nature genetics
volume 20 no. 3 november 1998

SNP attack on complex traits

Single nucleotide polymorphisms (SNPs) are major contributors to genetic variation, comprising some 80% of all known polymorphisms, and their density in the human genome is estimated to be on average 1 per 1,000 base pairs. Although SNPs are mostly biallelic—and consequently less informative than microsatellite markers—they are more frequent and mutationally more stable, making them suitable for association studies in which linkage disequilibrium (LD) between markers and an unknown variant is used to map disease-causing mutations. In addition, because SNPs have only two alleles, they can be genotyped by a simple plus/minus assay rather than a length measurement, making them more amenable to automation.

These are good reasons to develop SNPs as useful markers, but hardly sufficient to explain the momentum that the SNP movement has recently acquired, which stems from the hope that SNP-based approaches will lead to progress in the search for genetic variation associated with common diseases or sensitivity to drugs. At a recent meeting*, advances in SNP technology and SNP-based approaches to tackle complex traits as well as questions of human origin and prehistory were discussed. Frustrated with linkage analysis, which has had little success in identifying genes involved in determining complex traits, many geneticists have turned towards association studies which might be better suited to detecting genetic effects of low penetrance with higher resolution. For such studies, many more markers will be required—in addition to better statistical tools and high-throughput low-cost genotyping technology to analyse large marker sets in many samples.

Increasing amounts of sequence data available in public and private databases, (within which SNPs can be discovered *in silico*; Pui-Yan Kwok, Macdonald Morris), efforts underway to re-sequence DNA stretches from several individuals, and the use of 'SNP discovery' technology (such as denaturing high performance liquid chromotography; Peter Underhill), have led to the rapid accumulation of catalogued SNPs. So far, no SNP has been patented, but a number of applications are pending (Christian Stein), and it seems likely that many will end up in proprietary collections. Even with the best tools, understanding complex traits and human variation will be a challenge, to say the least; sharing resources will help. Two publicly available SNP databases as well as several SNP collections exist at present (see box)—and researchers are encouraged to submit any SNP that they discover.

The technological and economic goal is accurate, easy, cheap and fast large-scale SNP genotyping. Several methods are currently being developed, and it is unclear which one(s) will turn out to be the best. Examples based on minisequencing on DNA arrays (Ann-Christine Syvänen, Andres Metspalu), dynamic allele-specific

*First International Meeting on Single Nucleotide Polymorphism and Complex Genome Analysis. Skokloster, Sweden, 29 August–1 September, 1998, organized by Anthony Brookes, Ulf Landegren, Ann-Christine Syvänen, Anders Issacson and Ulf Gyllensten, Uppsala University.

SNP databases

- **HGBASE (http://hgbase.interactiva.de)** collects intragenic SNPs and contains approximately 2,700 entries. It is searchable by sequence and, at the moment, the only database where information can be deposited and retrieved.
- **dbSNP (http://www.ncbi.nlm.nih.gov/SNP/)**, a joint effort by the NHGRI and the NCBI, is now accepting submissions. Its curators are still working on making content available; the database will be searchable by STS accession number and fully integrated with GenBank.

SNP websites

- The **MIT SNP database (http://www-genome.wi.mit.edu/SNP/human/index.html)** contains over 3,000 SNPs (approximately two thirds of them mapped) and is searchable by genomic region or internal STS identifier.
- The **WashU SNP database (http://www.ibc.wustl.edu/SNP/)** contains several hundred SNPs which are currently being integrated into dbSNP.

hybridization (DASH, Anthony Brookes), microplate array diagonal gel electrophoresis (MADGE, Ian Day), pyrosequencing (Pål Nyrén), oligonucleotide-specific ligation (according to Ed Southern, the most sensitive assay) as well as the Whitehead/Affymetrix SNP chips (Jian-Bing Fan) and the TaqMan system (Ken Livak) were discussed. All of them require target amplification of each SNP by PCR. Even in the light of encouraging progress in multiplexing PCR (Michelle Cargill), a large number of individual reactions is required and the cost is considerable (James Weber). Ideally, one would like to determine the genotype directly from genomic DNA. Methods based on the generation of small signal molecules by invasive cleavage followed by mass spectrometry (Timothy Griffin) or immobilized padlock probes and rolling-circle amplification (Ulf Landegren) might eventually eliminate the need for PCR.

Apart from the challenges of generating SNP maps and efficient genotyping, how easy will it be to determine which SNPs are suitable for a particular question and how best to analyse the data? In the absence of understanding what makes complex traits complex, classical mendelian concepts (two alleles, normal *versus* abnormal) are usually imposed onto a more complicated reality. Joseph Terwilliger warned that only if the genes underlying complex diseases have one wild-type and one (or one major) susceptibility allele—that is, when allelic heterogeneity is low—is statistical analysis likely to detect association of the causative allele (or linked markers) with the disease phenotype. Intuitively, more markers should allow increased accuracy, but in statistical reality, this also means larger samples will be necessary or the risk of obtaining false positive results will increase. Skeptical about the use of SNPs in disease genetics, Terwilliger is nonetheless enthusiastic about their potential use in population genetics and genetic epidemiology. By way of contrast, Marta Blumenfeld and Nik Schork described a strategy by which they can overcome many of the statistical obstacles of SNP-based association studies. By sequencing DNA from a minimum of 100 individuals to establish SNP allele frequency, calculating LD strength in a region of interest prior to determining how many markers are needed, and analysing haplotypes (2–6 SNPs together) instead of individual markers, they have been able to identify new genes associated with complex traits—unfortunately the identities of the genes were not disclosed, and so proof of principle is yet to be provided.

Although the jury is still out on whether SNPs will provide easy answers to complex questions, they are increasingly popular with disease and population geneticists. While the former mainly concentrate on SNPs within or close to genes, the latter often prefer markers outside of genes (to avoid selection) and in areas of the genome devoid of recombination. Several approaches using SNPs on the Y chromosome (Chris Tyler-Smith, Francesc Calafell) and in a low-recombination interval on the X (Svante Pääbo) provide interesting leads on human history, as well as data about age, frequency and population distribution of SNPs. Of course, this is information directly relevant to disease geneticists, and underscores the need for more interaction between population and disease geneticists (Andrew Clark, Rosalind Harding). Knowledge about population evolution and history will reveal suitable populations for genetic studies and aid in study design and interpretation of results.

Time—or rather data—will tell whether SNPs live up to expectations. As Aravinda Chakravarti stated in his abstract, "Each genetic approach, considered either optimistic or pessimistic, has its underlying assumptions. Human geneticists have to begin to test these assumptions not by computer simulations and theoretical arguments but by empirical observations".