Even small SNP clusters are non-randomly distributed: is this evidence of mutational non-independence?

William Amos*

Department of Zoology, Cambridge University, Downing Street, Cambridge CB2 3EJ, UK

Single nucleotide polymorphisms (SNPs) are distributed highly non-randomly in the human genome through a variety of processes from ascertainment biases (i.e. the preferential development of SNPs around interesting genes) to the action of mutation hotspots and natural selection. However, with more systematic SNP development, one might expect an increasing proportion of SNPs to be distributed more or less randomly. Here, I test this null hypothesis using stochastic simulations and compare this output with that of an alternative hypothesis that mutations are more likely to occur near existing SNPs, a possibility suggested both by molecular studies of meiotic mismatch repair in yeast and by data showing that SNPs cluster around heterozygous deletions. A purely Poisson process generates SNP clusters that differ from equivalent data from human chromosome 1 in both the frequency of different-sized clusters and the SNP density within each cluster, even for small clusters of just four or five SNPs, while clusters on the X chromosome differ from those on the autosomes. In contrast, modest levels of mutational non-independence generate a reasonable fit to the real data for both cluster frequency and density, and also exhibit the evolutionary transience noted for ‘mutation hotspots’. Mutational non-independence therefore provides an interesting new hypothesis that appears capable of explaining the distribution of SNPs in the human genome.

Keywords: mutation hotspot; heterozygote instability; SNP clusters; mismatch repair; human

1. INTRODUCTION

Single nucleotide polymorphisms (SNPs) are not distributed at random across the genome, but instead are clustered (Lindblad-Toh *et al.* 2000; Hellmann *et al.* 2005; Tenaillon *et al.* 2008) and associated preferentially with recombination hotspots (Lercher & Hurst 2002; Myers *et al.* 2008). This clustering is widely interpreted as reflecting mutation hotspots (Rogozin & Pavlov 2003), though the forces responsible remain poorly understood. Particularly puzzling is the observation that clusters appear to be short-lived, with many or most of those present in humans not being present in chimpanzees (Ptak *et al.* 2005; Winckler *et al.* 2005; Jeffreys & Neumann 2009). Nonetheless, the distribution of SNPs along a chromosome is often used to infer the action of natural selection (Voight *et al.* 2006; Wang *et al.* 2006; Oleksyk *et al.* 2008). Here and elsewhere, a better understanding of how and why clusters form is clearly desirable.

Besides genuine mutation hotspots, SNP clustering can arise in several ways. First, natural selection can modulate local variability along a chromosome to create non-randomness; balancing selection tends to create regions of increased variability (Charlesworth *et al.* 1997; Bubb *et al.* 2006), while purifying and directional selection tend to reduce variability (Oleksyk *et al.* 2008) and make neighbouring regions appear to have increased variability. Second, the time to most recent common ancestor (TMRCA) of genes within a population has a high variance. Consequently, each chromosome can be

thought of as a linear patchwork of the products of recombination (Hudson & Kaplan 1995; Eriksson *et al.* 2002). Some regions will have deep ancestry and carry many SNPs, while those with shallow ancestry may carry fewer. Local recombination rate determines the grain of this patchwork and hence can potentially impact on cluster distribution and size. Third, some regions of the genome are likely to be relatively refractory to mutation or simply to have received less attention during SNP development, again causing other regions to seem to carry above-average variability.

In addition to molecular processes, distribution of SNP markers can be influenced by ascertainment biases in the discovery process used to generate them (Kuhner *et al.* 2000). The primary problem relates to the non-random development of SNPs with higher-than-average levels of polymorphism, due either to a discovery process based on maximally dissimilar sequences or to the use of very few individuals (reducing the chance of finding low heterozygosity markers; Nielsen 2000). Consequently, perceived levels of heterozygosity may be inflated (Clark *et al.* 2005), and perceived patterns of population differentiation (Wakeley *et al.* 2001), linkage disequilibrium (Akey *et al.* 2003) and influence of natural selection (Soldevila *et al.* 2005) may be distorted. In response, methods of correction are being developed (Ramírez-Soriano & Nielsen 2009). The discovery process may also impact on SNP marker distribution (for example, by projects that endeavour to find all possible SNPs in a given region) and may generate some large SNP clusters. However, such regions are based on a small subset of well-characterized, usually disease-associated genes.

