WEB WATCH

Sequence data wanted!

- http://www.gene.ucl.ac.uk/nomenclature/workshop/virtual.html
- http://hgvbase.cgb.ki.se

Can you help annotate the human genome? There are ~1,500 human genes for which the chromosome location is known but the actual coding sequence has not yet been identified. The HUGO Gene Nomenclature Committee (HGNC) is now looking to the scientific community to provide the sequences of these genes. To find a list of the genes in question and to submit data, visit their Virtual Gene Nomenclature Workshop web site. The deadline for data submission is 30 June 2003. This initiative is part of the HGNC's valiant project to provide unique symbols for the estimated 30,000 human genes. Now is your chance to take part in this quest!

The Human Genome Variation Database (HGVbase) provides another opportunity for researchers to contribute data to help annotate the genome. It aims to provide a comprehensive catalogue of the variation in the human genome. All types of polymorphisms can be submitted online regardless of chromosomal location, allele frequency or phenotypic effect. The web site is curated using both manual and automated curation tools and the quality of data is checked before entry. The database is run by an EU consortium, consisting of teams from the European Molecular Biology Laboratory, Germany, the European Bioinformatics Institute, UK, and the Karolinska Institute, Sweden. It provides a resource for those interested in the effect of genetic variation on drug responses and disease susceptibility. Indeed, data from HGVbase have already been used to identify genetic polymorphisms that are associated with benzene poisoning.

Catherine Baxter

CANCER GENETICS

Making sense of missense

BRCA1 mutations have been linked to an increased risk of breast and ovarian cancer. Most of the highly penetrant alleles identified encode truncated proteins but missense alleles have also been recorded — their impact on cancer susceptibility is particularly difficult to assess because of low penetrance and the lack of adequate functional assays. In their paper, Fleming *et al.* describe an evolutionary approach to identify the missense alleles that are most likely to be associated with a disease phenotype.

The rationale behind this study is that mutations in functionally important amino acids are most likely to be associated with an increased risk of cancer. These key amino acids can be identified on the basis of their conservation in mammals, or from evidence of recent positive selection in the human lineage. To find conserved *BRCA1* regions, the authors aligned Genbank sequences for exon 11 of *BRCA1* from 57 mammals. Considering five amino-acid sites at a time, regions were defined as conserved if the first and last residues were fixed or conservative (that is, identical in all species or all residues sharing similar biochemical properties) and at least 80% of the sites were also fixed or conserved. Seven out of eight conserved regions were located in regions known to interact with other proteins. In addition, a conserved stretch of amino acids was identified in a region of unknown function that is also conserved in *BRCA1* homologues from clawed frog and chicken.

If site conservation is a good indicator of its functional importance, one would predict that missense mutations affecting fixed sites, or resulting in non-conservative substitutions at conservative sites, are most likely to be associated with a disease phenotype. In fact, 38 of the 139 documented missense alleles

HUMAN GENETICS

LD orienteering

Alleles at different loci are not always inherited independently — those in linkage disequilibrium (LD) occur together in populations more often than is expected. The extent to which LD occurs in the human genome, and how this affects variation, will determine how easy it will be to map complex-trait loci through whole-genome association studies. Now, Stumpf and Goldstein use computer simulations to examine the block-like structure of LD and conclude that it is time to abandon the general idea of an average extent of LD in the human genome.

Initial assessments of how to use LD in genetic-association studies assumed a uniform recombination rate and an idealized demographic population history. But, theory is seldom exactly like real life. There has been much evidence that recombination rates vary across the genome, sometimes resulting in a block-like genome structure. Stumpf and Goldstein set out to model the effects of recombination hotspots, as well as demography, on the extent of the LD. To assess the interactions between recombination hotspots and demography the authors used simulations of populations that undergo several bottle-necks that create LD and observed how the associations decayed under different intensities of recombination hotspots. The results show that the probability of the LD block-like structure is intimately linked with demography — severe bottlenecks delay the block-like genome structure whereas relatively high intensities of hotspots maintain it for long periods of time.

Not satisfied with simulations alone, Stumpf and Goldstein turned to real data. Having considered models of populations with different demographic histories — Europeans, Finns and Georgian Jews — the authors conclude that as a result of a strong interaction between demography and hotspots, the block-like structure of LD might be present in some