Adjusting the focus on human variation

Studies of nuclear sequence variation are accumulating, such that we can expect a good description of the structure of human variation across populations and genomic regions in the near future. This description will help to elucidate the evolutionary forces that shape patterns of variability. Such an understanding will be of general biological interest, but could also facilitate the design and interpretation of disease-mapping studies. Here, we integrate the results from surveys of nuclear sequence variation. When nuclear sequences are considered together with mtDNA and microsatellites, it becomes clear that neither the standard neutral model, nor a simple long-term exponential growth model, can account for all the available human variation data. A possible explanation is that a subset of loci are not evolving neutrally; even so, more-complex models of effective population size and structure might be necessary to explain the data.

One corollary of the interest in the genetic dissection of common diseases is the need to explore polymorphic variants in the human genome. Single nucleotide polymorphisms (SNPs) in particular have gained attention as promising tools for genome-wide mapping efforts. SNPs might contribute directly to disease susceptibility, or can be used as markers to identify the neighborhood of a disease susceptibility variant by linkage disequilibrium (LD) mapping[1,2]. The successful design and implementation of disease mapping approaches based on SNPs will benefit from a detailed knowledge of levels of variation and LD between marker alleles. Thanks to a recent flourishing of studies of human sequence variation, a good description of the structure of human variation in different ethnic groups and genomic regions should soon be available.

Beyond simple description, such studies might elucidate the factors that have shaped human variation. These include genomic factors (for example, the distribution of recombination and mutation rates) as well as evolutionary factors, such as the history of population size and structure, and natural selection. An understanding of these processes will lead to inferences about human histories, which might, in turn, facilitate disease-mapping studies. For example, susceptibility genes for some diseases are hypothesized to have evolved under positive natural selection (e.g. the thrifty genotype hypothesis for type 2 diabetes[3]). If so, detecting the signature of selection might help narrow down the candidate region or the set of variants contributing to disease susceptibility.

The efficient detection of loci affected by natural selection requires a characterization of the variation expected at neutrally evolving loci. However, neutral evolutionary processes are highly stochastic: a great deal of variation is expected across independent realizations of evolution, even if the genomic and evolutionary factors remain unchanged. Thus, robust inferences from polymorphism data will require the combined analysis of many unlinked loci. This is particularly true if the effects of population

history are to be distinguished from those of natural selection. While population history has genome-wide effects, locus-specific forces such as natural selection will only shape genetic variability at loci that are tightly linked to the selected locus.

Here, we compile a number of surveys of sequence variation at nuclear loci that experience recombination to provide a preliminary synthesis based on multi-locus information. An important feature of these data sets is that they are 're-sequencing' studies: every individual was sequenced, thus providing information about the entire frequency spectrum (rather than pre-selected variants). We outline the methods and sampling schemes of these studies in Box 1.

Descriptions of human sequence variation

Summary statistics and the standard neutral model

Several descriptive statistics (defined in Box 2) are commonly used to summarize the polymorphism data and to compare data sets. Information about the history of a locus can be gleaned from two (non-independent) aspects of the data: levels of polymorphism and LD. For example, in Table 1, we use nucleotide diversity, π, and the number of polymorphic sites, S, to describe the polymorphism levels at nuclear loci that experience recombination. An exact comparison of diversity levels across the 16 sequencing studies is complicated by the fact that they are of different sizes, consider varying proportions of coding and non-coding regions and use different schemes for sampling populations (see Box 1). If we ignore these caveats, the average nucleotide diversity across loci is 0.081%, roughly similar to previous estimates[4].

As can be seen in Table 1, levels of diversity estimated by variant detection arrays (or VDAs) are slightly lower than the average obtained from sequencing studies. Average nucleotide diversity is estimated to be 0.046% for a European sample[5] and 0.051% for a worldwide sample[6],

Molly Przeworski
mfprzewo@
midway.uchicago.edu

Richard R. Hudson*
rhudson1@
midway.uchicago.edu

Anna Di Rienzo‡
dirienzo@
genetics.uchicago.edu

Committee on
Evolutionary Biology,
1101 E. 57th Street,
University of Chicago, IL
60637, USA.
*Department of Ecology
and Evolution, 1101 East
57th Street, University of
Chicago, IL 60637, USA.
‡Department of Human
Genetics, 920 East 58th
Street, University of
Chicago IL 60637, USA.

BOX 1. Survey methods

Data collection

A problem in integrating different surveys is that the techniques for detecting variation and constructing haplotypes can differ in sensitivity and accuracy, with common frequency alleles usually being more easily detected than rare ones. The method most commonly used for variation detection in the surveys reviewed here is automated sequencing of PCR products. While the results might be approximately comparable with regard to the accuracy of variation detection, data quality and validation are rarely discussed explicitly[23,48]. A notable exception is single-strand conformational polymorphism (SSCP) analysis, used in two worldwide surveys of variation (Dys44 and ZFX). While SSCP is believed to be less sensitive than sequencing, its sensitivity is reported to be well above 80%, thanks to the use of different gel conditions and the analysis of partially overlapping PCR fragments[48].