*w.amos@zoo.cam.ac.uk

Hence, the proportion of the genome affected is small and the overall impact on clustering will be minimal, particularly in terms of smaller clusters.

One further mechanism has recently been suggested. In yeast, it is well established that heterozygous sites are recognized during meiosis and attract gene conversion-like events (Borts & Haber 1989; Borts *et al.* 1990; Collins & Newlon 1994). Such added DNA replication might provide an added source of mutations around existing SNPs (Giver & Grosovsky 1997). Moreover, implicated enzymes in the mismatch repair pathway, such as PMS1 and PMS2, appear common to all higher organisms (Borts *et al.* 1990; Baker *et al.* 1995; Vallente *et al.* 2006). This 'heterozygote instability' (Amos 2010) has been invoked to explain why microsatellite length varies highly predictably with heterozygosity across global human populations (Amos *et al.* 2008) and why human Y-chromosome microsatellites are not longer than their chimpanzee homologues in the way autosomal loci are (Kayser *et al.* 2006), and appears consistent with the recent observation that sites carrying heterozygous deletions appear to cause locally elevated mutation rates (Tian *et al.* 2008). If polymorphic sites indeed act as a focus for gene conversion events, the extra round of DNA replication could cause SNPs to occur preferentially near to pre-existing SNPs. In other words, SNP clusters could form largely or entirely through a tendency for non-independence rather than some local factor that increases local mutation rate.

Thus, there appear three competing but non-exclusive hypotheses to explain the clustering of SNPs: (i) mutations occur more or less at random but SNPs appear clustered owing to variation in ancestry depth and the action of natural selection; (ii) SNP clusters form at genuine mutation hotspots caused by, for example, an unusual structural feature in the DNA; and (iii) SNPs attract further mutations to their vicinity through heterozygote instability. Distinguishing between these possibilities is hampered by observation biases accruing during marker development (Wakeley *et al.* 2001; Clark *et al.* 2005) and the existence of genuine mutation hotspots (Jeffreys & May 2004) that may be either the exception or the rule. Despite this, with millions of SNP markers now developed, one might expect the majority to reflect the underlying mutation pattern. I therefore decided to explore the patterns of SNP clusters generated by a semi-realistic population of chromosomes under both a random mutation process, where all mutations occur independently, and a non-independent model in which the presence of one SNP slightly increases the chance of another mutation. Both the size of any resulting clusters and their density were monitored and compared against real data from the HapMap database (see below).

2. MATERIAL AND METHODS

(a) Data

SNP data were downloaded from the HapMap website (www.hapmap.org). As representative of the human genome, I selected all non-redundant SNPs in build 36 of chromosome 1, phase II + III, genotyped in the European population CEU ($N = 314\,024$ SNPs). SNP clusters were identified by taking the ordered locations of all SNPs and

constructing clusters according to the rule that all SNPs within a cluster lie within 1 kb of another SNP on either side. In other words, a new cluster was started whenever a gap of 1 kb or more was encountered. This is an arbitrary but pragmatic rule.

Although difficult to quantify, the biggest potential problem with interpreting patterns of cluster size could be ascertainment bias. However, the SNP discovery process is largely similar across the genome (perhaps with the exception of the Y chromosome). Since the X chromosome is largely haploid in males, the non-independent model predicts that autosomal and X-chromosome SNPs will be distributed differently. Consequently, I downloaded data for all the autosomes plus the X chromosome, and analysed the cluster spectrum for each chromosome separately.

