The Limitation of Genome Wide Association Studies

Final Report – Computational Molecular Biology
Graham Dow
December 13, 2009

Genome wide association mapping has become prevalent in the genomic era – the ability to sequence large amounts of DNA makes it possible to link genetic aberrations with phenotypic traits. More poignantly, this technology has been co-opted with the express purpose of identifying genetic loci that are responsible for human disease; in an effort to both understand the biological factors at work and as a means for risk diagnosis. Nonetheless, significant challenges – technically, statistically, and conceptually – have limited the success of this approach. In this report, the common methodology of genome wide association mapping will be explained, the challenges and shortcomings of this approach will be addressed, and finally, alternative technologies that hold promise for the future will be highlighted. While genome wide association studies can contribute to our understanding of disease mechanisms, their focus on identifying common variants does not substantially improve our ability to predict individual risk.

Embedded within the genomes of individuals of any population lies inherent variability. Indeed, variability amongst individuals provides the phenotypic platform for biological evolution. While this variability is quite apparent amongst humans at the phenotypic level -- simply take note of the diversity between people you pass on the sidewalk -- the differences at the genetic level are actually quite subtle. Differences in the genetic code within a species are mainly the result of single nucleotide polymorphisms (SNPs). These SNPs comprise the allelic variants that, along with environmental influence, are responsible for the phenotypic diversity we observe in the human population. SNP

variants that are commonplace are the result of mutations that took place many generations ago, and spread throughout human genealogy either through genetic drift or selection. Rare SNPs, on the other hand, have arisen from recent mutations, even some within the current generation, with little time to spread. These SNPs are also the basis for genome wide association mapping.

When multiple SNPs occur relatively close to one another, typically distances of 30 kB in the human genomes (1), they predictably segregate together over generations. This effect is known as linkage disequilibrium (LD), where an individual that carries a SNP allele at one site usually carries other specific alleles at nearby SNP locations. A particular combination of SNPs along a chromosome is considered a haplotype. Because the human genome is approximately $3x10^9$ long, and due to the haplotype nature of chromosomes, one could survey the genetic variability in a genome by simply genotyping 100,000 carefully selected SNPs across the genome (2). In essence, the principle of a genome wide association study is to correlate these specific SNPs, and their associated haplotypes or genes, with diseases by comparing the genomes of afflicted and control individuals. The stages of a genome wide association study can be broken down into five major steps: (1) experimental design, (2) genotyping and cleaning of SNP information, (3) statistical association between SNPs and phenotype, (4) independent replication, and finally, (5) linkage of SNPs to casual disease genes.

The initial stage of a genome wide association study is critical to the success of the research endeavor, as careful planning of the study components can increase the chance of finding significant results. First of all, the phenotype for investigation must be well defined,

so that discrete populations can be compared, and easily measured. The chosen phenotype can have large impacts on the power of the study, as the outcome (defining SNPs and then

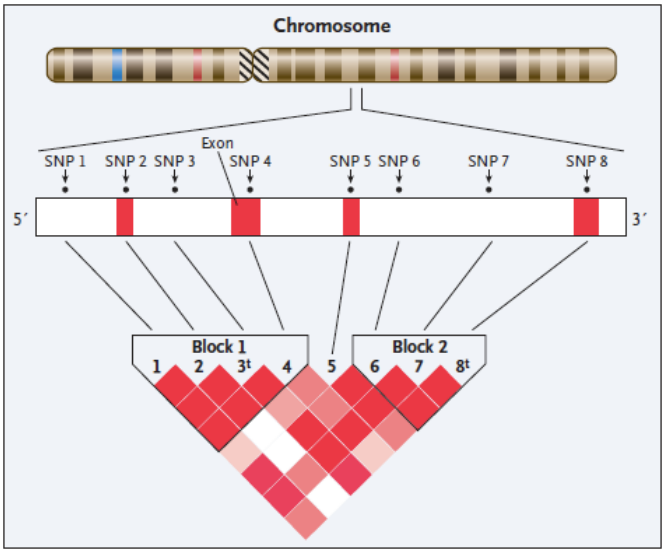

Figure 1 (3): The association between SNPs is represented by varying intensities of red, with the deepest red indicating the strongest association. Triangular blocks would represent a genetic haplotype, where all SNPs in a block normally associate with each other across generations. One of these SNPs (3^t or 8^t) could then serve as a proxy for the entire suite of SNPs in a genotyping assay. In this way, you could limit the number of SNPs to test in surveying the genetic landscape of large genomes.