The haplotypes were determined for each individual (with the exception of Dys44 in Table 3). However, as for variation detection, the methods for resolving the linkage phase varied across studies. Most investigators surveyed X-linked regions because it is possible to observe actual haplotypes in males. (The ZFX region was surveyed also in females and the haplotypes of the few multiple heterozygotes were either resolved by SSCP or inferred based on the unequivocally determined haplotypes.) For autosomal regions, the two chromosomes of an individual were separated experimentally by allele-specific PCR in the β-globin survey and by cloning for MC1R. An iterative approach, combining inference with empirical resolution of haplotypes, was implemented in three autosomal surveys (ACE, APOE and LPL).

Sampling problems

An important feature of the studies (except for the two SSCP surveys) is that estimates of variation were based on sequencing every individual in the sample. Past studies of human variation often used a 'reference' population sample (often of Caucasian origin) to discover polymorphisms; these would subsequently be scored in a worldwide sample to estimate allele frequencies. It was recognized that this scheme for sampling populations and polymorphisms could underestimate the amount of variation in the ethnic groups not represented in the reference sample, and consequently distort the reconstruction of human history. In addition, even if the reference sample were representative, this sampling approach did not provide estimates of the entire spectrum of allele frequencies in the worldwide sample. In the studies reviewed here, sampling schemes varied between two extremes: from worldwide collections of one or two individuals per population (e.g. Xq13.3) to large collections of individuals from two or more populations (e.g. LPL). A question worthy of further investigation is whether different sampling designs lead to contrasting pictures of human diversity. All of the sequencing studies include samples from Sub-Saharan Africa (or African Americans) and one or more of the major ethnic groups (Asians and Europeans). Because most high-frequency variants in humans appear to be shared across all major ethnic groups, the results of these studies might give a reasonably accurate reconstruction of the distribution of common variants; they provide more limited information regarding the rarer and population-specific variants.

BOX 2. Glossary

Neutral theory

This posits that the vast majority of polymorphisms within species and fixed substitutions between species are the result of the random drift of neutral mutations, rather than of natural selection[10]. Deleterious mutations are also assumed to occur, but are quickly eliminated.

The infinite site model

This assumes that each new mutation occurs at a site that has not previously mutated[49]. It is a good approximation when mutation rates at all sites are relatively low.

The standard neutral model

We refer to this as the assumption of a random-mating population of constant size, where mutations are neutral and occur according to the infinite site model.

Effective population size (N_e)

This term usually refers to the size of an ideal population with the same rate of genetic drift of gene frequencies as the actual population[50]. For example, if the population size fluctuated, the effective population size is equal to the harmonic mean of the population size.

Population mutation parameter (θ)

This term denotes $4N_e\mu$ (or $3N_e\mu$ if X linked), where μ is the neutral mutation rate per generation. This can be estimated from the number of nucleotide differences fixed between two species (or divergence), given an estimate of the time to the common ancestor.

Population recombination parameter (C)

Similarly, $C = 4N_ec$ where c is the recombination rate per generation.

Number of polymorphic sites in the sample (S)

S depends on the sample size. Therefore, comparing S across surveys requires a sample size correction. Under the standard neutral model,

$$E(S) = \theta \sum_{j=1}^{n-1} 1/j$$

where $E(S)$ denotes the expectation of S and n is the sample size. This equation leads to a commonly used estimate of θ based on S, denoted θ_w

Nucleotide diversity per site (π)

This is the frequency with which any two sequences in the sample differ at a site[13]. Under the standard neutral model, the means of π and θ_w are equal; their values will vary in accord with the expectation that, for example, higher mutation rates and/or larger population size will lead to higher polymorphism levels.

Frequency spectrum

The distribution of allele frequencies at polymorphic sites, specified by the proportion of alleles in different frequency ranges.

Minimum number of recombination events (RM)

Under the infinite site model, a recombination event between a pair of polymorphic sites can be inferred if all four haplotypes are observed. RM is the maximum number of (non-overlapping) such pairs[51].

whereas $\theta_w = 0.083\%$ in the combined African and European sample of Halushka et al.[7]. VDA studies might have less accuracy and less detection sensitivity than sequencing studies. Cargill et al.[6] verify all detected SNPs by sequencing; they estimate that 90% of all variants are detected by their method. Halushka et al. verify a subset; they estimate the rate of false positive to be 17% and the rate of false negative to be 8%.