(b) Simulations

Stochastic population simulations were written in C++, using the Mersenne Twister pseudorandom number generator (R. J. Wagner, <http://www-personal.umich.edu/~wagnerr/MersenneTwister.html>). Each population size $N = 1000$ individuals was initiated with $2N$ identical, mutation-free chromosomes (length = 20 Mb), each individual carrying a single pair. At each generation, a cycle of mutation, recombination and reproduction were implemented. To improve algorithm efficiency, only individuals selected for reproduction were subjected to mutation, these individuals being chosen at random. Each new individual was the product of two randomly selected parents who each contribute one chromosome, which is itself the product of recombination between the parental homologues (rate = 1 cM per Mb, no more than one event per chromosome). Individuals were not assigned sexes and I did not prevent an individual mating with itself. In any given simulation, mutations occurred either randomly or non-independently (see below). Simulations were run for 6000 generations to approximate mutation-drift equilibrium, which is on average achieved after $4N = 4000$ generations (Hartl 1988). Mutation rate was selected to achieve the desired SNP density as often as possible, a value of 1.5×10^{-9} per base per generation being typical. Population and chromosome size were strictly limited by the maximum allowable array size of approximately 25 million elements (2000 chromosomes $\times$ 12 500 mutations per chromosome).

Non-independent mutations were modelled as follows. In each parent, the total number of sites that are heterozygous, H , was determined. A mutation then occurred either at a random location or near to a randomly selected heterozygous site with probability, p , that scales linearly with the relative sizes of these regions

$$p = \frac{H}{L/W + H},$$

where L is the length of the chromosome and W is the 'sphere of influence' of the heterozygous site, defined as the size of a symmetrically located region around the heterozygous site in which a mutation is more likely. For these simulations, I used four values of W : 0.5, 1, 2 and 4 kb. If a non-independent mutation was selected, one heterozygous site was selected at random and the location of the new mutation was decided according to a flat distribution within the sphere of influence. Thus, if the average mutation rate is μ , an individual with only one heterozygous site experiences a mutation rate of approximately 2μ in a

Figure 1. How the frequencies and densities of SNP clusters vary with cluster size under different mutation models. Data series are: large black circles, data from human chromosome 1; large white circles, simulated randomly occurring mutations; smaller symbols, simulated mutation non-independence in which mutations are more likely to occur in regions of size 0.5 (grey triangles), 1 (white squares), 2 (grey diamonds) or 5 kb (crosses) around any existing heterozygous site. Simulated data are culled from 100 replicate simulations, accepting only those in which the terminal overall SNP density was one SNP every 700 ± 25 bases, yielding approximately 60 runs per set of conditions. Frequencies are normalized to the size of human chromosome 1. For full details of the simulations (see §2). (a) How the frequencies of different cluster sizes vary; (b) how cluster density varies for the same data.

region extending $W/2$ bases either side. Here, non-independence does not change μ , it merely biases where mutations occur. No upper limit was placed on the mutability of a site, multiple neighbouring heterozygous sites acting independently and additively. Full details of the algorithm and a compiled executable are available on request.

At the end of each simulation, the distribution of ‘SNPs’ was determined by looking for sites that are polymorphic in a sample of X individuals, where X was varied from 5 to 200 in steps of 10. In this way, the total number of SNPs identified was varied, making it possible to pick a density that matches as closely as possible the density on human chromosomes. SNP clusters were defined as above (see §2a) and recorded both in terms of their density (mean separation between members within a cluster) and size (number of SNPs in the cluster). Equivalent data were collected from human chromosome 1. Since centromeres, telomeres and related features often carry few or no SNPs, thereby distorting the ‘normal’ SNP density, regions greater than 100 kb that lacked any SNPs were excluded from the analysis, yielding a final density of approximately 700 bp between SNPs.

