Hypervariability, suppressed recombination and the genetics of individuality

M. V. Olson^{*}, A. Kas, K. Bubb, R. Qui, E. E. Smith, C. K. Raymond and R. Kaul

University of Washington Genome Center, Department of Medicine, University of Washington, Seattle, WA 98195, USA

We define 'genetic individuality' as intraspecies variation that has substantial heritability and involves traits that are sufficiently common that they can be observed in any modest-sized sampling of individuals. We propose that genetic individuality is largely shaped by the combinatory shuffling of a modest number of genes, each of which exists as a family of functionally and structurally diverged alleles. Unequivocal examples of such allele families are found at the O-antigen-biosynthetic locus in *Pseudomonas aeruginosa* and the human leucocyte antigen locus in humans. We examine characteristic features of these allele families and explore the possibility that genetic loci with similar characteristics can be recognized in a whole-genome scan of human genetic variation.

Keywords: balancing selection; human leucocyte antigen; O-antigen; *Pseudomonas aeruginosa*; coalescent

1. INTRODUCTION

In this paper, we take a modern look, informed by the recent accumulation of genome sequences, at an old and fascinating question. How is it that individuals of the same species can differ so greatly from one another? More specifically, our interest is in the genetic basis of individual differences that have significant heritability and are common enough to be observed in any modest-sized sampling of a population. We refer to intraspecies variation of this type as 'genetic individuality'.

The study of genetic individuality is likely to become a central preoccupation of genetics in the years ahead. After a century of model-organism genetics, we know a great deal about how genes work and how they influence phenotype. However, we have only a rudimentary understanding of genetic individuality. Even in human genetics, with its intrinsic focus on analysing individual differences in a natural population, the emphasis has been on disease phenotypes rather than on the types of common, highly variable traits that are found in any small sampling of humans.

Human geneticists have favoured the study of deleterious traits largely out of a practical impulse to ameliorate the effects of genetic diseases through improved knowledge of pathogenic mechanisms. However, there is also the concern that many of the traits that underlie genetic individuality, as defined above, may be too complex to analyse. This concern may be overblown. Parent-child resemblances in humans suggest that phenotypic differences in many biologically complex traits (e.g. the morphological features that allow easy facial recognition of

individuals) are often controlled by a small number of genes. If more than a few genes had large effects, instances of strong resemblance between a child and a particular parent, which are actually quite common, would be infinitesimally rare. Experience with plant and animal breeding leads to similar conclusions. For example, parental phenotypes are largely recapitulated in a few percent of the progeny of maize-teosinte crosses even though only a botanist would recognize that the two plants are closely related (Szabo & Burr 1996).

There have been few studies of the inheritance of complex traits in natural populations, or even in artificial crosses in which one of the parents comes from a natural population (e.g. the maize-teosinte example or similar studies in pigs (Knott *et al.* 1998) and mice (Guenet & Bonhomme 2003)). Hence, geneticists lack well-developed strategies for identifying the genes that modulate common, highly variable traits. Forward-genetic methods that start with phenotype are unattractive since genetic mapping becomes imprecise as soon as more than one or two genes have major effects. In principle, these difficulties can be overcome in experimental systems by carrying out sequential crosses with large numbers of progeny. However, forward-genetic strategies are particularly poorly suited to studying complex traits in humans. Human genetics relies on the analysis of existing, relatively small families. In this setting, genetic heterogeneity, which is likely to be a prominent feature of common, biologically complex phenotypes, poses major difficulties. For these reasons, we favour a reverse-genetic approach that starts by identifying genes that are likely to play important roles in specifying genetic individuality. Once identified, these genes would become candidates in genotype-phenotype association studies.

The potential attractiveness of this strategy depends on the number of relevant genes and the reliability with which they can be identified. Our conjecture that the number of

^{*} Author for correspondence (mvo@u.washington.edu).

One contribution of 18 to a Discussion Meeting Issue 'Replicating and reshaping DNA: a celebration of the jubilee of the double helix'.

genes with major effects on any given trait is small is based on the observation, discussed above, that strong parent-child resemblances are common in outbred populations. The fraction of genes involved in specifying major components of genetic individuality is also likely to be small since variation in most genes fits neutral expectations reasonably well. Our main focus is the possibility that the relevant genes may be recognizable simply on the basis of their unusual patterns of sequence variation. We explore this possibility by examining patterns of variation within loci that unequivocally contribute to genetic individuality. Then, we consider the more general case of genes that show similar, but less extreme, patterns.

2. EVOLUTIONARY MODEL

Any analysis of genetic variation must be based on an evolutionary model. A fully developed model would need to address both the origins and maintenance of the functionally diverged 'allele families' that we believe underlie genetic individuality. We postpone speculation about the origins of these allele families until the final section of this paper. We focus on the ways in which multiple, functionally diverged alleles can be maintained at high frequencies in a population. By definition, allele families that persist in populations must be maintained by some form of balancing selection (Richman 2000): the other major processes that influence allele frequencies—genetic drift, purifying selection and directional selection—all tend to eliminate, rather than maintain, genetic variation. We propose that balancing selection is the central evolutionary process that accounts for genetic individuality.

The idea that balancing selection could play a major role in shaping the characteristics of species is counter-intuitive to many molecular biologists. The traditional focus of molecular biology is on conserved molecular mechanisms. These mechanisms are conserved because evolution has produced singular, highly optimized solutions to specific functional requirements. When balancing selection is mentioned at all in molecular biology textbooks, it typically involves minor variations on the theme of evolutionary optimization. The 'balance', in the most commonly cited examples, is between the phenotypic effects of a highly optimized 'wild-type' allele and those of a mutant allele that confers a selective advantage under a special combination of genetic and environmental circumstances. The standard example is the sickle-cell mutation, an instance in which heterozygosity for wild-type and mutant alleles of the adult beta-globin gene offers a measure of resistance to childhood mortality due to malaria. In the absence of a pathogen, the wild-type allele represents an optimized solution to the functional constraints on a haemoglobin molecule. Homozygosity for the mutant allele is uniformly lethal in the absence of medical intervention. The balance achieved in heterozygotes involves a sacrifice in the functionality of the host in return for creating a diminished environment for the parasite. Because of the lethality of sickle-cell homozygotes, this type of balancing selection carries a high genetic load. If malarial selection were to remain strong over a long evolutionary period, other mechanisms of resistance with less genetic load would surely arise.