Assuming a standard neutral model, it is possible to use the average nucleotide diversity to calculate the expected density of polymorphic sites for any range of allele frequencies. The availability of markers with appropriate

TABLE 1. Summaries of nuclear sequence variation

Region	chr.	n^a	bp	S^b	π (%)	Divergence (%)	θ_W (%)^c	D	RM	cM/Mb^d	Ref.
LPL^e	8	142	9700	79	0.166	1.31	0.149	0.36	22	2.40	52
β-globin	11	349	2670	19	0.157	1.16	0.110	1.06	3	2.23	29
MC1R	16	242	951	6	0.114	1.58	0.104	0.19	0	1.61	53
ACE	17	22	24000	74	0.091	2.74^f	0.085	0.32	5	0.94	19
APOE^g	19	192	5491	22	0.053	1.21	0.069	−0.62	8	2.98	–
Dys44	X	250	7622	34	0.093	1.74^h	0.073	0.74	–^{i}	3.51	19
Xq13.3	X	70	10163	33	0.033	0.93	0.067	−1.62	1	0.17	31
PDHA1	X	35	4200	24	0.178	0.83	0.139	0.97	3	6.00	54
DMD44^j	X	41	3000	19	0.141	0.80	0.148	−0.15	7	3.51	12
ZFX	X	336	1089	10	0.082	1.38	0.144	−0.94	1	0.68	48, 55
DMD7	X	41	2389	9	0.034	1.55	0.088	−1.78^k	1	3.51	12
HPRT	X	10	2485	4	0.038	0.97	0.057	−1.24	0	1.23	45
PLP	X	10	769	2	0.095	0.65	0.092	0.12	0	2.51	45
GK	X	10	1861	1	0.019	0.64	0.019	0.01	0	2.41	45
IL2RG	X	10	1147	0	0	0.78	0	0	0	3.33	45
IDS	X	10	1909	0	0	0.26	0	0	0	0.29	45
VDA survey	–	14	2 000 000	2748	0.046	–^{l}	0.044	0.29^m	N/A	N/A	5
VDA survey	–	114	196 200	560	0.051^n	0.60^o	0.054	0.00	N/A	N/A	6
VDA survey	–	148	190 000	874	–^{p}	0.60^q	0.083	–^{l}	N/A	N/A	7

^{a}The number of chromosomes in the sample.
^{b}Excluding insertion/deletion polymorphisms.
^{c}

$$\theta_W = S / \sum_{j=1}^{n-1} \frac{1}{j}$$

^{d}Estimated from a comparison of Genethon and GB4 maps, using the two closest unambiguously located microsatellite markers. Chromosome-specific conversion factors for cR/Mb were used^{56}.
^{e}The missing information was filled in with type of the most common allele.
^{f}A mutation rate of 2.74×10^{-9} per year was estimated from mouse/human divergence [assuming 80 million years ago (Mya)]. To obtain an estimate of the rate of divergence between chimpanzee and human, we assume a divergence time of 5 Mya.
^{g}Nickerson, D.A., Taylor, S.L., Weiss, K.M., Clark, A.G., Fullerton, S.M., Stengard, J., Boerwinkle, E. and Sing C.F., unpublished.
^{h}Divergence based on 2637 bp.
^{i}Cannot be determined from the results presented.
^{j}Non-overlapping sequence from the same intron as Dys44.
^{k}Significant departure at 0.05 level from the standard neutral model with no recombination (two-tailed). Significance levels were assessed by simulation.
^{l}Not reported.
^{m}Calculated from the reported values for n, S and π.
^{n}Based on 420 of the 560 polymorphisms (i.e. excluding polymorphisms identified by DHPLC alone).
^{o}Based on 136 kb of chimpanzee sequence.
^{p}Reported to be roughly similar to θ_w.
^{q}Based on 27 kb of chimpanzee sequence (human specific primers were used for the study of 30 kb, but some templates failed to amplify).

TABLE 2. Expected average density of polymorphic sites

Minor allele frequency	Expected density (bp)
0.01–0.10	1:515
0.10–0.20	1:1522
0.20–0.30	1:2290
0.30–0.40	1:2794
0.40–0.50	1:3045

allele frequencies is important for efficient disease mapping, by linkage analysis or LD-based approaches[8,9]. Polymorphic sites with minor allele frequencies in the 0.25–0.50 range, thought to be informative for mapping studies, are expected to occur on average every 1124 bp in regions of the genome that evolve neutrally. However, as shown in Table 2, the density of markers with minor allele frequencies of 0.40–0.50 is expected to be on average only one every 3045 bp.

The meaning of descriptive statistics can go beyond that of a mere summary of the data: for specific population models they allow estimates of population parameters. For example, under the standard neutral model[10], the nucleotide diversity is an estimate of the population mutation parameter, θ, which contains information about the long-term effective population size (Box 2). Similarly, LD statistics, such as the minimum number of recombination events, RM, can be used to estimate the population recombination rate C (Box 2). Under other models (e.g. of natural selection) the relationship between summaries of the data and population parameters can be much more complex and the summaries cannot be so readily interpreted.

Deviations from the standard neutral model

The standard neutral model is unlikely to apply without modifications to human populations but, nonetheless, it represents a useful first null hypothesis from which to test departures in the data. Under this model, higher neutral mutation rates lead both to higher polymorphism and to higher divergence levels. This property is the basis of a statistical test (known as the HKA test) that uses polymorphism and divergence data at two independent loci. In effect, the test asks whether the relative polymorphism levels are compatible with the relative divergence levels at the two loci[11]. Applying the HKA test to the DMD7 and the DMD44 regions in a comparison of non-African populations led to rejection of the neutral model[12]. Similarly, the β-globin region has been employed in the HKA test as a 'neutral' locus to reject neutrality at the PDHA1 region in the non-African sample. However, until neutrally evolving loci can be distinguished with greater confidence, significant departures in the HKA test will not lead to the unambiguous identification of loci affected by natural selection.