Real SNP clusters exhibit remarkable evolutionary transience, many (or even most) clusters seen in humans not being present in chimpanzees and *vice versa*. To test whether clusters generated by mutational non-independence show similar evolutionary transience, a simulation with population size set at $N = 1000$ was run for 80 000 generations and the distribution of SNP clusters output every 4000 generations. Since, at any given neutral locus, the TMRCA of a population is on average $4N\mu$ generations, this should allow approximately 20 complete cycles in which the entire population of chromosomes is founded from a single ancestor.

3. RESULTS

The distributions of real and simulated SNP cluster sizes are presented in figure 1a. The two extremes are represented by the real data and those for simulations in which mutations occur at random (‘random SNPs’),

real SNPs tending to form fewer small and many more large clusters, the crossover point being around cluster size 15. Also, while the simulated random SNPs show an almost perfect linear decline in frequency with cluster size, real SNPs exhibit a curvilinear trend. Simulated non-independent SNPs yield intermediate trends, with sphere of influence 500 bases being similar to random and sphere of influence 5000 approximating quite closely to real SNPs.

Comparisons of how mean SNP separation within a cluster varies with cluster size in real and simulated data are summarized in figure 1b. As with cluster frequency, real SNPs and random SNPs show contrasting patterns, with random SNPs rising in mean separation up to a plateau of around 360 bases, while real SNPs reach a maximum separation of around 300 bases at a cluster size around 10, this then declining as cluster size increases further. Again, simulated non-independent SNPs reveal intermediate patterns, approximating real SNPs most closely with a large sphere of influence. For SNP density, the fit to the real data is arguably less convincing than for cluster frequency, although the largest sphere of influence does produce a peak at the same cluster size as real SNPs and also exhibits a decline in mean separation with the same slope as that of real SNPs, even if average density is slightly lower than for real SNPs.

SNPs on the X chromosome occur at about half the abundance (mean separation = 1220 bp) than on the autosomes (mean separation = 684 ± 89 bp, $n = 22$ chromosomes), an interesting observation in itself. Consequently, to compare cluster size and cluster density with the autosomes, I selected random subsets of SNPs from each autosome, reducing the mean SNP separation on each to within 1 bp of the X-chromosome value. Comparisons between the X, the autosomes and simulated independent mutations for both cluster size and cluster density are given in figure 2. In terms of cluster size frequency, the X and the autosomes are virtually indistinguishable, but both have fewer small and more large clusters compared with a Poisson process.

Figure 2. How the frequencies and densities of SNP clusters vary with cluster size between the X chromosome and the autosomes, with simulated random data for comparison. Data series are: large white circles, simulated randomly occurring mutations; large black circles, the X chromosome; grey diamonds, mean value for the autosomes, calculated for each chromosome separately and then averaged. Error bars are one standard error of the mean. Simulated data are culled from 100 replicate simulations, accepting only those in which the terminal overall SNP density was one SNP every 1220 ± 100 bases, yielding approximately 40 runs. For full details of the simulations see §2. (a) How the frequencies of different cluster sizes vary; (b) how cluster density varies for the same data. At all but six cluster sizes above four, the X-chromosome density is lower than for the autosomes.

Figure 3. How the locations of SNP clusters vary over time. A single simulation was run for 80 000 generations. At every $4N = 4000$ generations, the locations of clusters containing 25 or more SNPs were recorded and plotted on a separate line. Every individual carries a single pair of chromosomes, each 20 Mb long. For maximum comparability, SNP density was held as close as possible to 1 every 700 bases by varying the number of individuals on which SNPs were ascertained. A more or less random pattern is seen with little or no evidence of prolonged stability over time.

In terms of cluster density, clusters on both the X and the autosomes are much denser than those produced by simulation. However, even though the overall frequency of SNPs per kilobase is the same, clusters on the X chromosome are consistently denser than equivalent-sized clusters on the autosomes.

