Daniel Van de Mark
Biochem 218
June 6, 2010
Final Project

The Numerous Caveats of Designing, Implementing, and Interpreting Genome-Wide Association Studies

The information to be made available by genome-wide association (GWA) studies has long been heralded as the future of healthcare and personalized medicine. It has been suggested that knowledge linking particular DNA sequences to diseases will provide a strong foundation for preventative medicine. GWA studies have already proven successful in predicting loci associated with age-related macular degeneration, myocardial infarction, and type 2 diabetes [1]. In fact, more than 250 genetic loci that contribute to phenotypes observed as diseases or other quantitative traits (traits determined by the effect of a number of different loci), such as height, have been identified [2].

The general methodology of any association study is to identify the genetic variation responsible for a specific disease by looking at the frequency of single nucleotide polymorphisms (SNPs) in both disease and control populations. GWA studies extend this methodology to looking at whole genomes rather than just at specific loci of the subjects' genomes [3]. They utilize high-throughput genotyping methods, such as the use of SNP chips or high-throughput sequencing methods, to assay hundreds of thousands of SNPs and then try to link that data to diseases or other quantitative traits such as height [4]. GWA studies have the advantage over other types of association studies in that they make no biased decisions as to which genetic loci might be linked to a trait beforehand and instead just assay the entire genome; previous methods usually required some guesswork about which candidate regions might harbor causal agents [5].

There are still a number of technical issues that need to be overcome with GWA studies, and it should be noted that the method itself has a number of inherent limitations. On the technical side, studies are only valuable when the phenotypes of the individuals' genotyped are properly characterized [1]. Additionally, the number of cases and controls needed to carry out a successful study is not insignificant: by some estimates, thousands of cases and controls are required in order to have a sufficient sample sizes [1]. Also, bioinformatics and statistical methods that can adequately deal with such large amounts of data, as well as with the arising complexities of the data, are still under development [1]. There are a number of caveats to be taken into consideration in the design and implementation of these studies and in the analysis of the data from them. These caveats and various options to tackle inherent limitations in GWA studies are the subject of this report.

General Methodology of GWA studies

Genome-wide association studies match genetic variation in the form of SNPs in an individual's DNA to a specific disease phenotype or other phenotypic trait (note: from here on, I will discuss only linking to disease phenotypes, but it is implied that the same methods could always be used to link variants to other quantitative traits). These studies involve genotyping members of a population on the basis of genetic markers that are spread throughout the genome, and they rely on using populations that have an observable disease phenotype that can be tied back to the presence of specific markers in only a subset of the population. The basic steps of a GWA study are as follows: 1.) Define the design of the study and identify disease and control groups, 2.) Genotype the members of the disease and control groups, 3.) Employ statistical tests to associate diseases with SNP variants, and finally 4.) Replicate the results in an independent population or follow up on the mechanism of disease transmission in an experimental context in a laboratory [4]. The remainder of this report will be broken up into sections related to each of these steps in order to discuss what is involved in each at greater depth as well as some potential concerns and issues at each stage of the study.

Defining the Study Design and Selecting Subjects

There is a number of different study designs commonly applied to carry out GWA studies. The most common is the case-control design whereby groups of disease patients and disease-free (control) individuals are genotyped, and SNP frequencies compared between the groups to identify associations [4]. Other study designs commonly employed include the trio design, whereby disease patients are genotyped along with their parents to trace inheritance of causal SNPs, and cohort studies, whereby much larger groups are genotyped and extensive histories are taken for each member of the group so that genetic variants can be linked to a number of different diseases after the fact [4].

Each of these designs has its strengths and weaknesses. Case-control designs are usually cheaper, but carry a high risk for biased results: only individuals actually needing treatment for the disease are included in the disease group (as opposed to individuals who have a more tolerable form of the disease and are able to go without treatment), subjects who in fact have a mild or undiscovered case of the disease could be included in the control unknowingly, and it may be difficult to find proper control subjects because of population stratification (i.e. finding controls that have similar ethnic or geographical backgrounds as the disease group) [4]. Trio designs benefit from having a built in control group in the form of unaffected parents usually present within the study. However, trio designs are particularly sensitive to errors in genotyping because they aim at following the transmission of variants from parent to offspring; a misgenotyped SNP, in either parent or offspring, makes this effort futile [4]. Cohort studies benefit from large numbers of individuals to draw proper controls from and they can be used to study linkage for multiple conditions; however, they are extremely expensive and data intensive to carry out as they typically need to juggle very large numbers of subjects and fairly in depth histories of each subject in order to be effective [4].