FIGURE 1. Pairwise linkage disequilibrium (I*D'*I) plotted as a function of physical distance

Levels of linkage disequilibrium between all possible pairs of polymorphic sites (as measured by I*D'*I) as a function of physical distance for five unlinked genomic regions (see Table 1). We included only polymorphic sites at which the minor allele frequency was 25% or greater. I*D'*I varies between 0 and 1 (Ref. 57). Pink diamonds represent the average I*D'*I in a window of 1 kb. Three regions (ACE, β-globin, and PDHA1) show high levels of LD extending over at least 2.5 kb, while the remaining two regions show a more rapid decline.

An alternative approach to testing the standard neutral model is based on the widely used statistic Tajima's D, which is a summary of the frequency spectrum[13]. Under the standard neutral model, the expectations of θ_W and of nucleotide diversity, π, are equal; the mean of the statistic D, which considers their difference, is approximately 0. Because rare alleles contribute less to π than to θ_W, a positive value of D reveals a relative excess of intermediate frequency alleles, as expected under a model of population subdivision[14] or an old balanced polymorphism[15]. By contrast, a negative value of D reflects a relative excess of rare variants, as might be expected after exponential growth[16] or after a 'selective sweep' (in which a new or rare variant was favored and quickly fixed in the population[17]). A previous analysis of sequence variation at a small sample of nuclear loci yielded only positive D values[18]. By contrast, only 6 out of 11 large data sets in Table 1 (arbitrarily defined as containing at least six polymorphic sites) have a positive value of D, while one has a significantly negative D value. Significantly positive D values are observed in several subsamples, even though the results for the corresponding total samples are not significant (the Sumatran subsample at β-globin, and the non-African subsample at Dys44). As is common, we assume no recombination to calculate significance levels for D. The p values will be smaller if recombination is properly taken into account. Because recombination events are detected in most data sets, p values based on the assumption of no recombination should be regarded as conservative.

In interpreting the sequence data available so far, it should be noted that most regions encompass or are associated with functional genes. Whether this choice leads to a bias is unknown. Interestingly, one of the regions with a sharply negative D value (Xq13.3) was selected because there were no known coding regions in the vicinity. Future studies should concentrate on determining whether coding regions consistently yield a different picture of human sequence variation compared with those that are not associated with coding regions.

The rate of decay of linkage disequilibrium

The rate of decay of LD with distance is critical for disease-mapping efforts because it leads to estimates of the distance at which disease associations can be detected and, in turn, of the density necessary for SNP maps. As with other aspects of sequence variation, LD is sensitive to the effects of natural selection and factors related to population history. Thus, the integration of empirical observations with theoretical investigations based on models of population histories might provide insights into the distribution of LD. The haplotype information provided by the surveys reviewed here make it easier and more reliable to draw inferences concerning LD and allow a glimpse at the
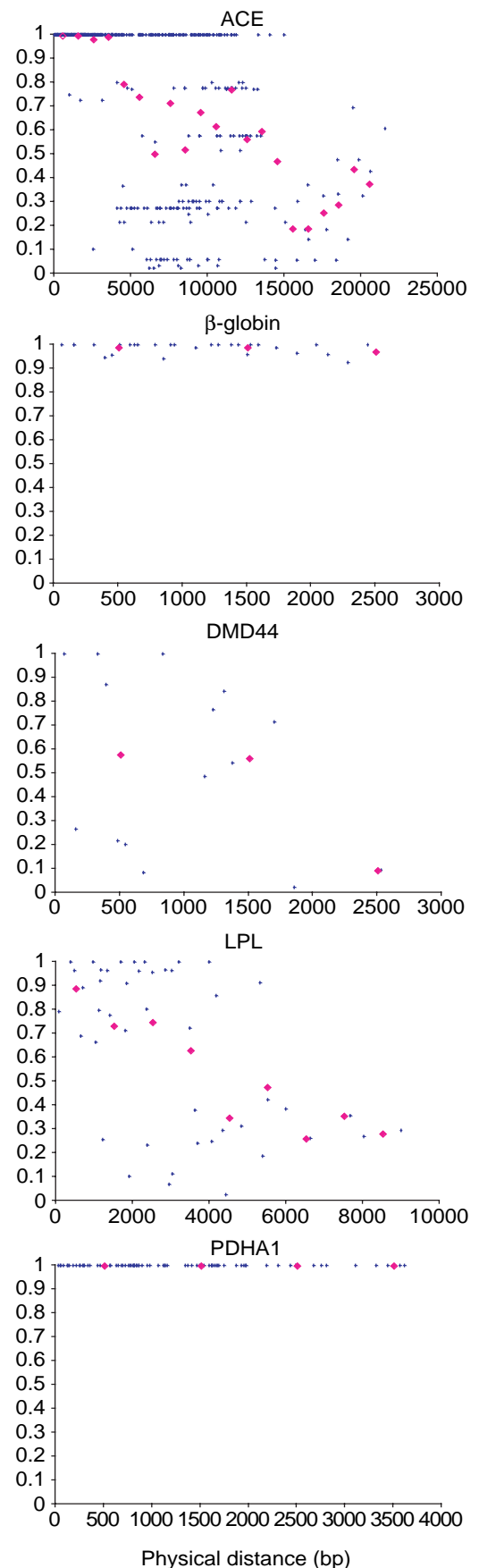
TABLE 3. Comparison of African vs non-African samples					
Locus	*n*	*S*	θ_W (%)	π (%)	D
Lpl Afr.	48	70	0.163	0.178	0.344
Lpl non-Afr.	94	53	0.107	0.150	1.296
β-globin Afr.	103	16	0.115	0.098	−0.423
β-globin non-Afr.	246	14	0.086	0.165	2.251
Ace Afr.	10	66	0.097	0.100	0.155
Ace non-Afr.	12	41	0.057	0.074	1.414
APOE Afr.	48	14	0.057	0.044	−0.736
APOE non-Afr.	144	16	0.053	0.055	0.101
Dys44 Afr.	115	32	0.079	0.099	0.776
Dys44 non-Afr.	135	21	0.050	0.085	1.950
Xq13.3 Afr.	23	24	0.064	0.035	−1.703
Xq13.3 non-Afr	47	17	0.038	0.031	−0.586
Dmd44 Afr.	10	15	0.177	0.188	0.297
Dmd44 non-Afr.	31	16	0.134	0.144	0.275
Dmd7 Afr.	10	6	0.089	0.080	−0.409
Dmd7 non-Afr.	31	4	0.042	0.011	−1.889

