Manhattan versus Reykjavik

Where is it best to hunt for genes that underlie cancer and heart disease? Isolated populations such as Iceland's, or ethnic melting pots like the United States? And what are the technological challenges, asks Alison Abbott.

According to some, Iceland is selling its soul to the devil, and the devil is a company called deCODE Genetics. Based in Reykjavik, deCODE's product is genetic information linked, anonymously, to medical records for the country's 270,000 inhabitants — or, at least, for the majority who have not asked to be excluded from the company's database.

The relative genetic homogeneity of Iceland's population, it was thought, should make it a good place to investigate the genetic factors involved in conditions such as heart disease. But the Icelandic parliament's decision to grant deCODE privileged access to the necessary data has angered some citizens, who object to their country's gene stock being used to profit a single company.

Among geneticists, however, this ethical controversy is taking second place to a more fundamental scientific debate. Do isolated, genetically homogeneous populations provide any real advantage in untangling the multiple factors, genetic and environmental, that contribute to common killers such as cancer, diabetes, heart disease and stroke?

There is a growing feeling that they might not. But if larger studies in genetically mixed populations, such as those of Western Europe and the United States, are the way forward, some difficult questions must be answered. What methods of analysis should be used? And can existing technologies cope with the huge amounts of genotyping that may be required? "We need a more elegant solution than just forcing large numbers through," says Eric Lander, whose genomics centre at the Massachusetts Institute of Technology's Whitehead Institute for Biomedical Research is one of the leaders in the field.

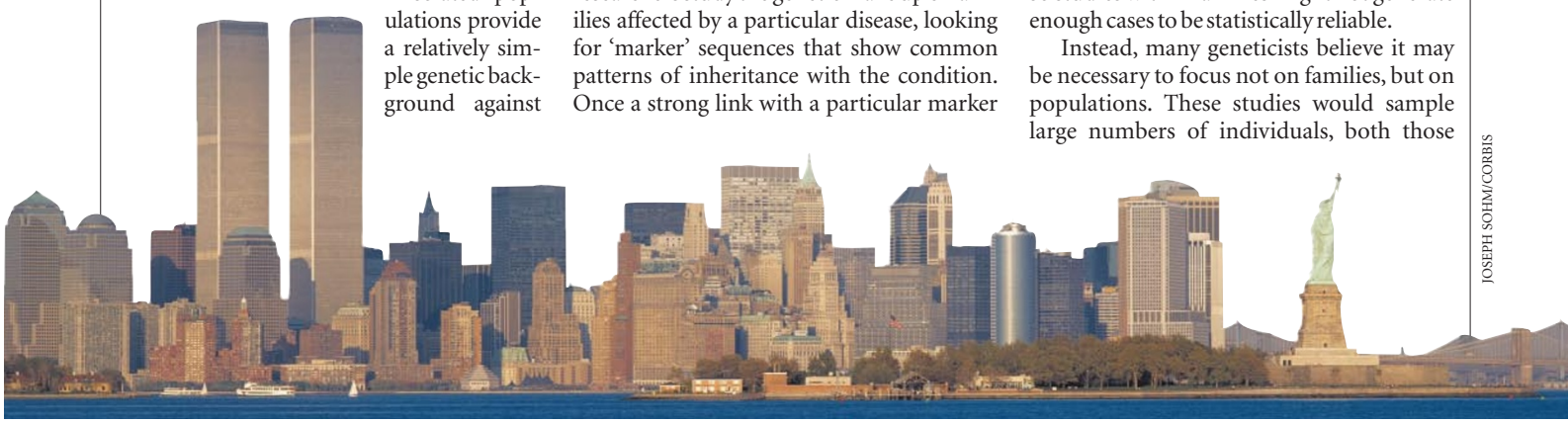
MICHAEL NICHOLSON/CORBIS

Spectrum of diversity: Andres Metspalu (right) contrasts the extreme genetic heterogeneity of Manhattan's population (below) with the homogeneity of isolated Reykjavik (above).

