

# Genome-wide variation in the human and fruitfly: a comparison

## Charles F Aquadro\*, Vanessa Bauer DuMont and Floyd A Reed

Average levels of nucleotide diversity are ten-fold lower in humans than in the fruitfly, *Drosophila melanogaster*. Despite this difference, apparently as a result of a lower population size, patterns of genomic diversity are strikingly similar in being correlated with local rates of recombination, and influenced by similar interactions between positive natural selection and recombination. Both species also show lower levels of variation on average in non-African compared to African populations, reflecting a similar evolutionary history and perhaps both natural selection and founder effects in new environments.

### Addresses

Department of Molecular Biology and Genetics, Biotechnology Building, Cornell University, Ithaca, New York 14853, USA  
\*e-mail: cfa1@cornell.edu

**Current Opinion in Genetics & Development** 2001, 11:627–634

0959-437X/01/\$ – see front matter  
© 2001 Elsevier Science Ltd. All rights reserved.

### Abbreviation

**MHC** major histocompatibility complex

### Introduction

Studies of genomic diversity got a head start in the fruitfly, *Drosophila melanogaster*, fueled in part by this species' relatively small, compact genome, prevalence of single-copy genes, and a wealth of genetic data that could be linked to a physical polytene chromosome map. Early glimpses of genetic variation in humans were often motivated by an interest in human history or the study of particular genes underlying human disease. A general picture of the landscape of genomic diversity for *D. melanogaster* began to emerge by the mid-1990s. These results motivated a significant body of both theoretical and experimental results that have significantly enhanced our understanding of the relative contributions and interactions of selective, mutational, and demographic forces in shaping this genetic variation. These results also provide a rich framework in which to interpret the rapidly emerging picture of human polymorphism which has been driven by the Human Genome Project, commercial efforts, and motivation provided by medicine and pharmaceuticals. Here, we briefly review and compare the general features of genome diversity in humans and *D. melanogaster*. There are striking similarities, and differences, in the levels and patterns of observed variation. Some can be accounted for by common evolutionary processes, some by fundamental differences between the two species.

### Measures and determinants of genetic variation

Several statistics are available for measuring genetic variation (see Przeworski *et al.* [1] for a brief summary). For the purpose of comparisons across gene regions and organisms, nucleotide variation is often summarized as

nucleotide diversity,  $\pi$ , the average probability that two nucleotides will differ between two randomly chosen sequences. Not only does nucleotide diversity provide a summary of variation, as a function of both the number of segregating (variant) sites and their frequencies in the population, but, at mutation drift equilibrium,  $\pi$  is also a direct estimate of  $4N_e\mu$  commonly symbolized as  $\theta_\pi$ , where  $N_e$  is the effective population size and  $\mu$  is the per generation mutation rate. The product  $4N_e\mu$  ( $= \theta_s$ ) can also be estimated from the total number of segregating variant sites in a sample, assuming a steady state allele frequency distribution resulting from a balance of the introduction of new alleles by mutation and their removal by drift. The comparison of these two estimates of  $\theta$  leads to a useful statistic to detect selective and demographic events through their disruption in mutation drift equilibrium [2].

A species with a large effective population size retains more variability than a species with a small population size. Differences in mutation rate can have similar effects. Regions of the genome with high mutation rates will have more variation both within, and between, species, than those regions with low mutation rates. Therefore, differences in variability between species can be a result of differences in either the mutation rate or the effective population size or a combination of the two factors.

The introduction of natural selection, where different variants result in differential contributions to future generations, will modulate levels of variation estimated by  $\theta_s$  or  $\pi$  away from the mutation-drift equilibrium level of variation. A new advantageous mutation can rapidly sweep through a population to fixation eliminating all linked variation in a process termed a 'selective sweep'. In contrast, balancing selection can maintain variation in a population beyond its expected time to fixation or loss by drift. The magnitude of these effects of selection on variation in regions of the genome depends on the degree of independence between adjacent nucleotides measured as rates of recombination per physical distance (e.g. cM/Mb). For example, directional selection at one site can fix a region of the chromosome around the selected site proportional to the strength of selection,  $s$ , and the rate of recombination,  $c$  [3]. With higher recombination, the region affected by the selective sweep will be smaller. In contrast, a larger 'footprint' of selection is predicted in areas of lower recombination. Balancing and other forms of diversity enhancing selection will lead, over a sufficiently long period of time, to greater variation than expected under neutrality in a window of size proportional to the local rate of recombination.

Within the last decade, the theoretical impact of background selection as a result of the inefficient removal of weakly deleterious mutations by natural selection has

also come to be appreciated. The predicted effect of background selection depends on the deleterious mutation rate and the strength of selection on those mutants. The selective removal of these mutations can lead to a regional reduction in  $N_e$ , and therefore also a proportional reduction in nucleotide diversity, particularly in regions of low recombination [4].

Thus, we can think of the effect of most forms of selection on linked neutral variation as a result of a reduction (or increase) in the regional effective population size. As a consequence, this effect only influences levels of linked neutral variation within, and not between, species, and thus can be distinguished from mutation rate differences.

In order to interpret the effect of selection on a region of the genome, estimates of recombination rates from integrated genetic and physical maps are required. Average rates of recombination per Mb are similar for both *D. melanogaster* and humans. In humans, recombination rates average 1.3 cM/Mb but varies from 0 to 8.8 cM/Mb across the genome [5•]. For flies, the average is 1.5 cM/Mb and varies from 0 to 5.3 cM/Mb [6]. Humans also have a particularly heterogeneous recombinant landscape with frequent large changes in recombination rates as one moves along the chromosome [5•,7], in contrast with *D. melanogaster*'s relatively smooth 'recombinational landscape' [8,9]. This contrast is consistent with documented hot spots of recombination in humans (e.g. upstream of  $\beta$ -globin), but none in *D. melanogaster*. It is also important to note that these estimates are largely a measure of crossing-over; the distribution of rates of gene conversion remain largely unknown for both organisms. Finally, recombination rates are not static, but are variable both within and between species [10,11].

