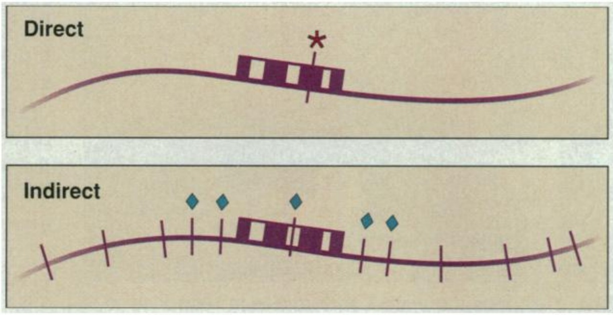
Variations on a Theme: Cataloging Human DNA Sequence Variation

Francis S. Collins, Mark S. Guyer, Aravinda Chakravarti

Genetic factors contribute to virtually every human disease, conferring susceptibility or resistance, or influencing interaction with environmental factors. Much research in both the public and private sectors is driven by the expectation that understanding the genetic contribution to disease will revolutionize diagnosis, treatment, and prevention. Understanding the role played by genetic factors in disease is also expected to increase understanding of the nongenetic, environmental contributions.

The genetic maps, physical maps, and technologies for gene identification that have emerged from the Human Genome Project (HGP) have already had a significant effect on the research community's ability to discover genes underlying Mendelian disorders. Positional cloning (*1*), in which a disease gene is identified by virtue of its location in the genome rather than by using knowledge of its biochemical function, was successfully applied in the search for human genes for the first time in 1986. By 1990, when the HGP began, only a handful of additional successes had accrued. By 1997, however, close to 100 disease loci have been identified with this strategy.

This dramatic progress has encouraged efforts to apply the same strategy to the study of genes underlying common disorders whose inheritance is much more complex, including diabetes, hypertension, asthma, common cancers, and the major neuropsychiatric diseases. Not surprisingly, progress in analyzing complex genetic disorders has been more modest. What success there has been has basically come from one of two approaches: (i) Identification of a subphenotype (such as younger age at onset of disease) in pedigrees in which the disease behaves in a near-Mendelian fashion and in which, therefore, the positional cloning strategy can be applied (*2*). Only a small fraction, usually less than 10%, of the occurrence of a common disease can be accounted for in this way. (ii) Genetic studies in isolated human populations, which characteristically harbor reduced genetic variation and consequently reduced genetic complexity of the disease trait. In such populations, a chromosomal segment shared by affected individuals can often be identified, and the genes within this shared segment can then be analyzed as disease gene candidates (*3*). However, except for cases in which these two approaches have been used, there have been no published successes in identification of genes underlying polygenic diseases by a pure positional cloning approach. The successful identification of the multiple genes underlying complex human diseases is likely to be much more difficult than some had originally envisaged.



Whole-genome association studies. (**Upper panel**) The direct study detects an increased prevalence (association) of a particular functional variant (*) in affected individuals within the coding region of a candidate gene. (**Lower panel**) The indirect study depends on linkage disequilibrium and detects an increased prevalence of a particular characteristic set (♦) of SNP alleles in affected individuals.

This experience has rekindled interest in the strategy of association studies on candidate genes (*4–6*). Association studies do not involve analysis of large family pedigrees but compare the prevalence of a particular genetic marker, or set of markers, in affected and unaffected individuals (*7*). A greater prevalence of a marker in affected individuals is considered evidence of association between the disease phenotype and the marker. Early association studies were limited by the modest number of polymorphisms available. More recent technological developments have enabled the identification and scoring of individual variations in DNA and have led to a large increase in the number of available polymorphisms and, therefore, to a renewed interest in association methods (*4*, *8*).

The set of common variations in DNA sequence in the estimated 80,000 human genes is finite, albeit large. If association studies could be extended to include a systematic search through the entire list of common variants in the human genome, the strategy should in theory reveal the identity of the gene or genes underlying any phenotype not due to a rare allele. Until recently, obtaining such a data set has been assumed to be many years off and probably not achievable until after the completion of the first reference human sequence (which is anticipated by the year 2005). Three recent developments, however, have indicated that the time is now right to begin the systematic cataloging of human sequence variation.

First, recent quantitative analyses of the power expected from whole-genome linkage studies versus whole-genome association studies suggest that the association studies should be particularly efficient for identification of genes with relatively common variants that confer a modest or small effect on disease risk—precisely the type of gene expected in most complex disorders (*6*). This increased analytical power translates into a vast reduction in the number of DNA samples needed to identify a gene that contributes to a particular disease. Furthermore, traditional pedigree linkage studies cannot be based on families with one affected member; with association studies, this is possible.