Isolated populations provide a relatively simple genetic background against which to investigate the genetics of a disease. They have already proved their worth in studies of conditions caused by single defective genes. Work in Finland, for instance, led to the discovery of the genes underlying a large number of rare hereditary conditions, including specific forms of dwarfism, epilepsy and eye disorders[1]. Such conditions tend to be perpetuated in the restricted gene pool of an isolated population.

Many of these discoveries have relied on a technique called linkage analysis, in which researchers study the genetic make up of families affected by a particular disease, looking for 'marker' sequences that show common patterns of inheritance with the condition. Once a strong link with a particular marker has been found, the geneticists focus on the surrounding chromosomal region to search for the disease gene itself. But for common diseases, which might involve many genes, each with a relatively small role to play in the overall condition, linkage analysis may not be so powerful[2,3]. Where individual disease genes exert only small effects, people carrying these genes do not always develop the disease, so studies within families might not generate enough cases to be statistically reliable.

Instead, many geneticists believe it may be necessary to focus not on families, but on populations. These studies would sample large numbers of individuals, both those



JOSEPH SOHM/CORBIS

with the disease and healthy controls. They would either look for an association between the disease and particular genetic markers, or would test hypotheses about mutations in 'candidate' genes whose normal function suggests they might be involved in disease.

Mix and match

Isolated populations such as those in Iceland and Finland were again thought to offer advantages for these 'association studies'. Many geneticists assumed that such populations would show relatively high levels of linkage disequilibrium (LD). This is the tendency for variants, or alleles, of two genetic sequences — such as a genetic marker and a nearby disease gene — to occur together within individuals more often than would be expected by chance. Researchers reasoned that in a population such as Iceland's, which expanded from a relatively small number of founders and has not experienced significant immigration, there should have been fewer opportunities for particular markers and disease genes to have become separated down the generations.

Recently, this assumption has been challenged. Last year, for instance, Leonid Kruglyak of the Fred Hutchinson Cancer Research Center in Seattle produced a theoretical model of populations with different levels of heterogeneity, and concluded that isolated populations are unlikely to be very different from more mixed populations in terms of LD[4].

Two papers published this month in *Nature Genetics*[5,6] provide data to support Kruglyak's conclusions. Researchers led by John Todd of the Wellcome Trust Centre for Molecular Mechanisms in Disease in Cambridge studied a region of chromosome 18 in Finnish, Sardinian, British and American populations. They measured LD for genetic markers known as microsatellites and found no significant differences between the various groups. The researchers suggested that "genetic isolates like Finland and Sardinia will not prove significantly more valuable than general populations for LD mapping of common variants underlying complex disease". Another team, headed by Pui-Yan Kwok of Washington University in St Louis, Missouri, conducted a similar study of the X chromosome using genetic markers called single nucleotide polymorphisms (SNPs) and came to the same conclusion. Both teams stress that their findings cannot necessarily be applied to the entire genome. "It shows the urgent need for much, much more data," says Kwok.

Population counts

Kári Stefánsson, chief executive officer of deCODE, says that it is still too early to tell whether homogeneous or heterogeneous populations will prove superior. "The proof of the pudding is in the eating," he says. Stefánsson adds that Iceland's meticulous medical records and extensive genealogical data, the latter of which extend back for some 1,000 years, represent a unique resource that will stand deCODE in good stead even if homogeneity proves not to be crucial. By tapping into this information, the company will be able to select healthy controls that are well matched to people affected by a given disease.

But the results from Kwok and Todd are adding to a growing impression that genetically isolated populations will not be a panacea. In Britain, for example, the Medical Research Council (MRC) and the Wellcome Trust are designing a study of some 500,000 middle-aged people. Blood samples for DNA analysis, together with lifestyle information, will be collected and correlated with the onset of diseases including cancer, diabetes and heart disease. "Our heterogeneous population, with its large ethnic minority groups, will be an advantage," says Tom Meade of the MRC's Epidemiology and Medical Care Unit at Queen Mary and Westfield College in London, who chairs the expert working group designing the study. "Our results will be representative of the population as a whole."

deCODE

Sorted: researchers prepare to genotype samples.