### Levels and types of genetic variation

The study of genetic variation in *D. melanogaster* has come largely from the analysis of individual genes or regions using complete sequencing of alleles sampled from natural populations throughout the world-wide range of the species. Sample sizes have generally been on the order of a few to ten or so chromosomes per locality, with the US being most intensively studied. Increasingly, additional locations including Ecuador, Europe, China, Australia, and, importantly, Africa are being included. In contrast, the study of human genetic diversity has been a more mixed combination of sampling and detection strategies. For nucleotide polymorphism, early (but remarkably accurate) estimates of average genomic diversity came from the comparison of cDNA sequences available in GenBank [12], with no attention to sampling. A limited but informative number of complete sequencing studies of introns have focused on diverse samples of modern humans [13]. Recently estimates have become available from comparisons of whole-genome sequences of several individuals [14,15•,16], to targeted complete sequencing of gene regions of candidate loci of human disease, phenotype or

pharmacogenetic interest [17–25,26•] many in large samples of individuals representing diverse geographic locations and/or ethnically defined populations. We focus our discussion on estimates of nucleotide diversity in humans that are likely to be unbiased by method of detection or screening of known polymorphic sites.

There is a roughly ten-fold difference in average nucleotide diversity between humans and *D. melanogaster*. It is striking that the average nucleotide diversity between *D. melanogaster* individuals (0.011 for non-coding regions) encompasses not only the average non-coding nucleotide polymorphism within modern humans (0.001), but is on the order of the average divergence (0.01–0.02) between humans, chimpanzees and gorillas [14,16,27,28•,29,30].

These averages belie a wide range of levels of diversity, however: for *D. melanogaster*, total (coding and noncoding) diversity ranges from 0–0.010, synonymous diversity from 0–0.032, and non-coding diversity from 0–0.028 [27,28•]. Humans also show a wide range of nucleotide diversity, ranging for noncoding introns for example from 0–0.0015 (for *Lpl*, *PDHAI*, and an intron of *Dmd* [1,18,31,32]). Nucleotide diversity in the MHC region in humans is perhaps the highest in either species (>0.10), but this high diversity is clearly associated with uniquely strong balancing selection [33].

As nucleotide diversity is a product of both mutation and effective population size, the order of magnitude difference in average nucleotide diversity between *D. melanogaster* and humans begs the question, which of these factors are different? Available evidence indicates the nucleotide mutation rate per generation appears to be approximately equal between humans and flies [34]. The ten-fold difference in average nucleotide diversity between the two species appears to be as a result of a roughly ten-fold larger long-term effective population size in *D. melanogaster* compared to modern humans ( $N_e \approx 300,000$  versus 20,000, respectively [13,35]).

Results from most genes also indicate more nucleotide diversity in African, compared to non-African populations for both flies and humans [28•,36]. For example, the average nucleotide diversity for *D. melanogaster* is 0.0063 for total coding regions in African samples and 0.0049 in non-African, and average synonymous sites diversity is 0.0184 in African and 0.0142 in non-African samples [28•]. For human African versus European–American samples, these contrasts are 0.00068 versus 0.00054, and 0.0012 versus 0.0009, respectively [36]. Other studies give similar results on average [1,26•]. Although this lower variation has been attributed largely to founder effects in the expansion of both species out of their ancestral ranges in Africa, there is some evidence that selection in the new, non-African environments could have contributed to the lower variation at some genes in *D. melanogaster* [28•].

An interesting trend emerges when we contrast estimates of diversity for nonsynonymous (amino acid changing) versus synonymous changes (Table 1). Assuming for the moment that synonymous changes are selectively neutral, the ratio of these two types of variants provides a measure of the amount of constraint on amino acid changes. For flies, this ratio is 0.049 (mean  $\theta_s$  values compiled for 19 genes from the literature [V Bauer DuMont, unpublished data]). This would imply that in flies, 95% (= 1 minus 0.049) of the nonsynonymous mutations are deleterious and removed by selection. In contrast, for humans this ratio is 0.358 [36], more than seven times higher than in flies, suggesting that 64% of the nonsynonymous mutations are deleterious. Similar results have been observed by Halushka *et al.* [36], Stephens *et al.* [26•] and others.

Could this difference be a result of more efficient selective constraint in flies? One approach is to compare the frequency distribution of different types of variants by comparing the ratios of diversity based on  $\theta_s$  to those based on  $\pi$ , which takes into account the frequency of the variants. For  $\pi$ , these ratios are 0.041 versus 0.258 for flies and humans, respectively (V Bauer DuMont, unpublished data; [37]). Table 1 illustrates nearly identical ratios for  $\pi$  and  $\theta_s$  in flies, versus a lower ratio for  $\pi$  compared to  $\theta_s$  for humans. One interpretation of this contrast is that there are more amino acid variants that behave as mildly deleterious (and thus segregate as rare variants) in humans yet are selectively removed in flies. This is compatible with less effective selection resulting from a smaller effective population size in humans in which genetic drift can overpower weak selection. This fits with the estimation by Fay *et al.* [38•] that 80% of amino acid variants are deleterious, with 20% being very slightly so (see Fay and Wu, this issue, pp 642–646). A lower  $N_e$  may not alone account for the relatively higher level of non-synonymous polymorphism in humans. Other factors such as reduced constraint as a result of a greater redundancy associated with multi-gene families in humans need to be investigated. We must also be very cautious in our assumption of synonymous sites as neutral (as discussed below).