The locations of SNP clusters and how they vary over time are presented in figure 3. Each horizontal line represents the simulated chromosome and dots indicate the centre of each SNP cluster carrying 25 or more SNPs. The overall pattern appears random, with no obvious tendency for any given cluster to be preserved over multiple consecutive time slices. On average, only 7 per cent of clusters containing greater than 25 SNPs lie within 10 kb of a similarly sized cluster 4000 generations later (range = 0–13%). Quantifying transience is not easy, but since some coincidence is expected by chance, $4N$

generations appear long enough for most clusters to emerge, grow and be lost.

4. DISCUSSION

I have analysed the size and density distribution of SNP clusters in the human genome, and also in semi-realistic stochastic simulations in which mutations either occur at random or are biased to occur near pre-existing mutations. Even for cluster sizes as small as three or four, the frequency and density of clusters generated by a Poisson process fail to match those seen on real chromosomes. In contrast, when some level of mutational non-independence is introduced, much-improved fits for both cluster frequency and SNP density within clusters can be achieved. Clusters on the X differ in density, but not frequency, from those on the autosomes even when

the number of SNPs per kilobase is normalized. SNP clusters formed by mutational non-independence also exhibit the evolutionary transience seen for real SNPs.

A simulated Poisson mutation process generates SNP clusters that differ in both frequency and mean SNP-to-SNP distance from real human SNPs. Given the number of possible factors that may influence the distribution of SNPs, I take a pragmatic approach and assume that most of the genome can be classified as one of three classes: (i) regions where SNPs are largely absent; (ii) clusters formed through natural selection, mutation hotspots and/or related factors; or (iii) all other regions. I assume that the majority of SNPs in clusters containing 10 or fewer SNPs fall into class (iii) and are affected rather little by strong selection, mutation hotspots, etc., and use simulations to show that such SNPs are not distributed as they would be if mutations occur randomly.

It could be argued that my assumptions are wrong and that the observed difference between small clusters on real and simulated chromosomes are due to some form of observation bias (Kuhner *et al.* 2000; Ramirez-Soriano & Nielsen 2009), fine-scale variation in recombination rates (Ptak *et al.* 2005), the isocore structure of the genome (Clark *et al.* 2005) or other factors (Chen *et al.* 2009). As a further test, I therefore compared SNP clusters on the X with those on the autosomes, finding both a large difference in overall SNP frequency and, when frequencies were normalized, a tendency for denser small clusters to occur on the X. Since an observation bias should operate similarly on the X chromosome and the autosomes, any unusual clustering of SNPs generated by an interaction between randomly distributed mutations and an observation bias should create very similar patterns on both chromosome types. That it does not provides support for the notion that mutations on one or both chromosome classes occur non-randomly. Of course, the sex chromosomes differ in many ways from the autosomes, not just heterozygosity. Nonetheless, the observed clustering differences are consistent with the idea that heterozygosity and clustering are somehow related, and provide a feature that needs to be accounted for by alternative models. Finally, it should be pointed out that the dominant form of observation bias impacts mainly on the distribution of major allele frequencies and should not affect the clustering of SNPs unless SNPs with similar allele frequencies tend to occur near each other; so, while the possibility of observation bias should not be discounted, if a bias does operate it is probably not the one discussed most in the literature (Nielsen 2000; Clark *et al.* 2005; Soldevila *et al.* 2005).

In contrast to randomly placed mutations, mutations that occur preferentially near to one another tend to produce a better—if still imperfect—fit to the real data, both in terms of the frequency spectrum of different cluster sizes and the density of the SNPs within a cluster. Indeed, at the largest sphere of influence size of 5 kb, the fit to real data is rather good on both counts. Trials with larger spheres on influence do not yield an appreciable improvement on this (data not shown). Naturally, generating a good match between simulations and real data by no means proves that similar processes produced both patterns, but these simulations do at least suggest that a non-independent process is both consistent with reality and capable of explaining the origin of even the

largest SNP clusters. Interestingly, the cluster size–frequency profile of real SNPs seems part of a continuum of states, with a smooth profile across all cluster sizes. Such an even pattern would seem unlikely if SNP clusters were generated by a mixture of several more or less unrelated processes (such as a Poisson process) and a more active process based around mutation hotspots, and suggests more that a single process is responsible for creating a majority of all SNP clusters.