For all designs though, proper classification of the phenotypes of the individuals is of the utmost importance to the success of the study. Misclassified individuals can reduce the

ability of the study to draw linkage between a condition and a genetic variant [4]. This is true both when an unknown, and potentially large, number of traits is to be assayed in cohort studies and when a single trait is assayed in a case-control or trio study. The cohort subjects have a greater number of phenotypes that need to be assayed correctly (and it is not always clear which phenotypes those are), but in all cases individuals need to be accurately assigned as having a condition or not having it.

Genotyping Subjects

When talking about genotyping subjects, what is really meant is ascertaining which SNPs are present in individual subjects. As discussed in a perspective piece in the New England Journal of Medicine by Christensen and Murray (2007), GWA studies are greatly facilitated by work from the International HapMap Project. SNPs closer together to one another in genetic distance are more likely to segregate with each other than are SNPs that are far apart from one another because of their high degree of linkage disequilibrium. As a consequence of this, GWA studies typically use a subset of the total number of SNPs to act as genetic markers. The SNPs in this subset are typically referred to as 'tagging SNPs' and they serve as proxies for the presence of other known SNPs that are close in genetic distance to the tagging SNP in question. The International HapMap Project identified the relationships between the tagging SNPs and other SNPs such that it is now easy to use tagging SNPs in this manner [1]. Thus, tagging SNPs are typically the only SNPs needing to be genotyped in the studies, greatly reducing both the cost and statistical burden.

Genotyping is frequently a high cost stage for GWA studies and, thus far, primarily SNP chips have been used to collect data, though it seems likely that if sequencing technology becomes increasingly more inexpensive, it too could be used in the future. SNP chip platforms typically assay between 500,000 to 1,000,000 SNPs, though higher density ones can be used to try to pick up copy number variants [4]. A frequently used strategy to try to overcome some of the cost concerns and statistical issues is to utilize a multistage design. In such a design, SNPs with disease associations are identified first in a GWA study at a set P-value using data from a limited number of samples; this subset of SNPs is then subsequently retested with more inexpensive, but limited, technology and with more samples [5].

A chief concern at the genotyping stage is that errors are not made; errors hold the potential to confound linkage results. A number of checks are available to try to identify errors in genotyping data. Among others, these include checking that the alleles picked up by the genotyping results do not severely violate Hardy-Weinberg equilibrium or, in trio studies, Mendelian inheritance patterns [4]. However, these are after-the-fact checks, and great care must be taken in collecting and processing samples to ensure the quality of the data and the reliability of the results.

Data Analysis

In the traditional GWA study where single SNPs are linked to a disease, genotypes linked to diseases or traits are usually presented as odds ratios (the ratio of the probability of a disease occurring in the disease group to the probability of it occurring in the control group) or population attributable risks (essentially the degree to which a disease in a population can be attributed to the genetic variant in question, which is in part determined by the odds ratio) [4]. Odds ratios in GWA studies typically are not very high (they are usually on the order of 1.2-1.3), which is indicative of the modest effect of most individual disease-linked SNPs' contribution to their corresponding diseases [4].

Because of the multiple testing carried out in GWA studies, there is a high likelihood of false positive results, and, due to the large number of SNPs assayed, a standard p-value of .05 is much too high to give stringent results [4]. To compensate for this, researchers typically apply the Bonferroni correction, which divides the standard p-value by the number of tests performed, in determining the p-value to use [4]. Replication studies in independent samples are then needed to identify true- and false-positives [4].

Confounding Factors in Data Analysis and Methods to Address Them

The traditional analysis of trying to link a single SNP to a single disease is not always sufficient to try to understand the causation of a particular disease. Many biological phenotypes are not inherited in a Mendelian fashion, but instead are quantitative traits that show great variation in phenotypes and result from the effect of several genes and from the impact of the environment on the individual [3]. GWA studies are set up better to analyze the contribution of individual SNPs to diseases or traits rather than the contribution of multiple SNP's combined impact. However, a number of groups have recently developed bioinformatics strategies to identify such effects. A few of these methods, but by no means all, are discussed below.

Pathway-Based Approaches

It is likely that a number of SNPs that affect genes in the same pathway have epistatic interactions and that all of these SNPs contribute in part to the development of a particular disease. To get at this problem of identifying multiple SNPs affecting different genes in a pathway, groups have employed what is called a pathway-based approach [6, 7]. The approach involves comparing GWA results to a null distribution derived by one of various permutation methods in order to identify pathways that are enriched in the sample analysis [6, 7].