Given the order of magnitude lower nucleotide diversity (and inferred population size) in humans compared to *D. melanogaster*, it is striking to note that variation for simple sequence repeats (in particular, dinucleotide ‘microsatellites’) is comparable, on average between humans and flies ( $4N_e\mu$  estimates of 5 and 8, respectively assuming a stepwise mutation model; [35,39]). However, this appears to be due to a significantly higher slippage mutation rate in humans compared to *D. melanogaster* [35] which appears to compensate for the approximately order of magnitude lower effective population size in humans.

### Variation and recombination

In both flies and humans, nucleotide diversity correlates positively with local rate of recombination (e.g. [9,40] for flies, and [1,13,41••] for humans). The correlation does not

**Table 1**

Ratio of nonsynonymous to synonymous nucleotide diversity.		
	Fruitfly	Human
$\theta_s$	0.0006 / 0.0122 = 0.049	0.000359 / 0.001003 = 0.358
$\pi$	0.0005 / 0.0122 = 0.041	0.000275 / 0.001067 = 0.258

appear to be explained by mutational differences across the recombinational landscape. Resolving the contributions of positive selection versus background selection continues to be a difficult problem. For *D. melanogaster* from Africa, Andolfatto and Przeworski [42•] show an excess of rare variants in low recombination, which suggests a dominant role of hitchhiking. A similar, though not significant, trend is seen for humans [41••]. The comparison of X-chromosome to autosome variability [9,28••,43•] also supports a role for hitchhiking in non-African populations of *D. melanogaster* (and *D. simulans*). Payseur and Nachman [44] argue that a lack of a correlation between variation at rapidly mutating microsatellites and recombination in humans is further support against background selection as the dominant influence in shaping the correlation. However, the effect of mutational rate and process heterogeneity for microsatellites [45], and a large variance in estimates of microsatellite variation, may obscure any correlation. The observation of a positive correlation between microsatellite variation and recombination in one study of *D. melanogaster* may be associated with a lower slippage rate in flies compared to humans [35]. The growing number of instances of positive selection (summarized below) clearly indicate that selective sweeps and associated hitchhiking can occur.

### Levels and patterns of linkage disequilibrium

While the issue of linkage disequilibrium is discussed at length elsewhere in this issue (see Wall, this issue, pp 647–651), it is worth, in our comparative context, noting one particularly interesting contrast to emerge from the studies of genomic diversity in humans and flies. In both species, levels of linkage disequilibrium tend to decay with distance between the sites being compared (as expected by basic theory). There is a great deal of variation in the size of the regions showing linkage disequilibrium, however. This appears to be due both to differences in rates of recombination in different parts of the genome but also to the impact of recent selection in some cases [46,47,48•]. Demography and population structure probably also has contributed to heterogeneity in this pattern for both humans and flies [49] and the higher average levels seen compared to simple predictions [50].

A striking contrast between humans and flies emerges from comparisons of local linkage disequilibrium (in regions <10 kb in length). Andolfatto and Przeworski [51] report a general excess of linkage disequilibrium across loci surveyed for variability in *D. melanogaster*. Estimates of

recombination rate relative to mutation rate ( $c/u$ ) based on observed levels of linkage disequilibrium tend to be lower than  $c/u$  estimates based on laboratory crosses. This pattern suggests that there is less evidence of recombination (more linkage disequilibrium) than expected given the direct laboratory estimates at these loci. Humans show the opposite pattern. Although there is an excess of long-distance linkage disequilibrium, there is an apparent deficiency of short-distance (<10 kb) linkage disequilibrium compared to the expectation from empirical rates of recombination and a population size of 10,000 [18,32,49,52].

Simple demographic models with migration rates that have been estimated for *D. melanogaster* are shown not to lead to the pattern observed [51]. Andolfatto and Przeworski [51] conclude that the excess linkage disequilibrium in *D. melanogaster* is largely a consequence of past and current natural selection acting across the genome. The deficiency of local linkage disequilibrium in humans may be resolved by the inferred occurrence of surprisingly frequent tracts of gene conversion [53]. Gene conversion may also explain the surprisingly low level of linkage disequilibrium in two regions of very low recombination on the *D. melanogaster* X chromosome [54].

### Contrast of variability between the X-chromosome and autosomes

Less variation is observed on the X-chromosome in North American compared to African samples of *D. melanogaster*. In contrast, the level of variability on the autosomes is similar between African and non-African populations [28•,55]. These data suggest that a simple bottleneck hypothesis may not be sufficient to explain the previous observation of less variability in non-African populations [56]. Although the effect of inversions and corrections for X-chromosome effective population size [28••] needs more investigation, one explanation for these results is that selective sweeps have been frequent within the derived non-African populations (see Andolfatto, this issue, pp 635–641).

In humans, average levels of nucleotide variation are also lower on the X-chromosome compared to autosomes [15••,26•,41••]. A lower effective population size for X-chromosomes does not appear to be sufficient in itself to account for this lower variation. Consideration of the lower mutation rate on X chromosomes, due to male driven molecular evolution (apparently lacking in flies [57]), brings them closer, but the X is still slightly less variable than expected. For example, Stephens *et al.* [26•] estimate average heterozygosity per site for the X as 0.00045 and for autosomes as 0.00096, a two-fold difference. We can adjust the X values as follows: multiply by  $4/3$  to account for the effective population size difference, and by 1.29 to account for the lower mutation rate on the X (based on a five-fold higher rate in males [58]), leading to a 'corrected' X estimate of 0.00077. The remaining 2.5-fold difference between X and autosome variation may simply be due to

chance or to errors in correction. However, as argued for *Drosophila*, it may also represent the consequence of selective sweeps, which will be more efficient in hemizygous males, and thus lead to a greater hitchhiking effect on the X chromosome [7,43•,56]. Comparison of genes with equivalent rates of recombination will be required to evaluate this relationship further.