A second impetus is the development of improved methods for the discovery and genotyping of single-nucleotide polymorphisms (SNPs). Past methods for SNP discovery have depended primarily on gel-based sequencing of DNA from several individuals and have therefore been relatively slow and expensive. Recent novel approaches to assessing DNA sequence differences between individuals offer considerable promise for reducing the cost and increasing the rate at which large numbers of SNPs can be discovered. We anticipate that it should be possible to generate SNPs in several thousand genes per year at a roughly estimated cost between $100 and $1000 per SNP. These rates and costs can be expected to drop substantially as technologies for high-throughput SNP discovery mature. Several promising methods for the semi- or fully automated discovery or scoring (or both) of SNPs in very large numbers are

F. S. Collins and M. S. Guyer are at the National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892–2152, USA. Collins e-mail: fc23a@nih.gov. Guyer e-mail: mg25m@nih.gov. A. Chakravarti is in the Department of Genetics, Center for Human Genetics, Case Western Reserve University, Cleveland, OH 44106, USA. E-mail: axc39@po.cwru.edu

also being developed, including minisequencing strategies (9), multiplex reverse dot blots (10), DNA chips (11), and the TaqMan approach (12).

Third, because they are potentially such valuable research tools, SNPs need to be made freely available as quickly as possible, so that the widest possible array of researchers in the public and private sectors can begin to use them immediately. Although it may seem odd that common variation in the human genome could be claimed as intellectual property, some patent experts consider SNPs (particularly those found in protein-coding regions, or cSNPs) to have sufficient defining features of novelty, utility, and non-obviousness to be patentable. If SNP development continues without guidance or public funding support, substantial numbers of SNPs and cSNPs could be generated in private collections. Although some of these private collections may be "publicly available," a tangled web of restrictive intellectual property attachments might well arise, inhibiting many researchers from using these powerful tools. For randomly chosen SNPs, the concern is somewhat less, as there should be several million of these in the genome, and there is thus no immediate danger of having them all discovered privately; nonetheless, even these are not an infinite set. About the estimated 200,000 cSNPs, which lie in coding regions, the concern is much greater.

There are two different yet complementary ways to perform whole-genome association studies. The most direct is to catalog and test directly all of the common functional variants (see the figure, upper panel). There should be relatively few in each gene; theoretical arguments and limited observations suggest that, in a species as young as the human, only two or three variants in any coding sequence will be frequent (~10% or greater) (4). However, not all of the functional variants will be in the coding regions. Furthermore, full-length sequence is currently only available for a small fraction of human genes (perhaps 5000), so this approach will not be broadly applicable for some time.

A second approach, which should be pursued in parallel, involves the use of a very dense map of SNPs arrayed across both coding and noncoding regions. A dense panel of SNPs from such a map can be tested in affected individuals and in controls to identify associations that narrowly locate the neighborhood of a susceptibility or resistance gene (see the figure, lower panel). This strategy is based on the hypothesis that each sequence variant that causes disease must have arisen in a particular individual at some time in the past, so the specific array of polymorphisms (haplotype) in the neighborhood of the altered gene in that individual must be inherited in all of his or her descendants. The presence of a recognizable ancestral haplotype therefore becomes an indicator of the disease-associated polymorphism. The size of this region (in which the genetic markers are said to be in "linkage disequilibrium") will vary with the age of the variant. Comparing the DNA sequences on two chromosomes in the vicinity of a variant that is 1000 to 10,000 generations old, which is roughly the age of the human population, the size of the chromosomal region shared by the two chromosomes and defining the ancestral haplotype will on average be as small as 10 to 100 kb. This is well below the resolution of current genetic maps, which is why a map of much higher marker density is necessary.

Considering the above arguments and the exciting possibility of rapid acceleration of our understanding of the most common human diseases, the National Advisory Council on Human Genome Research recently concluded that immediate action to develop a catalog of human sequence variation is warranted (13). Because this new opportunity is likely to require substantial resources, a consortium among institutes at the National Institutes of Health and other federal research agencies should be assembled to fund the effort to discover SNPs and cSNPs and to place them in public databases (such as GenBank or GDB) without restrictions on their use. The National Institute of Environmental Health Sciences has already announced its intention to fund the identification of common variants in a number of genes involved in gene-environment interactions (14), and this same strategy of identifying a list of high-priority candidates for cSNP determination might usefully be adopted by other agencies and organizations.

