SCIENCE REPORT

Gee, What's GWAS? A Look at Genome-Wide Association Studies

Alan Hsu

With the advent of genome sequencing, and the development of technological methods to analyze genomes quickly and accurately, a new type of whole-genome analysis has emerged to analyze the contribution of genetic variation towards disease. Genome-wide association studies (GWAS) are a tool to understand complex diseases ranging from diabetes to lupus. At the root of the functionality of GWAS are the multiple polymorphisms shared within the human population and the extent and frequency which they appear in relation to the incidence of certain pathologies. By merging these fundamental aspects of biology with cutting edge technology, researchers have utilized GWAS to uncover potential mechanisms of disease, and in turn, potential pathways for treatment. This paper discusses both the background and the detail of genome-wide association studies and highlights their current role in the field of lupus research.

On April 15th, 2003, a consortium of leading scientists and academic announced that the human genome had been completely and accurately sequenced. In a conference headlined by Dr. James D. Watson, who fifty years earlier had helped discover the structure of DNA, scientists of the multi-institutional Human Genome Project revealed their findings: the total human genome, composed of 3.1 billion base pairs, 30,000 genes, at a cost of nearly 3 billion dollars.[1] Even in the decades leading up to its realization, the sequencing of the full human genome had been lauded for its potential in advancing both science and medicine. Now, with the entire genome sequenced, studies to realize these advances were being rapidly initiated.

Among the first studies, which utilized findings from the newly discovered genome, were those that attempted to discover the genetic basis of disease. With the entire genome mapped, scientists could now analyze the sequence to select a set of single-nucleotide polymorphisms (SNPs), which span the entire genome, and utilize these SNPs to assess genetic variation between individuals. Using these SNPs, researchers could analyze patients with complex diseases in order to shed light upon the genetic, and potentially the biochemical, causes of their conditions. In the ensuing years after the sequencing of the genome, an array of genetic studies analyzed a multitude of diseases ranging from schizophrenia to stroke.[2, 3] This research continues today where, six years later, whole-scale genome analysis still remains, both in methodology and applicability, very much in its infancy.

GENETIC BASIS OF DISEASE AND GWAS

The usefulness of SNPs stems from the commonality of the polymorphisms among individuals. 90% of all sequence variation within the human population is a result of the 10 million common SNPs that have been found in humans, an average of nearly one SNP per 300 bases.[4] One specific nucleotide in the human genome sequence can take four possible "allelic" forms, corresponding to the four main nucleotide bases (adenine, cytosine, guanine, and thymine), but generally only two such bases become significantly present in the population, with one, generally the evolutionarily oldest, often becoming the most common variant. This single most common variant does not always have to be the SNP that confers a survival advantage, and often it is the case that the second, largely benign, SNP, has arisen more recently out of random mutation and has become stabilized in the population over time.

Disease-causing mutations that have arisen from single nucleotide changes, often resulting in either nonsense or frame-shift mutations, could thus logically be considered a SNP. However, given that these "high-risk" alleles often limit the reproductive fitness of their carriers, they should remain rare variants over time. For these mutations, it is often the case that only one single nucleotide change is necessary to cause full clinical manifestation of the disease. One example of such a disease is the Wiskott Aldrich Syndrome (WAS), wherein one nucleotide change in the sequence coding for the WASP-interacting protein (WIP)-binding region of the Wiskott Aldrich Syndrome protein (WASP) can lead to full degradation of the protein and subsequent disease symptoms.[5] Given the relatively clear mutation-to-disease correlation, genetic diseases such as WAS, with the potential for onset caused by one single nucleotide mutation, can be traced with relative ease through traditional linkage studies and hereditary analysis.

