Brookfield *BMC Biology* 2010, **8**:41
http://www.biomedcentral.com/1741-7007/8/41

BMC Biology
incorporating Journal of Biology

QUESTION & ANSWER

Q&A: Promise and pitfalls of genome-wide association studies

John FY Brookfield*

Why do we need genome-wide association studies?

To answer that, we first need to look at the kinds of genetic changes that have previously been studied by medical geneticists. These have usually been 'single-gene disorders', which result from mutations in single genes, where an individual with a mutant allele of the gene (in the homozygous state for a recessive disorder) has the disease with a hundred percent probability. Thus, an individual homozygous for the sickle-cell allele of the beta-globin gene will always have sickle-cell anemia (Figure 1). When all individuals with the disease genotype have the disease, we describe such a mutation as one hundred percent penetrant. When the penetrance is less, there are individuals who have the predisposing genotype, but do not have the disease. This is because other genes play a role in the determination of the disease, or because of the effects of the environment. This makes the mapping of the gene causing the disease using pedigree information (as illustrated in Figure 1) more difficult.

Where the penetrance is very low, it is virtually impossible to map genes using pedigrees, and here we enter the world of multifactorial disorders, where the presence or absence of the disease is influenced by many genetic differences and also by the environment. The role of genes in determining whether individuals have the disease can still be important, and this is measured by the 'heritability' of the disease, which is the proportion of the determination of the disease that is caused by genetic rather than environmental differences. Heritability for such disorders is measured through the correlations between relatives, most powerfully using monozygotic (identical) and dizygotic twins. Single-gene disorders tend to be rare, whereas many important multifactorial diseases, including, for example, hypertension, diabetes and schizophrenia, have much higher frequencies in the population, but still have high heritabilities. The goal of genome-wide association studies (GWAS) is to understand common multifactorial diseases and the genes that predispose us to them.

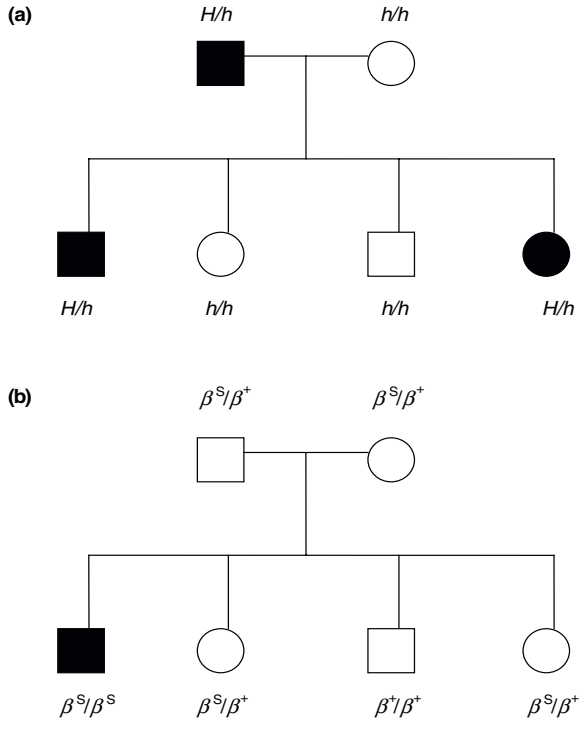

Figure 1. Pedigrees. Here a square represents a male, and a circle a female. **(a)** For a dominant disorder, such as Huntington's disease, represented by the allele symbol *H*, an individual need only inherit the mutant allele from one parent (that is, be heterozygous) to be affected. In such cases, affected individuals almost always are in fact heterozygous for the mutant allele, and they transmit it, and the disease, to half the offspring. **(b)** For a recessive disorder, in this case sickle cell anemia, shown by the β^S allele symbol, the disease is seen only in individuals homozygous for the allele (that is, who have inherited the allele from both parents); they are typically the offspring of two heterozygotes (carriers of the disease).

Do common multifactorial diseases result from the combined effects of common alleles of predisposing genes?

That is indeed thought to be likely, and is the basis for the so-called 'common disease-common variant model' for

*Correspondence: John.Brookfield@nottingham.ac.uk
School of Biology, University of Nottingham, University Park, Nottingham, NG7 2RD, United Kingdom

multifactorial diseases. The suggestion is that, since the disease is common, its presence may arise from a set of predisposing alleles at multiple loci, each of which is itself common in the population.

If an allele predisposes us to a disease, how does it become common? Why has selection not operated to eliminate it?