### Codon bias and GC content

Difference in  $N_e$  across a species' genome and between species can also affect the level and patterns of variability by modulating the effectiveness selection (see Akashi, this issue, pp 660–666). There do appear to be mutations that result in nearly neutral fitness effects. The evolutionary trajectory that these mutations follow, and thus their effect on variability, is greatly affected by changes in  $N_e$ . Amino acid variants can have a spectrum of fitness effects ranging from advantageous to lethal. Evidence also exists that synonymous mutations (those that do not alter the amino acid) may not be completely neutral. For example, codon bias (the nonrandom use of codons within a synonymous family) exists in both humans and flies. However, the cause of the bias potentially differs between the species.

There is evidence that selection for optimal translational accuracy or efficiency has shaped codon usage in *Drosophila*. This has been illustrated, for example, by a negative relationship between synonymous site divergence and the level of codon bias at a locus, and by a positive relationship between the level of codon bias at a locus and its level of recombination [59,60]. In *D. simulans* the ratio of polymorphism to divergence is significantly higher for non-optimal mutations than optimal, and optimal mutations segregate at significantly higher frequencies than non-optimal [61,62]. But codon bias appears to not be correlated with gene expression level in mammals and was thought to depend solely on a gene's location relative to the GC-rich isochores, which are thought to be caused by regional mutation bias [63]. However, recent studies propose that codon bias in humans is not due simply to mutational bias. For example, Smith and Eyre-Walker [64] show that GC content is regionally elevated above its mutation drift equilibrium level at both coding and non-coding regions possibly as a result of selection or GC-biased gene conversion. Indeed GC-biased mismatch repair has been observed in primates [65]. Also it is worth mentioning, as recombination intermediates create mismatches that result in gene conversion [66], that this GC bias of mismatch repair may also explain the intra- and inter-chromosomal correlation of recombination rate with GC content in humans [67]. Karlin and Mrázek's [68] data suggests that a genomic level signature of dinucleotide combinations may be responsible for human codon bias. Iida and Akashi [69] show that within alternatively spliced loci, constitutive exons use GC-ending codons significantly more than alternatively spliced exons. Although their result is not conclusive, it does suggest that some form of translational selection is involved because constitutive exons should experience such selective pressure more often.

The consequence of codon and GC bias in flies and humans is thus that subtle selection appears to exist even on synonymous sites. These sites must be used only cautiously as a proxy for strictly neutral levels of diversity in the genome.

### Evidence for targets of positive selection

Evidence is accumulating that recent positive selection is a relatively common occurrence in both humans and fruitflies. There are few examples of long-term balancing selection being a major force in shaping variability in much of the genome in either *Drosophila* or humans. *Adh* is the only strong example in *Drosophila* [70]. Variation at the MHC region is a corresponding strong example in humans of long-term balancing selection preserving polymorphism even through human speciation [33].

As the number of loci surveyed and sample sizes grow, so do the number of studies indicating that some form of recent positive selection has influenced the levels of variation within and/or between species at individual loci. In numerous cases in *D. melanogaster*, selection appears to have pushed individual haplotypes into higher than expected frequencies, given their apparent evolutionary 'youth' (e.g. *Pgm* [71,72\*], *Su(H)* [73], and *Fbp2* [74]). Also, some recent studies reveal an excess of amino acid fixations between *D. melanogaster* and *D. simulans* relative to synonymous fixations. These results have been interpreted as evidence of positive selection pushing amino acid mutations to fixation (*Rel* [75], *Hex-t1* [76], *mth* [77], *Cid* [78]). There is also evidence of positive selection resulting in a local reduction of variability relative to divergence and/or relative to the level of variability in surrounding chromosomal regions (*Sdic* [79]; *Notch* [V Bauer DuMont, JC Fay, CF Aquadro, unpublished data]). It should be noted that not all loci surveyed for variability show strong evidence of the action of positive selection (*Hex-A*, *Hex-C*, and *Hex-t2* [76]; *Dras* 1, 2, and 3 [80]) stressing that the above mentioned results are truly locus specific.

In humans there is also evidence of selection rapidly increasing the frequency of 'young' haplotypes. Two cases appear to be in response to exposure to malaria (e.g. Duffy [22]; G6PD [48\*]). Remarkably, the classic example of a balanced polymorphism (sickle cell anemia at the  $\beta$ -globin gene), does not show a general departure from neutrality [17], though this probably reflects again the recent origin(s) of the sickle cell variant. Disease has also been argued to have played a role in the relatively recent origin and increase in frequency of the CCR5- $\Delta$ 32 allele in northern European populations [81]. At the lactase gene, the haplotype that is associated with lactose persistence appears to have been driven by selection to high frequency in northern Europeans [82]. Variation at the melanocortin 1 receptor gene is also implicated as a candidate for recent selection. It remains controversial, however, whether the increased number of nonsynonymous versus synonymous polymorphisms in humans from northern latitudes is

caused either by a relaxation of selection for epidermal melanin compared to equatorial regions with high UV exposure, or by selection for reduced melanin and the need for UV for vitamin D metabolism [23–25]. A regional reduction in variation within the *Dmd* gene in humans is also a candidate 'footprint' of recent directional selection, though the target is unknown [32]. As for *Drosophila* and many other organisms, numerous genes involved in reproduction also appear to be driven by positive selection [83,84].

Consideration of the time frame for which tests of neutrality have maximal statistical power to detect the effects of natural selection may explain in part the large number of cases of recent selection. Tests of selection that depend on allele frequency distribution, such as Tajima's *D*, typically only have significant statistical power within  $N_e$  generations after a sweep [85]. The large effective population size (300,000) and generation rate (10 generations per year) of *D. melanogaster* predict that these types of tests are sensitive to selection occurring within the last 30,000 years. For humans, the expected sensitivity range is 500,000 years (assuming  $N_e = 20,000$  and 25 years per generation). Modern humans and *D. melanogaster* share a similar demographic history of a recent range expansion out of Africa, ~100,000 and 10,000 years ago, respectively [86,87], within the predicted time of sensitivity of detecting positive selection for each species. As discussed in the contrast between X-chromosomes and autosomes, lower X-linked variation in non-African samples of *Drosophila* may be a signature of this selection. The existence of a large tract of depressed nucleotide polymorphism in the Xq25-q29 region from human non-African samples [88], as well as the previously discussed examples of selection at the melanocortin 1 receptor and lactase genes are also suggestive of recent directional selection in non-African populations associated with novel environments.

