

Missing Heritability and the Future of GWAS

by Christophe Lambert, CEO & President of Golden Helix

“Where is the missing heritability?” is a question asked frequently in genetic research, usually in the context of diseases that have large heritability estimates, say 60-80%, and yet where only perhaps 5-10% of that heritability has been found. The difficulty seems to come down to the common disease/common variant hypothesis not holding up. Or perhaps more accurately, that the frequency of the assayed markers are not in line with the frequency of the disease (or specific sub-phenotype thereof). Most of the technologies directed towards finding the genetic links to diseases – e.g., the first generation of major microarray platforms used in genome-wide association studies (GWAS) – were developed using this hypothesis as a premise.

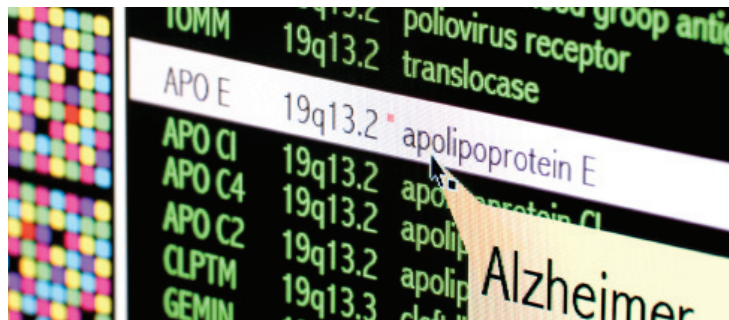
Limitations of First Generation Microarrays

One major limitation is that the microarrays used in most major GWAS efforts to date employ common genetic variants originally identified in a rather small number of presumably healthy people (HapMap Phase I). Many high-profile and heavily researched diseases, such as Type 1 diabetes, are really not so common, appearing in 1 person out of, perhaps, 500-800. Why, then, should we expect that common genetic polymorphisms found in a handful of HapMap individuals would be linked to the causes of disease in the relatively small proportion of people who have Type 1 diabetes? The assumption that common single nucleotide polymorphisms (SNPs) will reliably tag such variants is shaky.

Admittedly, we will find some additional signal if we use massive sample sizes, but we will still be missing the bulk of the heritability because of one important mathematical fact: correlation does not obey a transitive relationship. If A is correlated with B, and B with C, then A is not necessarily correlated to C, unless the correlation is perfect. The first generation of microarrays operated off the premise that nearby SNPs in linkage disequilibrium will be sufficiently correlated with the causative SNP to get in the ballpark of the causative variant. That is, they assume transitivity. However, if the causative variants are rare in the healthy population, then they are unlikely to be highly correlated with common variants typed in healthy individuals, and larger sample sizes are going to give, at best, diminishing returns.

Alzheimer’s Irony

Alzheimer’s is often given as an example of the success of the common disease/common variant hypothesis, yet, ironically, it provides an excellent example to illustrate the failure of tagging SNPs. We were recently involved in quality control and analysis



of the GenADA Alzheimer’s study posted by GlaxoSmithKline on dbGaP. The study used the Affymetrix 500K array, which does not assay the specific common polymorphism for APOE, a gene in chromosome 19 that has been linked to Alzheimer’s in several linkage and association studies. Separately, that SNP, rs429358, was genotyped using low-throughput methods. When we ran the association, there was no appreciable association in the APOE region from the 500K SNPs, but rs429358 was significant at almost a 1e-60 level in a sample size of 1577 individuals. We then searched chromosome 19 to find the SNP most correlated with rs429358 and found one correlated with a trend test p-value below 1e-13. However its association with case/control status was 0.0036 – a value not even considered nominally associated in a genome-wide context. Interestingly, we imputed this 500K data set up to the ~900k Affymetrix 6.0 density and still did not see a genome-wide significant signal. Could vast sample sizes have found this signal? Perhaps, but how many samples would it take? In this case, typing the correct marker made all the difference.

The Marriage of Next-Gen Sequencing and Microarrays

I believe the way around this is to locate variants in the diseased individuals and then run all of the machinery of GWAS that has served us well so far. One way is to use next generation sequencing on modest numbers of cases for a given disease (possibly pooling samples) to find markers that are rare in the overall population but common and more highly penetrant in the disease population, and use those markers in association studies – perhaps with custom microarrays. Given the Alzheimer’s associated SNP was commonly present in Alzheimer’s patients, one might have sequenced even a dozen Alzheimer’s patients and used those SNPs in a GWAS to find the highly significant signal.

