The search for association

The list of human genetic variations is expanding; but an understanding of how they contribute to disease is still patchy.

BY MONYA BAKER

Everyone carries dangerous genetic mutations. But only in the past five years or so have researchers begun to use genome-wide association studies (GWAS) to scour human genetic samples for the signals of individual variations. Typically, such studies assess hundreds of thousands of genetic variants in thousands of individuals sorted by traits: a certain height, perhaps, or asthma or obesity. Genetic variants that occur more frequently in one group than in another are subjected to stringent statistical analyses to determine whether associations between them and the traits are the result of biology or mere chance.

As of 1 October, an online catalogue of GWAS contained nearly 700 publications linking some 3,000 variants to about 150 traits. The list of traits begins with abdominal aortic aneurysm and ends with YKL-40, a protein used as a biomarker for cancer. Other GWAS have identified correlations between genetic variants and smoking behaviour, sleep duration and general self-reported health. The catalogue is growing swiftly: 9 out of 16 research articles in the October issue of *Nature Genetics* report GWAS.

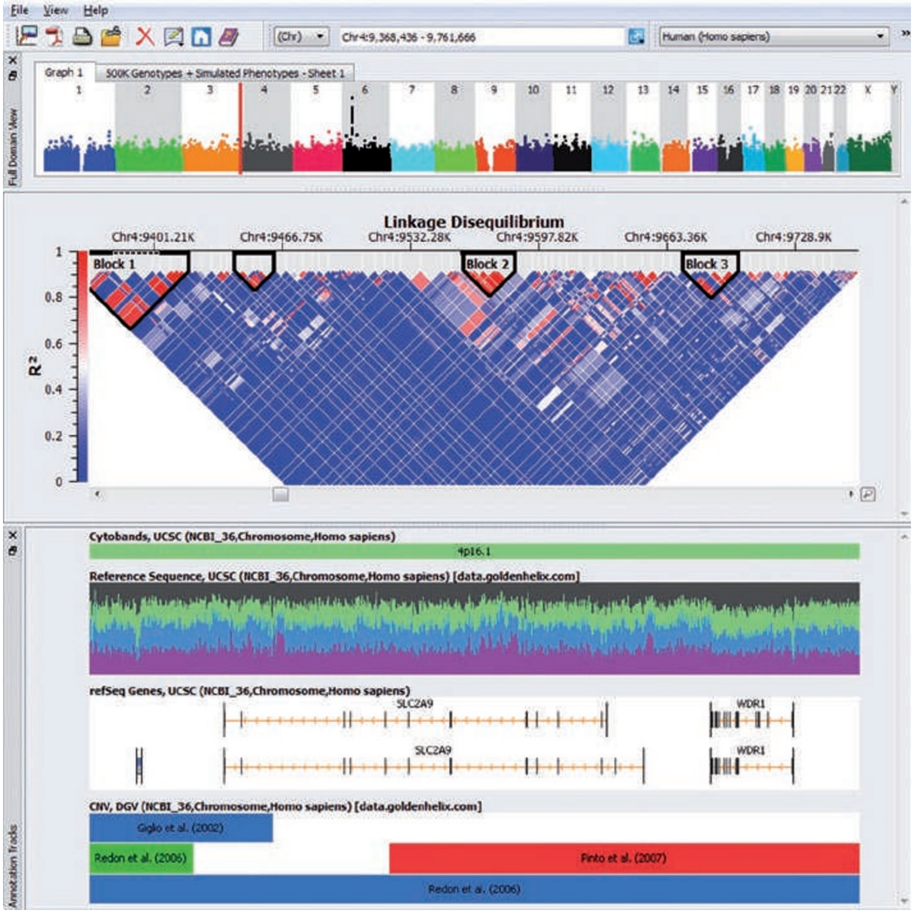
In their current incarnation, GWAS are running into a problem of diminishing returns. By collecting ever-larger samples, researchers are able to find more and more variant–trait associations, but these tend to have smaller and smaller effects. In fact, small effect sizes have been a hallmark of GWAS ever since the studies began. Originally, researchers hoped to find associations in which people carrying one variant would be several times more likely to have a trait than those carrying another. Instead, the effects found have been much more modest. An analysis published in June 2010 (ref. 1) pooled findings from several published GWAS that had each associated given traits with single nucleotide polymorphisms (SNPs) — the simplest and most common type of genetic variant, in which one DNA letter is changed to another. Extrapolating from previous findings, the researchers calculated that 201 SNPs associated with height could explain about 16% of genetic variance, 142 SNPs associated with Crohn's disease could explain about 20%, and 67 SNPs could explain about 17% of genetic variance in each of three common cancers.