Researchers in Estonia, meanwhile, are preparing to launch a similar study that aims to examine one million people — three-quarters of the country's population. Scientists behind the plan had originally argued that the country's relatively homogeneous population — "somewhere between Manhattan and Iceland", according to Andres Metspalu of the University of Tartu, one of the project's organizers — would make the study easier. "But we don't think this will be our main strength now," he says. Instead, the project's organizers are putting their faith in the use of health questionnaires, carefully designed to determine who is affected by which disease, coupled with the sheer size of their sample.

Big is better

Indeed, size is the fundamental advantage offered by the heterogeneous populations found in most of Western Europe and North America. The world's largest study of cancers is the multi-centre European Prospective Investigation into Cancer and Nutrition (EPIC), which started recruiting in the early 1990s and now has 500,000 participants. EPIC's principal investigator, Elio Riboli at the International Agency for Research on Cancer (IARC) in Lyons, was wise enough to insist that blood samples be gathered from participants from the beginning. This has allowed the project to move into the study of genetics, in addition to lifestyle factors, and it has just started genotyping. So far, its sample includes 350 cases of colon cancer and nearly 2,000 of breast cancer.

L. WILLATT/SPL

Hidden gems: identifying genes that cause common diseases will mean screening the entire genome.

In the United States, the National Cancer Institute in Bethesda, Maryland, has commissioned its director of epidemiology and biostatistics, Robert Hoover, to investigate the feasibility of organizing a consortium of large-scale population studies in cancer. "'Large' means all studies which are likely to generate 1,000 to 1,500 particular cancers in a reasonable time," says Hoover. This could include up to 15 studies over the next decade. Hoover is discussing whether genotyping for these projects should, and could, be coordinated, perhaps even using a single high-throughput facility. He is also considering how clinical and lifestyle information should be gathered to allow for comparisons between the different studies.

For most of the proposed work, researchers intend to focus on SNPs. These are single-base substitutions found scattered throughout the genome. They account for most of the genetic variability that helps make us all different. Large-scale identification of SNPs began only recently[7]. But since April last year, thanks to the efforts of the SNP Consortium — a collaboration funded by the Wellcome Trust together with many of the world's biggest drugs companies, and involving leading academic genomics centres — some 300,000 SNPs have already been put into a public database. That figure could top two million by next summer.

But association studies that use SNPs as markers and then fish randomly for genes that predispose to common diseases could be prohibitively time-consuming and expensive. The number of SNPs required depends on the degree of LD in the population. A thorough study could need hundreds of thousands of different SNPs from tens, or even hundreds, of thousands of individuals, sending the total number of individual SNPs to be genotyped into the billions. So far, even the biggest labs can only genotype several thousand SNPs per day. Indeed, their throughput would need to increase by some three orders of magnitude to make such mammoth studies feasible.

Prime candidates

That makes some geneticists argue that a candidate-gene approach might be the best option, at least for now. Some SNPs appear within genes of known function, and for some of these, there are already reasons to suppose that mutations in the genes might predispose people to particular diseases. For example, one study of breast and prostate cancers, being done under the EPIC umbrella, is analysing SNPs in genes involved in the synthesis of steroid hormones, which frequently influence the initiation or growth of such tumours. The research team, led by Federico Canzian at IARC, will analyse an average of five SNPs in each of 20 genes for a sample of some 1,000 cancer cases, and an equal number of matched controls. That gives a manageable total of 200,000 SNPs to be genotyped.

Meanwhile, at the French National Centre for Genotyping in Evry, near Paris, researchers led by Mark Lathrop are embarking on a larger study. They are sequencing 500 candidate genes for heart disease in 100 individuals to identify SNPs that can later be used in a study of about 8,000 individuals.