Some who espouse the rare variant hypothesis say that there will be many, many rare variants that add up to explaining the missing heritability. I’m inclined to think that for most diseases there will

be relatively few, and that we just haven't found them yet. I also think we will have to modify our view of heterogeneous "common diseases" like heart disease – there are probably dozens of ways the cardiovascular system can go wrong, many of them being characterized as "heart disease," but each being its own unique disease. Perhaps the only problem with the common disease/common variant paradigm is that there are hardly any truly common diseases, when we consider sub-phenotypes.

What About That Height Paper?

How then do we reconcile all of the above with the fascinating recent paper that came out in Nature Genetics by Jian Yang, et al., "Common SNPs Explain a Large Portion of the Heritability for Human Height" where 45% of human height variance can be explained with considering all 294,831 common SNPs used in the study simultaneously? Previous to this paper, GWAS studies on tens of thousands of individuals found approximately 50 genome-wide significant SNPs, and determined that these only accounted for around 5% of height variability. The authors write,

There are two logical explanations for the failure of validated SNP associations to explain the estimated heritability: either the causal variants each explain such a small amount of variation that their effects do not reach stringent significance thresholds and/or the causal variants are not in complete linkage disequilibrium (LD) with the SNPs that have been genotyped.

After showing in a regression framework that 294,831 SNPs together account for 45% of the variation, the authors conclude that both explanations hold – there are many causal variants of weak effect, and that the remaining variation is accounted for by untyped variants that are not sufficiently correlated with the typed variants to explain the remaining heritable variability.

For me there is a disconnect between demonstrating that thousands of variables in a regression model together have a high correlation with height, and concluding that therefore there must be thousands of weakly penetrant causes. Rather the data also seems consistent with there being many weak correlations with potentially quite few untyped causal variants. I think the paper shows a very nice result, in that the heritability is not missing, but I disagree that it must be a huge number of weak effects. In any case they do ultimately invoke untyped variants as the explanation for the other ~1/2 of the variability, which is consistent with the view that we just haven't typed the variants that matter for our phenotypes of interest.

I have a fundamental cognitive dissonance with the view that it will require the incorporation of tens of thousands of low odds SNPs to explain disease variation. It clashes with the paradigm that drives scientists to search for simpler and more fundamental causes for effects, dating back to Newton who said "Natura valde simplex est et sibi consona", or "nature is exceedingly simple and harmonious with itself". The evolution of a field of knowledge towards becoming a science begins with classification (e.g. taxonomies), followed by correlation, followed by causation. GWAS is still firmly entrenched in the correlation phase, and correlation only gives us potential directions towards locating cause. When we consider a disease that presents a consistent phenotype – a single effect, it is difficult for me to posit thousands of causes. Thousands of weak correlations, on the other hand, are totally expected due to the highly interconnected nature of human biology.

The Future of GWAS

I believe success in explaining the "missing heritability" will come if we use clinical data, proteomics, gene expression and metabolomic biomarkers to define subphenotypes (known as deep phenotyping), use next-generation sequencing to sample the variants in those disease subgroups, and then follow those with disease-focused GWAS based on custom arrays. Custom arrays with 10,000 SNP and indel markers are currently priced in the ~\$50/sample range for large studies. Whole-exome next-gen studies are in the ~\$3000/sample range and falling, and whole-genome sequencing is below \$10,000/sample and falling. In the coming years, I expect to see many cost-effective and productive studies consisting of sequencing 50-100 cases to find variants to design a custom chip, followed by a GWAS of a few thousand samples. The genotyping work for a 10,000 person study of this sort could be done for under \$1M, and this will only go down over time. I believe this marriage of next generation sequencing and custom microarrays is likely to be a long and fruitful one, and is an important part of our own product development direction. ...*And that's my two SNPs.*

About Golden Helix

We are inspired by significance. Not only statistical, but technological, scientific, and personal significance. It's embodied in everything we and our customers do. And we believe the only way to achieve significance is by transcending the status quo. Every day we strive for extraordinary analytic and technological advancements that empower scientists around the world to pursue that which is truly significant: from uncovering the genetic causes of disease and transforming drug discovery to developing genetic diagnostics and advancing the quest for personalized medicine. To learn more about Golden Helix, visit www.goldenhelix.com.