

## NEWS &amp; VIEWS



## GENOMICS

# Understanding human diversity

David B. Goldstein and Gianpiero L. Cavalleri

**The first edition of a massive catalogue of human genetic variation is now complete. The long-term task is to translate these data into an understanding of the effects of that variation on human health.**

When, as a medical student, the author W. Somerset Maugham could not find a nerve where it was supposed to be, his anatomy instructor uncovered the hidden nerve and explained that “the normal is the rarest thing in the world”<sup>1</sup>. On page 1299 of this issue<sup>2</sup>, in a paper from the International HapMap Consortium, we find a detailed account of one of the primary reasons why there is no ‘normal’ human type.

The human genome has about 10 million ‘polymorphisms’, defined as genetic variants in which the minor gene forms occur at least once out of every 100 forms. Any two unrelated humans have millions of genetic differences, making them look and even behave differently. This variation is the magnificent legacy of our evolutionary past, but it comes at a price. Along with making us different in benign and interesting ways, genetics also influences health.

Modern geneticists have been hugely successful at tracking down the genetic abnormalities that lead to diseases that are inherited in a simple way in families, such as cystic fibrosis or Tay–Sachs disease. These abnormalities, however, are comparatively uncommon. The genetic disorders to which most of us will

succumb are more complex ones, such as cancer, or cardiovascular and neurodegenerative diseases. Indeed, there seems to be a grim inevitability about these common genetic diseases — as some of the ones that kill early in life are increasingly better treated, certain later killers, such as dementia, seem set to skyrocket.

The complexity of these common diseases has made them, until now, largely impervious to genetic analysis. That genetics plays an important role, however, is beyond question — as illustrated by countless studies demonstrating, for example, that genetic variation explains more than 40% of the variation of most common diseases in a population and more than 70% for some disorders such as schizophrenia<sup>3–5</sup>. The difficulty is that gene variants predispose us to, but do not invariably cause, common diseases. And they predispose in combination with other gene variants and with the environment.

How do we determine which of the 10 million polymorphisms influence disease? Thankfully, it is not necessary to directly assay all 10 million sites and assess their associations with disease. This is because polymorphisms in the human genome are

often not independent of one another. When a mutation arises, it is associated with particular variants present on the same chromosome (variants that associate together are known as a ‘haplotype’). For this and other reasons, there are often strong statistical associations between polymorphisms, such that the presence of a particular variant at one site on a chromosome can predict or ‘tag’ the presence of a particular variant at another site.

The principal goal of the HapMap Consortium was to discover these associations among variants (hence the name of the project), and it has succeeded in a spectacular way. The first phase of the project, reported in this issue<sup>2</sup>, consisted of compiling data on the genetic make-up, or genotype, of groups of individuals representative of four populations for more than a million single nucleotide polymorphisms, or SNPs. (SNPs are one of the most common types of variant in our genome, as opposed to, for example, insertions and deletions.) The aim was to ensure that one SNP was assayed for every 5,000 bases of sequence. The second phase, not yet complete, will considerably increase the SNP coverage.

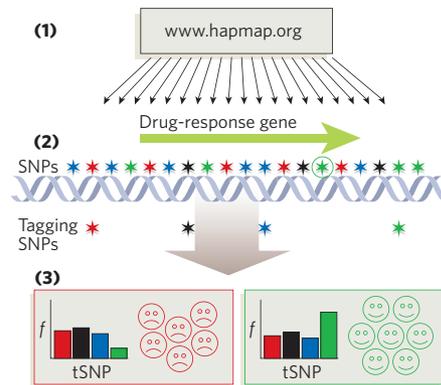
Using a conceptually easy and reliable way to select tags, the HapMap group has

**BOX 1 HapMap in practice**

For any research group interested in relating human genetic variation to human health, the HapMap project is an unprecedented gift. The data make it possible to select reliable, tagging single nucleotide polymorphisms (SNPs) for genes or genomic regions in a matter of minutes. **(1) Download genotype data.** Data for a gene (or region) of interest are freely available from the HapMap website.

**(2) Assess the associations among SNPs in the data to select tagging SNPs.** In the figure, SNPs that associate closely with each other have the same colour. The overall set of SNPs can be compressed to a subset of tagging SNPs, here depicted as one of each colour.

On average, a single SNP has between three and ten other SNPs (depending on the population) that are perfectly associated with it in the samples considered (although not necessarily in another sample from the same population). It is thus possible to greatly reduce the number of SNPs that need to be typed by excluding all but one from such sets of associated SNPs. More sophisticated approaches are likely to be even more cost-effective, and various statistical methods have been proposed. Although there is not yet a consensus on the overall best approach, one or more is



likely to emerge soon that will lead to efficient sets of tagging SNPs for the whole genome.

**(3) Genetic association.** The tagging SNPs are genotyped in a population sample in which individuals vary for some trait of interest, for example a good and bad reaction to a medicine. A tagging SNP that correlates with a certain reaction indicates that one of the SNPs with which it is associated influences that reaction — in this case, one of the green SNPs (circled) causes patients to have a good response to the medicine, leading to association with the relevant tagging SNP. **D.B.G. & G.L.C.**

estimated that the number of tags necessary to represent common variants across the genome may be less than one-tenth of the total number of such sites. More sophisticated approaches are likely to do even better.

