# It's raining SNPs, hallelujah?

Aravinda Chakravarti

*Department of Genetics and Center for Human Genetics, Case Western Reserve University School of Medicine and University Hospitals of Cleveland, Cleveland, Ohio 44106, USA. e-mail: axc39@po.cwru.edu*



In August 1974, I arrived for graduate study in the US in the midst of a raging debate. The initial availability of protein sequences in multiple species and protein polymorphisms within species raised fundamental questions regarding how sequence variation is created and maintained through evolutionary time. The world was divided into two parts: those who believed that almost all extant variation has been vetted by natural selection and those who advocated a new hypothesis—that the majority of variation has been selectively neutral throughout evolution[1]. Even ignorant graduate students were not spared: they were assigned either to the 'selectionist' or 'neutralist' camps. The arguments have dissipated with time, not because the principals have settled the score, but because, as with most debates, there has been and still is, little data to sway new recruits in either direction. The central question of how the nature and causes of *sequence* variation impact *phenotypic* variation is still largely unresolved. The impending human genome reference sequence and its variation in health and disease will bring these questions to the forefront once again, but this time, in the context of data in hand[2]. The answers will cheer the hearts of evolutionary biologists and every human geneticist who is trying to understand the nature of genetic variation underlying the common and genetically complex diseases. The study reported by Deborah Nickerson and colleagues[3] on page 233 of this issue and a companion article by Andrew Clark *et al.* appearing in August's issue of *American Journal of Human Genetics*[4], however, give reason to ponder how patterns of genetic variation and the dissection thereof are best matched. The authors of these manuscripts have determined the DNA sequence diversity in 9.7 kb of the cardiovascular disease candidate gene that encodes lipoprotein lipase (LPL) and discovered extensive molecular variation.

The rise of molecular genetics has witnessed a fascination with 'markers', polymorphisms in anonymous DNA segments, as these are ideal for tracing meioses in families, thereby gaining information by which to positionally clone rare disease genes. However, we still have scant knowledge about the extent and nature of sequence variation in human genes and adjacent regulatory signals. We understand even less of what such variation means; our grand plans to explain the genetic basis of development, disease and evolution need to be on a much firmer footing.

## Variety is the spice of DNA

Nickerson *et al.*[3] used contemporary DNA sequencing technology to obtain the nucleotide sequence of a contiguous 9.7 kb region of *LPL* in 71 individuals (24 African Americans from Jackson, Mississippi, 23 individuals of mixed European ancestry from Rochester, Minnesota, and 24 Finns). They describe 88 variants, of which 79 are single nucleotide polymorphisms (SNPs). The DNA segment studied includes exons 4 through 9 of *LPL*, is largely intronic, and contains a host of identifiable repeated sequences such as Alu, L1, MER1/2 and microsatellites. Of all variants, seven are in coding sequences, of which four (57%) alter the protein sequence. The most striking feature of these data, however, is the distribution of variant sites (Fig. 1; see also, Table 2 on page 237). Three features are evident: the total sequence diversity is 0.002, the diversity is fourfold less in coding (0.0005) than in noncoding DNA (0.0021), and twofold higher at repeated (0.0032) compared with unique (0.0017) DNA. (These numbers indicate the density of DNA variants across the map of *LPL* in terms of sequence diversity, measured as the average number of nucleotide differences per base pair between any two chromosomes. A normalized measure is obtained, as numbers are dependent both on the size of the region and the number of chromosomes surveyed[5].) Thus, variation in sequence diversity across the 9.7 kb may largely reflect sequence composition differences across this region.

Sequence diversity in any natural population is not only the consequence of mutation but also the ultimate result of evolutionary forces that shape this variation. Consequently, the nucleotide diversity map (Fig. 1) might be explained either by differences in mutation rates according to sequence composition or by differential effects of natural selection on coding, unique-noncoding and repeated DNA, or other yet unrecognized sequence features. Another major influence on the pattern of diversity is meiotic recombination. Consider a situation in which an advantageous mutation increases in frequency in a population or in which selec-

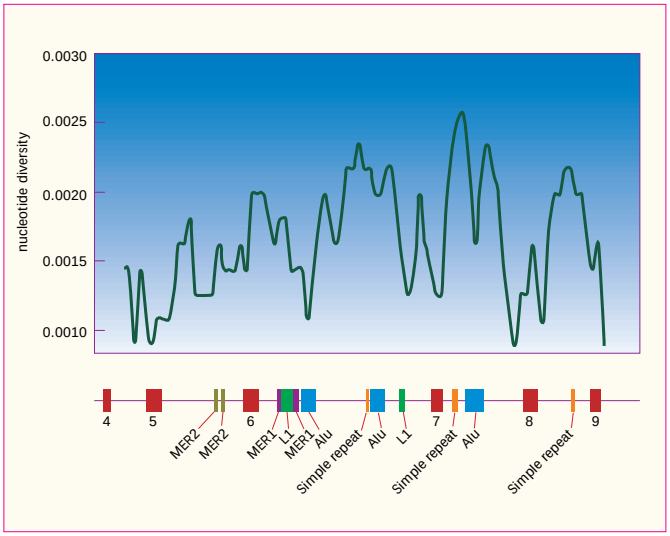


**Fig. 1** Variation in sequence diversity of *LPL versus* a feature map of the 9.7-kb segment studied. The map shows the locations of each exon and several identified repeated DNA motifs. Sequence or nucleotide diversity was measured over 1-kb segments and using a sliding window of 100 base pairs by correcting the number of variant sites for the number of samples examined[5]. Sequence diversity varies by over fourfold across this region.