associated genes) can depend on the effect of genetic variation to the disease or phenotype of interest. Choosing the test population(s) is also an important factor. The standard approach is to have case and comparison subjects from a common population. Using samples from distinct populations, say Africa versus North America, could introduce population substructure that may produce irregularities in the association between SNPs and phenotypes. Sample size is another consideration; in association studies a greater number of samples can produce a more robust linkage between SNPs and phenotypes, however, there is a technical limit to the number included because of physical and logistical constraints. Generally, participants are preferred in the thousands, as opposed to the hundreds (4), but the acquisition and organization of that many samples is challenging. Finally, the library of SNPs to be tested is chosen. Human SNPs can be genotyped using a chip platform, whereby sequence probes that contain different SNPs are hybridized to digested genomic DNA to assess sequence complementation. SNP-chips are commercially

produced by two major competitors, Illumina and Affymetrix. Depending on the investigator, different numbers of SNPs can be placed on a chip, but as mentioned previously, 100K is a good minimum for complete coverage of the genome.

The genotyping and verification process of the genome wide association study is the actual component of molecular biology in the study. Using the microarray-based SNP-chip technology, raw hybridization data is computed by measuring the signal produced by binding with the probes. Algorithms are then used to designate these signals as three possible genotypes (2 homozygote possibilities, one heterozygote possibility) at any given SNP. The algorithm used by the Wellcome Trust Case Control Consortium in their landmark study was called CHIAMO (5), but there is also a standard one supplied by Affymetrix, called BRLMM. Quality control is a significant issue in genome wide association studies, due to the high-throughput nature of the SNP-chip assays and data assimilation, the probability of spurious data increases. This step of the process can be the most troublesome because of the size of the data sets, and also the most dangerous, because any oversight can lead to misinterpretation of the results in later steps. Some of the checks taken to "clean the data" include tests for sample contamination, sample duplication or swaps, false identification of gentoypes, major deviations from Hardy-Weinberg equilibrium, and even relatedness between samples (5).

The third step in the process of genome wide association studies is the most computationally intensive. There are a few different approaches to determine significant association between SNPs and phenotypes, but the simplest method is a single-point analysis where the frequency of each allele in cases or controls is compared. This can be summarized in a simple chi-square statistic that highlights any deviation from the null

expectation (no frequency difference) and calculates a p-value to determine significance (6, Figure 2). The complexity in the analysis arises from the fact that so many statistical tests must be completed (100K and above), which precludes a large amount of false positives at the standard statistical significance level of $p < 0.05$. For example, if 100K SNPs are tested in a given study, then by chance alone, 5K would be expected to be significant, a number far greater than any reasonable estimation of culpable genes. Consequently, a much lower p-value must be used at this stage of analysis, on the order of $1x10^{-7}$ or lower (6). Other approaches use Bayes' factors, which assume an a priori probability of association to calculate a posterior probability of association in place of a p-value (7) or the false-positive report probability, which uses the observed p-value, the prior probability of association, and the statistical power of the test to determine if an association is a true positive (8). The typical representation of SNP associations is a Manhattan plot, where the negative log of the p-value from the chi-square test is plotted against chromosome position (Figure 2).

The fourth step in a genome wide association step is critical for validating any findings – independent replication of the association between SNP and disease is necessary to prove association beyond reasonable doubt. There are two main approaches to accomplish the step of association validation: (1) exact replication of the study on a different data set, and (2) fine mapping of the region of interest (9). To date, there are 2091 SNP-trait associations in the catalog of published genome wide association studies (SNPs $\geq$ 100K in the study) at a p-value of $1x10^{-5}$ or less, but only 439 are published at a p-value of $1x10^{-8}$ or less (10). As mentioned earlier, such a high p-value like $1x10^{-5}$ is susceptible to false-positive associations, and therefore, supplementary confirmation is absolutely necessary.

