How to read a genome-wide association study

18/07/2010
Categories: Background
Written by Jeff Barrett

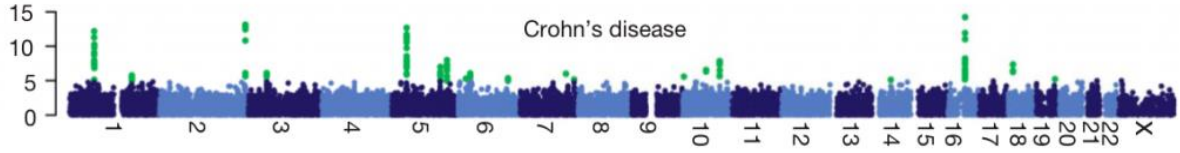
As any avid follower of genomics or medical genetics knows, genome-wide association studies (GWAS) have been the dominant tool used by complex disease genetics researchers in the last five years. There's a very active debate in the field about whether GWAS have revolutionized our understanding of disease genetics or whether they were a waste of money for little tangible gain. No matter where you fall in that spectrum, however, you need only to browse the table of contents of any recent issue of *Nature Genetics* to see how ubiquitous they are. Since GWAS provide so much of the fodder for unzipping your genome, and in order to help you cut through the hype in the mainstream press coverage of GWAS, I've put together a quick primer on how to go straight to the original paper and decide for yourself whether it's a landmark finding or a dud.

The basic GWAS approach is to look at approximately a million positions in the human genome (called 'SNPs') where different people carry different versions of the genetic code (so at some particular position I might have an 'A' and you might have a 'C'). I'm going to focus here on the most common GWAS design, called case-control, where the goal is to compare the frequencies of these different versions between a group of healthy individuals (controls) and another group of people with a specific disease (cases). The places where the frequencies between cases and controls are significantly different are therefore associated with risk of developing the disease.

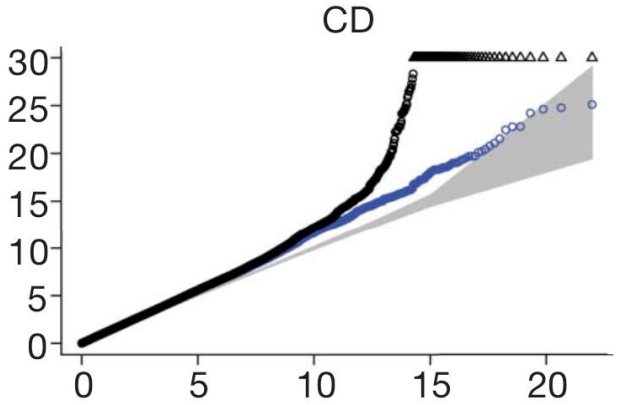
It's not always that easy, though! Listed below are five issues raised by almost every GWAS and how you can try to zero in on the key details about them in the paper. A few example figures are taken from the WTCCC GWAS, which also serves (in my biased view!) as an excellent example of the right way to carry out one of these studies.

1. **Sample size.** One key thing to look for early in the paper is how many samples the study has managed to collect. GWAS are generally aimed at finding very small effects (increasing your risk by, say, 15%) so they need lots of samples to confirm such small differences with statistical confidence. If a paper has fewer than a thousand cases and controls you should be suspicious. There are some exceptions (like big genetic effects on severe side effects from drugs), but these are rare.
2. **Quality control.** The biggest challenge to successfully carrying out a GWAS is getting good, clean genotype data. Pay close attention not only to the standard QC metrics (genotype call rate, Hardy-Weinberg equilibrium, etc) but also to whether extra attention was focused on the genotypes of the most associated SNPs. Most GWAS practitioners go to great lengths to find lab problems that might create false positive associations, but even after years of these best practices being understood, there are sometimes still GWAS published where authors, reviewers and journal have all missed possible genotype QC issues. Good QC should filter out artifacts, and yield a 'manhattan plot' like the one below. Each point is a SNP laid out across the human

chromosomes from left to right, and the heights correspond to the strength of the association to disease. You'll see that the strongest associations (highlighted in green) form neat peaks where nearby correlated SNPs all show the same signal. Any manhattan plot with points all over the place should be viewed as highly suspicious (as raised by Daniel in this excellent post).



3. **Confounders.** Be on the lookout for any variable in the study which could be different between cases and controls *other than the disease itself*. For instance, if the disease is more common in one part of the world than another, and this effect isn't accounted for, then the naturally arising genetic differences between those groups of people will look like they're associated with disease. This 'population structure', as it's called, is the most commonly discussed confounder, but many others exist, such as whether cases and controls were genotyped in the same laboratory, or the DNA was collected by the same method.One statistical tool, called the 'QQ plot' is a common way for GWAS to show that confounders aren't at work. The QQ plot shows the expected distribution of association test statistics (X-axis) across the million SNPs compared to the observed values (Y-axis). Any deviation from the X=Y line implies a consistent difference between cases and controls across the whole genome (suggesting a bias like the ones I've mentioned). A clean QQ plot (see below), on the other hand, should show a solid line matching X=Y until it sharply curves at the end (representing the small number of true associations among thousands of unassociated SNPs). The blue points in this figure show what's left after removing the validated associations, which shows that most of that tail was, in fact, due to true disease variants, but also that more interesting results might still be lurking in the data.



4. **Replication.** The ultimate arbiter of a GWAS result is whether it can be replicated independently. It's important to remember that this doesn't just mean independent samples (though that's crucial) but also using an independent technology. That way, any QC problems or confounders which affected the original study won't affect the replication.
5. **Biology.** Given that a GWAS has some firm results, there's almost always some speculative comment about why these regions of the genome are important to this disease. Take this section with a grain of salt, since it's surprisingly easy to dig up a paper published at some point in history to support almost any functional hypothesis!