### Conclusions

Nucleotide diversity is on average 10 times lower in humans compared to flies (average nucleotide diversity is ~0.001 in humans versus 0.01 in *D. melanogaster*). This contrast appears largely as a result of differences in effective population size, and not mutation rate. In contrast, microsatellite variability shows a very different pattern with humans being equally or even more variable, the smaller effective population size in humans being offset by a higher microsatellite mutation rate in humans compared to flies.

Both species show striking differences in the level of nucleotide diversity across the genome, varying well over an order of magnitude. In both species a significant correlate of variability is the local rate of recombination in which the assayed genes reside. The presence of a correlation between microsatellite variation and recombination rates is unresolved. A significant positive correlation has been reported for some surveys in *D. melanogaster*. However, a correlation has not been seen for humans.

Linkage disequilibrium extends farther, on average, in humans than flies, though there is great variation within both species. A particularly interesting contrast is that the local level of linkage disequilibrium is less than expected in humans, and more than expected in flies, based on direct estimates of recombination rates. The relative influences of population structure, admixture, selection, and gene conversion to these patterns remain unresolved.

Flies and humans both had their origins in Africa and probably spread world-wide within the last 10,000 to 100,000 years, respectively. Levels of nucleotide variation tend to be higher in African samples compared to non-African samples for both species, consistent with a loss of variation associated with founder effects and/or selection in new biotic and physical environments. But this pattern may be restricted to X chromosome genes in flies. Whether this reflects footprints of adaptive fixations and not founder effects versus the complication of inversion polymorphism on autosomes remains unresolved for flies. Such contrasts are not available for humans at present, but the footprints of adaptive change are becoming increasingly evident from studies of DNA sequence variation both within and between species for both humans and *Drosophila*. The general pattern to emerge for both species is that selection has acted mainly to reduce variation, rather than to enhance it: most evidence for selection indicates some form of recent directional selection, with few, though important, examples of long-term balancing selection.

The prospect of more thorough sampling, both of genomic and geographic regions, with large sample sizes promises to further our understanding of the ways that natural selection, recombination, demography and other factors have shaped genomic diversity in both flies and humans. This understanding will contribute significantly both to the understanding of our own evolutionary history, and to our ability to understand the often complex links between genomic and phenotypic diversity.

### Acknowledgements

Our work is supported primarily by a grant from the National Institutes of Health.

### References and recommended reading

Papers of particular interest, published within the annual period of review, have been highlighted as:

- of special interest
- of outstanding interest

1. Przeworski M, Hudson RR, Di Rienzo A: **Adjusting the focus on human variation.** *Trends Genet* 2000, **16**:296-302.
  2. Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989, **123**:585-595.
  3. Kaplan NL, Hudson RR, Langley CH: **The 'hitchhiking effect' revisited.** *Genetics* 1989, **123**:887-899.
  4. Charlesworth B, Morgan MT, Charlesworth D: **The effect of deleterious mutations on neutral molecular variation.** *Genetics* 1993, **134**:1289-1303.
  5. Yu A, Zhao C, Fan Y, Jang W, Mungall AJ, Deloukas P, Olsen A, Doggett NA, Ghebranious N, Broman KW *et al.*: **Comparison of human genetic and sequence-based physical maps.** *Nature* 2001, **409**:951-953.
- The most comprehensive assessment of the distribution of recombination across the human genome based on careful integration of genetic and physical mapping data. Evidence is summarized for hot spots of recombination and sex differences.
6. Kindahl EC: **Recombination and DNA polymorphism on the third chromosome of *Drosophila melanogaster*** [PhD Thesis]. Ithaca, NY: Cornell University; 1994.
  7. Nagaraja R, MacMillan S, Kere J, Jones C, Griffin S, Schmatz M, Terrell J, Shomaker M, Jermak C, Hott C *et al.*: **X chromosome map at 75-kb STS resolution, revealing extremes of recombination and GC content.** *Genome Res* 1997, **7**:210-222.
  8. Ashburner M: **Mapping and exchange.** In *Drosophila: a Laboratory Handbook*. Edited Ashburner M. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1989:451-501.
  9. Aquadro CF, Begun DJ, Kindahl EC: **Selection, recombination, and DNA polymorphism in *Drosophila*.** In *Non-Neutral Evolution: Theories and Molecular Data*. Edited by Golding B: Chapman and Hall; 1994.
  10. Broman KW, Murray JC, Sheffield VC, White RL, Weber JL: **Comprehensive human genetic maps: individual and sex-specific variation in recombination.** *Am J Hum Genet* 1998, **63**:861-869.
  11. Hamblin MT, Aquadro CF: **High nucleotide sequence variation in a region of low recombination in *Drosophila simulans* is consistent with the background selection model.** *Mol Biol Evol* 1996, **13**:1133-1140.
  12. Li W-H, Sadler LA: **Low nucleotide diversity in man.** *Genetics* 1991, **129**:513-523.
  13. Nachman MW, Bauer VL, Crowell SL, Aquadro CF: **DNA variability and recombination rates at X-linked loci in humans.** *Genetics* 1998, **150**:1133-1141.
  14. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al.*: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
  15. Sachidanandam R, Weissman D, Schmidt SC, Kakol JM, Stein LD, Marth G, Sherry S, Mullikin JC, Mortimore BJ, Willey DL *et al.*: **A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms.** *Nature* 2001, **409**:928-933.
- Results of a comprehensive study of the genome-wide level and distribution of single nucleotide polymorphisms (SNPs) in humans. Of particular note is the striking heterogeneity in variability across the genome.
16. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al.*: **The sequence of the human genome.** *Science* 2001, **291**:1304-1351.
  17. Harding RM, Fullerton SM, Griffiths RC, Bond J, Cox MJ, Schneider JA, Moulin DS, Clegg JB: **Archaic African and asian lineages in the genetic ancestry of modern humans.** *Am J Hum Genet* 1997, **60**:772-789.
  18. Clark AG, Weiss KM, Nickerson DA, Taylor SL, Buchanan A, Stengård J, Salomaa V, Vartiainen E, Perola M, Boerwinkle E *et al.*: **Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase.** *Am J Hum Genet* 1998, **63**:595-612.
  19. Nickerson DA, Taylor SL, Weiss KM, Clark AG, Hutchinson RG, Stengard J, Salomaa V, Vartiainen E, Boerwinkle E, Sing CF: **DNA sequence diversity in a 9.7-kb region of the human lipoprotein lipase gene.** *Nat Genet* 1998, **19**:233-240.
  20. Rieder MJ, Taylor SL, Clark AG, Nickerson DA: **Sequence variation in the human angiotensin converting enzyme.** *Nat Genet* 1999, **22**:59-62.
  21. Fullerton SM, Clark AG, Weiss KM, Nickerson DA, Taylor SL, Stengard JH, Salomaa V, Vartiainen E, Perola M, Boerwinkle E *et al.*: **Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism.** *Am J Hum Genet* 2000, **67**:881-900.
  22. Hamblin MT, Di Rienzo A: **Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus.** *Am J Hum Genet* 2000, **66**:1669-1679.