distribution of LD in the human genome based on sequence data. In Fig. 1, five large data sets with more than one minimum recombination event are used to illustrate the decay of LD with distance; only variants occurring at more than 25% frequency were included in this analysis. At least three (ACE, β-globin, and PDHA1) regions show high levels of LD extending over at least 2.5 kb, while the remaining two regions show a more rapid decline. Three surveys in Table 1 show little or no evidence for recombination over ~1–2 kb. The results for the ACE region are notable because strong LD extends beyond 10 kb and absolute pairwise LD (i.e. only two of the four possible haplotypes are present) is observed over 3.8 kb, involving 17 out of the 78 polymorphic sites. This pattern is unusual under the standard neutral model[19]. It could result from natural selection or from population history factors, such as population subdivision or a bottleneck. In the latter case, one would expect to see a similar pattern at a substantial fraction of other loci that experience comparable levels of recombination.

Role of population history in shaping patterns of variability

One of the main conclusions to emerge from studies of human variation is a greater variability in Sub-Saharan Africa, owing to a larger proportion of population-specific polymorphisms in Africa. The higher level of African diversity has been observed for mtDNA[20], microsatellites[21,22] and minisatellites[21,22] and is true for θ_W in most of the loci in Table 3, as well as for the VDA studies. In addition, the analysis of Dys44 (Ref. 23) showed that the population-specific variants have low frequencies of the non-ancestral alleles (defined relative to an outgroup sequence). This initial observation also applies to the sequence data sets in Table 3. As rare variants contribute more to θ_W than to π, the difference in diversity levels between African and non-African samples is more notable in θ_W than in π.

One explanation for the higher levels of African diversity is a history of population subdivision with a larger effective population size in Africa. For example, under a simple island model, higher-frequency variants are more likely to be shared among African and non-African populations because they are older and more likely to be sampled in the migration process and in population surveys. This scenario would result in a more homogeneous geographic distribution of common variants while rare variants would tend to be population-specific.

As shown in Table 3, non-Africans seem to have fewer rare variants than Africans, as manifested by the larger values of *D* outside Africa for many of the genomic regions examined. One possibility is that non-African populations experienced a phase of population size reduction, during which the rare variants were lost more quickly than the common ones. The apparent deficit of rare variants outside Africa, combined with the fact that most non-African variation is a subset of that found in Africa, is consistent with a recent origin of non-African populations from Sub-Saharan Africa. This migration from Africa is often thought to have been accompanied by population growth.

While it is clear that the human population size has increased dramatically over time, there is no consensus about the time of the expansion. mtDNA has a sharply negative *D*, consistent with either a selective sweep or exponential growth of human populations starting 50 000–100 000 years ago[24]. Microsatellite evidence points to an ancient expansion as well, but different studies have examined distinct models of population expansion and of the mutation process, and used different test statistics. Perhaps as a result, they come to conflicting conclusions about the time to the onset of growth, and about which populations were involved[25–28]. There is also considerable uncertainty about the mutation rate and the distribution of mutation sizes of microsatellites, leading to great variability of estimates and, possibly, to a bias towards older dates for the onset of growth if the variability in the mutation process is not properly taken into account[28].