The multiple allelic states of genes associated with gen-

etic individuality are presumably maintained by a form of balancing selection that differs greatly from the sickle-cell paradigm. We envision situations in which balancing selection provides a long-term response to persistent environmental heterogeneity. Host-pathogen 'arms races' provide one important example. In these situations, genetic variation in the host provides a heterogeneous environment for the pathogen and vice versa (Potts & Slev 1995). At any point in evolutionary time, pathogens prosper by targeting major genotypes in the host population. In the process, they decrease the frequencies of those genotypes and increase the frequencies of more resistant ones. The cycle continues as pathogen genotypes that target the previously rare host genotypes are positively selected. In the steady state, critical genes in both pathogen and host exist in a wide variety of functionally diverged alleles.

A similar dynamic occurs within any single species. Depending on their genotypes, individuals prosper to different degrees in the various micro-environments that are accessible to the species as a whole. Major genotypes are optimized for the most commonly encountered environments. However, there is constant selection for unusual genotypes that are optimized for exploitation of under-used environmental niches. In situations of this type, 'frequency-dependent' balancing selection (i.e. dependence of the selective advantage of an allele on its frequency in the population) becomes important.

Frequency-dependent balancing selection is likely to play a particularly prominent role in the evolution of social animals. In populations of these animals, the selective advantage of an allele of any gene that modulates interactions between individuals will depend both on its own frequency and the frequencies of other alleles with different phenotypic effects. An idealized example, for which a quantitative theory has been developed, involves 'freeloader' alleles of a gene that controls a social behaviour that produces an essential community resource (Aviles 2002). When present at low frequencies, freeloader alleles are favoured since individuals that bear them can exploit the abundant resources produced by others without sharing in the costs of creating them. However, when present at high allele frequencies, freeloader alleles become disadvantageous since even one individual's contribution to creating the now-scarce community resource enhances the overall environment sufficiently to improve that individual's reproductive success. In this model, one expects a low, steady-state abundance of freeloaders. A notable feature of the model is that, unlike standard explanations of altruistic behaviour, it does not depend on kin selection.

The general lesson from these diverse examples of frequency-dependent balancing selection is that families of functionally diverged alleles are likely to exist at many genetic loci that affect pathogen resistance, social behaviour and other traits that influence adaptation to the heterogeneous environments in which all organisms live. Less clear is whether these allele families have a 'variation signature' that distinguishes them from genes with other selective histories. To explore this question, we have examined two extreme examples of loci that have unequivocal histories of frequency-dependent balancing selection. The first involves the locus that controls the synthesis of the O-antigen of the opportunistic bacterial pathogen *Pseudomonas aeruginosa*, while the second

involves HLA, a locus that encodes critical components of the human immune system.

3. THE O-ANTIGEN BIOSYNTHETIC LOCUS OF *PSEUDOMONAS AERUGINOSA*

Pseudomonas aeruginosa is a Gram-negative bacterium in the same broad family of gammaproteobacteria as *Escherichia coli*. It can be cultured from diverse natural environments such as soil, water and wet slimy surfaces. *Pseudomonas aeruginosa* is an extremely versatile bacterium—an opportunistic pathogen of both plants and animals and a ubiquitous free-living organism. Because of its intrinsic resistance to disinfectants and antibiotics, *P. aeruginosa* is an increasingly important human pathogen.

Like all Gram-negative bacteria, *P. aeruginosa* has a hydrophilic carbohydrate coating that comprises the outer segment of LPS molecules whose proximal ends are anchored in the bacterium's outer member. This surface forest of sugar molecules is *P. aeruginosa*'s principal physical interface with its environment (Rocchetta *et al.* 1999). At least 20 different O-antigen types have been defined serologically. Some of the serological differences reflect dramatic differences in carbohydrate structures, while others involve only subtle variations. The extreme variability in antigen structure is a population-level phenomenon—there is no evidence of an antigen-switching system in *P. aeruginosa*. Most of the genes that specify O-antigen structure are clustered at a particular site on the 6 Mbp *P. aeruginosa* chromosome. The genes at this locus encode enzymes involved in the modification of sugars, the synthesis and transport of polysaccharides, and LPS assembly.

Recently, we undertook an exhaustive study of the O-antigen-biosynthetic locus in the 20 different *P. aeruginosa* serotypes (Raymond *et al.* 2002). The genomic structure of this locus is remarkable in that the sequences of different serotypes are often so highly diverged that there is no recognizable sequence similarity across the whole locus. Analyses based on comparisons of inferred protein sequences reveal that there is considerable overlap in the types of genes (e.g. those encoding glycosyltransferases, epimerases, dehydrogenases and translocases), if not the actual DNA sequences, present in loci associated with different groups of serotypes. The 20 serotypes cluster cleanly into 11 groups. Within a group, the sequences are nearly identical, whereas between groups they are fully diverged. Most remarkably, these highly divergent sequences are embedded within a segment of the *P. aeruginosa* chromosome that shows the same low level of strain-to-strain variation in DNA sequence—a few tenths of a percent—that is observed across most of *P. aeruginosa*'s 6 Mbp genome (Spencer *et al.* 2003). Although they have not been studied in as much detail, the O-antigen-biosynthetic loci of other Gram-negative bacteria appear to have similar characteristics (Curd *et al.* 1998; Reeves & Wang 2002).

The boundaries between the low levels of variation that surround the O-antigen-biosynthetic locus and the non-alignable sequences within the locus are quite sharp (Raymond *et al.* 2002). A typical situation is shown in figure 1. This figure illustrates the level of sequence divergence in the vicinity of the O-antigen-biosynthetic cluster

Figure 1. Sequence divergence between two strains of *Pseudomonas aeruginosa* in the vicinity of the O-antigen-biosynthetic-gene cluster. The reference strain is PAO1 (Stover *et al.* 2000), while the comparison strain was isolated from CF 'Patient 2', as described in Spencer *et al.* (2003). The bin size for the histogram is 2000 bp. In the region demarcated by the double-headed arrow, the two sequences are not alignable. The non-alignable region spans 20.0 kbp in PAO1 and 15.8 kbp in the Patient 2 isolate.

between the PAO1 reference strain (serotype O5) and a strain isolated from the airways of a CF patient (serotype O1).

The O-antigen-biosynthetic loci of *P. aeruginosa* strains have anomalously low G+C content relative to the bulk of chromosomal DNA (Stover *et al.* 2000). This anomaly is commonly taken to indicate that horizontal gene transfer played an important role in distributing loci that control the synthesis of different O-antigen structures among diverse species of Gram-negative bacteria. However, our interest is in the end result, not the mechanism, of the evolutionary processes that led to the current patterns of O-antigen variability. We believe that these loci have characteristics, albeit manifested in an extreme form, that are broadly relevant to the genetic variation responsible for genetic individuality in all species. These characteristics are enumerated below.

