Biostatistics: Revealing analysis

Reprinted by permission from Macmillan Publishers Ltd: *Nature* **482**, 263-265 (8 February 2012) | doi:10.1038/nj7384-263a, copyright 2012

Erika Check Hayden

David Alexander's job didn't exist ten years ago. He works for Pacific Biosciences in Menlo Park, California, writing software that can analyse the data generated by DNA polymerase enzymes, which sequence DNA in real time. A decade ago, it took scientists weeks to sequence DNA, one base at a time, using a seemingly endless series of reactions. Back then, they also thought that they would be able to find the roots of major diseases just by identifying the common genetic variants shared by affected individuals.

Both the technology and the hypotheses have changed greatly since then. In the mid- to late-2000s, while Alexander was working towards his PhD, scientists were using genome-wide association studies (GWAS) — searching genomes for known genetic variants that are shared by people with a particular disease or trait. But by the time he graduated, last June, GWAS had mostly been superseded by techniques that sequence entire genomes. The machines designed to do this sequencing are pouring out huge amounts of data, thereby creating a huge need for mathematics and statistics experts. So Alexander, and many others working on statistical genetics, now have many more opportunities. “Scientifically, there are much richer questions to ask, and there are still a lot of deep discoveries to be made; it's an interesting time,” he says. His career track reveals just how much opportunities in the field have changed.

Career variation

It was not for a lack of trying that GWAS didn't pan out. The completion of the Human Genome Project in 2003 spurred major funders from around the world to invest millions of dollars to build an international haplotype map, a catalogue of all the common human variants at single bases, called single nucleotide polymorphisms (SNPs), to be used in GWAS. The SNP map should have helped researchers to identify genes that are associated with disease. But instead, it showed that SNPs don't account for much of the heritability of disease.

Researchers now think that many rare variants play a part in causing disease, but rare variants are much harder to find than the common SNPs. As a result, statistical geneticists are now mining sequence data for directly causative mutations, rather than for SNPs. And geneticists are starting to combine data from different types of studies, using a method called integrative genomics — for instance, studying combinations of SNPs, the protein-coding genes surveyed in exome studies, epigenetic factors (heritable information not found in the DNA sequence), gene-expression factors and environmental interactions. “This field has ballooned and changed to a ridiculous degree in the past ten years, because there have been multiple waves of technological revolution,” says Gilean McVean, a statistical geneticist at the University of Oxford, UK. “As genomics becomes a much more integrated part of health care, things are going to change again and new opportunities will open up, so it's a good time to be a statistical geneticist.”

Bag of tricks

Statisticians will be kept busy for years by the problems raised by analysing these huge data sets. They will need to find the best ways to grapple with studies that combine multiple methods, each of which yield millions of data points. The challenge is to find true associations within the huge volumes of data without getting duped by the errors that tend to affect data sets of this magnitude, says Lucia Hindorff, an epidemiologist at the US National Human Genome Research Institute (NHGRI) in Bethesda, Maryland. “The answers aren't straightforward,” she says. “That's one of the reasons why statisticians have a lot of work to do.” And statistical geneticists are needed at universities, at genome centres and in industry alike.

However, a survey of statistical geneticists by a working group from US National Institutes of Health in Bethesda has suggested that trainers are having difficulty recruiting enough qualified trainees into their programmes. Alexander Wilson, head of genometrics at the NHGRI, who organized the survey, says that although the number of genetic variants available to be analysed has grown significantly since the 1980s, the number of people available to analyse them has remained relatively constant. According to Suzanne Leal, a genetic epidemiologist at Baylor College of Medicine in Houston, Texas, many biologists eschew significant statistics training. And because only a handful of statistical geneticists are trained each year, “these positions are difficult to fill”, says Michael Boehnke of the University of Michigan in Ann Arbor. So, although job demand outstrips supply in many fields, the market remains promising for statistics specialists, not least because they can help funding agencies to make good on their research investments.

And unlike other fields, many academic jobs in statistical genetics require only a doctoral degree, so PhD holders don't tend to find themselves stuck on an extended treadmill of multiple postdoc positions. “You're going to have many job opportunities; it's not like with other biological sciences where you do six or seven years of postdocs,” Leal says. “You can do a two-year postdoc and then go on to a faculty position if you're any good.”

With the plummeting cost of equipment, sequencing is becoming more feasible for many labs. However, the analytical problems are becoming so complex and expensive that disease-focused centres are starting to create joint analysis positions with larger hubs of genome expertise.