The ability for polymorphisms to cause disease is often measured by the "relative risk" (RR) of disease which the variation may incur, which is measured relative to the risk of disease found in the general majority of the population. Thus, if a polymorphism confers disease in 90% of its carriers relative to the 3% of disease in non-carriers then the relative risk would be 30. For diseases such as WAS, a SNP can have a tremendously high relative risk of disease, given that most of the individuals with one unique mutation can exhibit the symptoms of WAS. The capability of this particular mutation to effect disease can also be considered under its "effect size" or the scope in which one particular mutation contributes to disease onset and the penetrance of the particular mutation, which, in medical genetics, describes the proportion of people with the polymorphism which actually exhibit disease symptoms.[6,7] Thus a SNP in WASP which disables essential WIP-WASP binding, leading to WASP degradation, would

Author Contact: A.H. University of Pennsylvania, 2009. Address correspondence to A.H. at alanhsu2@gmail.edu

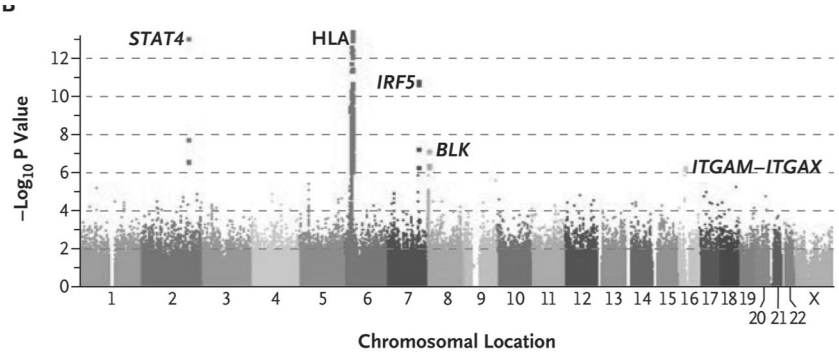

Figure 1. Identification of major loci associated with SLE in a GWAS of 1,311 SLE cases and 3,340 controls characterized for 502,033 SNPs. Adapted from 10.

have a very large effect size, confer a high relative risk of disease, and likely have very high penetrance within the population of its carriers.

But SNPs also operate subtly, contributing to disease with a small effect size and a small increase in relative risk of disease, often only 1.2 to 2 (hypertension and smoking confer similar RRs towards coronary artery disease).[6] With such small effect sizes for individual SNPs, full disease onset can be dependent on multiple polymorphisms across multiple genes, each one conferring varying degrees of risk of disease. The subtleties of ascertaining the genetic bases of complex diseases thus confound the applicability of traditional linkage and hereditary analyses in researching these complex diseases. However, with the advent of genome sequencing, and the development of technological methods to analyze genomes quickly and accurately, a new type of whole-genome analysis has emerged to analyze the contribution of genetic variation towards complex disease. These genome-wide association studies (GWAS) are not unlike a fisherman casting out multiple lines, as they utilize complementary sequences proximal to the SNPs (the hooks) to identify SNPs (the fish) that may elucidate certain characteristics, and potentially the disease state, of a patient (the body of water). These complementary sequences, thousands upon thousands in number, are hybridized to microchips and used to probe subject DNA.[6]

The applicability of genome-wide association studies is grounded in the concept of linkage disequilibrium, which describes the association between multiple alleles at multiple loci within the genome. The importance for linkage disequilibrium (LD) arises due to the association between various SNPs in a certain region of the genome. Two or more SNPs with strong linkage disequilibrium are closely linked, often traveling together in blocks of genome sequence through inheritance and evolutionary history. These strong associations between proximal SNPs allows for the use of a tagging SNP to indicate any potential variation among all SNPs within a linkage disequilibrium block, which is often representative of a specific genomic sequence at that loci.[4,6] Thus, while nearly 10 million SNPs are considered to be present in the human genome, strong linkage disequilibrium allows for the use of only a few hundred thousands tagging SNPs to identify trends in variation within the entire genome. Utilizing these tagging SNPs, GWAS attempt to identify variations in loci which may significantly contribute to disease. The majority of GWAS are case control studies, in which the DNA from a select group of afflicted patients and non-afflicted controls are taken and probed for variation within the frequency of certain SNPs.