As well as describing these associations, the HapMap project has made other advances. Three years ago, fewer than 1.7 million polymorphisms were known. Today, thanks to the project and related efforts, that number is more than 8 million. Knowing most of the polymorphisms offers tremendous advantages. We are now in a position to apply bioinformatics tools to these data and to prioritize polymorphisms in terms of potential functionality — focusing, for example, on those that change protein sequence, or that are located in functionally important chromosomal regions (as defined by the degree of conservation of that region across species), or that seem to be relevant using other genomic criteria.

What does this mean for the study of human disease and variable response to treatment? Imagine that you wanted, before the HapMap project, to assess whether common genetic variation in the molecular target of a medicine influences patient response. You would have to re-sequence the gene in a group of individuals representative of your study population in order to identify a sufficient number of the polymorphisms. Four years ago, our group did this for a gene known as *SCN1A*, which encodes a target of antiepileptic drugs; it took us two years to identify the common polymorphisms and appropriate tags. Today, the

same job can be accomplished with simple computer algorithms, in minutes, using the HapMap data (Box 1).

The potential of whole-genome association using HapMap data is highlighted by results presented in another paper in this issue (page 1365)<sup>6</sup>. Cheung and colleagues restudied 27 genes whose expression had been found by family-based studies to be influenced by genetic variants. They showed that many of these would have been detected through genome-wide association, and that in one case HapMap data facilitated the identification of a variant shown by *in vitro* analysis to be responsible for altering gene expression.

The key question, then, is the relevance to human health. Because these tools are wholly new, previous failures to identify the genetic contributions to common human diseases need not presage future failures. But equally, there are no guarantees. Although we know that genetics contributes to common diseases, we do not know what sorts of gene variants are responsible. If they are common (that is, one of the 10 million), the new tool-kit will greatly accelerate the identification of disease-related genetic variation. But if the responsible variants are rarer, they will be more difficult to find. Similarly, it is not known how well other kinds of variants (for example, repetitive elements or insertions and deletions) might be represented by SNP tags. Another complexity not yet adequately addressed concerns how well the four population groups studied in the HapMap project represent variation in other human populations.

We must also admit that there is little direct evidence that the identification of risk factors for common human diseases will improve health. In those rare cases where clear risk factors have been identified, they typically have not been helpful in treatment or prevention. For example, a variant of the *APOE* gene is a strong predictor for late-onset Alzheimer's disease, but there are no known alterations in lifestyle or diet that ward off the disease. The best hope may be that risk factors will suggest new pathways for therapies, but here too there are few successful examples.

A stronger case might be made for the near-term clinical relevance of identifying genetic predictors of a patient's responses to treatment. Polymorphisms can have big effects on such responses, and identification of these effects can suggest alternative treatments. For example, there is little difference in overall effectiveness between different antipsychotic drugs, even when comparing new-generation medicines with a class introduced decades ago<sup>7</sup>. But there are huge differences among patients in response to the different medicines. If we could identify the gene variants that predict for example whether newer medicines cause unacceptable weight gain, or whether older ones cause a severe movement disorder, treatment options could be adjusted accordingly. The HapMap project is likely to greatly facilitate the search for such variants.

The current state of genomic sciences may be considered as a sort of awkward adolescence. The power of the modern genomic tool-kit is breathtaking. In a few years we have gone from knowing almost nothing that could be characterized as genomic (that is, focused on whole genomes rather than particular genes) to having complete genome sequences for many organisms, and now a nearly complete catalogue of the common genetic differences among people. Technical prowess is not in itself a mark of maturity in science, however. The next phase for genomics research requires a greater focus on both biological understanding and clinical utility. It is time for genomicists to turn their attention from technology to application. ■

David B. Goldstein and Gianpiero L. Cavalleri are at the Institute for Genome Sciences and Policy, Center for Population Genomics and Pharmacogenetics, Duke University, 103 Research Drive, DUMC Box 3471, Durham, North Carolina 27710, USA. e-mail: d.goldstein@duke.edu

1. Meyers, J. *Somerset Maugham: A Life* (Knopf, New York, 2004).
2. The International HapMap Consortium *Nature* **437**, 1299–1320 (2005); www.hapmap.org
3. Pedersen, N. L., Posner, S. F. & Gatz, M. *Am. J. Med. Genet.* **105**, 724–728 (2001).
4. Sullivan, P. F., Kendler, K. S. & Neale, M. C. *Arch. Gen. Psychiat.* **60**, 1187–1192 (2003).
5. Zdravkovic, S. et al. *J. Intern. Med.* **252**, 247–254 (2002).
6. Cheung, V. G. et al. *Nature* **437**, 1365–1369 (2005).
7. Lieberman, J. A. et al. *N. Engl. J. Med.* **353**, 1209–1223 (2005).