Participation by the private sector in this public effort would be particularly desirable. There are several recent examples of such public-private partnerships in genome research, including the Washington University–Merck & Co. collaboration to generate several hundred thousand human expressed sequence tags (ESTs) (15), the contribution made by Sandoz Pharmaceuticals to the human gene mapping efforts of the Whitehead Institute and Stanford University Genome Centers (16), and the Washington University–Howard Hughes Medical Institute partnership for generation of mouse ESTs (17).

Further improvements in the technologies for the discovery and detection of SNPs should also be immediately and aggressively pursued. Two goals should be targeted simultaneously: (i) the development of a dense map of at least 100,000 SNPs and (ii) the identification of common cSNPs in as many genes as possible. Although not initially an exhaustive list, these cSNPs will be of immediate biological use and importance. Both of these proposed resources will be a boon to investigators engaged in the genetic dissection of complex biological phenomena.

References and Notes

1. F.S. Collins, *Nature Genet.* **1**, 3 (1991); A. Ballabio, *ibid.* **3**, 277 (1993); F. S. Collins, *ibid.* **9**, 347 (1995).
2. Prominent successes of this approach include the BRCA1 [Y. Miki *et al.*, *Science* **266**, 66 (1994)] and BRCA2 [Wooster *et al.*, *Nature* **378**, 789 (1995)] genes for breast and ovarian cancer; the DNA mismatch repair genes for hereditary nonpolyposis colon cancer (HNPCC) [C.E. Bronner *et al.*, *Nature* **368**, 258 (1994); N. Papadopoulos *et al.*, *Science* **263** (1994)]; the MODY1, MODY2, and MODY3 genes for diabetes [K. Yamagata *et al.*, *Nature* **384**, 458 (1996); N. Vionnet *et al.*, *ibid.* **356**, 721 (1992); K. Yamagata *et al.*, *ibid.* **384**, 455 (1996)]; and six genes involved in salt and water metabolism in families with syndromic hereditary hypertension [R. P. Lifton, *Science* **272**, 676 (1996)].
3. A mutation in the endothelin receptor B gene, which is one of several genes leading to Hirschsprung disease with a complex phenotype in Mennonites, was identified within such a shared segment [E. G. Puffenberger *et al.*, *Cell* **79**, 1257 (1994)].
4. E. S. Lander, *Science* **274**, 536 (1996).
5. ——— and N. J. Schork, *ibid.* **265**, 2037 (1994).
6. N. Risch and K. Merikangas, *ibid.* **273**, 1516 (1996).
7. Classic examples of the success of this strategy are the human leukocyte antigen complex and a long list of autoimmune diseases such as type 1 diabetes [G.T. Nepom and H. Ehrlich, *Annu. Rev. Immunol.* **9**, 493 (1991)]. More recently, examples of disease associations have been found between the ApoE4 allele and Alzheimer's disease [W. J. Strittmatter and A. D. Roses, *Annu. Rev. Neurosci.* **19**, 53 (1996)], the Factor V Leiden variant and venous thrombosis [J. Voorberg *et al.*, *Lancet* **343**, 1535 (1994) and B. Zoller *et al.*, *ibid.*, p. 1536], a promoter polymorphism in the insulin gene and type I diabetes [S. T. Bennett *et al.*, *Nature Genet.* **9**, 284 (1995)], an Sp1 binding site in the first intron of the COL1A1 gene and osteoporosis [S. F. A. Grant *et al.*, *ibid.* **14**, 205 (1996)], and a deletion in the gene for the chemokine receptor CCR5 and AIDS resistance [M. W. Smith *et al.*, *Science* **277**, 959 (1997)].
8. R. Spielman *et al.*, *Am. J. Hum. Genet.* **52**, 506 (1993).
9. T. Pastinen *et al.*, *Genome Res.* **7**, 606 (1997).
10. A. P. Shuber *et al.*, *Hum. Mol. Genet.* **6**, 337 (1997).
11. S. P. Fodor *et al.*, *Science* **251**, 767 (1991).
12. K. J. Livak *et al.*, *PCR Methods Appl.* **4**, 357 (1995).
13. E. Marshall, *Science* **277**, 1752 (1997).
14. http://www.niehs.nih.gov/dirosd/policy/egp/
15. D. Gerhold and C.T. Caskey, *Bioessays* **18**, 973 (1996).
16. G. D. Schuler *et al.*, *Science* **274**, 540 (1996).
17. J. Kaiser, ed., *ibid.* **271**, 749 (1996).
18. We thank E. Jordan and M. Boehnke for helpful discussions and J. Ades for help in preparing the figure.