Although genetic variants with small effects can still help to uncover fundamental biology with therapeutic implications, researchers hunting for those with larger effects are pinning their hopes on several advances: an onslaught of newly discovered simple polymorphisms, the ability to assess more complicated variants (see 'The tough new variants') and multiplying applications of sequencing. If the human genome were an archaeological site, these options would be equivalent to canvassing continents with metal detectors of varying convenience and reliability, or picking a handful of sites for a full excavation.

RARER SNPS, BIGGER EFFECTS?

GWAS are only as good as the SNPs they sample. Rather than directly finding mutations responsible for an effect, the standard technique identifies SNPs that tend to co-occur with it. And the SNPs that have already been profiled in GWAS may be the ones least likely to be linked with large effects. Because the most common variants were the first to be catalogued, they were also the first that vendors put on genotyping microarrays. To make sure that SNP microarrays could identify variants in the greatest possible number of samples, vendors chose variants that occurred across several geographic populations. These tended to be the SNPs that evolved first, so natural selection has had time to weed out harmful mutations that might have occurred nearby in the genome.

But multinational projects are now discovering and characterizing younger, rarer genetic variants. The 1000 Genomes Project aims to sequence 2,500 individuals, who represent an equal distribution from the continental regions of Africa, the Americas, Europe, and east and south Asia. The goal is to identify most of the variants that exist at frequencies of 1% or more in each of the populations studied, says David Altshuler, the project's co-leader and a human geneticist at the Broad Institute in Cambridge, Massachusetts. Similarly, the International HapMap Project has identified millions of SNPs and characterized their occurrence across populations. Sequencing data from ten geographic populations indicated that more than half of the



Genome-wide association studies require several forms of statistical analyses.

GOLDEN HELIX

identified genetic variants occur at frequencies of less than 5 per cent. More than one-third of newly discovered SNPs with frequencies of less than 0.5% were observed in only one population. Such discoveries mean that many more variants can be added to microarrays for assay, and so tested in GWAS, says David Bentley, chief scientific officer at Illumina, a genetics company in San Diego, California. "There is a new generation of GWAS that are fundamentally different from previous studies, because they capture a new fraction of variations that have previously been uncharted," he says.

L. RUBENSTEIN/BROAD INST.

David Altshuler: no one approach can explain heritability.

Illumina and other commercial vendors have been modifying their microarrays in response to releases of data. Illumina unveiled its HumanOmni2.5-Quad DNA Analysis BeadChip in June this year — letting researchers assay 2.5 million SNPs and other variants — and plans to launch the Omni5 next year, for 5 million SNPs. Using the Omni5, researchers will be able to combine one set of comprehensive SNPs with specialized sets tuned to emerging sequencing data. Illumina's competitor Affymetrix, in Santa Clara, California, has in its catalogue products geared towards Chinese, Japanese, European and African ethnicities. A new microarray design allows researchers to design custom arrays containing 50,000 up to a planned 5 million SNPs using a database stocked with proprietary and public SNP data.

Nonetheless, it is not clear how effective adding to the available SNPs from healthy populations is going to be in finding SNPs associated with disease, says Christophe Lambert, chief executive of Golden Helix, a genetic-analysis company in Bozeman, Montana. This year, his company worked on an association study for Alzheimer's disease that failed to detect a signal from a variant known to boost risk for the condition. The variant, in the gene *APOE*, wasn't included on the commercial assay used in the test. Although a custom-designed array found the variant's association with the disease to be extremely significant ($P < 10^{-60}$), the standard array did not pick up its signal. "None of the SNPs on the standard chip was correlated strongly enough with the risk variant to detect it," says Lambert. Even when Lambert's team used data from the 1000 Genomes Project to 'impute' the presence of one SNP by detecting another, the analysis did not pick up on the association. Sampling more individuals or using denser microarrays might have helped, but identifying variants in diseased individuals would produce the most-informative SNPs for genotyping across populations, says Lambert.

