Comment

Guilt by association

Gregory A Petsko

Address: Rosenstiel Basic Medical Sciences Research Center, Brandeis University, Waltham, MA 02454-9110, USA.
Email: petsko@brandeis.edu

Published: 29 April 2009

Genome ***Biology*** 2009, **10:**104 (doi:10.1186/gb-2009-10-4-104)

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2009/10/4/104

It's a favorite device of politicians who wish to smear a rival candidate. The late, unlamented Democratic Senator Joseph McCarthy employed it, successfully, against leftists whose careers he wished to destroy. The Republicans tried, unsuccessfully, to use it against Barack Obama. In its simplest form, it involves branding someone a Communist, or a terrorist, or a criminal, because they have family or friends, or possibly just casual acquaintances, who are Communists, or terrorists, or criminals. It's called guilt by association.

It's been in the news lately because it's also a favorite tactic of genome biologists, but in this case its purposes are not sinister. Scientifically, it goes by the name of genome-wide association studies, though guilt by association is just as apt. It's an attempt to find connections between simple changes in the coding sequence of genes and the risk for developing complex diseases. A product of the human genome sequence (and, one could almost say, a means of ensuring job security for the hordes of sequencers who were responsible for that project), genome-wide association studies represent the first comprehensive attempt by the genomics community to demonstrate a big payoff, in terms of benefits to human health, for the enormous amounts that were spent on that original project.

If there were world enough and time, as Andrew Marvel (or was it Francis Collins?) would say, we would perform such studies simply by sequencing the complete genomes of large cohorts of people with, say, type II diabetes, or lung cancer, or schizophrenia, or Alzheimer's disease, and then letting the computational folks sift through the resulting reams of data to sort out the varying combinations of simple spelling mistakes in many different genes that give rise to autism, or stroke, and so on - and it's quite likely that, when sequencing costs come down sufficiently, this is exactly what we will do. In the meantime, though, the effort is more restricted.

The current approach relies on data from the International Human Haplotype Mapping (HapMap) Project, which aims to determine the prevalence of common polymorphisms in the human genome, and on the fact that genetic variance at one locus can predict with high probability genetic variance at an adjacent locus, typically over distances of 30,000 base pairs of DNA, making it possible to map the common variability - and, as we shall see, the key word here is 'common' - associated with the risk of a given disease simply by genotyping approximately 500,000 judiciously chosen markers in the genome of several thousand case subjects and comparing the frequency of those markers with genotypes of control subjects. Consequently, it has become relatively routine to identify common variants (for example, those that are present in more than 5% of the population) that confer not a certainty but rather a risk of disease, typically with odds ratios of 1.2 to 5.0.

But now, in a series of articles in the 15 April 2009 online issue of the *New England Journal of Medicine*, a debate is taking place between proponents of this approach and those who argue that it will never succeed in revealing the genetic basis for complex, polygenic disorders. The reason for this scrutiny is the failure of most of the ongoing studies to find any convincing link with common diseases. It had been expected that the risk of getting cancer, diabetes, and so on would be largely controlled by a relatively small number of common variants, each of which conferred a significant risk, but only a small number of disease-associated variants have been found thus far, and, with a few exceptions, the risk they seem to confer is quite small. As Hardy and Singleton frame the question in one of four papers on the subject, "...discussion has centered on evaluating how far such studies will take us in understanding the risks and causes of disease - and thus the time and resources that should be invested in genotyping more case subjects with any one disease to garner what many see as diminishing genetic returns."

Because they rely on the HapMap, current studies identify loci, not specific genes. Moreover, we currently have haplotype maps only from single nucleotide polymorphisms

(SNPs) present in at least 5% of chromosomes of each of just three groups of defined ancestry: Yoruban, Northern and Western European, and Asian (Chinese and Japanese), so by definition the markers are for rather common variants (present at > 5% frequency) in the human population. Underlying the project at present, then, is the assumption that common diseases are associated with common variations. A further assumption is that, even if individual alleles have only a small effect on one's risk for a disease, each contribution is large enough that a manageably small number will sum to a significant effect. But what if each contribution is very small?

