

COMMENTARY

Deeper into the genome

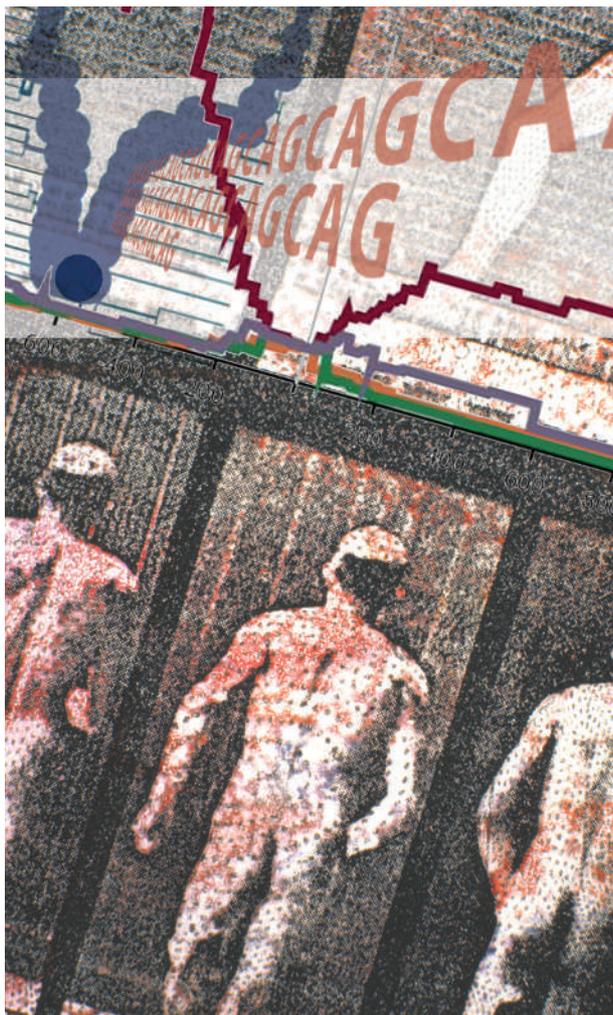
The next large-scale human genome project after HapMap should catalogue inherited variation in the general population that directly affects gene function, argues **Richard Gibbs**.

The International HapMap Project has catalogued the patterns of more than 1 million single-base changes (known as single nucleotide polymorphisms, SNPs) in the genome sequences of 269 people drawn from four diverse human populations¹. Most of these SNPs do not directly influence gene function, but the data provide valuable information about the overall pattern of chromosome organization and offer new tools for finding disease-causing genes in humans.

A complementary but more direct way to discover disease-causing genes is 'medical resequencing' (MRS). Key parts of suspect genes are sequenced and compared between patients and controls to identify genetic variations that may contribute to disease². This approach has been popular and fruitful, although it relies heavily on picking the right candidate gene at the outset. MRS activity is increasing, with a few sequencing centres now analysing hundreds of individual genes. Using MRS to analyse all known human genes (currently more than 20,000) in selected sets of patients and controls is also under discussion. This would greatly enhance the chances of successful disease-gene searches.

Rare finds

Such large-scale projects favour a centralized effort, similar to that used to decipher the human genome. Here, we argue that MRS activities should be accelerated, but the goal should be to discover genetic variation in the general population (not just patients and controls) that can potentially affect gene function directly. During this sequencing phase, we would not need to know the disease status of the individuals sampled. This rich catalogue of genetic changes, here called 'functional variants', would include SNPs that alter amino acids in proteins, and possibly gene-splicing or expression levels. Most would be rarer than those pursued by HapMap. This functional-variant database would be immediately available for all



Directly analysing genes could shed light on the causes of disease.

researchers and would have considerable impact on future disease-gene studies.

Some of the functional variants that cause disease have already been catalogued thanks to studies of mendelian diseases — human conditions with simple (one-gene) inheritance³. Polymerase chain reaction (PCR), fluorescent DNA sequencing and other techniques have enabled the discovery of about 1,700 mendelian disease genes, most of which have multiple functional variants. These diseases are rare, affecting about 1 in 10,000 to 1 in 100,000 individuals, and the frequency of the mendelian functional variants in the general

population is often less than 0.05%.