- (i) Balancing selection maintains multiple, functionally diverged alleles in the population.
- (ii) Divergence in function is associated with high sequence divergence over regions the size of genes or gene clusters.
- (iii) Recombination within the diverged region has been strongly suppressed. Because of this suppression, it is invariably the intact functional unit that is passed from one generation to the next. (In the case of the O-antigen-biosynthetic locus of *P. aeruginosa*, we saw no examples of gene clusters that could be explained as recombinant products of ancestral types. Indeed, since the observed loci appear to represent highly optimized solutions to the problem of synthesizing a particular O-antigen structure, we presume that recombinant loci would rarely survive in natural environments.)

Although the ability of *P. aeruginosa* to thrive in its diverse environmental niches is undoubtedly enhanced by O-antigen variability, we do not invoke 'species-level'

selection to explain the maintenance of this allele family. A particular *P. aeruginosa* strain, with its own O-antigen type, prospers or declines on its own. The other members of its own species, which display a spectrum of O-antigen types, provide a critical component of any particular strain's environment. If most other *P. aeruginosa* lineages have a particular O-antigen type, there may be selective advantages to an individual with a minor type. For example, potential hosts are more likely to have innate or acquired immunity to the major type. Alternatively, even if this classical form of host–pathogen interaction is unimportant, environmental niches for which the major type is well suited (e.g. because it is particularly proficient at excluding a niche-specific toxic substance) will be over-used, while other niches, to which the minor type may be better suited, will be under-used. Any of these scenarios would lead to frequency-dependent balancing selection and the maintenance of multiple, functionally diverged alleles of the O-antigen-biosynthetic locus.

In contrast to the locus itself, the surrounding DNA sequences have the same low levels of variation that are characteristic of most of the genome (Spencer *et al.* 2003). This background level of variation presumably reflects the population history of *P. aeruginosa* and the usual effects of mutation, recombination, genetic drift, migration and purifying selection. Recombination within low-variation regions of the chromosome presumably leads to combinatorial shuffling of the modest number of loci in *P. aeruginosa* whose genetic characteristics are similar to those of the O-antigen-biosynthetic locus. Other examples of such loci include those that control the biosynthesis of alternative flagellar structures, alternative iron-uptake systems, and alternative anti-bacterial proteins and associated immunity factors (Spencer *et al.* 2003). Although every single-nucleotide variant between two *P. aeruginosa* genomes has the potential to contribute to the phenotypic differences between the strains, it is reasonable to suppose that major features of genetic individuality are due to the combinatorial shuffling of this modest repertoire of loci that exist as families of highly diverged alleles.

While available data are limited, the model presented here for *P. aeruginosa* appears to be relatively typical of bacteria that occupy diverse environmental niches. For example, the pattern of variation in *E. coli* is generally similar to that of *P. aeruginosa* (Welch *et al.* 2002). By contrast, *Yersinia pestis* (Deng *et al.* 2002) and *Helicobacter pylori* (Alm & Trust 1999) appear to have lost much of their variation during recent selection for a highly specialized lifestyle. In the bacterial literature, the ideas presented here are often addressed in terms of the existence of 'pathogenicity islands' or 'substitution islands' in particular bacterial strains. We have adopted a more general vocabulary to emphasize the conceptual similarities between the evolution of bacteria and metazoan organisms. Because of their huge populations and high reproductive rates, bacteria are a good place to look for extreme illustrations of evolutionary effects that have more subtle manifestations in multicellular organisms.

4. HLA

HLA is the most variable known locus in the human genome. Indeed, enough low-resolution whole-genome-

variation scans of multiple humans have now been carried out that it is unlikely that other loci of comparable size (i.e. hundreds of thousands to millions of base pairs) and levels of single-nucleotide variation will be found (Sachidanandam *et al.* 2001). Hence, HLA is a promising place to examine the signature of strong balancing selection in the human genome. While far from typical, HLA is a better model for other human loci than are bacterial examples.

The class I and class II regions of HLA encompass much of the known, functionally important variation at HLA. In both regions, the protein-coding sequences that appear to be subject to strong balancing selection are confined to small segments of a handful of genes (Hughes & Yeager 1998). In the class I region, *HLA-A*, *-B* and *-C* show the highest variability, while the comparably variable genes in the class II region are *HLA-DQA1*, *-DQB1* and *-DRB1*. Although their detailed roles in antigen presentation differ, all six genes encode cell-surface proteins that bind peptide antigens in a molecular context suitable for recognition by T-cell receptors (Klein & Sato 2000*a,b*). Most of the genetic variation in the class I and class II genes is found in the short segments of the coding regions that encode the peptide-binding pockets of the class I and class II proteins. The peptide-binding pockets of many of the alleles are highly diverged from one another, presumably because of long-term evolutionary optimization of a particular peptide-binding capability. The molecular phylogenies of the alleles are extremely deep: for example, many of the deep branches are shared between chimpanzees and humans, clearly demonstrating that substantial amounts of genetic variation can flow through speciation bottlenecks (Klein 1987).

Our major interest in HLA variation involves the relationship between the small-scale variation in the coding regions of the class I and class II genes, which is concentrated in regions of obvious functional specialization and the broader patterns of variation across these gene clusters. Anecdotal samplings of long-range variation at HLA suggest that extensive sequence divergence exists far beyond the limits of the short protein-coding segments that are thought to be the 'business end' of HLA variation. For example, Guillaudoux *et al.* (1998) sequenced cosmid clones that were derived from the two different class I haplotypes of a single individual. Across much of the region between *HLA-B* and *HLA-C*, which spans tens of thousands of base pairs and contains no known genes or other functional elements, sustained sequence divergence of 3–7% was observed. This level of variation is more than 50-fold higher than that typically seen when comparing two random samplings of the same segment of the human genome. Several other anecdotal instances of high pairwise variation over long portions of the class I and class II regions have been reported (Horton *et al.* 1998). In some cases, the variation shows a 'block' structure, which is already evident in the first report by Guillaudoux *et al.* (i.e. a sub-segment of the region displays several percent variation, while the variation collapses to near-zero in an adjacent sub-segment). If the primate molecular clock runs at a typical rate in these regions, 5% divergence would correspond to a coalescence time (i.e. time back to the last common molecular ancestor of two sequences) of *ca.* 20 Myr, a period preceding the divergence of the entire

variable position	73	75	80	85	98	105	119	136	148	164	167	174
individual												
14660	A/C	T/G	G/T	A/T	A/G	G	T/C	T/G	A/G	A/G	T/C	G/A
14661	C	G	T	T	G	T/G	T	T/G	G	G	C	A
14663	A/C	T/G	G/T	A/T	A/G	G	T	T/G	A/G	A/G	T/C	G/A
03715	A/C	T/G	G/T	A/T	A/G	T/G	T	T	A/G	A/G	T/C	G/A
00576	C	G	T	T	G	G	T	G	G	G	C	A
0037F	C	G	T	T	G	G	T	T/G	G	G	C	A
00470	C	G	T	T	G	T	T	T	G	G	C	A
01018	A/C	T/G	G/T	A/T	G	G	T	T/G	A/G	G	T/C	A
10923	A/C	T/G	G/T	A/T	A/G	G	C	T/G	A/G	A/G	T/C	A
12062	C	G	T	T	G	G	T	G	G	G	C	A
04535	A/C	T/G	G/T	A/T	A/G	G	T	T/G	A/G	A/G	T/C	G/A
01960	C	G	T	T	G	T/G	T	T/G	G	G	C	A
chimpanzee	C	G	T	T	G	T/G	T	T	G	G	C	A
gorilla	C	G	T	T	G	G	T	T	G	G	C	A
human haplotype 1	C	G	T	T	G	G(T)	T(C)	G(T)	G	G	C	A
human haplotype 2	A	T	G	A	A(G)	G	T(C)	T	A	A(G)	T	G(A)

↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑ ↑

positions that are invariant or nearly invariant on each haplotype

Figure 2. Analysis of polymorphic sites at an intergenic sequence between the *DQA1* and *DRB1* genes in the class II region of HLA. The rows show the genotypes of different human individuals, as well as a chimpanzee and a gorilla. The columns correspond to the 12 variable positions in this 102 bp region. The complete sequence of the region shown has the following sequence in the April 2003 release of the human-genome sequence: (C)T(G)GATC(T)TACT(T)GAAAAGGGGAGA(G)TTCTTG(G)CAATCCACACAGC(T)ATCACCTGTTTCTTTC(G)ACAAATTTTCAG(G)TATTATATCTTTCTC(G)CA(C)CACATA(A). Variable positions are in parentheses. The current human-reference sequence corresponds to haplotype 1 at all the invariant or near-invariant sites that define this haplotype.

human–great ape lineage from the lesser apes and monkeys (Goodman 1999).

For balancing selection to maintain haplotype integrity across tens of thousands of base pairs for tens of millions of years, recombination between different haplotypes would need to be almost completely suppressed. This situation is reminiscent of, although less extreme than, that observed in the O-antigen biosynthetic loci of *P. aeruginosa*. The occasional collapses of variation observed in particular pairwise comparisons between HLA haplotypes are likely to represent the ‘frozen accidents’ of rare, ancestral recombination events between highly diverged haplotypes.

We carried out systematic sampling of non-coding variation at HLA to determine how much of the pairwise variation that had been reported could be explained in terms of a small number of deeply diverged, ancestral haplotypes. An alternative explanation for the observations would be high mutability in the region, which would manifest itself both in extreme haplotype variability among humans and in unusual levels of interspecies divergence.

A simple test of these alternatives is to sample highly variable sequences in the class II region from an ethnically diverse panel of humans and from great apes. The data shown in figure 2 for one short segment of the class II region rule out high mutability as an explanation for the high pairwise divergences and suggest a remarkably simple haplotype structure. In the panel of 12 humans, extreme pairwise divergences are observed: there are 12 polymorphic sites in 102 bp. At six of these sites—increasing to eight if one ignores private polymorphisms seen on only one out of the 24 chromosomes analysed—all the variation can be explained in terms of two haplotypes. Furthermore, the single chimpanzee and gorilla samples were homozygous for one of the same two haplotypes. At sites that were not polymorphic in the human panel, there were no

differences in the gorilla and only one difference in the chimpanzee across the 102 bp region. This low interspecies divergence rules out high mutability at the locus as an explanation for its hypervariability.

A second short segment, 25 kbp closer to *DQA1* than the one shown in figure 2, was also analysed. The results were qualitatively identical: high pairwise divergences were again observed, and they could all be explained in terms of two deeply diverged haplotypes. Furthermore, the variation at this second locus was in strong linkage disequilibrium with that at the first: in nearly all cases, individuals heterozygous for the two basic haplotypes at the first locus were also heterozygous at the second, while other individuals were homozygous for a single basic haplotype at both loci. Hence, the deep haplotype divergences have evolved at this locus with minimal historical recombination between them despite the vast time-scale required to accumulate the observed haplotype differences.

Typical segments of the human genome have been diverging from a common ancestral sequence for less than a million years (Harpending & Rogers 2000). Hence, studies of variation in the contemporary population are uninformative about evolutionary processes that occurred on longer time-scales. By contrast, balancing selection has maintained multiple molecular lineages at HLA for at least 20 million years. A full accounting of the effects of that long history will require extensive resequencing of the region from many individuals of each of several primate species. However, the early indicators are that the deep history of HLA variation is dominated by a small number of highly divergent haplotypes. Correlation of this deep haplotype structure with the known functional characteristics of polymorphic genes in the region may provide clues as to precisely which combinations of functional variants were under selection. Of equal interest will be insights into

the mechanistic basis for the strong suppression of recombination between divergent haplotypes.

With respect to our broader theme, we regard HLA as an intermediate case between the O-antigen-biosynthetic locus of *P. aeruginosa*, in which haplotype divergence and functional specialization are both complete, and more typical sources of genetic individuality in humans. These more typical cases, which are expected to have much shallower histories than HLA variation, will be challenging to recognize against the background of neutral variation. Our ability to recognize these loci will depend on the length of time that multiple allelic lineages have been protected from extinction through genetic drift by balancing selection, and on the extent to which recombinational suppression has maintained linkage disequilibrium over unusual distances.

5. ANONYMOUS SEGMENTS OF THE HUMAN GENOME THAT EXIST IN UNUSUALLY DIVERGENT ALLELIC STATES

We have carried out computational screens of large, publicly accessible datasets that provide whole-genome views of human genetic variation. There are two main sources of useful data. One of these involves overlap regions between clones in the tiling path that was constructed during the physical mapping phase of the Human Genome Project (McPherson *et al.* 2001). Since a high proportion of this tiling path incorporated clones from a BAC library constructed from the DNA of a single individual, the probability that two overlapping BACs will come from different human haplotypes is approximately 0.5. Indeed, when complete BAC sequences from adjacent clones exist, the overlap regions are approximately equally likely either to be nearly identical or to differ by roughly one nucleotide in a thousand. Our interest has been in authentic overlap regions that differ at an unusually high number of sites over thousands of base pairs.

A formally similar source of data comes from a large SNP discovery project based on random sampling of sequencing 'reads' from an ethnically diverse pool of individuals (Sachidanandam *et al.* 2001). These 'reads', which are typically several hundred nucleotides long, can be aligned with the reference sequence and discrepancies noted. Discrepancies that are supported by high-quality portions of the 'read' are likely to reflect true SNPs. Our interest has been in 'reads' that are highly discrepant with respect to the reference sequence.