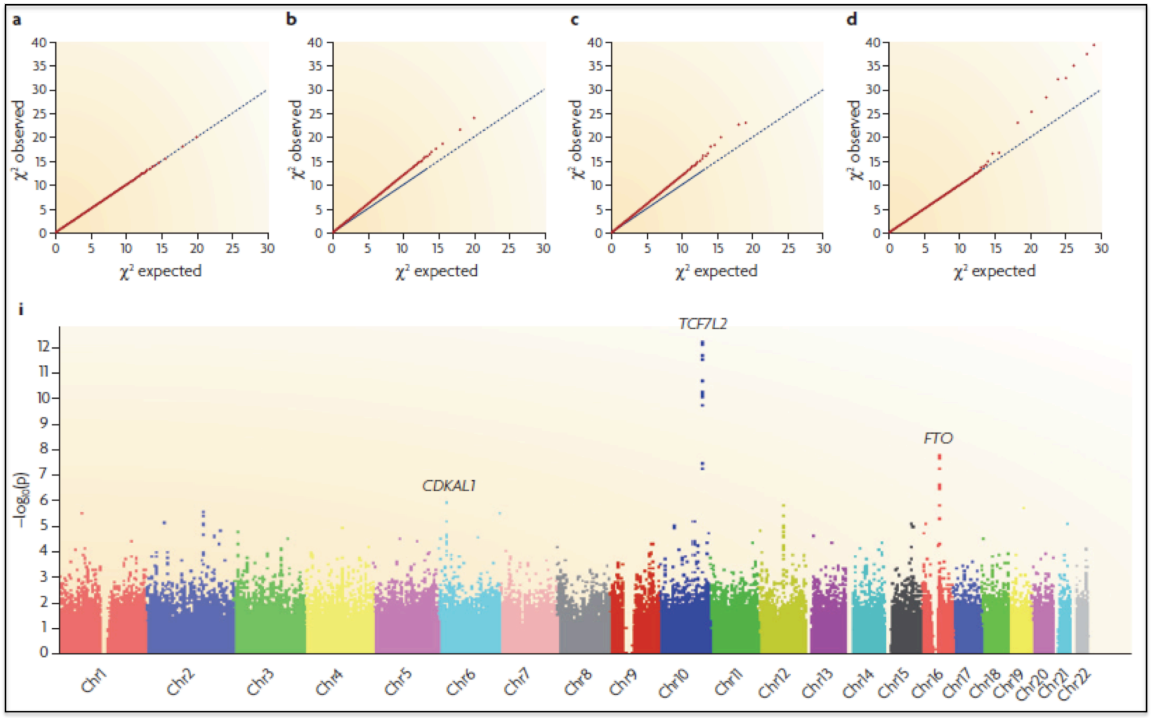

Figure 2 (Adapted from 5): Quantile-quantile plots of the distribution of observed test statistics in a genome-wide association study (a-d), blue line denotes null expectation and red circles indicate idealized test results of four scenarios. (a) little evidence of association, (b) inflation of association, indicative of population substructure or relatedness of samples, (c) excess of strong association, also possibly from population substructure, (d) convincing evidence of disease association with SNPs. Genome-wide association findings in a Manhattan plot with respect to genome location, highlighting loci of high association (i).

The first approach mentioned in step four is pretty self-explanatory; one uses the same SNP-chip platform to facilitate a duplicate study on a different sample population. If a secondary data set is not readily available, which is often the case because of the large sample sizes required, the investigator might try the second approach. Fine mapping entails sequencing the haplotype block where the SNP(s) reside to find a non-redundant set of polymorphisms that can be used to duplicate the association. To accomplish this task, one can retrieve a common SNP set from the HapMap database (11), or if their region of interest is underrepresented (the HapMap database only contains about 30% of the common SNPs present in the genome (9)) they can sequence the region of interest from a

large sample of the population to uncover a novel set of SNPs in that area. This option is becoming easier with the increased accessibility to improved sequencing technologies. This step can also be instrumental in narrowing down region of the genome that holds the strongest association with the disease of interest, a process that will be informative for the final step.

The last step is the most biologically relevant: making a connection between the SNP region and the gene(s) that increase the risk of disease. Seeking the culprit loci may inform the scientific community of a biological mechanism for the disease, which in turn, could lead to putative therapeutics. This step can often be bypassed if the lone goal of the study is to seek genetic fingerprints that increase the risk of disease, which basically creates a means of screening for those susceptible. However, most studies take interest in the underlying mechanism and will pursue a casual relationship. Nonetheless, if the effect size of the SNP is small, it may be very difficult to conclude what variant is responsible. Two major approaches are used in dissecting the SNP of interest, one is computational and the other is experimental. Computational analysis of the base change can lead to functional characterization, for instance: a new base may impair a known binding site in the protein, may alter a crucial folding motif, or change the binding of regulatory partners at a transcriptional or protein level. If the region where the SNP resides does not have thoroughly annotated genes, efforts could be directed to uncover the functional aspects by sequence comparison with protein and motif databases like MyHits motif scan or BLOCKS+. In an experimental light, perturbation of the genes of interest could introduce pathologies in cell culture that lead to disease. Techniques such as RNAi or over-expression would be feasible in this regard. Ultimately, finding the mechanisms of disease is the overarching

goal of biomedical research. Genome wide association studies are a useful tool for narrowing in on the responsible components (12), but monumental effort in subsequent studies is absolutely necessary for defining a model of disease.