“Biology is now a science in which large data sets are central, but bioinformatics and statistical genetics are getting to a point where there are many specialized roles — data handling, processing, quality control, interpreting — that cannot all be done well by one person,” says McVean. Analysts working on moving genomics technologies into health care at the University of Oxford's Biomedical Research Centre, for instance, are made honorary members of a bioinformatics and statistical genetics core at the Wellcome Trust Centre for Human Genetics in Oxford, run by McVean. They have access to the pipelines for sequencing data as well as to bioinformatics and statistical genetics expertise, but are funded separately from the centre.

Although statisticians in these positions can expect to have their own students and develop new methods, the roles are more inherently collaborative than many academic jobs, says McVean.

“It's not the traditional academic route of going off to form your own little group and working in isolation, but rather going off to support diverse groups in a centre,” he says. He is preparing to recruit for similar positions at the Ludwig Institute for Cancer Research and the Kennedy Institute of Rheumatology, both in Oxford. Both institutions, says McVean, would find it difficult to amass the personnel needed for independent, dedicated bioinformatics support.

Increased competition between new sequencing technologies — and companies hoping to make sense of the data — also means opportunities for computational and statistical experts in genetics in industry. Companies such as Pacific Biosciences, Illumina in San Diego, California, and Life Technologies in Carlsbad, California, are developing new methods for sequencing and need people who can come up with ways to analyse the new forms of data that will be produced.

Another track, which might be called clinical genomics, is relatively small, but growing. Companies in this field are developing ways to interpret individuals' genomic data for either medical or drug-discovery purposes, and are looking for individuals with a suite of talents. For instance, Omicia, based in the San Francisco Bay area of California, is developing a platform to help physicians and clinical labs to interpret genomic data. In just the past few months, it has hired three people: a Silicon Valley engineer who specializes in quick analyses of large data sets; an application engineer to help the company develop interfaces that are fast and easy for customers to use; and a medical researcher who has a bachelor's degree in genetics and hopes to attend medical school. Omicia's chief executive and co-founder, Martin Reese, says that the company is looking to hire more people in these specialities, especially analysts.

Rowan Chapman, a partner at Mohr Davidow, a venture-capital firm in Menlo Park that funds companies such as Pacific Biosciences, says that the firms are always looking for analysis experts. “There's a massive amount of data being generated, particularly by next-generation sequencing platforms, and the cost of the analysis is now greater than the cost of the data generation,” she says. “Finding the right people to analyse those data is a challenge.”

Strong background

Succeeding in statistical genetics requires a good grounding in both statistics and genetics, which can be gained through academic work as part of any doctoral programme that allows students to take classes in both disciplines. But two other skills are increasingly necessary: expertise in computer-programming languages designed to aid manipulation of large data sets, such as R, Perl or Python, and the ability to use these languages to analyse large amounts of data quickly. Expertise in distributed computing and writing code for various operating systems is particularly desirable.

Most researchers say that these skills can be gained through hands-on experience working with large data sets, or during doctoral or postdoctoral work on a specific project. And that work doesn't have to be in biology. Stefano Lise, an analyst recently hired by the Oxford Biomedical Research Centre, did his undergraduate, graduate and postdoctoral work in physics before switching to bioinformatics and next-generation sequencing; and McVean sees many recruits enter the field from banking and finance.

Statistician Yun Li joined the faculty of the University of North Carolina in Chapel Hill after earning her doctoral degree in biostatistics at the University of Michigan in 2009. In her undergraduate degree, Li had minored in computer science; she then earned a master's in statistics before starting her doctorate. While working on her PhD, Li developed data-analysis methods for the 1000 Genomes Project, a multinational study in which more than 1,000 individuals' genomes are being sequenced. She says that the hands-on experience working with what she calls “dirty” data — raw data whose characteristics and limitations have not been fully explored by researchers — has been invaluable in her current position.

“A typical genetic study nowadays will need to analyse millions or tens of millions of variants in tens of thousands of individuals,” says Li, who is now developing ways to work with large data sets and applying these and other methods to disease-focused studies. “This entails skills both to identify problems — which is important because many issues are typically not defined for data from cutting-edge research — and to solve problems.”

Whether trainees are interested in an academic or industrial job, it is computer-science skills that will help them to secure it. By far the most successful candidates are those who can not only write software, but also work with distributed computing systems, and computer operating systems such as Linux and Unix, say those in the field. “The more you understand software and computer science, the better off you are; writing software is 90% of what we're doing,” says Alexander.

For a field that is likely to continue its rapid change, the only sure thing is that data sets will continue to get bigger, and those who know how to handle them will be in high demand.

Erika Check Hayden reports for Nature from San Francisco.