Analysis of these highly discrepant 'reads' is fraught with challenges since authentic examples of high allelic variation are intermixed with numerous artefacts. Most prominently, these artefacts arise from unrecognized data errors and false alignments (i.e. alignments of sequences that are similar to one another but not allelic). For these reasons, SNPs were not harvested during the SNP-discovery project from the high-variation tail of the SNP-density distribution (Sachidanandam *et al.* 2001). Our approach has been to carry out automated filtering of the data by a variety of stringent criteria (e.g. insisting on particularly high data quality and rejecting sequences that can be aligned with known repeats). We then determined actual variation within genome segments that survived this

filtering by PCR-based resequencing of a panel of 10 individuals. Full analysis of these data is still underway. However, we report initial examples in which the variation detected by computational analysis of the whole-genome datasets was again observed upon resequencing of the corresponding interval in the 10 person panel. Furthermore, we required that the SNPs observed during resequencing be in approximate Hardy-Weinberg equilibrium. The latter test is a simple, powerful way of rejecting instances of non-allelic variation that arise when PCR primers amplify more than one site in the human genome.

A striking feature of many of the high-variation loci detected in this way is that, just as we observed in the HLA class II region, haplotype diversity is remarkably limited. Figure 3 shows two examples in which the resequencing data on the 20 chromosomes present in the 10-person panel can be explained in terms of just two haplotypes. This pattern would be expected if the two haplotypes had not recombined with one another, and had been maintained by balancing selection as they diverged. However, selection should never be invoked until explanations based on neutral processes have been ruled out. Hence, we developed a simple theoretical model for the expected behaviour of neutral sequences sampled from the high-variation tail of the SNP-density distribution.

6. COALESCENT MODELS FOR HIGH-VARIATION SEGMENTS OF THE HUMAN GENOME

Coalescent theory provides the simplest framework within which to explore the expected behaviour of neutrally evolving segments of the genome (Fu & Li 1999; Rosenberg & Nordborg 2002). In this model, sequence variation among a set of allelic sequences is interpreted in terms of the statistical properties of idealized molecular phylogenies that trace all the sequences back to a common ancestral sequence (i.e. the root, or 'coalescence point'). Particularly if one excludes rare SNPs and SNPs that are predominantly found in particular human sub-populations, variation at most human loci fits the expectations of the coalescent model reasonably well if one assumes a random-mating population with a constant size of *ca.* 10 000 (Harpending *et al.* 1998). This model is sensitive to the average properties of the ancestral human population, as it existed in some part of Africa over the period 0.1–1 Myr ago. The deeper limit is set by the coalescence time of most neutrally evolving segments of the human genome and the shallower one by expansion of the human population and its dispersal across the globe. Mutations that arose after the expansion and dispersal of the human population are not expected to be common in any broad geographical sampling of human variation.

Figure 4 shows a simulated coalescent tree for a neutrally evolving segment of the human genome. The defining characteristic of coalescent trees is that the time intervals between coalescence events are statistically independent and increase exponentially as the number of branches decreases. For the 20-leaf tree in figure 4, the first coalescence, which was between leaves 7 and 8, occurred in 13 generations (*ca.* 260 years), while the last coalescence—leading to the root of the tree—required 22 405 generations (*ca.* 450 000 years). In a neutral tree with a constant mutation rate, two sequences diverge in

Figure 3. Two examples of short segments of the human genome chosen for analysis because of a high local density of SNPs. In (a) and (b) a region of a few hundred base pairs was identified as unusually polymorphic by computational analysis of a pair of sequences that had been sampled from different individuals in a large-scale study of human polymorphism (Sachidanandam *et al.* 2001). Almost the same region was then resequenced from a panel of 10 individuals. The coordinates of polymorphisms observed on resequencing and the genotypes observed at these sites are indicated over the bar that defines the resequenced region. In both cases, only three genotypes were observed on resequencing, and all 20 chromosomes in the 10 person panel could be explained in terms of two haplotypes that display approximate Hardy–Weinberg equilibrium. The locations of the resequenced regions in the genome are indicated relative to the nearest annotated exon.

proportion to the sum of the branch lengths that lead back to their coalescence point. Hence, the tree in figure 4 would be expected to lead, if mutations accumulated at equal rates on all branches, to two major haplotypes (encompassing leaves 1–14 and 15–20). Any representative of either basic haplotype is separated from any representative of the other by twice the full depth of the tree ($2 \times 43\,181 = 86\,362$ generations or *ca.* 1.7 Myr). Given that human–chimpanzee differences in neutral sequences are *ca.* 1% and that speciation is thought to have occurred *ca.* 5 Myr ago, these two haplotypes would be expected to differ from one another by *ca.* 0.17%.

The tree in figure 4 was chosen as a relatively typical example of a neutral tree because its basic characteristics (e.g. overall depth and major-branching pattern) are typical of trees in the middle of the broad distribution of tree types produced by this highly stochastic model. However, the challenge in distinguishing between neutral trees and those reflecting the long-term effects of balancing selection involves the deep tail of the distribution of neutral-tree depths. We will discuss the properties of these deep neutral trees once we have adapted the coalescent model

to explore the properties of trees shaped by strong balancing selection. Our balancing-selection model presupposes that two functionally diverged alleles have been maintained by balancing selection at equal allele frequencies in a random-mating population with an effective size of 10 000 for 100 000 generations (2.3-fold the depth of the tree in figure 4; *ca.* 2 Myr). To keep the model simple and free of gratuitously chosen parameters, we assume no positive selection on either allele and complete suppression of recombination. Hence, the selectively important functional divergence is envisioned as having arisen rapidly at the root of the tree and to have been maintained thereafter by balancing selection. Furthermore, most of the sites on the non-recombining segments associated with the two alleles are envisioned to have evolved neutrally. Under these assumptions, a representative tree is shown in figure 5a.

The hallmark of this balancing-selection model is that it leads to deeply diverged haplotypes with atypically little neutral variation. Since the model presupposes that the two alleles evolve independently without recombination, there are only N copies of each allele in a population of

Figure 4. A simulated molecular phylogeny with properties similar to typical neutral trees. The phylogeny was simulated based on a coalescent model assuming a random-mating population with constant effective size $N = 10\,000$. The tree simulates the evolution of an autosomal locus. The calibration bar is in number of generations. The overall depth of the tree is 43 182 generations (*ca.* 860 000 years). Approximately half of that depth (20 777 generations) is required to reduce the number of branches from 20 to 2; then, another 22 405 generations are required for the final coalescence. These values are close to the average properties of neutral trees with substantial numbers of leaves (i.e. such trees are expected to have a total depth of approximately $4N$ generations, split equally between the last coalescence and all preceding coalescences). This tree, as well as those shown in figure 5, was displayed using the program TreeExplorer (http://evolgen.biol.metro-u.ac.jp/te/te_man.html; Kumar *et al.* 2001).

effective size N , as opposed to $2N$ for typical autosomal alleles that are evolving neutrally. For this reason, the functionally equivalent, and freely recombining, variants of each allele coalesce rapidly and accumulate little variation.