In a companion perspective, Goldstein argues that very many very small contributions is exactly the case and questions the wisdom of continuing this strategy. He points out that there are examples of its successful application: "For example, when exposed to the anti-HIV drug abacavir, a hypersensitivity reaction develops in more than half the carriers of the HLA-B*5701 allele, whereas such a reaction occurs in less than 5% of patients without this allele. Similarly, just three common variants are sufficient to explain 14% of the population variation in HIV-1 viral load." But, he continues, "with traits such as height or type 2 diabetes, it seems that an inordinate number of common SNPs would be needed to account for a sizable fraction of heritability... The apparently modest effect of common variation on most human diseases and related traits probably reflects the efficiency of natural selection in prohibiting increases in disease-associated variants in the population." In other words, common diseases might well be caused by many different combinations of a large number - probably hundreds - of very rare variants, which would even eliminate the utility of the SNP hunt in identifying pathways leading to disease. "In pointing at everything," Goldstein writes, "genetics would point at nothing."

Similar concerns are expressed by Kraft and Hunter in a companion piece (although they favor continuing the common variant hunt). "First, the relative risks that are found to be conferred by common risk genotypes account for only a small proportion of the sibling recurrence risk (or the risk that a sibling will also have the disease of interest). Second, in multivariate analyses of large epidemiologic data sets in which a family history of a disease is a risk factor, the inclusion of data regarding which subjects carry the known associated variants only minimally reduces the risk associated with a family history of the disease. Third, in the case of diseases that have been the focus of several genome-wide association studies, some alleles have been detected more than once, but each study has identified multiple alleles that were not identified in other studies, suggesting that many more alleles remain to be discovered. These factors suggest that many, rather than few, variant risk alleles are responsible for the majority of the inherited risk of each common disease."

One truly surprising result from the studies thus far is that the majority of loci identified as associated with disease risk do not map to the coding regions of individual genes. Instead, they possibly affect either the splicing of the messenger RNA or the sequences of microRNAs that regulate gene expression. Deducing the effects of non-coding changes on the level of active protein(s) in the cell is simply not possible from first principles at the moment; it will require huge experimental efforts in multiple laboratories.

So all this really seems very discouraging, but I have a modest proposal for a somewhat altered approach that I think could yield exciting results rapidly. The problem with most fishing expeditions, which is what genome-wide association studies are, is that one is never sure that one is fishing where the fish are. My proposal is to focus on where we know there are fish (or, to use another analogy, to look for the keys under the lamppost because that's where the light is). That seems unlikely to provide new information, but hear me out. I think the mistake we're making is in looking at the association between SNPs, many of which mean little or nothing, and disease. What we should be doing is looking at the association between diseases.

There are literally hundreds of inherited metabolic disorders, most of which are autosomal recessive - they require mutation in both copies of the gene in question to produce the disease. In many cases there are dozens or even more than a hundred known alleles in the gene in question, any two of which suffice. Carriers for these diseases have just a single variant and are usually free from symptoms of the disease. But it is slowly becoming clear that for at least some of the inborn errors of metabolism, the carriers are at altered risk for something else. It may be that a carrier has a reduced risk for an infectious disease, but often I think a problem with a metabolic enzyme will produce haploinsufficiency in some pathway that is involved in a very different disorder. Thus, carriers for the recessive, lysosomal storage disorder Gaucher disease are almost an order of magnitude more likely to develop Parkinson's disease. Is the connection through lysosomal dysfunction? Maybe, and that's testable: it suggests that carriers for other lysosomal storage diseases such as Niemann-Pick, Tay-Sachs, Anderson-Fabray, and Pompe's diseases should also be at increased risk for Parkinson's and perhaps other neurological disorders.

That's exactly the sort of thing a genome-wide association study could determine, and it would shed valuable light on the causes of a class of common diseases. I think it's almost guaranteed to turn up things, because metabolism is tied into all the other pathways in the cell, and because by definition the carrier alleles for a recessive disorder are mutations that must have some definite effect on the expression or function or stability of the protein in question.

In short, I think we should look at rare diseases for less rare genetic loci (the carrier frequency for Gaucher's is estimated to be 1 in 100 in the general population, and around in 1 in 20 among Ashkenazi Jews) that we know have physiological consequences, and ask whether they are associated with the risk for other, more common diseases. I think that's where the interesting connections are most likely to be found, at least until we can sequence lots of whole genomes very cheaply. After all, in a real criminal case, the police usually focus on suspects they know are likely to be guilty, because they have already been proven guilty of other things in the past. That's guilt by association, to be sure, but it tends to work.