Still, the ability to look more deeply within populations has intriguing possibilities. In a study published this September[3], researchers at deCODE Genetics in Reykjavik found that the same SNP was associated with glaucoma risk in Chinese and Icelandic populations, but in the former it was much rarer and indicated a much higher risk. And if different susceptibility variants show up near the same gene in different populations, researchers will have independently implicated that genomic area in the disease.

DUKE UNIV.

David Goldstein: you have to choose what to pursue.

Working across populations and with rarer variants can get complicated, says Augustine Kong, head of statistics at deCODE. SNPs specific to a particular population could be difficult to replicate, and the lower the frequency of an allele, the larger the number of samples needed to detect an association. However, if rarer SNPs have stronger effects, larger sample sizes might not be necessary. Researchers are keen to find out whether a substantial number of the new variants discovered by genome-mapping projects will be associated with large effects. "Before, we just didn't have the technology to interrogate these low-frequency variants comprehensively," he says. "It gives you chances that you didn't have before to make discoveries."

SEQUENCING STRAIGHT TO CAUSAL VARIANTS

Some experts think that it is time to skip array-based GWAS that find SNPs associated with causative variants, and to hunt for contributing variants directly. Mary-Claire King is a geneticist at the University of Washington in Seattle, whose work in family studies identified the breast-cancer genes *BRCA1* and *BRCA2*. She says that even the rarer variants discovered by the 1000 Genomes Project are unlikely to be highly associated with disease. New variants

The tough new variants

When single nucleotide polymorphism (SNP) studies failed to explain much of the heritability of diseases, researchers began pinning their hopes on a trickier source of variability: copy number variation (CNV). Whereas SNPs — changes of one DNA letter into another — are relatively easy for microarrays to detect and for databases to compile and sort, CNVs are a headache to identify and classify. Certain stretches of DNA are duplicated, inverted or repeated in some individuals and missing from others. "It's more complicated and the data will always be a little more dirty," says Stephen Scherer, director of the Centre for Applied Genomics at the Hospital for Sick Children in Toronto, Canada. In some cases, researchers can detect CNVs using microarrays designed for detecting SNPs. Others use products designed to identify CNVs directly, from companies such as Agilent Technologies in Santa Clara, California, and Roche Nimblegen in Madison, Wisconsin. One Agilent array, designed with the Wellcome Trust Case Control Consortium, detects about 11,000 common CNVs.

Measuring whether a nucleotide at a particular spot is A or G is easier than detecting how many times a certain sequence occurs. That concerns Peter Donnelly, director of the Wellcome Trust Centre for Human Genetics in Oxford, UK. "Because there was a long history of GWA studies that didn't replicate, the field insists on strong criteria for declaring an association," he says. "Yet when it moves to CNVs, which are harder to measure, the standards the field requires are weaker."

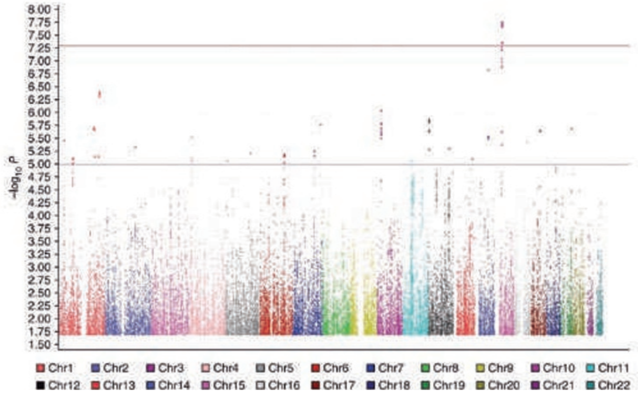
The jury is out on how much CNVs matter for common diseases. A study this year[8] profiled 3,423 CNVs, or perhaps half of all those larger than 500 base pairs. It found that most not only don't explain much disease, but are also so closely associated with common SNPs that they've already been explored, albeit indirectly.

Scherer is not so sure. He was part of a team that resequenced a human genome and compared it to a reference. It found that the genome differed from the reference in only 0.1% of SNPs, but in 1.2% of CNVs. The analysis indicated that up to one-quarter of CNVs are not associated with SNPs, and so are likely to be missed by SNP studies[9].

As with SNPs, larger effects may be found in rarer and harder-to-measure variants. Scherer has done studies showing that people with autism-spectrum disorders carry more rare CNVs than do controls. To be certain that the CNVs were correctly typed, he and his colleagues ran subsets of samples through calling algorithms that convert an instrument's signals into a sequence of base pairs, and used two platforms (by Illumina, of San Diego, California, and Agilent) to identify them[10].

