nature genetics
volume 29 no. 4 december 2001

Challenges for the 21st Century

The International Single Nucleotide Polymorphism and Complex Genome Analysis meeting has a history of bad timing and a growing reputation for stoic resilience. Two years ago, the meeting coincided with Hurricane Floyd. This year's meeting* took place less than a month after the World Trade Center tragedy. Thus it was to the organizers' credit and a testament to the enthusiasm of the participants that 80-odd scientists gathered to discuss and debate the current and future status of SNP analysis.

Identifying SNPs. Finding nucleotide variants is not a difficult task: many molecular geneticists come across them unintentionally every day in their investigations. Directed SNP discovery can entail parallel sequencing strategies either on a genome-wide scale (Craig Venter, Celera Corp.), for single chromosomes such as the Y chromosome (Peter Oefner, Stanford Univ.) or within genes that seem to be reasonable candidates for genetically complex disorders such as cardiovascular disease (Francois Cambien, INSERM), inflammatory bowel disease (Gilles Thomas, CEPH) and type 2 diabetes (Alan Schafer, Incyte Genomics Ltd.). However, a number of bioinformatics groups are using the high level of redundancy in deposited sequence data to search for SNPs *in silico*. Christopher Lee (UCLA) described how EST databases provide representation of thousands of individuals, enabling the discovery of coding SNPs through sequence alignment. The same data mining approach has been applied to overlapping BAC clones. The cumulative result of these efforts is greater than 2 million candidate SNPs in dbSNP and HGBASE. Now the task is to validate these apparent SNPs, as spurious variation can result from sequencing error, duplicated regions and alternative splicing. Moreover, wet lab work will be required to determine allele frequencies and to discriminate SNPs restricted to single populations from 'cosmopolitan' SNPs—those found in a variety of populations.

Putting SNPs to work. The gold rush mentality towards SNP discovery and analysis stems from both their presumed potential for investigating the molecular genetic basis of complex disorders and the amenability of the SNP to automatic genotyping. The proposed approach to mapping loci using SNPs is the case–control study, which aims to identify statistically significant associations between specific alleles at a locus and defined phenotypes. The design of a genome-wide association study depends on one's assumptions regarding the variation that underlies complex traits. It is argued that *Homo sapiens* are unusual among primates in their limited genetic variation (Svante Pääbo, Max-Planck-Inst.). However, with respect to the variation that underlies disease phenotypes common to all human populations, two competing hypotheses have emerged, the validity of which dictates design of investigations.

**The Fourth International Meeting on Single Nucleotide Polymorphisms and Complex Genome Analysis Stockholm, Sweden, 10–13 October 2001.*

The common disease/common variant (CD/CV) hypothesis. The CD/CV hypothesis[1] holds that alleles that existed prior to the global dispersal of humans or those subject to positive selection represent a significant proportion of susceptibility alleles for common disease. These can be expected to confer moderate risk and occur at relatively high frequencies (>1%) in extant populations[2]. Their high frequency implies that association studies in large population cohorts will be fruitful for identifying risk alleles. Eric Lander (Whitehead Institute) cites the APOE*4 allele[3], which confers increased susceptibility to Alzheimer disease, and the CCR5-Δ32 allele, which prevents infection by HIV-1 (ref. 4) as examples of common variants causing a common phenotype in separated human populations. Under the CD/CV hypothesis, the major issue is testing the pool of common variants—either by direct assesment of each variant or by indirect testing of ancestral segments. The latter approach is limited primarily by the extent of linkage disequilibrium (LD) surrounding the causative variant. One model of human population growth predicts that significant LD around common variants in most human populations would not, on average, span more than 3 kb of the genome[5]. More recent empirical studies have given reason for greater optimism, however, with regions of 60 kb showing significant LD in a US population of northern-European descent[6]. A series of papers published in October's issue of *Nature Genetics*[7–9] shows that large genomic regions of LD interrupted by recombination hot spots are a feature of the human genome. A limited number of common haplotypes, defined by representative SNPs, would seem to account for the majority of haplotypes[9,10]. This suggests that association studies using representative SNPs should identify common haplotypes associated with increased susceptibility to disease and forms the basis of calls to construct a genome-wide haplotype map that identifies all major haplotypes and their characteristic SNPs.

The common disease/rare allele (CD/RA) hypothesis. On the other side of the fence stand the proponents of the CD/RA hypothesis, who hold that there is no reason to expect that most common genetic diseases result from common alleles. This hypothesis has recently been formalized using population modeling that predicts extensive allelic heterogeneity at disease loci[11]. Andrew Clark (Pennsylvania State Univ.) has extended this work and argues that 99.9999% of mutations underlying common diseases in humans have occurred after the explosive expansion and divergence of populations. Additionally, many researchers believe that we should expect significant locus heterogeneity in complex disease. Looking at parallels from mendelian disorders, CD/RA advocates point out the genetic and allelic heterogeneity of retinitis pigmentosa and nonsyndromic autosomal recessive deafness, 'simple' mendelian disorders with multiple known loci and a plethora of disease alleles. If this scenario holds true for common disorders, a genome-wide association study of disease in a heterogeneous population would be an exercise in futility. Joseph Terwilliger (Columbia Univ.) argues that in case–control studies using outbred populations, the search for identical-by-descent alleles has no reasonable basis. Reiterating points raised in a commentary[12] last year, Terwilliger asserts that the only means we have of minimizing genetic and allelic heterogeneity is through the use of family studies and populations with unusual histories. Thus, to the extent that rare alleles account for common diseases, the odds are reduced that haplotype maps will be useful in analyzing outbred populations, as current plans for construction of a haplotype map focus only on common alleles.

Back to biology. Whether common or rare alleles are found to confer increased risk for a particular disorder, the next step is common to all investigations: once one has identified a haplotype associated with increased susceptibility to a disease all SNPs that define the haplotype are candidates for causality. It is at this point that one has reached the limit of genetics and must turn to biology. Politics aside, elucidating the relationship between genotype and phenotype is one of the most challenging and important tasks of the 21st Century.

1. Lander, E.S. *Science* **274**, 536–539 (1996).
2. Reich, D.E. & Lander, E. S. *Trends Genet.* **17**, 502–510 (2001).
3. Corbo, R. M. & Scacchi, R. *Ann. Hum. Genet.* **63**, 301–310 (1999).
4. Martinson, J. J., Chapman, N.H., Rees, D.C., Liu, Y.T. & Clegg, J.B. *Nature Genet.* **16**, 100–103 (1997).
5. Kruglyak, L. *Nature Genet.* **22**, 139–144 (1999).
6. Reich, D.E. *et al*. *Nature* **411**, 199–204 (2001).
7. Jeffreys, A.J., Kauppi, L. & Neumann, R. *Nature Genet.* **29**, 217–222 (2001).
8. Rioux, J.D. *et al*. *Nature Genet.* **29**, 223–228 (2001).
9. Daly, M.J., Rioux, J.D., Schaffner, S.F., Hudson, T. J. & Lander, E. S. *Nature Genet.* **29**, 229–232 (2001).
10. Johnson, G.C.L. *et al*. *Nature Genet.* **29**, 233–237 (2001).
11. Pritchard, J.K. *Am. J. Hum. Genet.* **69**, 124–137 (2001).
12. Weiss, K.M. & Terwilliger, J.D. *Nature Genet.* **26**, 151–157 (2000).