Strong statistical tests could be devised to distinguish between trees similar to the typical neutral tree shown in figure 4 and those similar to the balancing-selection tree in figure 5*a*. Indeed, Tajima's D -test is a test of the required type (Tajima 1989). This test becomes strongly positive—relative to a neutral expectation of 0—when the average pairwise differences between sequences are higher than expected for the number of variant sites. Trees such as that shown in figure 5*a*, as well as our actual data from the class II region of HLA or the anonymous human loci described in figure 3, have this property. Half of all pairwise comparisons that involve leaves sampled from opposite major branches of the tree lead to high average pairwise differences even though there are only two basic sequence patterns present and, hence, relatively few variant sites.

However, the real challenge in recognizing the variation signature associated with long-term balancing selection

Figure 5. Comparison between a simulated tree based on a model for long-term balancing selection and an unusually deep neutral tree. (a) The balancing selection tree. The model assumes immediate functional divergence between two alleles at the root of the tree and long-term maintenance of both alleles at equal allele frequencies. The effective population size is 10 000. Overall, tree depth was arbitrarily set at 100 000 generations. Within both major branches of the tree, all sub-lineages are assumed to have equal fitness. (b) An unusually deep neutral tree. As in (a), $N = 10\,000$; the total depth of the tree shown is 103 947 generations, 89 564 of which are associated with the final coalescence.

involves the quite different problem of distinguishing the balancing-selection signature from that of the long-coalescence-time tail of the neutral distribution. For example, figure 5*b* is a simulated neutral tree, sampled from the deep side of the distribution of tree depths: it is virtually identical in appearance and quantitative properties

to the balancing-selection tree simulated in figure 5a. Simulations based on our simple, balancing-selection model consistently produce trees with the general features of that in figure 5a, while most neutral trees look more like the tree in figure 4 than the one in figure 5b. However, when scanning the human genome for rare loci that show a putative balancing-selection signature (high variation concentrated in a small number of deeply diverged haplotypes), we are obviously at risk of simply recovering the long-coalescence-time tail of the neutral distribution.

This difficulty in distinguishing between molecular phylogenies that arise under balancing selection and unusually deep, neutral phylogenies is due to the statistical properties of genetic drift. As overall coalescence times get longer, the time associated with the last coalescent event becomes increasingly dominant. In a 20-leaf tree such as those simulated here, there are 19 coalescent events whose expected coalescence times increase exponentially as one goes farther back in time. The first 18 events involve the sum of many independent samplings from exponential distributions. As a result, their total contribution to the tree's depth has a relatively tight distribution around its expected value of approximately $2N$ generations. By contrast, the last sampling—whose expectation value is approximately equal to the sum of the first 18 coalescences—has a much broader distribution. This property of neutral trees is illustrated in figure 6, which shows the distribution of overall tree depths, as well as the expected fraction of total tree depth associated with the final coalescence. Neutral trees with depths as great as, or greater than, the one shown in figure 5b are expected to account for *ca.* 2% of genome segments. On average, in trees of the depth of the one shown in figure 5b (100 000 generations), the final coalescence is expected to account for 78% of the overall tree depth (in contrast to 86% for the one shown). Hence, situations in which the pattern of variation is dominated by two highly diverged haplotypes are not expected to be rare, even in the absence of balancing selection.

There are numerous caveats to this analysis. Of course, the model is simplistic, a problem that is likely to be particularly severe when analysing tails of various distributions. Data are accumulating rapidly that will allow refinement of the model for relatively typical neutral sequences (*i.e.* those that coalesce on time-scales comparable to the expected tree depth of $4N$ generations, or *ca.* 0.5–1.0 Myr). However, efforts to refine the model at longer times—for example, by detecting population bottlenecks and taking them into account—are likely to flounder for a lack of sufficient data. Most ancestral variation has been purged from our genomes by genetic drift during the last million years. That leaves a 4 Myr gap before the first available interspecies comparison provides another window on the selective forces that have shaped modern humans. Ancient DNA is unlikely to fill this gap: its intrinsic chemical instability appears likely to limit its informativeness primarily to 0.1 Myr or less (Hofreiter *et al.* 2001).

These rough time-scales are of interest with respect to the evolution of genetic individuality in humans. We do not know enough, either about human biology or human evolution, to know how many of our highly variable phenotypes have their roots in selective forces that were

Figure 6. Characteristics of neutral trees. (a) The cumulative distribution of tree depths; (i) has a linear frequency scale, while (ii) has a logarithmic scale in order to provide a detailed view of the tail of the distribution. The vertical axis, $1 - \text{cum_freq}$, indicates the probability that a neutral tree will have an overall depth as great as or greater than the time specified on the horizontal axis. Time is measured in the number of generations divided by the effective population size. Although the latter value has been fixed at 10 000 in the examples discussed, the results in this figure are general to all values of N . (b) The fraction of total tree depth associated with the last coalescence. The dominance of the last coalescence in determining tree depth implies that deep, neutral trees will typically display only two major haplotypes.

already acting deep in our primate past. The cumulative distribution in figure 6a, if it proves an even approximately reliable guide to the distribution of neutral tree depths, suggests that neutral trees dating back to the human–chimpanzee divergence (*ca.* 5 Myr, *ca.* 250 000 generations) will still exist in approximately one ‘genome segment’ in 100 000. For neutral sequences, the effective segment size is determined by the size of blocks that rarely undergo meiotic recombination on the relevant time-scale. Recent data on the hot-spot structure of human meiotic recombination suggest that segment sizes of a few kilobase pairs are a plausible estimate (Jeffreys *et al.* 2001; Ardlie

et al. 2002; Kauppi *et al.* 2003). At this segment size, there are roughly a million segments in the genome. If one segment in 10^5 has, simply by chance, been protected from fixation by genetic drift, there may be a modest number of such segments in the genome. If so, our analysis predicts that they will typically have a sequence divergence of *ca.* 1% (i.e. the same as the human–chimpanzee divergence) and be characterized by two highly divergent haplotypes.

The possibility that we may be able to discern the signature of balancing selection simply by carrying out variation scans across the genome will depend on other distinguishing characteristics in addition to high variation and the existence of a small number of divergent haplotypes. There are classical tests for selection based on the detailed pattern of variation at sites that are known, *a priori*, to have a high likelihood of being functionally important—most prominently, non-redundant sites in coding regions (Kreitman 2000). However, these sites comprise such a small fraction of the genome that they will often tell a statistically ambiguous story. Hence, it will be important to rely on the much larger number of neutral variants that are intermixed with sites that are under selection to call attention to regions of the genome that specify major components of genetic individuality.