Scherer says that many research groups are still learning about CNVs and don't fully realize the need to validate their data. "People are looking for low-hanging fruit; they see what they want to see and publish it," he says. The situation is improving, with the maturation of databases that collect diverse data on variation. "Now that we have much better data sets to compare to, it's becoming more accurate." M.B.

are literally born every generation, she says, so a frequency rate of even 0.5% means that a variant has persisted for a while. “The question, is how common can an allele become if there is selection against it and none for it? Not very,” says King. She advocates using sequencing within large families to find and track alleles that are inherited along with disease. Results from the 1000 Genomes Project will be useful, she says, not for finding SNPs to pursue but for filtering out variants that are not truly rare.



Plot showing an analysis of genetic variants’ association with myopia. A locus on chromosome 15 has a *P* value of 10^{-14}, a very robust association.

A. M. SOLOUKI *ET AL. NATURE GENET.* **42,** 897–901 (2010).

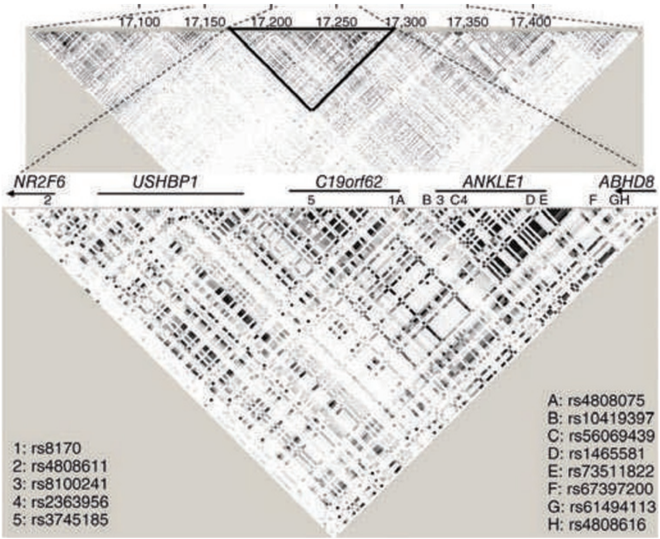
For David Goldstein, director of the Center for Genome Variation at Duke University in Durham, North Carolina, the main limitation for SNP-based GWAS is that they usually don’t allow identification of the precise causal variant that influences the trait, but instead implicate a genomic region within which the causal variants must reside. The priority for research now, he says, is to focus on identifying the precise variants that contribute to disease; doing so will provide much more information about the relevant biological processes.

Peter Donnelly: copy number variants are hard to assess.

This year[4], researchers reported the first use of massively parallel sequencing to identify the gene responsible for a Mendelian disease, one which is caused by mutations in a single gene. Researchers at the University of Washington took samples from just four individuals with the developmental disorder Miller syndrome, including two siblings, and sequenced all the coding regions of their genomes using a technique called whole-exome sequencing, which relies on genome-capture products from companies such as Agilent Technologies in Santa Clara, California, and Roche Nimblegen in Madison, Wisconsin, followed by next-generation sequencing. Although still labour-intensive, the method requires only about 5% as much sequencing as a whole genome. By filtering identified variants against publicly available SNPs and other human exome data, the researchers found that all four subjects carried previously unidentified mutations in a single gene involved in synthesizing nucleotides; follow-up studies in three further people with Miller syndrome revealed mutations in the same gene. Since then, those and other researchers have published a slew of papers tying a variety of other inherited conditions to variation in individual genes.

The signals for diseases caused by multiple genes will not be so clear, but bioinformatics techniques borrowed from SNP-based GWAS are being applied to sequencing data. To enable this, many types of variants must be placed into distinct categories so that they can be subjected to statistical analyses. Collapsing or binning methods count how many of a particular kind of variant are found in a gene or a predetermined stretch of base pairs, and then compare frequencies in individuals with the disease to those without it. But researchers are still learning how to weed out artefacts from the sequencing data, says Goldstein. With sequencing, “you have to pick which variants to pursue, and that’s prone to statistical abuse”, he says. “People could say the association is weak, but it makes sense.” (See ‘Seeing more SNPs’.)