It is arguably trivial to show that when new mutations occur near pre-existing mutations they tend to create clusters. For this reason, the match in both frequency and density across a wide range of cluster sizes is important. There are arguably many ways to produce some level of clustering, but it is much more challenging to produce a good fit in all aspects. Indeed, I deliberately kept the simulations as simple as possible so that any fit did not appear overly contrived. Having said this, there are many possible ways by which the actual fit could be improved. First, I used a uniform probability distribution for the sphere of influence, when it might seem more natural to use, for example, a Gaussian distribution such that mutation probability is greatest immediately adjacent to the existing SNP. Second, I assumed that neighbouring SNPs interact additively when any real interaction could be quite complicated, almost certainly including a maximum mutation rate above which further SNPs do not cause a further increase. Third, I did not include known elements of reality but, rather, largely unquantified ones, such as the action of various forms of selection or mutagenic sequences. Introducing these elements will be the subject of future research.

A persistently puzzling feature of real SNP clusters is their evolutionary transience (Ptak *et al.* 2005; Winckler *et al.* 2005). If a region of DNA has a structure such that it attracts a higher mutation rate, why should this change over relatively short tracts of time? Naturally, any one ‘hotspot’ may change, but comparisons between humans and chimpanzees suggest many or even most do. However, mutational non-independence seems to capture this transience well. Such transience appears at odds with the idea that clusters of mutations keep attracting ever more mutations: intuitively this should lead to long-term stability. However, heterozygosity is a feature of a chromosome pair, not a single chromosome. SNP clusters are generally ascertained, both in life and in the simulations, by assaying tens of chromosomes to find polymorphic sites. Consequently, any single chromosome will carry far fewer substitutions than the number of SNPs recorded for the population in which it exists. In other words, SNP clusters are much more a feature of the population than an individual and can be viewed as an emergent property of the mutation process rather than as a directly heritable trait.

In conclusion, most SNPs in the human genome reside near other SNPs to form small clusters. These small clusters differ in both their size–frequency profile and their mean SNP–SNP separation compared with simulations based on randomly distributed mutations. However, a good match is produced to a model in which mutations occur preferentially near heterozygous sites. Such a pattern could be expected from the gene conversion events that occur during meiosis, is consistent with observations that heterozygous deletions raise the local mutation rate

(Tian *et al.* 2008), accounts for the evolutionary transience of SNP clusters, helps explain the difference in cluster density between the X and the autosomes, and produces an interesting new model for how point mutations accumulate in the genome. If this model proves to be correct, then it has implications for many areas of evolutionary biology. In a practical sense, mutational non-independence would undermine a key assumption in phylogenetic reconstruction; namely that mutations occur randomly and independently, and will tend to create spurious clusters that might otherwise imply balancing selection. Interestingly, mutational non-independence could actually benefit organisms in which it occurs, because mutations will be attracted towards regions that are already polymorphic and will occur less in genomic regions that are monomorphic, corresponding broadly to regions where mutations are more and less likely to be beneficial, respectively. This mechanism should be particularly effective at increasing the generation of diversity at highly polymorphic immune-related genes such as the major histocompatibility complex. Elucidating exactly how mutational non-independence operates in real genomes will provide an exciting challenge for future research.

