4. A study of this magnitude would be logistically difficult (if not impossible) without meta-analysis. Several software packages are available to facilitate meta-analysis, including STATA products and METAL[5][6]

A fundamental principle in meta-analysis is that all studies included examined the same hypothesis. As such, the general design of each included study should be similar, and the study-level SNP analysis should follow near-identical procedures across all studies[7]. Quality control procedures that determine which SNPs are included from each site should be standardized, along with any covariate adjustments, and the measurement of clinical covariates and phenotypes should be consistent across multiple sites. The sample sets across all studies should be independent – an assumption that should always be examined as investigators often contribute the same samples to multiple studies. Also, an extremely important and somewhat bothersome logistical matter is ensuring that all studies report results relative to a common genomic build and reference allele. If one study reports its results relative to allele A and another relative to allele B, the meta-analysis result for this SNP may be non-significant because the effects of the two studies nullify each other.

With all of these factors to consider, it is rare to find multiple studies that match perfectly on all criteria. Therefore, study heterogeneity is often statistically quantified in a meta-analysis to determine the degree to which studies differ. The most popular measures of study heterogeneity are the Q statistic and the I^2 index, with the **I^2** index favored in more recent studies. Coefficients resulting from a meta-analysis have variability (or error) associated with them, and the I^2 index represents the approximate proportion of this variability that can be attributed to heterogeneity between studies I^2 values fall into low (<25), medium (>25 and <75), and high (>75) heterogeneity, and have been proposed as a way to identify studies that should perhaps be removed from a meta-analysis. It is important to note that these statistics should be used as a guide to identifying studies that perhaps examine a different underlying hypothesis than others in the meta-analysis, much like outlier analysis is used to identify unduly influential points. Just as with outliers, however, a study should only be excluded if there is an obvious reason to do so based on the parameters of the study – not simply because a statistic indicates that this study increases heterogeneity. Otherwise, agnostic statistical procedures designed to reduce meta-analysis heterogeneity will increase false discoveries.

V. CONCLUSIONS

Genome-wide association studies have a huge impact on the field of human genetics. They have identified new genetic risk factors for many common human diseases. On the horizon is whole-genome sequencing. Within the next few years we will see the arrival of cheap sequencing technology that will replace one million SNPs with the entire genomic sequence of three billion nucleotides. Challenges associated with data storage and manipulation, quality control and data analysis will be manifold more complex, thus challenging computer science and bioinformatics infrastructure and expertise. Genome-wide association studies in humans have already proven a resounding success in providing a framework for unraveling the genetic basis of complex traits. The results have provided unprecedented views into the contribution of common variants to complex traits, illuminated genome function, and have opened new possibilities for the development of therapeutic interventions. Trait architecture conforms to a roughly exponential distribution of effect sizes: the majority of common complex trait-associated variants studied thus far have modest effects (OR < 2), and for most traits, substantial heritability remains to be explained. Identifying the genetic basis of the remaining trait variance will require additional discoveries, particularly of rare trait-associated variants and better characterization of genetic modes of action and interaction and refined estimates of heritability. DNA sequencing will play a key role in the next generation of GWAS, through candidate locus resequencing in large cohorts, whole-exome sequencing, and eventually whole-genome sequencing of large numbers of individuals. Also, integration of functional biological knowledge into association analyses promises to point directly to putative functional variants. Importantly, identification of causal variants and expansion of these studies into populations of diverse ancestry will facilitate further biological understanding and population genetics of complex traits. Continued accelerated pace of discovery of medically important trait-associated variants in humans will depend on implementation of new technologies and analytic approaches to integrate diverse data types, but also, critically, on the lessons learned from the burst of discovery that has been the result of the first round of genome-wide association studies in humans.

References

[1] Genome Wide Association Studies and BeyondJohn S. Witte, Institute for Human Genetics, Departments of Epidemiology and Biostatistics and Urology, University of California, San Francisco, San Francisco,California 94158¬9001;

[2] Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES. Highresolution haplotype structure in the human genome. Nat. Genet. 2001;29:229–232. [PubMed]

[3] Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. A second generation human haplotype map of over 3.1million SNPs. Nature. 2007;449:851–861. [PMC f ree article] [PubMed]

[4] Huang L, Li Y, Singleton AB, Hardy JA, Abecasis G, et al. Genotypeimputation accuracy across worldwide human populations. Am. J. Hum. Genet. 2009;84:235–250. [PMC free article] [PubMed]

[5] Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen WM, et al. (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. Nat Genet 40:198203doi: 10.1038/ng.74[PMCfreearticle] [PubMed]

[6] Willer CJ, Sanna S, Jackson AU, Scuteri A, Bonnycastle LL, et al. (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. Nat Genet 40:161169doi: 10.1038/ng.7[PubMed]

[7] Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies.Pharmacogenomics 10:191–201doi: 10.2217/14622416.10.2.191 [PMC free article] [PubMed]

[8] Progress and Promise of Genome-Wide Association Studies for Human Complex Trait Genetics Barbara E. Stranger, Eli A. Stahl, Towfique Raj GENETICS February 1,2011 vol.187 no.2, 367-383; DOI: 10.1534/genetics.110.120907

[9] Genomewide Association Studies and Assessment of the Risk of Disease Teri A. Manolio, M.D., Ph.D.N Engl J Med 2010; 363:166¬176 July 8, 2010 DOI: 10.1056/NEJMra0905980