The HapMap data will help to reveal functional variants that cause more common diseases, such as adult cardiac diseases, cancer and schizophrenia. These disorders mostly have complex (multi-gene) underlying genetics and the HapMap tools work best when the functional variants involved occur in the diseased population with a frequency greater than 5% (ref. 4).

The process of detecting the presence of a particular SNP is called 'genotyping' and the HapMap project has identified a subset of the estimated 3 million common SNPs that can be identified in genotyping assays. The project also identified a further subset of 'tag' SNPs that are most useful in genetic association studies, because they link common haplotypes — larger blocks of DNA that are inherited together. Using these tags, the screening of whole genomes in patient sample collections is a much more realistic proposal than it was just three years ago.

Beyond HapMap

When disease-causing functional variants fall below 5% in the diseased population, approaches aided by HapMap lose power. These limitations of HapMap are entirely expected, as the overall distribution of SNPs seen in human populations follows the shape of a power-law distribution, with the rarer SNPs

accounting for most overall variation (see left side of graph overleaf)¹. Rare functional variants corresponding to the mendelian disorders are in this category. More common SNPs appear on the right of the graph and include the few disease-causing functional variants that have been identified.

Markers that fall in the frequency range 0.05–5%, are less well known, primarily because technical hurdles prevent their discovery. Nevertheless, they are likely to contribute to human disease — theoretical modelling supports their importance^{5,6} and direct evidence from disease examples is

emerging. Cohen *et al.* recently studied individuals who are genetically susceptible to cholesterol deposition and therefore heart disease, and who were further characterized by the mean levels of low-density lipoprotein (LDL) in their blood⁷. They reasoned that functional variants in key genes would be found most easily in the individuals with the highest and lowest levels of LDL. Using MRS, Cohen and his colleagues discovered a pair of functional variants in a key gene, *PCSK9*, that conferred low LDL levels. Notably, the functional variants they discovered to be associated with low LDL levels were in the category of 'very rare mutations', but were found at sufficiently elevated frequency in the general population to be considered 'low-frequency variation'. These functional variants would probably not have been identified using HapMap genotyping.

Bigger and better

A global search for putative functional variants provides a logical framework for extending the HapMap resource. We propose a large-scale genome project to catalogue rare genetic variation — beyond what HapMap has achieved. Some limited public and private initiatives already exist along these lines⁸, but a larger and more ambitious programme that targets diverse human populations, independently of disease or phenotype, is required to accelerate discovery of the functional variants present at low frequency.

This project would first build a functional-variant database by using MRS targeted to the coding regions of all 20,000 known human genes, plus an additional segment of the presumed promoter (gene-control) region of each. Analysis of 2,000 individual DNA samples (covering 4,000 chromosomes) would offer a good chance of finding variation that is present at the 0.05% level in the general population. This is therefore a bold proposal that could capture most of the expected functional variants (perhaps 50,000 to 100,000)⁹ found in humans.

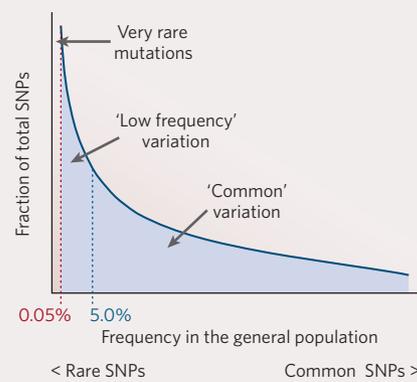
The populations sampled should include African, Asian and Caucasian representatives, in addition to Hispanic and smaller native groups. The optimal representation of different populations will be difficult to define without more data, but these choices may not be critical, as the large sample size will ensure that even large frequency differences between populations will not obscure individual variants.

The scale of this project would require improved technologies for MRS, which are already under way. MRS is easier, technically speaking, than *de novo* sequence determination because there is always a target sequence for comparison. This means that faster reductions in MRS costs are more likely than those observed for other genome projects.

Crucially, the functional-variant approach would avoid centralization of disease analyses, while exploiting the sequencing power of large genome centres to build the primary database. The overall goal would be to identify variation

GENETIC VARIATION IN HUMANS

Variation is measured by single nucleotide polymorphisms (SNPs).



that could subsequently be tested by the same kinds of genotyping methods used for HapMap. We can expect private companies to generate 'DNA probes' to test for these genetic markers in new populations, as has already occurred with the existing database of around 20,000 publicly available SNPs¹⁰. The probes could be used in a distributed fashion — in individual investigator laboratories — as part of each ongoing disease discovery effort.