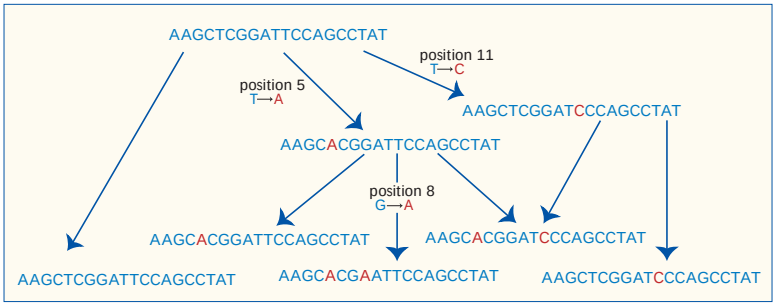


**Fig. 2** The mutual relationships of extant haplotypes varying at three nucleotide positions. A current sample of five haplotypes for a 20-bp segment is shown. These haplotypes are related by three mutation and one recombination events occurring over time. Note that extant haplotypes may be completely ancestral or derived from an ancestral sequence by simple mutation with or without recombination. Recombinant sequences bring multiple changes on a common haplotype, increasing diversity with respect to ancestral forms.

tion against a deleterious mutation decreases its frequency. In either case, neutral variation is eliminated at linked sites as a consequence of selection of one or few alleles over others, the size of the region affected in this manner (called the 'selective sweep') being determined by local recombination rates[5]. Surely crossing over across 10 kb must be infrequent? Although 10 kb represents 0.01% recombination per generation (on average), the many distinct lineages that have led to the modern *LPL* haplotypes can include very strong effects of crossing over in the past, that can be observed in haplotype data. Recombination can increase diversity in any genomic region (Fig. 2) and Clark *et al*. find substantial evidence that recombination has been as important as mutation in shaping the variation structure within the LPL gene. A likely scenario is that diversity in *LPL* was initially due to a constant rate of mutation and that recombination has limited the 'selective sweeps' to regions containing exons where selection has reduced variation, a scenario that remains in the realms of speculation without good methods for inferring the history and causes of such sequence variation.

## Plans of action

In the absence of knowledge of what each variant means, Nickerson and colleagues wonder whether we can make sense of this rich diversity: which site or collection of sites might influence variation in risk of cardiovascular or some other complex disease? How would association studies identify them? The answers to these questions depend on two different strategies recently suggested by a number of authors[2,6,7]. In the first, a catalogue of SNPs would be generated for all gene-coding sequences (cSNPs). This resource will allow direct tests of whether cSNPs are involved in the aetiology of a complex disease or not; the more complete the catalogue the greater the power of this approach. *LPL* itself has four candidate cSNPs that may be used in this way; indeed, one of them (291Asn→Ser) is known to be associated with premature atherosclerosis[3].

Not all functional changes, however, may lie within coding sequences. A second strategy would be to create a high-resolution human genome map of 100,000 (1 every ~30 kb) or more anonymous SNPs as proposed in the new five-year plan by the National Human Genome Research Institute. Nickerson *et al.* question the wisdom of this strategy as this tactic will mark genes such as *LPL* only once. Given the enormous diversity of *LPL* polymorphisms, how would one choose this SNP? Would this SNP adequately represent the variation in the gene? Probably not; Clark *et al*. show that it is unfeasible to randomly sample 3–4 SNPs that would adequately capture LPL gene diversity. The SNP-map approach, however, is not doomed, by any stretch of the polymorphic imagination. What is at issue, however, is the density of the SNP map required to derive meaningful information from it. It is currently unclear whether a 30-kb resolution is adequate for most association studies. It will probably suffice for studies of some isolated populations and would be adequate for many regions of the genome. But it is not a panacea. For regions such as the LPL gene, a much finer resolution may be required. More research is necessary to understand the pattern of sequence diversity across the genome and the degree of map resolution required for association mapping. This unexpected degree of *LPL* diversity may be misleading; we may not need to characterize all of the variation to define haplotypes that can define functional alleles. Defining cSNPs, associating them with functional differences and defining the haplotypes that would uniquely mark them will provide that answer. Until then, as is often observed, the advance of knowledge informs one of just how much one does not know; despite the wealth of SNPs that *LPL* has yielded, we are still embarassingly ignorant, although more knowledgeable, about the extent of that ignorance. □

1. Kimura, M. *The Neutral Theory of Molecular Evolution*. (Cambridge University Press, Cambridge, 1984).
2. Collins, F.C., Guyer, M.S. & Chakravarti, A. *Science* **278**, 1580–1581 (1997).
3. Nickerson, D.A. *et al. Nature Genet*. **19**, 233–240 (1998).
4. Clark, A.G. *et al. Am. J. Hum. Genet*. (in press).
5. Li, W.-H. *Molecular Evolution.* (Sinauer Associates, Sunderland, Massachusetts, 1997).
6. Lander, E.S. *Science* **274**, 536–539 (1996).
7. Risch, N. & Merikangas, K. *Science* **273**, 1516–1517 (1996).