REFERENCES

- Akey, J. M., Zhang, K., Xiong, M. & Jin, L. 2003 The effect of single nucleotide polymorphism identification strategies on estimates of linkage disequilibrium. *Mol. Biol. Evol.* **20**, 232–242. (doi:10.1093/molbev/msg032)
- Amos, W. 2010 Heterozygosity and mutation rate: evidence for an interaction and its implications. *BioEssays* **32**, 82–90.
- Amos, W., Flint, J. & Xu, X. 2008 Heterozygosity increases microsatellite mutation rate, linking it to demographic history. *BMC Genet.* **9**, 72. (doi:10.1186/1471-2156-9-72)
- Baker, S. M. *et al.* 1995 Male mice defective in the DNA mismatch repair gene PMS2 exhibit abnormal chromosome synapsis in meiosis. *Cell* **82**, 309–319. (doi:10.1016/0092-8674(95)90318-6)
- Borts, R. H. & Haber, J. E. 1989 Length and distribution of meiotic gene conversion tracts and crossovers in *Saccharomyces cerevisiae*. *Genetics* **123**, 69–80.
- Borts, R. H., Leung, W.-Y., Kramer, W., Kramer, B., Williamson, M., Fogel, S. & Haber, J. E. 1990 Mismatch repair-induced meiotic recombination requires the *PMS1* gene product. *Genetics* **124**, 573–584.
- Bubb, K. L. *et al.* 2006 Scan of human genome reveals no new loci under ancient balancing selection. *Genetics* **173**, 2165–2177. (doi:10.1534/genetics.106.055715)
- Charlesworth, B., Nordberg, M. & Charlesworth, D. 1997 The effects of local selection, balanced polymorphism and background selection on equilibrium patterns of genetic diversity in subdivided populations. *Genet. Res. Camb.* **70**, 155–174. (doi:10.1017/S0016672397002954)
- Chen, J.-Q., Wu, Y., Yang, H., Bergelson, J., Kreitman, M. & Tian, D. 2009 Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol. Biol. Evol.* **26**, 1523–1531. (doi:10.1093/molbev/msp063)
- Clark, A. G., Hubisz, M. J., Bustamante, C. D., Williamson, S. H. & Nielsen, R. 2005 Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res.* **15**, 1496–1502. (doi:10.1101/gr.4107905)
- Collins, I. & Newlon, C. S. 1994 Meiosis-specific formation of joint DNA molecules containing sequences from homologous chromosomes. *Cell* **76**, 65–75. (doi:10.1016/0092-8674(94)90173-2)
- Eriksson, A., Haubold, B. & Mehlig, B. 2002 Clustering of SNPs along a chromosome: can the neutral model be rejected? (<http://arxiv.org/abs/physics/0207024>)
- Giver, C. R. & Grosovsky, A. J. 1997 Single and coincident intragenic mutations attributable to gene conversion in a human cell line. *Genetics* **146**, 1429–1439.
- Hartl, D. L. 1988 *A primer of population genetics*. Sunderland, MA: Sinauer.
- Hellmann, I., Prüfer, K., Ji, H., Zody, M. C., Pääbo, S. & Ptak, S. 2005 Why do human diversity levels vary at a megabase scale? *Genome Res.* **15**, 1222–1231. (doi:10.1101/gr.3461105)
- Hudson, R. R. & Kaplan, N. L. 1995 The coalescent process and background selection. *Phil. Trans. R. Soc. Lond. B* **349**, 19–23. (doi:10.1098/rstb.1995.0086)
- Jeffreys, A. J. & May, C. A. 2004 Intense and highly localized gene conversion activity in human meiotic crossover hotspots. *Nat. Genet.* **36**, 151–156. (doi:10.1038/ng1287)
- Jeffreys, A. J. & Neumann, R. 2009 The rise and fall of a human recombination hotspot. *Nat. Genet.* **41**, 625–629. (doi:10.1038/ng.346)
- Kayser, M., Vowles, E. J., Kappe, D. & Amos, W. 2006 Microsatellite length differences between humans and chimpanzees at autosomal but not Y-chromosomal loci. *Genetics* **173**, 2179–2186. (doi:10.1534/genetics.106.055632)
- Kuhner, M. K., Beerli, P., Yamato, J. & Felsenstein, J. 2000 Usefulness of single nucleotide polymorphism data for estimating population parameters. *Genetics* **156**, 439–447.
- Lercher, M. J. & Hurst, L. D. 2002 Human SNP variability and mutation rate are higher in regions of high recombination. *Trends Genet.* **18**, 337–340. (doi:10.1016/S0168-9525(02)02669-0)
- Lindblad-Toh, K. *et al.* 2000 Large-scale discovery and genotyping of single-nucleotide polymorphisms in the mouse. *Nat. Genet.* **24**, 381–386. (doi:10.1038/74215)
- Myers, S., Freeman, C., Auton, A., Donnelly, P. & McVean, G. 2008 A common sequence motif associated with recombination hot spots and genome instability in humans. *Nat. Genet.* **40**, 1124–1129. (doi:10.1038/ng.213)
- Nielsen, R. 2000 Estimation of population parameters and recombination rates from single nucleotide polymorphisms. *Genetics* **154**, 931–942.
- Oleksyk, T. K., Zhao, K., de la Vega, F. M., Gilbert, D. A., O'Brien, S. J. & Smith, M. W. 2008 Identifying selected regions from heterozygosity and divergence using a light-coverage genomic dataset from two human populations. *PLoS ONE* **3**, e1712. (doi:10.1371/journal.pone.0001712)
- Ptak, S., Hinds, D. A., Koehler, K., Nickel, B., Patil, N., Ballinger, D. G., Przeworski, M., Frazer, K. A. & Pääbo, S. 2005 Fine-scale recombination patterns differ between chimpanzees and humans. *Nat. Genet.* **37**, 429–434. (doi:10.1038/ng1529)
- Ramírez-Soriano, A. & Nielsen, R. 2009 Correcting estimators of t and Tajima's D for ascertainment biases caused by single-nucleotide polymorphism discovery process. *Genetics* **181**, 701–710. (doi:10.1534/genetics.108.094060)
- Rogozin, I. B. & Pavlov, Y. I. 2003 Theoretical analysis of mutation hotspots and their sequence context specificity. *Mut. Res.* **544**, 65–85. (doi:10.1016/S1383-5742(03)00032-2)
- Soldevila, M. F., Calafell, F., Helgason, A., Stefansson, K. & Bertranpetit, J. 2005 Assessing the signature of selection