23. Harding RM, Healy E, Ray AJ, Ellis NS, Flanagan N, Todd C, Dixon C, Sajantila A, Jackson IJ, Birch-Machin MA *et al.*: **Evidence for variable selective pressures at MC1R.** *Am J Hum Genet* 2000, **66**:1351-1361.
24. Makova KD, Ramsay M, Jenkins T, Li WH: **Human DNA sequence variation in a 6.6-kb region containing the melanocortin 1 receptor promoter.** *Genetics* 2001, **158**:1253-1268.
25. Rana BK, Hewett-Emmett D, Jin L, Chang BH, Sambughin N, Lin M, Watkins S, Bamshad M, Jorde LB, Ramsay M *et al.*: **High polymorphism at the human melanocortin 1 receptor locus.** *Genetics* 1999, **151**:1547-1557.
26. Stephens JC, Schneider JA, Tanguay DA, Choi J, Acharya T, Stanley SE, Jiang R, Messer CJ, Chew A, Han JH *et al.*: **Haplotype variation and linkage disequilibrium in 313 human genes.** *Science* 2001, **293**:489-493.
- A large study of nucleotide diversity for 313 genes in 82 humans, distinct in its scope given that haplotypes are resolved at each locus. The distribution of shared and population-specific variants and haplotypes provides insight into human history, and supports the added value of haplotype over simple SNP data.
27. Moriyama EN, Powell JR: **Intraspecific nuclear DNA variation in *Drosophila*.** *Mol Biol Evol* 1996, **13**:261-277.
28. Andolfatto P: **Contrasting patterns of X-linked and autosomal nucleotide variation in *Drosophila melanogaster* and *Drosophila simulans*.** *Mol Biol Evol* 2001, **18**:279-290.
- The author presents a thoughtful and comprehensive summary of nucleotide diversity in *D. melanogaster* and *D. simulans*. He focuses on contrast between African and non-African samples, and those from the X-chromosome versus the autosomes. Levels of variation are comparable among populations on autosomes, but show less variation in non-African populations on the X-chromosome. Issues concerning appropriate corrections for X-chromosome versus autosome effective population size and the effects of autosomal inversions complicate the interpretation, but are given thoughtful discussion.
29. Goodman M, Koop BF, Czelusniak J, Fitch DH, Tagle DA, Slightom JL: **Molecular phylogeny of the family of apes and humans.** *Genome* 1989, **31**:316-335.
30. Kaessmann H, Heissig F, von Haeseler A, Pääbo S: **DNA sequence variation in a non-coding region of low recombination on the human X chromosome.** *Nat Genet* 1999, **22**:78-81.
31. Harris EE, Hey J: **X chromosome evidence for ancient human histories.** *Proc Natl Acad Sci USA* 1999, **96**:3320-3324.
32. Nachman MW, Crowell SL: **Contrasting evolutionary histories of two introns of the duchenne muscular dystrophy gene, *Dmd*, in humans.** *Genetics* 2000, **155**:1855-1864.
33. Gaudieri S, Dawkins RL, Habara K, Kulski JK, Gojobori T: **SNP profile within the human major histocompatibility complex reveals an extreme and interrupted level of nucleotide diversity.** *Genome Res* 2000, **10**:1579-1586.
34. Drake JW, Charlesworth B, Charlesworth D, Crow JF: **Rates of spontaneous mutation.** *Genetics* 1998, **148**:1667-1686.
35. Schug MD, Hutter CM, Wetterstrand KA, Gaudette MS, Mackay TFC, Aquadro CF: **The mutation rates of di-, tri- and tetranucleotide repeats in *Drosophila melanogaster*.** *Mol Biol Evol* 1998, **15**:1751-1760.
36. Halushka MK, Fan JB, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A: **Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis.** *Nat Genet* 1999, **22**:239-247.
37. Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Lane CR, Lim EP, Kalyanaraman N, Nemesh J *et al.*: **Characterization of single-nucleotide polymorphisms in coding regions of human genes.** *Nat Genet* 1999, **22**:231-238.
38. Fay JC, Wyckoff GJ, Wu CI: **Positive and negative selection on the human genome.** *Genetics* 2001, **158**:1227-1234.
- An analysis that published SNP variation within and between species, partitioning variants into frequency classes and type (nonsynonymous versus synonymous), allows the authors to estimate that 60% of nonsynonymous variants are strongly deleterious, and 20% are mildly deleterious. Further, they conclude that 35% of amino acid difference between humans and Old World monkeys have been fixed by positive selection.