In contrast to the rapidly evolving mtDNA and microsatellites, sequence variation data from several nuclear loci do not support a model of ancient exponential growth[18,29]. In fact, the values of *D* at β-globin, LPL, and Dys44 for the total samples are not compatible with a model of constant population size followed by more than 10 000 years of exponential growth[30]. The pattern is even more striking when the African and non-African subsamples are considered separately. In Table 3, four non-African samples are incompatible with the above model for any time of the onset of growth[30]. Similarly, the mean *D* from VDA studies is roughly 0; this value is highly unlikely under a model of long-term exponential growth (e.g. 100-fold growth for 50 000–100 000 years ago). A complicating factor for VDA studies is that the algorithms for analysing the VDA data might give higher certainty to alleles seen more than once (E. Lander, pers. commun.). If so, rare alleles might not be recovered accurately and the signal of population expansion could be weakened.

The ostensibly contrasting findings of mtDNA and nuclear loci can be reconciled by invoking the action of selection on a subset of loci, and/or a more complex demographic model. One possibility is that the significantly negative *D* values reflect recent selective sweeps at a site a short genetic (but not necessarily physical) distance away. In support of this, two loci with sharply negative values of *D* (mtDNA and Xq13.3) are all in areas of little[31] or no recombination, where the effects of natural selection are expected to extend over larger physical distances. The observed distribution of *D* values (including mtDNA and nuclear loci) might also be explained by invoking the

action of balancing or diversifying selection on a subset of nuclear loci.

If all loci are evolving neutrally, the observed D values might be compatible with a more complex model, such as a population contraction in non-African populations followed by recent population growth. Fay and Wu[32] investigate a model of population bottleneck (followed by a constant population size) and show that the patterns of polymorphism for mtDNA versus nuclear loci are expected to differ, as the smaller population size of mtDNA exacerbates the impact of a bottleneck and accelerates the rate of recovery after population size reduction. Whether this explanation is sufficient is unclear: while the introduction of a bottleneck followed by exponential growth increases the number of loci consistent with older onsets of growth, such a model cannot account for all the data[30].

Of course, even the complex models outlined above are likely to oversimplify the real history of human populations. Other demographic scenarios that might have affected patterns of sequence variation and LD include population subdivision with a change in migration rates over time[33] and admixture with archaic humans[34]. Interestingly, some ancient admixture models result in blocks of LD extending over much greater distances than expected under the standard neutral model[35]. Recent developments (e.g. admixture, population growth and founder effects) that have occurred in historical times are also likely to affect patterns of variation.

Evidence for the role of natural selection on shaping human variation

A fundamental question concerns the role that adaptation plays in molding the differences between species and in shaping patterns of population variation in the genome. Pursuing these issues could also help to identify regions of functional importance. To date, however, there is little unambiguous evidence for natural selection in sequence variation data. Whereas the D values for β-globin (Sumatran sample), Xq13.3 and DMD7 are unusual under the standard neutral model, natural selection is only one of many possible explanations. Similarly, alternative scenarios of either population history or a variety of forms of selection could potentially explain the significant departures observed in the HKA test (PDHA1 and DMD44/DMD7 in non-Africans).

Thus, while several loci in Table 1 exhibit some feature hinting at the action of natural selection, the identification of a particular locus as selected requires a more thorough characterization of population history. For example, if the correct demographic model turns out to be one of constant population size followed by exponential growth, several of the positive D values in Table 1 will be significantly high. Similarly, the patterns observed at PDHA1, MC1R and the Duffy blood group locus[36] might reflect population-specific selective pressures that could be analysed more fruitfully using information on the geographic structure of human populations.

A type of selection that has been extensively investigated is long-term balancing selection, which is expected to increase levels of variation at linked sites. It is interesting that the pattern of polymorphism at the β-globin region, the textbook example of heterozygote advantage, has not revealed this expected signature[18,29] – perhaps because the selection pressure was too recent to leave a

trace in sequence variation data. To date, the best known example of a locus evolving under long-term balancing selection in humans is the Major Histocompatibility Complex[37]. Similarly, in *Drosophila*, where the role of balancing selection has been extensively explored, there is little if any evidence for balanced polymorphisms[38].

Selection that reduces variability at neutrally evolving linked sites has also received considerable attention[39]. This broad class includes selective sweeps[40], background selection[41] (i.e. the continual elimination of strongly deleterious mutations), and selection with temporally varying selection coefficients[42]. Variation-reducing models are not expected to affect the rate of divergence between species at neutrally-evolving sites. The possibility that background selection is an important force shaping patterns of human variation was recently raised by the estimate of a high deleterious mutation rate per generation[43]. The signature of selective sweeps might also be common. In contrast to balancing selection, the signature of selection is stronger when the selective sweep occurred recently. It is noteworthy that the most dramatic environmental, including life-style, changes in humans are thought to have occurred over the past 10 000 years. This suggests that, as for sickle cell anemia, many selective pressures are relatively recent.