- in PRNP from polymorphism data: results support Kreitman and Di Rienzo's opinion. *Trends Genet.* **21**, 389–391. (doi:10.1016/j.tig.2005.05.001)
- Tenaillon, M. I., Austerlitz, F. & Tenaillon, O. 2008 Apparent mutational hotspots and long distance linkage disequilibrium resulting from a bottleneck. *J. Evol. Biol.* **21**, 541–550. (doi:10.1111/j.1420-9101.2007.01490.x)
- Tian, D. *et al.* 2008 Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**, 105–109. (doi:10.1038/nature07175)
- Vallente, R. U., Cheng, E. Y. & Hassold, T. J. 2006 The synaptonemal complex and meiotic recombination in humans: new approaches to old questions. *Chromosoma* **115**, 241–249. (doi:10.1007/s00412-006-0058-4)
- Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. 2006 A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72. (doi:10.1371/journal.pbio.0040072)
- Wakeley, J., Nielsen, R., Liu-Cordero, S. N. & Ardlie, K. 2001 The discovery of single-nucleotide polymorphisms—and inferences about human demographic history. *Am. J. Hum. Genet.* **69**, 1332–1347. (doi:10.1086/324521)
- Wang, E., Kodama, G., Balbi, P. & Moyzis, R. K. 2006 Global landscape of recent inferred Darwinian selection for *Homo sapiens*. *Proc. Natl Acad. Sci. USA* **103**, 135–140. (doi:10.1073/pnas.0509691102)
- Winckler, W. *et al.* 2005 Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308**, 107–111. (doi:10.1126/science.1105322)