39. Dib C, Fauré S, Fzames C, Samson D, Drouot N, Vignal A, Millasseau P, Marc S, Hazan J, Seboun E *et al.*: **A comprehensive genetic map of the human genome based on 5,264 microsatellites.** *Nature* 1996, **380**:152-154.
40. Begun DJ, Aquadro CF: **Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*.** *Nature* 1992, **356**:519-520.
41. Nachman MW: **Single nucleotide polymorphisms and recombination rate in humans.** *Trends Genet* 2001, **17**:481-485.
- An updated comparison of levels of human nucleotide diversity, where at least 10 chromosomes were sampled, demonstrates a genome-wide positive correlation with local rates of recombination. A tendency for low recombination genes to have an excess of rare variants supports an important role for positive selection.
42. Andolfatto P, Przeworski M: **Regions of lower crossing over harbor more rare variants in African populations of *Drosophila melanogaster*.** *Genetics* 2001, **158**:657-665.
- New data for X-linked genes, together with published data, reveal that regions of low recombination harbor significantly more rare variants than regions of high recombination in African *D. melanogaster*. Of the various hypotheses considered, hitchhiking associated with selective sweeps appears to be the best explanation.
43. Begun DJ, Whitley P: **Reduced X-linked nucleotide polymorphism in *Drosophila simulans*.** *Proc Natl Acad Sci USA* 2000, **97**:5960-5965.
- A comparison of synonymous variation within *D. simulans* and between *D. simulans* and *D. melanogaster* demonstrates both diversity, and the ratio of diversity to divergence, are lower on the X-chromosome compared to an autosome (3R). Positive selection is implicated by these results. The choice of *D. simulans* allows the complications of autosomal inversions present in *D. melanogaster* to be avoided.
44. Payseur BA, Nachman MW: **Microsatellite variation and recombination rate in the human genome.** *Genetics* 2000, **156**:1285-1298.
45. Ellegren H: **Heterogeneous mutation processes in human microsatellite DNA sequences.** *Nat Genet* 2000, **24**:400-402.
46. Huttley GA, Smith MW, Carrington M, O'Brien SJ: **A scan for linkage disequilibrium across the human genome.** *Genetics* 1999, **152**:1711-1722.
47. Huttley GA, Eastal S, Southey MC, Tesoriero A, Giles GG, McCredie MR, Hopper JL, Venter DJ: **Adaptive evolution of the tumour suppressor BRCA1 in humans and chimpanzees. Australian breast cancer family study.** *Nat Genet* 2000, **25**:410-413.
48. Tishkoff SA, Varkonyi R, Cahinhinan N, Abbes S, Argyropoulos G, Destro-Bisol G, Drousiotou A, Dangerfield B, Lefranc G, Loiselet J *et al.*: **Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance.** *Science* 2001, **293**:455-462.
- A thoughtful analysis using microsatellite and SNP variation to analyze two recent G6PD malaria-resistance alleles in humans. The results indicate independent origins and rapid rise in frequency for the two resistance alleles within the last 10,000 years, suggesting that malaria has only had a strong selective impact since the advent of agriculture.
49. Pritchard JK, Przeworski M: **Linkage disequilibrium in humans: models and data.** *Am J Hum Genet* 2001, **69**:1-14.
50. Kruglyak L: **Prospects for whole-genome linkage disequilibrium mapping of common disease genes.** *Nat Genet* 1999, **22**:139-144.
51. Andolfatto P, Przeworski M: **A genome-wide departure from the standard neutral model in natural populations of *Drosophila*.** *Genetics* 2000, **156**:257-268.
52. Przeworski M, Wall JD: **Why is there so little intragenic linkage disequilibrium in humans?** *Genet Res* 2001, **77**:143-151.
53. Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A: **Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels.** *Am J Hum Genet* 2001, **69**:831-843.
54. Langley CH, Lazzaro BP, Phillips W, Heikkinen E, Braverman JM: **Linkage disequilibrium and the site frequency spectra in the *su(s)* and *su(wa)* regions of the *Drosophila melanogaster* X chromosome.** *Genetics* 2000, **156**:1837-1852.
55. Begun DJ, Whitley P, Todd BL, Waldrip-Dail HM, Clark AG: **Molecular population genetics of male accessory gland proteins in *Drosophila*.** *Genetics* 2000, **156**:1879-1888.

