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).