Throughout the explanation of the methodology for genome wide association studies, I have tried to highlight challenges and shortcomings, as they are apparent in each stage. At this point, a more general discussion of the limitations of genome wide association studies is appropriate. The most glaring weakness of these studies is the use of common SNP variants for which disease associations are tested (13). Inherent in the fact that these SNPs are common in the human population, one can intuitively come to the conclusion that they should have relatively small effect on causing the disease. Indeed, if these common SNPs had large effects, then many more people would be suffering the ill affects. An example is illustrated in Goldstein's review (Figure 3), which shows that even the strongest associated SNP with type 2 diabetes increases the sibling relative risk to only 1.02, when the overall risk to siblings of affected individuals is three times that. The relative risk is reported in an odds ratio, where the value is calculated as the odds of an all-

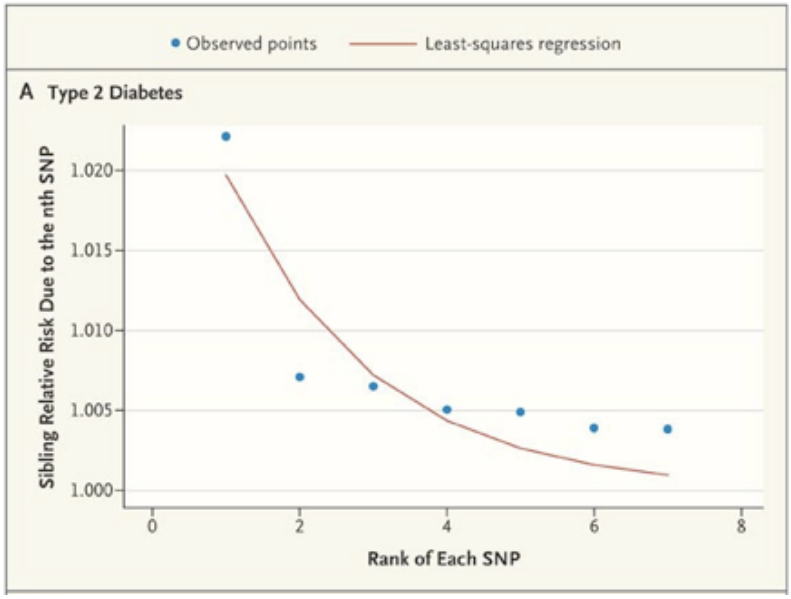

Figure 3 (adapted from 13): The effect of the most powerfully associated SNPs on the relative risk of siblings to contract type 2 diabetes. Notice the significant drop-off after the first SNP, diminishing the already small effect even more.

ele in cases divided by the odds of the allele in controls. An odds ratio of 1 indicates there is no difference in allele frequency between cases and controls. The odds ratio for most genetic variants fall in the 1.2-1.3 range or lower, while a few sterling examples may get between 3 and 4 (14). As a result, the power of using genome wide association studies, and their associated SNPs, in order to assess the genetic risk of individuals to disease is very small. Consequently, the possibility of using common SNPs as a diagnostic tool is very unlikely.

In contrast to the negative outlook of prediction power, proponents of genome wide association studies (14) praise their capability for finding insightful pathway components or uncovering new mechanisms that were not previously considered. While this is a meaningful endeavor, as I previously lauded, the translation from insightful component pieces to actual therapeutics is a long and difficult road. First off, association studies are effective in identifying general loci, not actual casual genes. If the SNP resides in a known expressed gene, then the relationship might be more apparent. However, SNPs are often located in non-coding regions and the culprit gene(s) need be sought in the surrounding genome. From an evolutionary perspective, the lack of strong, common variants in expressed genes makes sense; one should expect that such variants be selected against in the population. The area surrounding an intergenic SNP can be quite large depending on the number of SNPs used in the study, which complicates the identification process and requires further experimentation through *in vitr*o or *in silico* means. Once the candidate gene(s) have been identified, however, the toughest task still remains. In fact, critical proteins and genes are already known to function in countless disease models, but realizing that potential in a functional therapeutic is another story. As Goldstein pointed out:

"nearly a century and three Nobel Prizes separate the determination of the chemical composition of cholesterol from the development of statins" (13). With the vast improvement in recent years of sequencing technology and computational power, the potential for studies to uncover quantitative players in disease pathology has increased tremendously. However, the onus, and to most extent, the limiting factor, still remains in the laboratory.