Neutral polymorphisms in a region of low recombination are affected by selection over a greater physical distance than are those in regions of normal to high recombination. As a result, if variation-reducing selection is prevalent, regions of low recombination should exhibit low levels of diversity. This prediction is realized in *Drosophila*, where diversity levels are correlated with crossing-over rates[44]. An alternative mechanistic explanation is that crossing-over is somehow mutagenic, leading to a higher mutation rate in areas of high exchange. If mutations were neutral, this explanation would predict that divergence levels should also be higher in areas of higher crossing-over. As this was not observed[44], the correlation was taken as evidence for an important effect of variation-reducing selection at linked loci.

Similarly, it has been proposed that levels of diversity are lower in regions with lower rates of crossing-over in humans[45]. Consistent with this proposal, the data sets in Table 1 show a significant correlation between nucleotide diversity and our estimates of the rate of crossing-over ($p = 0.01$, one-tailed Spearman rank correlation test). The correlation between sequence divergence and crossing-over rates is not statistically significant ($p = 0.381$). (Note that standard correlation tests are not entirely appropriate because loci in areas of low recombination are more likely to show extreme values of diversity under the standard neutral model than are loci in regions of normal recombination.) The current focus on studies of nuclear sequence variation will soon permit a more extensive evaluation of these correlations. However, estimates of the crossing-over rates in humans are notoriously unstable and much cruder than in *Drosophila*. In addition, the extent to which recombination rates vary across the genome is unknown, as is the scale over which such changes occur. A comparison of sequence-based physical distance and genetic distance for chromosome 22 suggests that crossing-over rates might vary by an order of magnitude on the scale of several megabases[46]. In the human pseudoautosomal region, single sperm typing reveals a threefold rate variation on the scale of a few hundred kilobases[47]. Overall rates of recombination are hard to gauge, both in *Drosophila* and in

humans, as the contribution of gene conversion to the rates of exchange is poorly characterized. Alternative assumptions about the sources of error and the extent of rate heterogeneity across the genome could lead to different estimates of the local recombination rate and different qualitative conclusions.

account for all the data, or if natural selection will have to be invoked for a subset of the loci. With proper sampling and methods of ascertainment, surveys of sequence variation should bring into focus the salient features of human history. Within this framework, inferences about selection at individual loci can proceed more fruitfully.

Conclusion

The available data make it clear that the standard neutral model is inadequate; it cannot account for the different levels of variation in African and non-African populations, or for the frequency spectra that are observed in these populations. It remains to be seen if a demographic model can

Acknowledgements

References

1 Risch, N. and Merikangas, K. (1996) The future of genetic studies of complex human diseases. *Science* 273, 1516–1517
2 Collins, F.S. *et al.* (1997) Variations on a theme: cataloging human DNA sequence variation. *Science* 278, 1580–1581
3 Neel, J.V. (1962) Diabetes Mellitus: a 'thrifty' genotype rendered detrimental by 'progress'? *Am. J. Hum. Genet.* 14, 353–362
4 Li, W.H. and Sadler, L.A. (1991) Low nucleotide diversity in man. *Genetics* 129, 513–523
5 Wang, D.G. *et al.* (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. *Science* 280, 1077–1082
6 Cargill, M. *et al.* (1999) Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22, 231–238
7 Halushka, M.K. *et al.* (1999) Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22, 239–247
8 Kruglyak, L. (1997) The use of a genetic map of biallelic markers in linkage studies. *Nat. Genet.* 17, 21–24
9 Kruglyak, L. (1999) Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nat. Genet.* 22, 139–144
10 Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature* 217, 624–626
11 Hudson, R.R. *et al.* (1987) A test of neutral molecular evolution based on nucleotide data. *Genetics* 116, 153–159
12 Nachman, M.W. and Crowell, S.L. Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy gene, *Dmd*, in humans. *Genetics* (in press)
13 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595
14 Tajima, F. (1989) DNA polymorphism in a subdivided population: the expected number of segregating sites in the two-subpopulation model. *Genetics* 123, 229–240
15 Hudson, R.R. and Kaplan, N.L. (1988) The coalescent process in models with selection and recombination. *Genetics* 120, 831–840
16 Slatkin, M. and Hudson, R.R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129, 555–562
17 Braverman, J.M. *et al.* (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140, 783–796
18 Hey, J. (1997) Mitochondrial and nuclear genes present conflicting portraits of human origins. *Mol. Biol. Evol.* 14, 166–172
19 Rieder, M.J. *et al.* (1999) Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.* 22, 59–62