One problem is that it is unclear how to classify different kinds of mutations. It might make sense, for example, to lump together different mutations in the same gene that stop translation early. But what about apparently silent mutations, or others whose effects on protein products seem minor? Another issue is working out the parameters for determining effective controls; because sequencing studies have smaller sets of controls, researchers need to be more rigorous to make sure that, for example, control sets don’t include individuals with late-onset disease. But those problems won’t send researchers back to SNP-based microarrays, says Goldstein. “If you were launching a new study now on a common disease, you’d turn to sequencing-based studies.”



Studies must consider how variants are inherited together. Variants not directly measured (letters) can be imputed from those that are (numbers).

A. C. ANTONIOU *ET AL. NATURE GENET.* **42,** 885–892 (2010).

THE COMMON TOUCH

First-generation GWAS looked at common genetic variants, but many initial-sequencing studies concentrate on finding individuals with extreme forms of disease, because cost limits them to small sample sizes. Prices are changing quickly and vary by sample and technology, but at the moment, genotyping a sample for millions of SNPs costs about US$400, whereas sequencing an entire genome costs around $10,000. Finding the genetic basis for an extreme form of disease can shed light on more common forms: in the 1980s, well before sequencing, family-based studies of severe cholesteraemia led to the discovery of the low-density lipoprotein receptor gene, common variants of which are now associated with high cholesterol levels[5]. But Altshuler thinks that studies linking rare variants with strong effects need to be followed up to understand how the relevant genes contribute to common forms of the disease. Trying to study common diseases as if they were single-gene disorders will be of limited use, he says. “Mendelian forms of type 2 diabetes explain less than 1% of heritability, but GWAS explain about 10%,” he says. To fully understand both the inheritance and mechanisms of common diseases, he says, it will be necessary to study the diseases as they occur in the general population.

Another problem with sequencing is that it is slow — one reason why Illumina’s Bentley, whose company does both SNP genotyping and sequencing, says that he doesn’t expect to see a decline in demand for microarrays any time soon. “With our best efforts at the moment we are sequencing one genome every three days; we can genotype more than 50 to 100 samples every three days,” he says.

Even if GWAS continue to find only small effect sizes, they can still have a large impact, says Peter Donnelly, director of the Wellcome Trust Centre for Human Genetics in Oxford, UK. “There is real value in working out the genetic architecture of a disease, regardless of

Seeing more SNPs

As genome-wide association studies (GWAS) get larger, the technical challenges pile up, and an onslaught of dense microarrays is compounding the issue by encouraging researchers to combine data sets. Genotyping a few-dozen single nucleotide polymorphisms (SNPs) in a sample is not much cheaper than genotyping hundreds of thousands, says Peter Donnelly, director of the Wellcome Trust Centre for Human Genetics in Oxford, UK. So rather than designing a targeted follow-up study on a handful of SNPs, researchers are more likely to try to replicate an association through meta-analysis, using samples that have been fully genotyped elsewhere. “That needs care,” says Donnelly. Even in straightforward GWAS, everything that looks like a signal is probably an artefact, he says. Combining results typed on one platform in one lab and on another in a different lab creates more opportunities for artefacts.

Even when cases and controls are processed by the same group, all the cases can be on one set of microarray plates and all the controls on another. This introduces potential for systemic error that sometimes leads to up to 30% of the data being discarded, says Christophe Lambert, chief executive of Golden Helix in Bozeman, Montana, which provides software and analytical services for genetic research. “Everyone is running these experiments and asking the statisticians to fix the problems, when a simple block randomization at the beginning could have fixed it.”

Some problems occur before the sample is collected, says James Clough of Oxford Gene Technology, a genotyping-services firm. “Samples will be collected in multiple centres and multiple countries.” That can pose challenges when clinical standards vary. The best studies put more effort into collecting phenotypes than collecting samples, he says.

