significant differences at the 5T/6T site between England and Cameroon and between England and India: a pattern that indicates that positive selection might have increased the 5T allele frequency in northern Europe. In addition, haplotype analyses showed that there was a significant excess of a common 5T haplotype in a European–American sample but not in an African–American sample: further evidence of positive selection for the 5T allele in Europe.

Of course further studies would be desirable to confirm these findings; however, in themselves these results are certainly indicative that the 5T/6T polymorphism has an important functional effect. As coronary heart disease is a late-onset disorder, it is likely that the cardiovascular phenotypes associated with the 5T/6T polymorphism are pleiotropic effects of natural selection on other conditions that are currently unknown. The MMPs have important roles in many physiological processes in reproduction and development as well as the pathogenesis of a number of diseases.⁹ It is possible that there is an interaction between the *MMP3* gene and specific European environmental factors that influences one or more of these physiological or pathological processes, and so confers a selective advantage.

Similar recent studies have detected signature of natural selection at a number of other polymorphic sites,^{3,10–13} and we

should expect to see many more studies in this area in the near future. Evolutionary analyses like these not only provide insights into the human evolutionary history and the relative contributions of neutral and advantageous mutations to the wealth of polymorphisms in the human genome today, but also help narrow down the sets of genetic variants to be examined in studies that aim to identify causes of disease risk and response to medicine ■

Acknowledgements

We thank Professor Newton Morton for valuable comments

Shu Ye is at the Human Genetics Division,
School of Medicine, University of
Southampton, Southampton, UK.
E-mail: Shu.Ye@soton.ac.uk

References

- 1 Przeworski M, Hudson RR *et al*: Adjusting the focus on human variation. *Trends Genet* 2000; **16**: 296–302.
- 2 Nielsen R: Statistical tests of selective neutrality in the age of genomics. *Heredity* 2001; **86**: 641–647.
- 3 Akey JM *et al*: Interrogating a high-density SNP map for signatures of natural selection. *Genome Res* 2002; **12**: 1805–1814.
- 4 Rockman MV *et al*: Positive selection on *MMP3* regulation has shaped heart disease risk. *Curr Biol* 2004; **14**: 1531–1539.
- 5 Ye S *et al*: Progression of coronary atherosclerosis is associated with a common genetic variant of the human stromelysin-1 promoter which results in reduced gene expression. *J Biol Chem* 1996; **271**: 13055–13060.
- 6 Medley TL, Kingwell BA, Gatzka CD, Pillay P, Cole TJ: Matrix metalloproteinase-3 genotype contributes to age-related aortic stiffening through modulation of gene and protein expression. *Circ Res* 2003; **92**: 1254–1261.
- 7 Beyzade S *et al*: Influences of matrix metalloproteinase-3 gene variation on extent of coronary atherosclerosis and risk of myocardial infarction. *J Am Coll Cardiol* 2003; **41**: 2130–2137.
- 8 Lamblin N *et al*: Polymorphisms in the promoter regions of MMP-2, MMP-3, MMP-9 and MMP-12 genes as determinants of aneurysmal coronary artery disease. *J Am Coll Cardiol* 2002; **40**: 43–48.
- 9 Sternlicht MD, Werb Z: How matrix metalloproteinases regulate cell behavior. *Annu Rev Cell Dev Biol* 2001; **17**: 463–516.
- 10 Rockman MV *et al*: Positive selection on a human-specific transcription factor binding site regulating IL4 expression. *Curr Biol* 2003; **13**: 2118–2123.
- 11 Hahn MW *et al*: Population genetic and phylogenetic evidence for positive selection on regulatory mutations at the factor VII locus in humans. *Genetics* 2004; **167**: 867–877.
- 12 Zhou G *et al*: Haplotype structure and evidence for positive selection at the human IL13 locus. *Mol Biol Evol* 2004; **21**: 29–35.
- 13 Hull J *et al*: Unusual haplotypic structure of IL8, a susceptibility locus for a common respiratory virus. *Am J Hum Genet* 2001; **69**: 413–419.