The functional-variant proposal contrasts with similar ideas for large-scale MRS that focus on patient populations, such as the Cancer Genome Project¹¹, which aims to catalogue all major mutations that occur in the

"A larger and more ambitious programme is required to accelerate discovery."

most common human cancers. Such projects risk revealing patient identities, assuming that their DNA sequence data is publicly released (as occurred for the Human Genome Project).

A 2004 study showed that as few as 75 SNPs could be used to identify an anonymous patient¹². Because intense genetic analysis of disease samples requires simultaneous sequencing of several genes from each individual, identification is easy. In the functional-variant database, the free public release of the SNP data could not be used to identify the contributing individuals because only the mutations themselves would need to be fully available. The follow-up process of discovering these multiple variants in the same individual would only occur in the laboratories of individual investigators working on specific diseases, at which point patient consent could be obtained.

The functional-variant approach has other advantages over large-scale MRS projects on patient populations. Although there are thousands of well-characterized tissue samples in 'disease collections', most of these are dispersed among the laboratories of different clinical scientists. Little coordination over phenotyping

or storage exists, and collection protocols and ethical approvals are not standardized. Plans to centralize these collections, perform MRS and then publicly release the data will have to contend with complex logistics and Institutional Review Board issues. For clinicians, assigning credit for what might be years of sample collecting becomes a greater concern when all the attention is focused on new sequencing studies. The functional-variant database avoids most of these issues by separating the initial sequencing phase from the laboratory investigations of specific diseases.

One disadvantage of the Cancer Genome Project is that each new mutation discovered might be present in only a small fraction of the cells in a minute tissue sample. Non-cancer tissue from each patient must therefore be checked, to ensure that the mutation is related to the cancer and is not simply inherited rare variation. This complication does not, however, apply to the functional-variant database.

At current prices, with PCR Sanger fluorescent DNA sequencing, the cost of analysing the full complement of coding genes for 2,000 individuals would be about US\$750 million. We can, however, reasonably expect MRS costs to drop progressively, and if the work is spread over five years it is likely to cost less than \$500 million. This is more than five times the amount spent on HapMap, but is lower than the projected \$1.35 billion cost for the Cancer Genome Project.

Finally, any proposal for large-scale MRS raises a familiar dilemma in genomics — how to balance the efficient high-throughput sequencing of large genome centres with the powerful research approaches provided by the wider community. Among all possible future projects, the functional-variant proposal offers to enhance the efforts of established networks of biomedical investigators in a way that echoes the Human Genome Sequence project.

Richard Gibbs is at Baylor College of Medicine, Houston, Texas 77030, USA.

1. International HapMap Consortium *Nature* **437**, 1299–1320 (2005).
2. Gibbs, R. A. *et al.* *Genomics* **7**, 235–244 (1990).
3. Institute of Medical Genetics, Cardiff, UK www.hgmd.cf.ac.uk/hgmd0.html
4. Belmont, J. W. & Gibbs, R. A. *Am. J. Pharmacogenom.* **4**, 253–262 (2004).
5. Pritchard, J. K. & Cox, N. J. *Hum. Mol. Genet.* **11**, 2417–2423 (2004).
6. Reich, D. E. & Lander, E. S. *Trends Genet.* **17**, 502–510 (2001).
7. Cohen, J. *et al.* *Nature Genet.* **37**, 161–165 (2005).
8. www.sanger.ac.uk/genetics/exon/ and www.celera.com/celera/applera_genomics
9. Botstein, D. & Risch, N. *Nature Genet.* **33** (suppl.), 228–237 (2003).
10. www.ncbi.nlm.nih.gov/SNP/
11. Working Group on Biomedical Technology www.genome.gov/Pages/About/NACHGR/May2005NACHGR/Agenda/ReportoftheWorkingGrouponBiomedicalTechnology.pdf
12. Lin, Z., Owen, A. B. & Altman, R. B. *Science* **305**, 183 (2004).

Acknowledgements: Thanks to J. Belmont, D. Nelson and F. Yu for discussions and for reading the manuscript.