Careful characterization of phenotype could make genetic signals more apparent, says Greg Gibson, director of the Center for Integrative Genomics at the Georgia Institute of Technology in Atlanta. Many aspects of phenotype are extremely variable, so longitudinal measurements of factors such as blood-lipid levels, body-mass index or toxin exposure could control for transient effects and effectively boost genetic signals. GWAS could be more successful at implicating genes if they concentrate on qualities more closely tied to genetics, such as lipid levels or endophenotypes, he says. “Just mapping genotype to disease is several steps away from gene expression.” M.B.

what it turns out to be. For example, even if all the genetic components of a disease were based in very many common variants with small effects, it would be good to know that.” And even if the effects of variants of a gene in a general population are small, those of modulating that gene with a drug can be large. For instance, variations in the gene encoding 3-hydroxy-3-methylglutaryl coenzyme A reductase have been connected in GWAS with small effects[6] on cholesterol levels, but the statin drugs that modulate that gene product are effective and very widely prescribed. Although statins were not inspired by GWAS, such studies have turned up surprising connections with therapeutic implications, such as the role of the immune system in age-related macular degeneration, or of cell-cycle regulators in type 2 diabetes. In fact, says Altshuler, such results could be useful for focusing sequencing studies. “The genome-wide association paradigm might be that you find the gene using GWAS, and then sequence to find the rarer variants.”

One of the biggest GWAS so far assessed samples from more than 100,000 individuals for more than 2 million SNPs, and identified 95 loci associated with variation in cholesterol and triglyceride levels in blood, 59 of which had never been reported before, and many of which were not near genes known to be associated with lipid metabolism[5]. Follow-up experiments in mice not only showed that some newly implicated genes had direct effects on plasmid lipid levels, but also identified a new cell-signalling pathway that could be targeted for therapeutic intervention. Another study[7] examined four genes that had been implicated by GWAS as contributing to high blood-triglyceride levels. Common variants explained less than 10% of observed variation, so researchers sequenced the genes to identify rare missense and nonsense variants — two categories of mutations likely to change protein function. Nearly twice as many of these were found in affected individuals than in controls.

DIFFERENT STRATEGIES

The debate over the best approach for finding causal variants, says Altshuler, reflects researchers’ various options for studying disease, and their limited funds. The decision whether to sequence a handful of samples or genotype thousands depends on whether researchers believe that a disease will be explained by a few rare variants or many common ones.

The answer will vary by disease. Current GWAS, for example, explain more heritability for autoimmune disorders and late-onset diseases such as Alzheimer’s and heart disease than for mental conditions such as schizophrenia and autism. Natural selection suggests ready explanations, although they are hard to prove. Almost by definition, late-onset diseases tend to affect individuals in their post-reproductive years, and so are less likely to be selected against. And some genetic variants that contribute to one disease might actually be protective against others, and so could be favoured by natural selection. Genetic variants for sickle-cell anaemia, for example, can help to prevent carriers from contracting malaria, and there are hints that genes causing predisposition to some autoimmune diseases also confer resistance to infection.

In an effort to gather concrete evidence on which technologies are best suited to explaining the inheritance of common diseases, Altshuler has begun a study, with Mike Boehnke at the University of Michigan and Mark McCarthy at the University of Oxford, to compare the same population using several techniques. In this case, the study will compare what Altshuler calls “extremes of risk”: subjects who are at high risk for diabetes because of their age and weight but do not have the disease will be compared with slimmer, younger subjects who have been diagnosed with it. Presumably, individuals in the first group will carry relatively more protective variants, whereas those in the latter will have more susceptibility variants. About 2,600 people will be genotyped for 5 million SNPs, and be submitted to whole-exome and whole-genome sequencing.

Altshuler says that the study should not only uncover important information about diabetes, but also offer empirical data to help researchers choose the most appropriate technology, or combination of technologies. “We want to know what each approach finds that the others don’t,” he says. “Right now, no one actually knows which one is going to apply to which disease. Investigators have to take different bets.” ■

Monya Baker *is technology editor for* Nature *and* Nature Methods.

1. Park, J. H. *et al. Nature Genet.* **42,** 570–575 (2010).
2. International HapMap 3 Consortium *Nature* **467,** 52–58 (2010).
3. Thorleifsson, G. *et al. Nature Genet.* **42,** 906–909 (2010).
4. Ng, S. B. *et al. Nature Genet.* **42,** 30–35 (2010).
5. Teslovich, T. M. *et al. Nature* **466,** 707–713 (2010).
6. Burkhardt, R. *et al. Arterioscler. Thromb. Vasc. Biol.* **28,** 2078–2084 (2008).
7. Johansen, C. T. *et al. Nature Genet.* **42,** 684–687 (2010).
8. Wellcome Trust Case Control Consortium *Nature* **464,** 713–720 (2010).
9. Pang, A. W. *et al. Genome Biol.* **11,** R52 (2010).
10. Pinto, D. *et al. Nature* **466,** 368–372 (2010).