Genomics

Polymorphic landscape of the human genome

Vivian G Cheung

European Journal of Human Genetics (2005) **13**, 133–135.

doi:10.1038/sj.ejhg.5201331

Published online 10 November 2004

A cluster of recent studies show that variation in the number of copies of large-scale insertion and deletion is common in human populations.

In recent years, our understanding of the nature and the extent of genome variation has expanded greatly. With development of techniques to isolate and manipulate DNA, the focus of genetic variation has shifted from protein polymorphism to DNA sequence polymorphism.

The most common form of DNA sequence variation is single nucleotide

polymorphism (SNP) but, in addition, there are microsatellite repeat polymorphism and insertion/deletion polymorphism. SNPs and microsatellites have been very well characterized in terms of their genomic locations and their frequencies in different populations. In contrast, insertion/deletion polymorphism is less well characterized.

Inappropriate recombination events may lead to duplication and/or deletion of genomic regions, and in some cases resulting in genomic disorders such as charcot-marie-tooth and hereditary neuropathy with liability to pressure palsies (CMT1A/HNPP).^{1,2} However, until now it was not clear whether these large-scale chromosomal differences are rare events often associated with diseases or if they are more common phenomena in the genome landscape. In the last few months, several papers^{3–5} have described the identification of large (several kilobases to megabases) deletions and duplications of DNA fragments when genomes of normal individuals were compared. The identification and characterization of copy number polymorphism (CNP) show that in addition to single-nucleotide differences, genomes of unrelated individuals have large regions of thousands to millions of nucleotides that are different.

Previously, copy number determination was performed by hybridizing DNA probes onto chromosome spreads; such studies have limited resolution and would fail to detect copy number differences that are a few megabases or less in size. However, the development of microarray technology, the availability of complete human genome sequence and of mapped DNA clones corresponding to almost all segments of the genome make possible high-resolution studies of copy number. Consequently, these recent studies rely on new technologies that make it feasible to perform high-resolution copy number detection on a genome scale. Michael Wigler's group, for example, used microarrays that contain oligonucleotides representing *Bgl*II fragments of the human genome in a technique known as representational oligonucleotide microarray analysis (ROMA).^{3–6} Charles Lee's group also used microarrays but this time of BAC clones that span the genome at ~1-Mb intervals.⁴ Both groups hybridized labeled

DNA onto their microarrays and determined the copy number of genomic regions based on the hybridization intensity of each feature on the arrays. Tony Brookes *et al*⁵ took a different approach, they relied on the observation of unique SNP patterns when they genotyped markers found in duplicated and/or deleted genomic regions. All these groups validated their findings with fluorescent *in situ* hybridization and quantitative PCR. The Brookes group also examined DNA samples from complete hydatiform moles (CHM) that are monospermic, therefore homozygous at all sites. One would expect to detect one allele for all markers. However, signals from highly similar but not identical regions (such as segmental duplications and copy number polymorphisms) gave heterozygous like signals. In addition, since it is possible that these variants are not true polymorphisms but rather somatic mutations, Wigler's group also performed studies with DNA samples from multiple tissues of the same individuals and in general finding the same variants in the different tissues samples.

Despite differences in techniques, the three groups arrived at similar conclusions. They found that copy number polymorphism from several thousands to a few million nucleotides in length is common. One study, which surveyed 55 individuals, reported on average 12.4 CNPs between genomes of two unrelated individuals⁴ and the other study, which examined 20 individuals, reported on average 11 CNPs between two individuals.³ The frequency of these variants can be quite high, for instance, some of the CNPs are present in more than 10% of the individuals studied.⁴ Many (up to 56%) of them include coding regions of one or more genes. As many as one-third of the CNPs are found in known regions of segmental duplications.