56. Begun DJ, Aquadro CF: **African and North American populations of *Drosophila melanogaster* are very different at the DNA level.** *Nature* 1993, **365**:548-550.
57. Bauer VL, Aquadro CF: **Rates of DNA sequence evolution are not sex-biased in *Drosophila melanogaster* and *D. simulans*.** *Mol Biol Evol* 1997, **14**:1252-1257.
58. Huang W, Chang BH, Gu X, Hewett-Emmett D, Li W: **Sex differences in mutation rate in higher primates estimated from AMG intron sequences.** *J Mol Evol* 1997, **44**:463-465.
59. Sharp PM, Li WH: **On the rate of DNA sequence evolution in *Drosophila*.** *J Mol Evol* 1989, **28**:398-402.
60. Kliman RM, Hey J: **Reduced natural selection associated with low recombination in *Drosophila melanogaster*.** *Mol Biol Evol* 1993, **10**:1239-1258.
61. Akashi H: **Inferring weak selection from patterns of polymorphism and divergence at 'silent' sites in *Drosophila* DNA.** *Genetics* 1995, **139**:1067-1076.
62. Akashi H, Schaeffer SW: **Natural selection and the frequency distributions of 'silent' DNA polymorphism in *Drosophila*.** *Genetics* 1997, **146**:295-307.
63. Wolfe KH, Sharp PM, Li WH: **Mutation rates differ among regions of the mammalian genome.** *Nature* 1989, **337**:283-285.
64. Smith NG, Eyre-Walker A: **Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans.** *Mol Biol Evol* 2001, **18**:982-986.
65. Brown TC, Jiricny J: **Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells.** *Cell* 1988, **54**:705-711.
66. Harfe BD, Jinks-Robertson S: **DNA mismatch repair and genetic instability.** *Annu Rev Genet* 2000, **34**:359-399.
67. Fullerton SM, Bernardo Carvalho A, Clark AG: **Local rates of recombination are positively correlated with GC content in the human genome.** *Mol Biol Evol* 2001, **18**:1139-1142.
68. Karlin S, Mrázek J: **What drives codon choices in human genes?** *J Mol Biol* 1996, **262**:459-472.
69. Iida K, Akashi H: **A test of translational selection at 'silent' sites in the human genome: base composition comparisons in alternatively spliced genes.** *Gene* 2000, **261**:93-105.
70. Kreitman ME, Aguade M: **Excess polymorphism at the *Adh* locus in *Drosophila melanogaster*.** *Genetics* 1986, **114**:93-110.
71. Verrelli BC, Eanes WF: **Extensive amino acid polymorphism at the *pgm* locus is consistent with adaptive protein evolution in *Drosophila melanogaster*.** *Genetics* 2000, **156**:1737-1752.
72. Verrelli BC, Eanes WF: **Clinal variation for amino acid polymorphisms at the *Pgm* locus in *Drosophila melanogaster*.** *Genetics* 2001, **157**:1649-1663.
- This paper builds on the results from [71] to present a convincing case that one haplotype at the *Pgm* locus in North American *D. melanogaster* has been recently pushed by selection into high frequency. Within that haplotype class, variation is skewed towards rare variants, whereas other classes of variants fit neutrality. A geographic cline is also present, with the derived haplotype being in highest frequency in northern localities. Positive selection appears to have strongly influenced amino acid variation at this important metabolic enzyme.
73. Depaulis F, Brazier L, Veuille M: **Selective sweep at the *Drosophila melanogaster* suppressor of hairless locus and its association with the In(2L)t inversion polymorphism.** *Genetics* 1999, **152**:1017-1024.
74. Bénassi V, Depaulis F, Meghlaoui GK, Veuille M: **Partial sweeping of variation at the *Fbp2* locus in a west African population of *Drosophila melanogaster*.** *Mol Biol Evol* 1999, **16**:347-353.
75. Begun DJ, Whitley P: **Adaptive evolution of Relish, a *Drosophila* NF-kappaB/IkappaB protein.** *Genetics* 2000, **154**:1231-1238.
76. Duvernell DD, Eanes WF: **Contrasting molecular population genetics of four hexokinases in *Drosophila melanogaster*, *D. simulans* and *D. yakuba*.** *Genetics* 2000, **156**:1191-1201.
77. Schmidt PS, Duvernell DD, Eanes WF: **Adaptive evolution of a candidate gene for aging in *Drosophila*.** *Proc Natl Acad Sci USA* 2000, **97**:10861-10865.
78. Malik HS, Henikoff S: **Adaptive evolution of Cid, a centromere-specific histone in *Drosophila*.** *Genetics* 2001, **157**:1293-1298.
79. Nurminsky D, Aguiar DD, Bustamante CD, Hartl DL: **Chromosomal effects of rapid gene evolution in *Drosophila melanogaster*.** *Science* 2001, **291**:128-130.
80. Gasperini R, Gibson G: **Absence of protein polymorphism in the *Ras* genes of *Drosophila melanogaster*.** *J Mol Evol* 1999, **49**:583-590.
81. Stephens JC, Reich DE, Goldstein DB, Shin HD, Smith MW, Carrington M, Winkler C, Huttley GA, Allikmets R, Schriml L *et al.*: **Dating the origin of the CCR5-Delta32 AIDS-resistance allele by the coalescence of haplotypes.** *Am J Hum Genet* 1998, **62**:1507-1515.
82. Hollox EJ, Poulter M, Zvarik M, Ferak V, Krause A, Jenkins T, Saha N, Kozlov AI, Swallow DM: **Lactase haplotype diversity in the Old World.** *Am J Hum Genet* 2001, **68**:160-172.
83. Wyckoff GJ, Wang W, Wu C-I: **Rapid evolution of male reproductive genes in the descent of man.** *Nature* 2000, **403**:304-309.
84. Swanson WJ, Yang Z, Wolfner MF, Aquadro CF: **Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals.** *Proc Natl Acad Sci USA* 2001, **98**:2509-2514.
85. Simonsen KL, Churchill GA, Aquadro CF: **Properties of statistical tests of neutrality for DNA polymorphism data.** *Genetics* 1995, **141**:413-429.
86. Tishkoff SA, Dietzsch E, Speed W, Pakstis AJ, Kidd JR, Cheung K, Bonne-Tamir B, Santachiara-Benerecetti AS, Moral P, Krings M: **Global patterns of linkage disequilibrium at the CD4 locus and modern human origins.** *Science* 1996, **271**:1380-1387.
87. David JR, Capy P: **Genetic variation of *Drosophila melanogaster* natural populations.** *Trends Genet* 1988, **4**:106-111.
88. Miller RD, Taillon-Miller P, Kwok PY: **Regions of low single-nucleotide polymorphism incidence in human and orangutan Xq: deserts and recent coalescences.** *Genomics* 2001, **71**:78-88.