For single-gene disorders, we think of the frequency as typically depending on a 'mutation-selection balance'. Mutations at the disease locus arise all the time, and (in homozygotes for recessive mutations, or heterozygotes for dominant mutations) cause disease. The disease lowers an individual's ability to survive and breed (Darwinian fitness), and mutations are eliminated from the population by selection. Eventually, the population reaches an equilibrium frequency of the disease, where the rate of loss of the disease alleles by selection is exactly balanced by the rate of gain of disease alleles by mutation. It follows that diseases with high mutation rates (such as Duchenne muscular dystrophy) are more common than diseases with lower mutation rates. Also, a disease that has a small effect on Darwinian fitness, such as one that has its effects after reproduction, will have a higher equilibrium frequency than a disease with a lethal effect in childhood. As I have said, single-gene disorders are rare: the mutation rate is relatively low, and selective pressure against them relatively strong.

Multifactorial diseases, by contrast, can be common, for a number of reasons. First, it is important to remember that an allele that predisposes to a multifactorial disease is only affected by selection to the extent that the frequency of the disease is higher in individuals with that allele than in individuals lacking it, and if it has only a small effect on the probability of the disease, the selection against it is correspondingly reduced. There is also a difference in the relevant time scale for multifactorial against single-gene disorders, which are typically caused by relatively recent mutations. A common allele predisposing us to a multifactorial disease could have arisen tens or hundreds of thousands of years ago, and may have become common in an environment that is very different from that in which we now live. Perhaps the selection operating at that time, particularly on diseases of old age, was very different from that prevailing now. There can also be gene-environment interactions, where an allele might produce a disease only in people living in a modern environment. There is also the possibility that the allele that predisposes to a disease may have other, beneficial effects. (Such an allele is described as 'pleiotropic'.) The overall effect of the allele on fitness might then be very slight or could indeed be positive. Finally, the process of random genetic drift can raise the frequencies of alleles that predispose to disease, and this could have been common during the rapid increases in population size of modern humans as they spread from their origins in Africa during the last 150,000 years.

How exactly can genome-wide association studies (GWAS) make it possible to identify the predisposing genes in multifactorial diseases?

This approach is driven by the new technologies that allow tens or hundreds of thousands of polymorphisms, usually single-nucleotide polymorphisms (SNPs), to be assessed simultaneously. This technique is applied to a set of cases (individuals with the disease) and a set of matched controls, and differences in the frequencies of SNPs between the two groups are assessed in order to identify SNPs that may be associated with the disease. With so many SNPs being tested, the situation is a bit tricky statistically as there is a danger of false positive associations - that is, associations that occur purely by chance and not because the SNP is linked to the disease. So, generally, the significance is adjusted on the basis of getting a 'false discovery rate' of 5% - that is, of all the SNPs called as being associated with the disease, it is expected that only 5% will be truly unassociated SNPs showing an association by chance in the samples.

Wait! - What is a polymorphism? Is it a kind of mutation?

This is a question with a slightly complex answer. Population geneticists have long used the term for genetic variation where, at a particular genetic locus, or, ultimately, base pair, there are two or more genetic variants where the commonest has a frequency below around 95%. In other words, it is a situation where there is not a single normal (wild-type) allele with one or more rare variants - in which case the rare variant base pairs would be called mutations. What makes the answer slightly complex is that there has recently and occasionally been a subtle change in the use of the term. Thus, if, at a given base pair, 90% of alleles have a T and 10% have an A, we say that there is a SNP - a single nucleotide polymorphism - at that base. However, occasionally, some describe this in an asymmetrical way - in which the A is said to be a polymorphism and the T is not.

So the idea is that the SNPs identify predisposing alleles and thus the biological basis of the disease?

Ultimately, yes. But there are also practical benefits just to having a way of identifying individuals at risk without knowing the mechanism. For many multifactorial diseases, treatments and tests are available that are offered on the basis of calculated risk. Thus, a diagnostic test might be carried out on an individual whose lifestyle, age, family history and other factors added up to a 22% risk of a

condition, but such a test might not be offered to someone with an 18% risk. The additional information about risk that is supplied by genotype can allow a more precise targeting of tests to those individuals who are truly most at risk.

One hope is that there could be different treatments for a given disease, designed for those with differing underlying genetic causative factors. Thus, if one patient with a multifactorial disease has predisposing alleles A, C and E, while another patient with the same disease has predisposing alleles B, D and F, then it could be that the best drug treatments for these two patients are different even though their symptoms are not, because of the different etiologies of their diseases. This is what is known as personalized medicine.