Since the Wigler's and Lee's groups each reported a few hundred CNPs, and some of the CNPs were found in over 10% of the individuals studied, one might expect to find some of the same CNPs among the data from the two studies. However, as Nigel Carter points out⁷ the overlap turns out to be quite modest, there are only 11 loci (~5%) in common between the two studies. Among these are polymorphisms on 2p11.2, 4q22.3, 5q13, 14q12 and

14q32. Even some of the very high frequency alleles in one study are absent in the other. These apparent discrepancies might be due to population differences, or, as suggested in Carter's review, differences in selection and/or mapping of the probes on the arrays. The Wigler's group focused on parts of the human genome represented by *Bgl*II restriction fragments while the Lee's group studied the fragments represented by BAC clones that are distributed ~1Mb throughout the genome. Thus, it is possible that regions included in the *Bgl*II fragments were not present in the BAC clones and *vice versa*. Further studies with better genome coverage to examine the frequency and the distribution of these polymorphisms among different populations are needed. Additional validation with family material to demonstrate an inherited basis of these sequence variants would also be valuable.

What are the implications of these findings? Sequencing the human genome has prompted many technical advances and stimulated the discoveries of various genomic landmarks. One area is the development of a dense set of polymorphic markers for gene mapping. SNP and microsatellite markers are the major tools in genetic mapping studies for tracking DNA segments that are shared among related individuals. It is unlikely that CNPs will become the predominant markers used in mapping studies since relative to SNP and microsatellite markers; they are less abundant and more difficult to genotype. However, Brookes and co-workers showed one must be cautious in interpreting the genotyping results of markers that are found in duplicated and/or deleted regions of the genome. It is possible that some genotypes that were interpreted as errors might in fact be the result of deletion (or duplication) of an allele. Since several 'genomic disorders' are known to result from duplication/deletion of regions of the genome,^{1,2} a set of well annotated markers that are known to reside in such regions might enable the mapping of additional loci for genomic disorders. Such studies will elucidate why gains or losses of some regions lead to disease while others do not.

From a functional perspective, copy number differences might contribute to variation in gene expression. Copy

number differences of coding and regulatory regions are likely to affect expression of genes at the transcript and/or protein levels. For example, copy number polymorphism of the β -defensin gene cluster probably accounts for natural variation in the expression level of *DEFB4* which, in turn, might account for individual differences in immune defense.⁸ Other gene expression studies have shown that subtle differences in expression levels of genes have significant consequences. So it will be important to determine the overall effect of copy number polymorphism on gene expression, as this might significantly contribute to the variation that underlies individual differences in predisposition to complex diseases.

Advances have allowed us to examine and catalogue the polymorphic loci in our genomes, and have demonstrated at a

genotypic level how different we are from each other. With this knowledge as a foundation, we are equipped to study the effect of these differences and relate them to the complex expression of diseases and phenotypes ■

Vivian G Cheung, MD is at the
Department of Pediatrics, University of
Pennsylvania, 3516 Civic Center Blvd,
ARC 516G, Philadelphia, PA 19104, USA
and Department of Genetics, University of
Pennsylvania, Philadelphia, PA, USA.
Tel: +1 215 590 4950;
E-mail: vcheung@mail.med.upenn.edu

References

- 1 Shaw CJ, Lupski JR: Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Hum Mol Genet* 2004; **13**: R57–R64.
- 2 Emanuel BS, Shaikh TH: Segmental duplications: an 'expanding' role in genomic instability and disease. *Nat Rev Genet* 2001; **2**: 791–800.
- 3 Sebat J *et al*: Large-scale copy number polymorphism in the human genome. *Science* 2004; **305**: 525–528.
- 4 Iafrate AJ *et al*: Detection of large-scale variation in the human genome. *Nat Genet* 2004; **36**: 949–951.
- 5 Fredman D *et al*: Complex SNP-related sequence variation in segmental genome duplications. *Nat Genet* 2004; **36**: 861–866.
- 6 Lucito R *et al*: Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res* 2003; **13**: 2291–2305.
- 7 Carter NP: As normal as normal can be? *Nat Genet* 2004; **36**: 931–932.
- 8 Hollox EJ *et al*: Extensive normal copy number variation of a beta-defensin antimicrobial-gene cluster. *Am J Hum Genet* 2003; **73**: 591–600.