Three different bases of permutation for developing the null distribution have been proposed in the literature: sample randomization, gene randomization [7], and SNP randomization [6]. Each has its strengths and weaknesses as discussed in Guo et al. (2009) Sample randomization, whereby phenotypes present in the sampled population are shuffled and association statistics are recalculated appears to be the gold standard as this method preserves genome architecture and linkage disequilibrium. It is however computationally intensive. Gene randomization, whereby gene statistics are shuffled over the genome and association statistics recalculated, is much less computationally

intensive but also carries issues due to only genic regions and not the entire genome being used to generate the null data set. The SNP randomization method is largely similar to the gene randomization method, however it allows for genetic effects to occur throughout the genome and not just in genes, and it better takes into account the effect of having differing numbers of SNPs within single genes [6]. Both the gene and SNP randomization methods might potentially have issues due to breaking linkage disequilibrium structure in generating the null, however it is thought that this difficulty can be overcome by increasing the number of permutations. Thus, the SNP randomization method appears to offer the best option for both a computationally friendly and relatively accurate method [6].

An Example of A Learning Based Approach

Another method that is sometimes used to handle the issue of identifying epistatic interactions in genome-wide association studies is the learning based approach, an example of which is the SNPRuler [8]. The basic method of the SNPRuler approach involves two steps. In the first, a rule-searching algorithm searches genome-wide data for potential interactions by looking for possible rules contained within suspected interactions. The second step involves using a χ^2 test to evaluate the SNPs identified as representing potential epistatic interactors and to rule out false positives from the first step [8].

The SNPRuler method is advantageous in that it is not computationally intensive, and the rule learning algorithm isn't as sensitive to marginal effects of individual SNPs as some other methods are [8]. The SNPRuler is limited however to detecting epistatic interactions that have clear rules that can be learned by the algorithm [8].

Use of Previous Biological Knowledge

Clearly, handling multiple SNPs that have epistatic interactions in conferring susceptibility to disease or quantitative traits is going to be a strong focus for the bioinformatics community for some time to come as different strategies are developed and tested to do so. It has also been pointed out that such analysis can potentially benefit greatly from the use of previous biological knowledge to narrow down which SNPs to test for epistasis: information about biochemical pathways, Gene Ontology (GO) terms, protein-protein interactions etc… all can be used to narrow down which SNPs are likely to have epistatic interactions [9]. Doing so obviously reduces the unbiased character of GWA studies by again assuming candidate loci, and such efforts would only be helpful to the extent that the biological information used is accurate and complete, but it could still make processing the data much simpler while still offering a high likelihood for significant results.

Other Inherent Issues in GWA Studies

The difficulty in easily identifying epistatic interactions is only one limitation of GWA studies. There are numerous others, due to the studies' designs, that also need to be kept

in mind. One such example is that there are a number of cases where the gene impacted by the mutation may be difficult to identify, even in cases where the SNP can be mapped by a genome-wide association study: notably, this is likely to be the case when the mutation lies in a regulatory element for a gene (as regulatory elements themselves are not always readily identifiable or able to be linked directly to a particular gene of interest) [3]. In fact, it appears as though most SNPs linked to diseases or traits by GWA studies do not lie within coding regions of genes, but instead lie in regions that are more likely to affect either transcriptional regulation, RNA stability or splicing and translation efficiency [10]. With any of these types of SNPs, expression studies may also be needed to verify that the SNPs linkage to disease is due to its impact on the expression of a particular gene [10].

Copy number variation in response to genetic variation also represents a potential issue, though as previously mentioned in the section on genotyping, genotyping methods are getting better at dealing with copy number variants [4]. GWA studies also make the assumption that common diseases are caused by common variants in the population that will be detectable at a reasonable level; they are much less efficient at identifying rare variants that could be responsible for disease [4]. There could be many rare alleles of the same gene that all contribute to the same disease, which would be much more easily identified by studies in families [1].

And GWA studies, as currently carried out, do not look at the contribution of epigenetic variation [10]. As findings from GWA studies have shown that a number of SNPs contributing to disease and traits tend to impact gene expression, and since epigenetic regulation has a strong influence on gene expression, it makes sense that epigenetic variation is likely responsible for a large amount of the disease contribution unaccounted for by current GWA studies. New methods will need to be developed to look at epigenetic regulation efficiently on a genome-wide scale.

Replication of Results

There is often difficulty in reproducing results from these studies. Population stratification issues frequently come into play, especially since at times replication studies might be carried out with similar phenotypes in populations that are inherently different from that of the original study, in terms of ethnicity or geography [4]. As previously mentioned, genotyping errors can cause difficulties, particularly in trio design studies, and may make results difficult to reproduce [4]. Additionally, differences in measuring and identifying the phenotype linked to the variants between studies can cause great challenges to reproducibility of results [4].