But, more fundamentally, the identification of causative loci in GWAS can indeed give insight into the biology of the disease. An allelic difference detected in GWAS might only have a weak effect on the probability of getting the disease. But the modest effect seen may be slight not because the gene involved is unimportant in the pathway that leads to the disease, but because the alleles involved might both be functional alleles showing only subtle, quantitative differences in their action. The importance for treatment of the identification, through a disease association, of a gene or a pathway might well outstrip the importance of the effect of allelic differences at that gene on disease risk.

So are you implying that the SNPs associated with disease are directly causing defects in predisposing genes?

No. The simplest way in which a genetic variant, such as a SNP, can be associated with an increased risk of disease is indeed if such a variant directly causes the elevated risk. But it is much more likely, in any given case, that the SNP being investigated is associated with other genetic differences which, in fact, determine the risk. In population genetics terms, we say that the marker investigated (in this case the SNP) is in linkage disequilibrium with the genes causing the disease. Unfortunately, because of the ways that linkage disequilibrium can arise, this does not always help us to find the culprit gene.

So how does such linkage disequilibrium arise?

There are many ways in which this can happen. One simple way is through population substructure. Thus, if a population consisted of a mixture of individuals with African and European ancestry, for example, and the disease was commoner among those with European ancestry, then, if one took a random sample of cases and a random sample of controls, the cases would be enriched for people with European ancestry, and all the SNPs that showed differences in frequency between Europeans and Africans would also differ between cases and controls, even though almost all would be unlinked to any genes actually causing the disease.

So does that mean that if you are looking at GWAS across populations, you are likely to be led astray by genetic differences between the different populations?

This is a danger, but it can be prevented, in principle, by matching the ancestry of cases and controls. Thus, each time you include an affected individual (a 'case') who has a particular ancestry, you add a control with a similar ancestry. This means that the cases and controls will come from the same mix of ethnic groups, and differences in the frequency of the disease between groups will not create false positives. Even when you do this, however, it is possible you may be led astray by cryptic population stratification.

What is cryptic population stratification?

While it could be straightforward to ensure equal numbers of individuals with European versus African ancestry in the cases and controls, there will be subpopulations within these populations, which will be harder to match. Any SNP that shows a very great frequency variation between populations is at risk of being flagged up as being associated with a disease if the populations themselves show differing frequencies of the disease.

How else can linkage disequilibrium be generated?

Other ways in which linkage disequilibrium can arise involve physical linkage, where the marker is found at a chromosomal locus that is near the genetic difference actually causing the disease. This is the kind of linkage disequilibrium that GWAS is searching for. While linkage disequilibrium is not the same as physical linkage, variants that are linked in the sense of being close together on the chromosome are much more likely to be associated than are physically unlinked variants, because a chromosomal recombination event would be required to separate them, and this does not happen very often, especially if the SNP and the disease gene are very close to one another. New mutations can remain associated with physically linked variants for hundreds of generations.

So, if there is a strong association in a well matched sample between a SNP and the disease, the best guess is that there is a causative allele tightly linked physically to (and in linkage disequilibrium with) the SNP. The effect of the causative locus on the probability of getting the disease can be approximately estimated though the odds ratio associated with the SNP.

What is an odds ratio?

The odds ratio is simply the probability of having the disease given one genotype at a SNP (or other genetic variant) divided by the probability of having the disease given another genotype at the SNP. In a statistical sense, it is a measure of the effect size, rather than a significance value. So, as sample sizes go up, the odds ratios should become more accurate, and the evidence that odds ratios differ from one (with an odds ratio of one implying no genetic effect) should become more statistically significant. As very large numbers of cases and controls are now included in genome-wide association studies, more and more loci are found to show associations, and loci with low effect sizes (under 1.5) start to be detected with statistical confidence. Thus, for example, more than 30 loci have been identified as being associated with risk of Crohn's disease.

What is meant by the 'missing heritability' people seem to be talking about?

This is the mystery at the heart of results from genome-wide association studies. Each of the SNP loci showing a disease association has a frequency in the population and a genetic effect (measured by the odds ratio). From the frequency of the marker and the effect size it is possible to calculate the contribution that this locus would make to the total genetic determination (the heritability) of the disease. One can then sum the effects of all the loci discovered, to assess their combined genetic influence. But, almost always, this genetic influence is much less than the influence measured by the heritability. The 32 loci shown to affect Crohn's disease risk collectively explain only 20% of the heritability for the disease, for example. There must be some genetic explanation of the missing 80% of the heritability that is not being detected by the GWAS approach.

Where might the missing heritability be?