Despite the challenges, there are reasons for optimism about this approach. First, such regions—at least when the selection has remained strong for millions of years—may simply be more common than segments that acquired similar properties by chance. If there are several hundred genes or gene clusters that fit our balancing-selection model and have coalescence times of at least a few million years, the signal-to-noise ratio in variation scans may be quite favourable. This argument has particular force since population bottlenecks are likely to have purged neutral variation from the population more rapidly than is predicted by a constant-population-size model. Second, it is possible that many of the loci responsible for human genetic individuality will show the combined imprints of balancing and directional selection. These two processes need not be incompatible: directional selection may commonly act on each of the alleles in an allele family—perhaps in response to relatively recent events during human evolution—without destabilizing the balance between the alleles. Extreme examples of directional selection, albeit of unknown functional significance, are known in primates (Johnson *et al.* 2001). The effects of such selection are stunningly visible in genome sequences since they can even lead to an inversion of the usual effects of purifying selection: for example, coding regions under directional selection can change more rapidly than the surrounding non-coding DNA.

7. DISCUSSION

We have presented a broad framework for investigating the molecular basis of the phenotypic plasticity of individuals within natural populations. We propose that much of this plasticity is due to unusual patterns of functional variation that is maintained by balancing selection acting on a small proportion of the genome. Examples such as the O-antigen biosynthetic locus in the bacterium *P. aeruginosa* and the HLA locus suggest that such genes are

easily recognized when they have a long history of strong balancing selection. Such selection leads to the persistence of allele families that become increasingly divergent as time elapses. In extreme cases, this divergence carries along a great deal of neutral variation and is associated with extensive, or even complete, suppression of meiotic recombination. The resultant linkage disequilibrium between neutral and selected variation, classically known as ‘hitchhiking’, can have dramatic effects over intervals of at least tens of thousands of base pairs. Tian *et al.* (2002) recently suggested that the population genetics of *Arabidopsis* may be particularly suited to the detection of balancing selection by the methods advocated here. Studies in carefully selected organisms—particularly those such as *Arabidopsis* and *Drosophila*, which are both highly developed as model organisms and easily sampled from natural populations—represent an important path toward an improved understanding of genetic individuality.

The prospects of readily identifying the genetic loci in humans that fit this model are uncertain. The variation signature at HLA has evolved over tens of millions of years—that at the O-antigen biosynthetic locus in *P. aeruginosa* presumably over a still much longer time. The critical time period for the evolution of traits that are of specific importance to human genetic individuality is likely to be in the range of tens of thousands to millions of years. Quantitative arguments based on a simple model of neutral evolution suggest that functionally polymorphic genes under strong balancing selection will only be easily distinguishable from the high-divergence tail of the neutral distribution when the selection originated early in this period. However, other tests for balancing selection that examine the distribution of variation within genes provide complementary criteria that are less dependent on the time-scale over which selection has acted.

The question of how functionally diverged allele families arise and how individual alleles retain functional integrity deserves attention. One possible source of these families is occasional hybridization between sub-species that rarely mate. In extreme form, the O-antigen-biosynthetic locus is likely to have evolved in this way since horizontal gene transfer is a mechanism by which a species can collect alleles that evolved for long periods in other, reproductively isolated populations. HLA is also a good candidate for this model since many now-extinct lineages coexisted with human ancestors and may well have been conspecific with them (Curnoe & Thorne 2003). Periodic episodes of hybridization would be a simple mechanism by which alleles that became specialized during periods of positive selection in one lineage could be added to the allelic repertoire of another.

Regardless of the origins of functionally diverged allele families, their long-term maintenance is likely to depend on local suppression of recombination. We envision that functional specialization typically requires the coevolution of numerous sites within a gene or gene cluster. The O-antigen-biosynthetic locus is an extreme case in which DNA-level homology between the major groups of alleles has been completely lost. HLA is an intermediate case. The mechanism by which recombination between divergent HLA class II haplotypes is suppressed is unknown; however, the strong pattern of linkage disequilibrium between deeply diverged haplotypes indicates that the

suppression has been nearly absolute over at least tens of thousands of base pairs for 10–20 Myr. One possibility is that the suppression of recombination is simply due to the sequence divergence itself. If much of this divergence pre-existed, as might be the case in the hybridization model for the acquisition of members of deeply diverged allele families, there may have been little recombination from the time when the divergent haplotypes were first acquired. Other, more elaborate scenarios are also possible. For example, functionally critical sequences within hot spots, which appear to exert a major influence on the distribution of recombination events during human meiosis (Jeffreys *et al.* 2001; Kauppi *et al.* 2003), may be lost by mutation. Indeed, if the positions of crossovers are extensively determined by hot spots, positive selection may eliminate hot spots that would lead to frequent disruption of functionally diverged alleles. Finally, at least to some degree, allelic integrity may simply be preserved by purifying selection.

Our model for genetic individuality is verifiable even in humans. We predict that a small fraction of genes and gene clusters will display patterns of variation that are strongly suggestive of long-term balancing selection. When the balancing selection has its origins deep in our primate past, its effects may be easily recognized. Particularly if recombination that would disrupt functionally diverged alleles has been suppressed, much neutral polymorphism will be maintained along with the functionally important variants. For more recently evolved allele families, more sophisticated tests may be necessary. However, the larger challenge will be to correlate allele families discovered in any of these ways with phenotype. The magnitude of this challenge depends on the quantitative complexity of human genetic individuality. If much of the phenotypic plasticity of humans arises through the combinatorial shuffling of a modest number of loci, each of which exists in many functionally diverged allelic states, the association of these loci with particular traits may pose a more manageable challenge than is commonly assumed.

This work was supported by a grant from the National Institutes of Health under the Centers of Excellence in Genome Science Program (P50 HG02351). Many members of the University of Washington Genome Center staff participated in the acquisition and analysis of data discussed in this paper.

