Genetics by numbers

Genomewide association studies are starting to turn up increasingly reliable disease markers. Monya Baker investigates where we are now and what comes next.

Who would have thought that the future of human health would read like a list of car number plates? Last year, a suite of studies1–3 pinned an increased likelihood of developing heart disease on some mysterious culprits: seemingly incomprehensible strings of numbers such as rs10757274 and rs1333040. The number sequences, technically known as single nucleotide polymorphisms or SNPs, are located close to one another on chromosome 9. No one knows what they do, if anything. But carrying two copies of any of them boosts a person’s chance of developing heart disease by the same amount as smoking ten cigarettes a day. The effect is less than that brought on by diabetes or heavier smoking, but it is still one of the strongest risks so far identified by genomewide association studies, which attempt to find genetic variants that occur more frequently in one group of people than another.

The association is so robust that it is already being used as a positive control for further genomewide studies of cardiovascular disease. “If you don’t find it, you know something is wrong” with the analysis, says Ruth McPherson of the University of Ottawa Heart Institute in Canada and lead author on one of the two articles that first announced the association1 in May 2007. Researchers from deCODE Genetics in Reykjavik, Iceland, published the other2. Having one copy of any of the wrong SNPs boosts risk by about 40%. Having two copies of the SNP doubles one’s chances of having a heart attack early in life.

But despite the strength of the association, pegging meaning to it is difficult. Genomewide association studies find SNPs, single letter changes in the genome, that appear commonly and may correlate to more variation in nearby DNA. SNPs are not necessarily the causative mutations. They might not even be in gene-coding or gene-controlling regions, and thus may not contribute, even indirectly, to risk. The heart-risk SNPs indicate only that the culprit is located within a long genomic region on chromosome 9 known as 9p21.3. No protein-coding genes are apparent within the region, and investigations of two nearby genes — both tumour-suppressors — haven’t yielded an explanation.

Making sense of the numbers

Still, enthusiasts of genomewide investigations can’t wait to find more mysteries like rs10757274 and rs1333040. Scanning hundreds of thousands of SNPs for a disease or condition can turn up dozens of associations. Early studies gained notoriety for their unreplicable results, turning up many false positives. But that is changing with increasing statistical reproducibility with more diseases.

The associations are so robust, that if you don’t find them, you know something is wrong.” — Ruth McPherson

(see ‘SNP spotting’) leading some to call 2007 the year of genomewide association studies, as research groups pumped out dense lists of associated SNPs and gene regions like so many car number plates — meaningless without further information. Ask researchers what is next for 2008 and 2009, and they will gush about longer lists. As researchers scan larger populations for more SNPs, more SNPs will be associated more reproducibly with more diseases.

But ask when these obfuscated strings of digits and letters will be used to help find a drug, explain a disease, or recommend tailored treatments for patients, and you’ll get a sober response that years of work lie ahead. Researchers readily admit that SNP-scanning studies cannot find important contributors to disease, such as environmental factors or extra copies of genes. Moreover, identified variants, or alleles, contribute to only a small part of the overall risk. Still, by scanning the entire genome, association studies can uncover unsuspected connections between genes and disease. “Finding the initial SNP is not the same as finding the underlying biology,” says human geneticist David Altshuler of the Broad Institute in Cambridge, Massachusetts, “but it’s a big step forward.”

Companies are already selling or planning to sell tests that scan for genetic variants associated with disease. Last November, deCODE began selling a test to doctors that will reveal how many copies of the risk allele at 9p21.3 an individual carries. Those kinds of tests worry Muin Khoury, director of the National Office of Public Health Genomics at the Centers for Disease Control and Prevention in Atlanta, Georgia, who has written cautionary commentaries on the issue. “Even if the association is replicated in many, many studies,” he
saying, “it is still weak information.” Right now, family history is more predictive. There is no evidence that results of genetic tests encourage people to adopt healthier lifestyles, Khoury says, and such tests could even yield false hopes, as someone who lacks identified risk alleles could be carrying risk variants that are as yet unidentified. Evidence that they make a difference may be a long time coming. “The private sector doesn’t want to do the research and right now no one is telling them they have to,” Khoury says.

Cooperation is key

Nevertheless research continues to seek out new associations in the hope of developing a richer understanding of disease. Ultimately, common variants of even very important genes will have small effects; finding them depends on having enough data to sort through. Genomewide association studies are expensive. Although prices have plummeted, genotyping a few hundred thousand SNPs still costs a few hundred dollars per individual, not counting the complex and expensive tasks of collecting and managing patient data. “Sample size has been our big enemy here, and sharing data is the best way possible to address this issue,” says Lon Cardon, a statistical geneticist at the University of Washington in Seattle.

To boost sample sizes, the Wellcome Trust (WTCCC) and assembled 2,000 patient samples for each of several diseases along with a generic set of 3,000 controls, some 19,000 subjects in all. The rationale is that by comparing SNPs between diagnosed and undiagnosed individuals, genetic risk associations will be apparent.

Missing the point

Nilesh Samani, chair of cardiology at the University of Leicester, UK, and one of two lead investigators responsible for coronary heart disease with the WTCCC, explains that even studies with many samples will miss variants with modest effects. Suppose, he says, that there are 10 loci in a genome that each increase the likelihood of a condition by 20%. Statistically, an examination of 2,000 cases and 2,000 controls would pick up at most three of these loci. An independent group with similar sample sizes might also find two or three loci, but they might be different loci, and the plague of false positives would make results inconclusive. “It’s only when we pool all of these studies together that we have a realistic chance of picking up all of those loci,” Samani says.