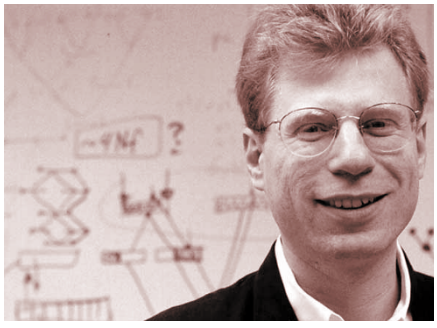
deCODE

Stefánsson: deCODE controls a unique resource.

But many genes linked to disease are likely to evade candidate-gene studies, so large association studies could still be necessary. "The candidate-gene approach is the only practical way at the moment," says Hoover. "But eventually we may want to do the whole-genome association studies with no *a priori* hypothesis." Geneticists hope that advances in genotyping technology will come to the rescue, increasing throughput and decreasing costs to the point that this becomes feasible.

At present, SNPs are genotyped by two main methods. In the 'chip' approach, fluorescently labelled 'complementary' DNA sequences that bind to particular SNPs are attached to a solid surface, usually a glass slide, which is then exposed to the sample DNA. The SNPs in the sample are then identified from the resulting fluorescence pattern on the DNA chip. Once the chips have been made, this method is very efficient as it allows large numbers of SNPs to be analysed in one go. But manufacturing and processing the chips is time-consuming, and adding new SNPs to the analysis means designing new chips.

The alternative is to use mass spectrometry. The sample DNA is treated so that the nucleotides in certain SNPs are substituted for nucleotides bearing additional chemical groups. The SNPs are then detected by analysing the molecular mass of fragments of the DNA. Only a small number of different SNPs can be genotyped at a time, but the analysis is extremely fast, and new SNPs can be incorporated into a study very easily.



Kruglyak: doubts the advantages of isolation.

Labs around the world are now scaling up both of these technologies, and working to make them more efficient. At the same time, small biotech companies are offering alternative methods, using innovative enzyme-based assays. But none of these technologies seems likely to provide the huge leap in throughput needed to conduct the thorough, genome-wide association studies that some gene hunters would like to launch.

Variety shows

Even without major technological advances, clever study design might help to reduce the amount of genotyping to a more reasonable level. Here, the focus would be on finding out more about variability in LD across the genome. Preliminary studies of chromosome 22, recently sequenced by a team led by Ian Dunham of the Sanger Centre at Hinxton, near Cambridge[8], suggest that this variability will be high. In genomic regions where LD is extremely low, there may be little point analysing SNPs, as associations with disease will be extremely difficult to detect. And in regions where LD is very high, geneticists might be able to spot associations using a relatively low density of SNPs.

Dunham suggests that geneticists would benefit from maps detailing variability in LD across the genome for different populations. They could then decide which SNPs to use on a study-by-study basis, to ensure the best balance between cost and the efficient capture of disease genes. Kwok agrees: "It would not be difficult, and LD maps could reduce the number of SNPs needed for a whole-genome association study to as low as 30,000." Kwok is talking with other researchers about the possibility of creating such maps. "As yet there is no assured funding," he says.

As the technology now stands, 30,000 SNPs would still make association studies involving thousands, or tens of thousands, of individuals extremely daunting. But gene hunters say that such studies lie within the grasp of foreseeable developments in genotyping. "The speed of technological advance, with the resultant cost reductions, are likely to make studies on this scale feasible in the next two years," concludes David Bentley, head of human genetics at the Sanger Centre. ■

Alison Abbott is *Nature*'s Senior European Correspondent.

1. de la Chapelle, A. *J. Med. Genet.* **30,** 857–865 (1993).
2. Risch, N. & Merikangas, K. *Science* **273,** 1516–1517 (1996).
3. Risch, N. *Nature* **405,** 847–856 (2000).
4. Kruglyak, L. *Nature Genet.* **22,** 139–144 (1999).
5. Eaves, I. A. *et al. Nature Genet.* **25,** 320–323 (2000).
6. Taillon-Miller, P. *et al. Nature Genet.* **25,** 324–328 (2000).
7. Wang, D. G. *et al. Science* **280,** 1077–1082 (1998).
8. Dunham, I. *et al. Nature* **402,** 489–495 (1999).