Replication of results in subsequent independent studies is obviously important and hopefully with time (and better technology, techniques, experimental setups, etc…) success rates of doing so will improve. Another option to verify results of a particular linkage study though is to also verify the linkage functionally in the laboratory [4]. Researchers can of course tease out mechanisms that lead from individual SNPs to a disease phenotype, by determining which protein is affected by the mutation (either the

protein itself or its level of expression) and then figure out what happens in cells or model organisms as a result of that change. Doing so is a viable verification alternative to repeated GWA studies.

Expert Outlooks on the Future Contributions of GWA Studies

David Goldstein, the director of the Center for Human Genome Variation at the Institute for Genome Sciences and Policy at Duke University, recently described a number of issues with GWA studies and where he feels they can prove to be useful in the future in a published perspective piece in the New England Journal of Medicine (2009) [11]. He pointed out that most SNP variants linked to variation in quantitative trait phenotypes are responsible for only a small fraction of that variation. He demonstrated that, assuming the variants that have the largest effect have been discovered by initial GWA studies, tens of thousands of additional common variants would need to be discovered in order to explain most of the variation in phenotype, and most of those would each contribute a minor amount to the phenotype [11]. Thus there is a built in system of diminishing returns to carrying out such studies.

He, however, also argues that additional GWA studies might be very useful if they aim at identifying variants that play a role in drug responses or susceptibility to infectious agents [11]. He also argues that, because most common variants don't account for the total variation we see in phenotypes, there could be rare variants that play a bigger role and that more attention should be focused on identifying them [11]. He points out that the rare variants will be more difficult to identify and their identification will require identifying prime populations to genotype and greater sequencing efforts [11].

A contrasting opinion from Joel Hirschhorn, Associate Professor of genetics at Harvard Medical School, was also published in the same issue of the New England Journal of Medicine (2009). In it, Hirschhorn addresses concerns that GWA studies may end up yielding too many loci and loci that are not biologically significant by pointing out that there has already been a number of findings from GWA studies that validate their use: GWA studies identifying loci implicated in traits such as lipid levels and type 2 diabetes turned up a significant number of implicated genes that were known to play a role in both of those pathways [2]. Thus, by turning up positive controls within data sets, the studies are validating their continued use.

Concluding Remarks

It seems clear from the various caveats discussed in this review that the design and implementation of these studies is both a work in progress, and even under the best of circumstances an uncertain matter. A number of decisions have to be made as to how to set up and carry out these studies. Which design should be used? Which subjects should be selected and what information should be collected from them? Which genotyping platform should be used? Should a multistage design be used and if so, how should it be implemented? How should the data be analyzed? Should SNPs be tested for epistatic interactions, and if so, which statistical or computational method should be used to do so?

These are just a handful of the questions that researchers have to ask themselves in designing and carrying out these experiments: questions that do not necessarily have clear-cut best answers at the moment. It seems likely that as technology becomes more inexpensive and more studies and larger studies become feasible, reproducibility of results will improve. And the continued development of computational approaches to handle the data will improve the quality of associations drawn from the studies.

1. Christensen, K., and Murray, J.C. (2007). What genome-wide association studies can do for medicine. N Engl J Med *356*, 1094-1097.
2. Hirschhorn, J.N. (2009). Genomewide association studies--illuminating biologic pathways. N Engl J Med *360*, 1699-1701.
3. Hirschhorn, J.N., and Daly, M.J. (2005). Genome-wide association studies for common diseases and complex traits. Nat Rev Genet *6*, 95-108.
4. Pearson, T.A., and Manolio, T.A. (2008). How to interpret a genome-wide association study. JAMA *299*, 1335-1344.
5. Hunter, D.J., and Kraft, P. (2007). Drinking from the fire hose--statistical issues in genomewide association studies. N Engl J Med *357*, 436-439.
6. Guo, Y.F., Li, J., Chen, Y., Zhang, L.S., and Deng, H.W. (2009). A new permutation strategy of pathway-based approach for genome-wide association study. BMC Bioinformatics *10*, 429.
7. Wang, K., Li, M., and Bucan, M. (2007). Pathway-Based Approaches for Analysis of Genomewide Association Studies. Am J Hum Genet *81*.
8. Wan, X., Yang, C., Yang, Q., Xue, H., Tang, N.L., and Yu, W. Predictive rule inference for epistatic interaction detection in genome-wide association studies. Bioinformatics *26*, 30-37.
9. Moore, J.H., Asselbergs, F.W., and Williams, S.M. Bioinformatics challenges for genome-wide association studies. Bioinformatics *26*, 445-455.
10. Hardy, J., and Singleton, A. (2009). Genomewide association studies and human disease. N Engl J Med *360*, 1759-1768.
11. Goldstein, D.B. (2009). Common genetic variation and human traits. N Engl J Med *360*, 1696-1698.