The data provided by the WTCCC are largely limited to genotype, age, gender and presence of disease. They are proving invaluable for certain studies, but finding how variants affect specific traits such as cholesterol levels or blood pressure requires richer data. A particularly rich vein of it came from the Framingham SNP Health Association Resource (SHARe), which genotyped 550,000 SNPs in more than 9,000 individuals participating in the Framingham Heart Study.

Sponsored by the National Heart, Lung and Blood Institute in Bethesda, Maryland, the 60-year-old Framingham study has followed three generations of individuals in Massachusetts. Data from it established smoking and high cholesterol as risks for heart disease. Although not all phenotypic data exist for every individual, the data set contains thousands of clinical variables from blood analyses to vascular imagery to lifestyle surveys, sometimes from the same individual over a span of years. Assembling the database required Framingham to undertake a sort of archaeology expedition of old medical records and publications, but, as the acronym implies, the data are available to other researchers.

As of 25 January 2008, all genomewide association studies funded through the US National Institutes of Health (NIH) are required to deposit their data in the Database of Genotype and Phenotype (dbGaP). Although researchers who collect the data have exclusive rights to publish their analyses for at least nine months, scientists who promise to uphold privacy safeguards will be able to access others’ data sets immediately. SHARe has a similar moratorium.

Such practices have become standard in genomics studies. Nevertheless, some worry that if scientists can publish analyses of downloaded data, they might have fewer incentives to collect the data necessary for answering new, interesting questions. They may also not properly...
handle the data they do get. Bruce Psaty, a cardiologist and epidemiologist at the University of Washington in Seattle, examined how researchers used data available from two large studies. He found that not only did some researchers go beyond the scope of original agreements, they failed to account for the study’s design in their analyses. For example, some analyses searching for predictors of a patient’s first heart attack did not exclude patients who had already had heart attacks. Poor analyses of large epidemiological studies could mean real associations are missed or false ones found. Resources are wasted either way.

Share and share alike

If scientists are careful and peer reviewers are rigorous, giving more researchers access to more data should mean better science at lower prices, says Altshuler. “No single group can make full use of such large data sets. The creativity of the whole world has to be greater than the creativity of the group that collected the data.” Kári Stefánsson, chief executive of deCODE Genetics, says that researchers are already doing a good job of finding collaborators but he resents what he calls the “Soviet flavour” of the NIH mandate. “I don’t want to share my data with anyone because the NIH decides I should,” he says. “I want to do it because I decide to do it.”

Teri Manolio, director of the Office of Population Genomics of the National Human Genome Research Institute in Bethesda, Maryland, acknowledges such concerns but says that science will adjust. The WTCCC and the Genetic Association Information Network (a public–private US collaboration with goals similar to those of the WTCCC) both grew out of the Human Genome Project. As that “big-science” project got under way, she recalls, there were “big concerns” that people’s careers would end when they put their data on the web. That didn’t happen. “It quickly became apparent that just putting a sequence up wasn’t a publication, and that that kind of data-sharing didn’t actually hurt people,” says Manolio. In fact, she adds, some initial contributors to the dbGaP have gained additional collaborators and recognition by sharing data, and several academic and corporate investigators plan to contribute data even though they are not required to.

With the need for large data sets and the danger of false-positives now both widely recognized, Cardon hopes that scientists can be relied on to do what is necessary to conduct interesting, convincing analyses. “I think it is really useful to put data up on dbGaP, but it is no substitution for collaborating with the people who collected them,” he says. “It’s certainly more powerful than just downloading and trying to go it alone.”

Strength in numbers

When multiple data streams come together, results are tangible. After surveying hundreds of thousands of SNPs from 2,758 individuals for associations with blood lipid levels, researchers working with data from the Diabetes Genetics Initiative had settled on 196 SNPs for further analysis. Then they were contacted by another group with data from two more studies that had measured SNPs and blood lipid levels in 6,058 people.

Sharing data helped both groups identify more SNPs for further analysis, resulting in publications from both groups1–7. One SNP identified in that collaboration, rs599839, on the short arm of chromosome 1, was found to be associated with lower levels of low-density lipoprotein cholesterol, and further study showed increased expression of three genes in the liver, including one involved in glucose uptake. Interestingly, that SNP had also been associated with heart disease just months earlier, in a study8 that confirmed the risk on 9p21.3.

Work to understand SNPs’ influence will mean chasing down more informative phenotypes and, eventually, lab work to confirm function. Recently, deCODE found that the SNPs on the 9p21.3 region associated with heart disease are also associated with brain and abdominal aneurysms9. McPherson’s group found the locus associated with arterial disease, suggesting that the region has something to do with the integrity of blood vessel walls. Nevertheless, figuring out the underlying mechanism is proving elusive for both groups. “We haven’t found anything new through thorough sequencing of the region,” says Stefánsson. McPherson is looking for RNA sequences that might affect gene expression, but the region is long, she says, and even when the responsible sequence is ferreted out, the mechanism might not be obvious.

Even if the effect of an identified variant is tiny, the importance of finding a new mechanism could be huge, and genomewide association studies can uncover these in places in the genome that no one would think to look. What many researchers are looking forward to most is surprises. “For the past 20 years, traditional risk factors were the A-to-Z of cardiovascular biology,” says cardiologist Heribert Schunkert at the University of Lübeck in Germany, who, along with Samani, confirmed the association of 9p21.3 with heart disease. Many loci being identified today have no link to known risk factors. “That’s exciting,” he says. “It tells me we know very little about the true biology.”

Monya Baker is the editor of Nature Reports Stem Cells and writes for Nature from San Francisco.