The realization that common SNPs are contributing very small amounts to disease formation leaves an unquenchable thirst to uncover more powerful genetic components. The source of such critical players in disease might be the exact variants that genome wide association studies overlook – rare variants. Rare variants that exist in the population at frequencies below 5% (cut-off for common variants in genome wide studies is 5% and above) that are related to disease causation could have significant impacts on both prediction of relative risk and understanding mechanism. If these rare variants impart a substantial increase in risk, say for example, an odds ratio of 5, they would allow the medical community to diagnose those individual who are highly susceptible. From an evolutionary perspective, one should expect that rare variants with large effects can subsist in the population because of their rare nature, often hidden in heterogeneity. The caveat, however, is obvious; the percentage of people diagnosed with the risk would be substantially reduced when compared with using common variants because so few people have the rare variant. Nonetheless, it offers a promising approach to improve the power of genome wide association studies. From a mechanistic viewpoint, discovering a potent modulator of the normal biological pathway would ease the difficulty in determining causation. The variant could highlight a critical regulation point, an important hub in cell

signaling, or a novel component of the disease pathology. Targeting that process would take guesswork out of drug development, although the challenge of producing the therapeutic would still remain.

Hopefully, the technology to tackle rare variants is not far off. Some possible avenues include the expansion of the genome wide association study to include more SNPs per chip, of which, some SNPs will be below the 5% threshold. One could also create a chip with only SNPs that ostensibly exist in low frequencies of the population. Alternatively, investigators could use the location of defined common variants to direct focused resequencing efforts in the region of interest. Committed sequencing of afflicted individuals could uncover more powerful rare variants that were not exposed in the previous genome wide scan. Finally, the last approach to propose has the most potential, but is also the most technically challenging. Dedicated full genome sequencing in case vs. control studies. This would require enormous capabilities on three fronts: high throughput sequencing, statistical comparisons at each base pair, and computationally intensive analysis. I will concede that this final approach is a daunting proposition, but it may be the future of biomedical research once technology catches up with our imagination.

Genome wide association mapping has made great strides in the post-genomic era of utilizing the potential of sequence analysis to uncover biological culprits of disease. Limitations notwithstanding, the success of such studies depends on the thorough quality control of sequence data and significant associations, and the utilization of computer technologies to facilitate analysis. Genome wide association studies can truly reveal mysteries in disease pathology, but the predictive power of such associations still remains an elusive trick.

Literature Cited

1. Hardy J, Singleton A. 2009. Genomewide Association Studies and Human Disease. *N Engl J Med* 360: 1759-68
2. Wang WYS, Barratt BJ, Clayton DG, Todd JA. 2005. Genome-wide association studies: Theoretical and practical concerns. *Nature Reviews Genetics* 6: 109-18
3. Christensen K, Murray JC. 2007. What Genome-wide Association Studies Can Do for Medicine. *N Engl J Med* 356: 1094-7
4. McCarthy M, Abecasis G, Cardon L, Goldstein D, Little J, et al. 2008. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature* 9: 356-69
5. Consortium WTCC. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661-78
6. Hunter DJ, Kraft P. 2007. Drinking from the Fire Hose -- Statistical Issues in Genomewide Association Studies. *N Engl J Med* 357: 436-9
7. Stephens M, Balding D. 2009. Bayesian statistical methods for genetic association studies. *Nature* 10: 681-90
8. Wacholder S, Chanock S, Garcia-Closas M, El ghormli L, Rothman N. 2004. Assessing the Probability That a Positive Report is False: An Approach for Molecular Epidemiology Studies. *J. Natl. Cancer Inst.* 96: 434-42
9. Ioannidis J, Thomas G, Daly M. 2009. Validating, augmenting and refining genome-wide association signals. *Nature* 10: 318-29
10. Hindorff L, Junkins H, Mehta J, Manolio T. 2009. A Catalog of Published Genome-Wide Association Studies.
11. Consortium TIH. 2005. A haplotype map of the human genome. *Nature* 437: 1299-320
12. Hirschhorn JN. 2009. Genomewide Association Studies -- Illuminating Biologic Pathways. *N Engl J Med* 360: 1699-701
13. Goldstein DB. 2009. Common Genetic Variation and Human Traits. *N Engl J Med* 360: 1696-8
14. Manolio TA, Brooks LD, Collins FS. 2008. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* 118: 1590-605