20 Cann, R.L. *et al.* (1987) Mitochondrial DNA and human evolution. *Nature* 325, 31–36
21 Bowcock, A.M. *et al.* (1994) High resolution of human evolutionary trees with polymorphic microsatellites. *Nature* 368, 455–457
22 Armour, J.A. *et al.* (1996) Minisatellite diversity supports a recent African origin for modern humans. *Nat. Genet.* 13, 154–160
23 Zietkiewicz, E. *et al.* (1998) Genetic structure of the ancestral population of modern humans. *J. Mol. Evol.* 47, 146–155
24 Rogers, A.R. and Harpending, H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* 9, 552–569
25 Di Rienzo, A. (1998) Heterogeneity of microsatellite mutations within and between loci, and implications for human demographic histories. *Genetics* 148, 1269–1284
26 Kimmel, M. *et al.* (1998) Signatures of population expansion in microsatellite repeat data. *Genetics* 148, 1921–1930
27 Reich, D. and Goldstein, D. (1998) Genetic evidence for a Paleolithic human population expansion in Africa. *Proc. Natl. Acad. Sci. U. S. A.* 95, 8119–8123
28 Gonser, R. *et al.* (2000) Microsatellite mutations and inferences about human demography. *Genetics* 154, 1793–1807
29 Harding, R.M. *et al.* (1997) Archaic African and Asian lineages in the genetic ancestry of modern humans. *Am. J. Hum. Genet.* 60, 772–789
30 Wall, J.D. and Przeworski, M. When did the human population size start increasing? *Genetics* (in press)
31 Kaessmann, H. *et al.* (1999) DNA sequence variation in a non-coding region of low nucleotide diversity on the human X chromosome. *Nat. Genet.* 22, 78–81
32 Fay, J.C. and Wu, C.I. (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Mol. Biol. Evol.* 16, 1003–1005
33 Wakeley, J. (1999) Non-equilibrium migration in human evolution. *Genetics* 153, 1863–1871
34 Nordborg, M. (1998) On the probability of Neanderthal ancestry. *Am. J. Hum. Genet.* 63, 1237–1240
35 Wall, J.D. (2000) Detecting ancient admixture in humans using sequence polymorphism data. *Genetics* 154, 1271–1279
36 Hamblin, M.T. and Di Rienzo, A. (2000) Detecting the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* 66, 1669–1679
37 Hughes, A.L. and Yeager, M. (1998) Natural selection at major histocompatibility complex loci of vertebrates. *Annu. Rev. Genet.* 32, 415–435
38 Hey, J. (1999) The neutralist, the fly and the selectionist. *Trends Ecol. Evol.* 14, 35–38

39 Hill, W.G. and Robertson, A. (1966) The effect of linkage on limits to artificial selection. *Genet. Res.* 8, 269–294
40 Maynard-Smith, J.M. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35
41 Charlesworth, B. *et al.* (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303
42 Gillespie, J.H. (1997) Junk ain't what junk does: neutral alleles in a selected context. *Gene* 205, 291–299
43 Eyre-Walker, A. and Keightley, P.D. (1999) High genomic deleterious mutation rates in hominids. *Nature* 397, 344–347
44 Begun, D.J. and Aquadro, C.F. (1992) Levels of naturally occurring DNA polymorphism correlate with recombination rates in *D. melanogaster*. *Nature* 356, 519–520
45 Nachman, M.W. *et al.* (1998) DNA variability and recombination rates at X-linked loci in humans. *Genetics* 150, 1133–1141
46 Dunham, I. *et al.* (1999) The DNA sequence of human chromosome 22. *Nature* 402, 489–495
47 Lien, S. *et al.* (2000) Evidence for heterogeneity in recombination in the human pseudoautosomal region: high resolution analysis by sperm typing and radiation-hybrid mapping. *Am. J. Hum. Genet.* 66, 557–566
48 Jaruzelska, J. *et al.* (1999) Spatial and temporal distribution of the neutral polymorphisms in the last ZFX intron: analysis of the haplotype structure and genealogy. *Genetics* 152, 1091–1101
49 Kimura, M. (1969) The rate of molecular evolution considered from the standpoint of population genetics. *Proc. Natl. Acad. Sci. U. S. A.* 63, 1181–1188
50 Wright, S. (1931) Evolution in mendelian populations. *Genetics* 16, 97–159
51 Hudson, R.R. and Kaplan, N.L. (1985) Statistical properties of the number of recombination events in the history of a sample of DNA sequences. *Genetics* 111, 147–164
52 Clark, A.G. *et al.* (1998) Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am. J. Hum. Genet.* 63, 595–612
53 Rana, B.K. *et al.* (1999) High polymorphism at the human melanocortin 1 receptor locus. *Genetics* 151, 1547–1557
54 Harris, E.E. and Hey, J. (1999) X chromosome evidence for ancient human histories. *Proc. Natl. Acad. Sci. U. S. A.* 96, 3320–3324
55 Jaruzelska, J. *et al.* (1999) Is selection responsible for the low level of variation in the last intron of the ZFY locus? *Mol. Biol. Evol.* 16, 1633–1640
56 Hudson, T.J. *et al.* (1995) An STS-based map of the human genome. *Science* 270, 1945–1954
57 Lewontin, R.C. (1964) The interaction of selection and linkage. I. General consideration; heterotic models. *Genetics* 49, 49–67

Letters to the Editor

We welcome letters on any topic of interest to geneticists and developmental biologists.
Please write to:
The Editor
TIG@current-trends.com

Trends in Genetics, Elsevier Science London, 84 Theobald's Road, London UK WC1X 8RR.
Fax: 020 7611 4470