The GWAS methodology is designed to detect the effects of causative genetic loci where the rarer allele still has a reasonable frequency in the population (greater than 5%). If there are genetic loci influencing the trait where the rare allele has a frequency under 5%, or even under 1%, the GWAS technique is unlikely to be able to detect these loci. One idea about the cause of the missing heritability is that this is supplied by mutant alleles at very many loci, the majority of which are very rare. In a sense, we are back in the world of single-gene disorders, at least to the extent that the individual predisposing loci have rare variants, created by fairly recent mutations and on the way to elimination by selection. It should be said that another possibility is that there are many other loci with common causative alleles, but alleles with low odds ratios (that is, small effects), which will only be detected in even larger samples of cases and controls, and it is these that supply the missing heritability. There will also be an underestimation of the genetic effect of the known loci since they are represented by their surrogate linked SNPs, and the true effects of the causative alleles themselves could be greater.

So is this what is meant by genetic heterogeneity?

Yes, exactly. In general terms, genetic heterogeneity in disease causation means that the disease may be caused by different genes in different individuals. In the case discussed above, if there are very many loci that have rare alleles that are causing the disease, there will be very great differences between the genotypes of affected individuals, and it will be hard to detect the individual causative loci.

But if the variants causing the disease are rare, why are the diseases common?

As I say above, fitness-lowering mutations at a locus in mutation-selection balance should not be common, because selection is quantitatively stronger than mutation. However, disease mutations are commoner in loci such as the dystrophin locus, which is very large and has a correspondingly high mutation rate, which explains the comparatively high incidence of Duchenne muscular dystrophy. There is only one gene that can mutate to alleles that cause Duchenne muscular dystrophy, but it may be that, for common multifactorial disorders, there are very many loci that can mutate to alleles that contribute to producing the disease symptoms. So the total mutation rate for some conditions, such as schizophrenia, may be high because so many loci can mutate to predisposing alleles. In effect, it is a question of target size.

What are the achievements of GWAS so far?

There are cases of important causative variants being identified by GWAS. The GWAS approach is hypothesis-free, in that it looks at very many SNPs simultaneously rather than focusing on loci whose biology suggests that a causal relationship to the disease is likely. The result of this is that, since each SNP tested constitutes a separate hypothesis test, very significant associations are needed in order to rule out false positives. Thus, sample sizes have to be large in order to find variants with low odds ratios. However, in a study of age-related macular degeneration, a sample of only 96 cases and 50 controls identified an important causative variant in the complement factor H gene. Two of the three most significant associations came from SNPs in an intron of this locus and they were themselves significantly associated with a tyrosine-histidine substitution encoded in exon 9 of the gene, which was inferred to be the causative SNP. The finding of the causative SNP in such small

samples was due to the intronic SNPs initially identified having high odds ratios - 7.4 in one case when homozygotes for a C base at the SNP were compared with other genotypes. The odds ratios were indeed so high that these intronic SNPs explained around half the total population risk. This case is unusual, in that a SNP with a massive effect on the odds ratio nevertheless showed a high population frequency for the disease-predisposing base. It could be that part of the cause is that the age of onset of the disease is one that would very rarely be attained by our ancestors, and the selection on the condition was probably minimal at the time when allelic frequencies were being determined.

Where can I find out more?

Articles:

Barrett JC, Hansoul S, Nicolae DL, Cho JH, Duerr RH, Rioux JD, Brant SR, Silverberg MS, Taylor KD, Barmada MM, Bitton A, Dassopoulos T, Datta LW, Green T, Griffiths AM, Kistner EO, Murtha MT, Regueiro MD, Rotter JI, Schumm LP, Steinhart AH, Targan SR, Xavier RJ; NIDDK IBD Genetics Consortium, Libioulle C, Sandor C, Lathrop M, Belaiche J, Dewit O, Gut I, *et al.*: **Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease.** *Nat Genet* 2008, **40:**955-962.

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, Manolio TA: **Potential etiologic and functional implications of genome-wide association loci for human diseases and traits.** *Proc Natl Acad Sci U S A* 2009, **106:**9362-9367.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, Bracken MB, Ferris FL, Ott J, Barnstable C, Hoh J: **Complement factor H polymorphism in age-related macular degeneration.** *Science* 2005, **308:**385-389.

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, Visscher PM: **Finding the missing heritability of complex diseases.** *Nature* 2009, **461:**747-753.

Published: 12 April 2010

doi:10.1186/1741-7007-8-41
Cite this article as: Brookfield JFY: **Q&A: Promise and pitfalls of genome-wide association studies.** *BMC Biology* 2010, **8:**41.