The scope for these studies, predictably, mirrors the expansiveness of the human genome itself. In order to find truly significant associations between certain SNPs and disease incidence, the studies require an enormous amount of significant power. This is accomplished through extremely large sample sizes, numbering in the tens of thousands, and strict statistical thresholds for significance, with a p value of 10-8 as an often-used point of difference.[6] But scales of this study inevitably present the logistical difficulties of obtaining an appropriately balanced and large enough sample population. Furthermore, there also exists a "Faustian bargain" between studying extremely large case-control populations and considering gene-environment effects.[8] Studies which utilize extremely large case-control populations, without controlling for multiple environmental exposures, are useful only for finding variations in pathways that function independently of the environment. Given the wide range of environmental exposures potentially within the study population, these large-scale GWAS may conceal important causal pathways dependent on gene-environment interaction, which have been shown to be integral in progression of multiple diseases, such as asthma and allergy.[8]

One example of how GWAS has been applied is the recent research into the genetic basis of Systemic Lupus Erythematosus (SLE). Given the wide clinical heterogeneity seen among patients, and the effects which the disease has on multiple systems of the body, SLE has emerged as a candidate disease upon which the principles of GWAS can be used to ascertain the variations potentially related to disease onset. As such, researchers in the field have placed much promise upon the use of GWAS to elucidate its causal pathways and reveal potential methods of treatment.

GWAS & SLE

SLE is a genetic disease characterized by significant production of auto-antibodies in multiple organ systems, most often in the form of anti-DNA antibodies, and an accumulation of auto-antibody-auto-antigen immune complexes, often leading to an inflammatory response and tissue damage.[9] Within the past two years, a multitude of genome-wide association studies have expanded the number of candidate susceptibility loci, from nine in 2007, to currently more than 20 loci that show significant association with SLE onset.[9] The discoveries have shed light on the potential causality of SLE as well as the applicability of GWAS in assessing complex diseases, in

general. In many cases, GWAS has reaffirmed findings found in traditional linkage studies, such as the link between disease onset to variation in regions encoding proteins related to the major histocompatability complex (MHC) and interferon production. But it is the novel genetic loci, discovered by recent GWAS, which have garnered attention within the research literature. These newly discovered associations have linked SLE to the function of the complement system, B and T cell activation, and apoptosis, among other systems.

Yet, as with all GWAS studies, the identification of association does not indicate causality. Genes within the HLA region have exhibited strong association with SLE through multiple GWAS, but the difficulty upon identification of candidate genes lies in identifying the causal pathway by which HLA function may affect SLE onset.[10] With such a clinical heterogeneity evident in SLE patients, the presence of multiple contributing variations at one or two loci within the HLA region may confound the ability to fully determine the exact contribution of the HLA region to SLE pathology. Determining the close association between SLE and regions with highly specific functions, however, may suggest potential specific explanatory pathways.

One such association is a functional polymorphism identified within the PTPN22 gene, which encodes for an intracellular protein tyrosine phosphatase called Lyp. The R620W polymorphism, which is an arginine to tryptophan substitution, disrupts the binding of Lyp to the SH3 domain of Csk, which is a suppressor of T-cell signaling.[10] Thus, such a polymorphism could disable T-cell suppression and contribute to over-activated T cells and autoimmunity seen within SLE. The finding of this polymorphism is an example of how GWAS can be supplemented with additional experimental data to suggest a biochemical pathway, from variation to disease. It is also likely, however, that a polymorphism such as R620W, on the basis of its specificity, would have a much larger effect size than one affecting the broader HLA pathway.