REFERENCES

- Alm, R. A. & Trust, T. J. 1999 Analysis of the genetic diversity of *Helicobacter pylori*: the tale of two genomes. *J. Mol. Med.* **77**, 834–846.
- Ardlie, K. G., Kruglyak, L. & Seielstad, M. 2002 Patterns of linkage disequilibrium in the human genome. *Nature Genet.* **3**, 299–309.
- Aviles, L. 2002 Solving the freeloaders paradox: genetic associations and frequency-dependent selection in the evolution of cooperation among nonrelatives. *Proc. Natl Acad. Sci. USA* **99**, 14 268–14 273.
- Curd, H., Liu, D. & Reeves, P. R. 1998 Relationships among the O-antigen gene clusters of *Salmonella enterica* groups B, D1, D2 and D3. *J. Bacteriol.* **180**, 1002–1007.
- Curnoe, D. & Thorne, A. 2003 Number of ancestral human species: a molecular perspective. *Homo* **53**, 201–224.
- Deng, W. (and 20 others) 2002 Genome sequence of *Yersinia pestis* KIM. *J. Bacteriol.* **184**, 4601–4611.
- Fu, Y. X. & Li, W. H. 1999 Coalescing into the 21st century: an overview and prospects of coalescent theory. *Theor. Popul. Biol.* **56**, 1–10.
- Goodman, M. 1999 The genomic record of humankind's evolutionary roots. *Am. J. Hum. Genet.* **64**, 31–39.
- Guenet, J. L. & Bonhomme, F. 2003 Wild mice: an ever-increasing contribution to a popular mammalian model. *Trends Genet.* **19**, 24–31.
- Guillaudeux, T., Janer, M., Wong, G. K. S., Spies, T. & Geraghty, D. E. 1998 The complete genomic sequence of 424,015 bp at the centromeric end of the HLA class I region: gene content and polymorphism. *Proc. Natl Acad. Sci. USA* **95**, 9494–9499.
- Harpending, H. & Rogers, A. 2000 Genetic perspectives on human origins and differentiation. *A. Rev. Genomics Hum. Genet.* **1**, 361–385.
- Harpending, H. C., Batzer, M. A., Gurven, M., Jorde, L. B., Rogers, A. R. & Sherry, S. T. 1998 Genetic traces of ancient demography. *Proc. Natl Acad. Sci. USA* **95**, 1961–1967.
- Hofreiter, M., Serre, D., Poinar, H. N., Kuch, M. & Paabo, S. 2001 Ancient DNA. *Nature Rev. Genet.* **2**, 353–359.
- Horton, R., Niblett, D., Milne, S., Palmer, S., Tubby, B., Trowsdale, J. & Beck, S. 1998 Large-scale sequence comparisons reveal unusually high levels of variation in the HLA DQB1 locus in the class II region of the human MHC. *J. Mol. Biol.* **282**, 71–97.
- Hughes, A. L. & Yeager, M. 1998 Natural selection at major histocompatibility complex loci of vertebrates. *A. Rev. Genet.* **32**, 415–435.
- Jeffreys, A. J., Kauppi, L. & Neumann, R. 2001 Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222.
- Johnson, M. E., Viggiano, L., Bailey, J. A., Abdul-Rauf, M., Goodwin, G., Rocchi, M. & Eichler, E. E. 2001 Positive selection of a gene family during the emergence of humans and African apes. *Nature* **413**, 514–519.
- Kauppi, L., Sajantila, A. & Jeffreys, A. J. 2003 Recombination hotspots rather than population history dominate linkage disequilibrium in the MHC class II region. *Hum. Mol. Genet.* **12**, 33–40.
- Klein, J. 1987 Origin of major histocompatibility complex polymorphism: the trans-species hypothesis. *Hum. Immunol.* **19**, 155–162.
- Klein, J. & Sato, A. 2000a The HLA system. First of two parts. *New Eng. J. Med.* **343**, 702–709.
- Klein, J. & Sato, A. 2000b The HLA system. Second of two parts. *New Eng. J. Med.* **343**, 782–786.
- Knott, S. A. (and 11 others) 1998 Multiple marker mapping of quantitative trait loci in a cross between outbred wild boar and large white pigs. *Genetics* **149**, 1069–1080.
- Kreitman, M. 2000 Methods to detect selection in populations with applications to the human. *A. Rev. Genomics Hum. Genet.* **1**, 539–559.
- Kumar, S., Tamura, K., Jakobsen, I. B. & Nei, M. 2001 MEGA2: molecular evolutionary genetics analysis software. *Bioinformatics* **17**, 1244–1245.
- McPherson, J. D. (and 112 others) 2001 A physical map of the human genome. *Nature* **409**, 934–941.
- Potts, W. K. & Slev, P. R. 1995 Pathogen-based models favoring MHC genetic diversity. *Immunol. Rev.* **143**, 181–197.
- Raymond, C. K., Sims, E. H., Kas, A., Spencer, D. H., Kutya- vin, T. V., Ivey, R. G., Zhou, Y., Kaul, R., Clendenning, J. B. & Olson, M. V. 2002 Genetic variation at the O-antigen biosynthetic locus in *Pseudomonas aeruginosa*. *J. Bacteriol.* **184**, 3614–3622.
- Reeves, P. P. & Wang, L. 2002 Genomic organization of LPS-specific loci. *Curr. Top. Microbiol. Immunol.* **264**, 109–135.

- Richman, A. 2000 Evolution of balanced genetic polymorphism. *Mol. Ecol.* **9**, 1953–1963.
- Rocchetta, H. L., Burrows, L. L. & Lam, J. S. 1999 Genetics of O-antigen biosynthesis in *Pseudomonas aeruginosa*. *Microbiol. Mol. Biol.* **63**, 523–553.
- Rosenberg, N. A. & Nordborg, M. 2002 Genealogical trees, coalescent theory and the analysis of genetic polymorphisms. *Nature Rev. Genet.* **3**, 380–390.
- Sachidanandam, R. (and 40 others) 2001 A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409**, 928–933.
- Spencer, D. H., Kas, A., Smith, E. E., Raymond, C. K., Sims, E. H., Hastings, M., Burns, J. L., Kaul, R. & Olson, M. V. 2003 Whole-genome sequence variation among multiple isolates of *Pseudomonas aeruginosa*. *J. Bacteriol.* **185**, 1316–1325.
- Stover, C. K. (and 40 others) 2000 Complete genome sequence of *Pseudomonas aeruginosa* PA01, an opportunistic pathogen. *Nature* **406**, 959–964.
- Szabo, V. M. & Burr, B. 1996 Simple inheritance of key traits distinguishing maize and teosinte. *Mol. Gen. Genet.* **252**, 33–41.
- Tajima, F. 1989 Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* **123**, 585–595.
- Tian, D., Araki, H., Stahl, E., Bergelson, J. & Kreitman, M. 2002 Signature of balancing selection in *Arabidopsis*. *Proc. Natl Acad. Sci. USA* **99**, 11 525–11 530.
- Welch, R. A. (and 18 others) 2002 Extensive mosaic structure revealed by the complete genome sequence of uropathogenic *Escherichia coli*. *Proc. Natl Acad. Sci. USA* **99**, 17 020–17 024.

GLOSSARY

BAC: bacterial-artificial chromosome

CF: cystic fibrosis

HLA: human leucocyte antigen

LPS: lipopolysaccharide

SNP: single-nucleotide polymorphism