Where GWAS may prove most useful, however, is linking SLE with systems rather than specific genes. GWAS have identified multiple genetics hits for polymorphisms in regions encoding for proteins related to the complement system (ITGAM/Complement Receptor 3 and C1q complex proteins) as well as B cell activation (BLK and BANK1).[9,10,11] While fleshing out the biochemical pathways for each of the individual genes may prove useful to identify their function in SLE, these findings have instant real-world applicability in that they can suggest potential novel therapeutic options in treating SLE by addressing the system affected. Indeed, researchers have suggested that symptoms of SLE may arise as a result of the inability of the complement system to clear dead cells or immune complexes, thus leading to inflammation and autoimmunity.[9] Identifying multiple genetic associations between SLE and the region encoding complement proteins, solidifies this hypothesis, and provides the justification for directly addressing the complement system in SLE therapy. One interesting additional point to note, is that a large amount of autoimmune disease show a clustering of polymorphisms around the same loci, indicating that similar systems and causal pathways may contribute to each disease.[12] Thus, identification of these systems in SLE research could have repercussive impacts on the analysis of other diseases.

GWAS: HERE TO STAY?

With a multitude of genome sequencing centers and millions in funding going towards GWAS and related studies, it does indeed appear as if GWAS will play a significant role in the future of disease-related research. Yet the optimism of scientific discovery must be checked by the realization that GWAS, and identifying candidate loci, is only the first step towards truly unearthing causal pathways for complex diseases. In order for full realization of findings made from GWAS, this data must be supplemented with laboratory and clinical studies. In addition, significant gains must be made in obtaining appropriate study populations, which ensure both validity as well as repeatability of findings. One concern to note currently is that the majority of studies conducted with GWAS have focused upon white European study subjects, with several additional studies suggesting that the lower degree of linkage disequilibrium in Africans and geographically-isolated populations may limit the efficiency of GWAS in these populations. As with much genetics research, the findings discovered via GWAS are likely to put forth a variety of ethical questions. Realization of the full research potential of GWAS would necessitate research upon populations of various ethnic backgrounds. As these methodological issues are addressed, however, and technological advances are made in improving experimental efficiency, GWAS has the potential to be a breakthrough tool in disease research.

References

1. Wade, Nicholas, "Once Again, Scientists Say Human Genome Is Complete." New York Times Online. April 15, 2003. Available at: http://www.nytimes.com/2003/04/15/science/once-again-scientists-say-human-genome-is-complete.html.
2. Stefansson H. et. Al. "Neuregulin 1 and susceptibility to schizophrenia." Am J Hum Genet. 2002 Oct 71(4):877-92. Epub 2002 Jul 23.
3. Gretarsdottir S. et. Al. "The gene encoding phosphodiesterase 4D confers risk of ischemic stroke." Nat Genet. 2003 Oct; 35(2):131-8. Epub 2003 Sep 21.
4. The International HapMap Consortium. "The International HapMap Project." Nature. 2003 Dec 18;426(6968):789-96.
5. Orange JS, Stone KD, Turvey SE, Krzewski K. "The Wiskott-Aldrich syndrome." Cell Mol Life Sci. 2004 Sep; 61(18):2361-85.
6. Mullen SA, Crompton DE, Carney PW, Helbig I, Berkovic SF. "A neurologist's guide to genome-wide association studies." Neurology. 2009 Feb; 72(6):558-65.
7. Hall JG. "Factors Affecting Gene Expression." The Merck Manuals: Online Medical Library. <http://www.merck.com/mmpe/sec22/ch327/ch327c.html>
8. Vercelli D, Martinez FD. "The Faustian bargain of genetic association studies: bigger might not be better, or at least it might not be good enough." J Allergy Clin Immunol. 2006 Jun; 117(6):1303-5.
9. Harley IT, Kaufman KM, Langefeld CD, Harley JB, Kelly JA. "Genetic susceptibility to SLE: new insights from fine mapping and genome-wide association studies." Nat Rev Genet. 2009 May; 10(5):285-90.
10. Criswell LA. "The genetic contribution to systemic lupus erythematosus." Bull NYU Hosp Jt Dis. 2008; 66(3):176-83.
11. Ardoin SP, Pisetsky DS. "Developments in the scientific understanding of lupus." Arthritis Res Ther. 2008;10(5):218. Epub 2008 Oct 10.
12. Lettre G, Rioux JD. "Autoimmune diseases: insights from genome-wide association studies." Hum Mol Genet. 2008 Oct 15; 17(R2):R116-21.