muscular dystrophy, neurofibromatosis and retinoblastoma), there is a large male excess. So, I believe the evidence is convincing that the male base-substitution rate greatly exceeds the female rate.

There are a number of reasons why the historical male:female mutation ratio is likely to be less than the current one, the most important being the lower age of reproduction in the earliest humans and their ape-like ancestors. More studies are needed. But whatever the outcome of further investigation, this unique transposition event and perhaps others like it are sure to provide important cytogenetic and evolutionary insights.

Acknowledgements

I have profited greatly from discussions with my colleagues B. Dove, B. Engels, C. Denniston and M. Susman. My greatest debt is to D. Page for a continuing and very fruitful dialog and for providing me with unpublished data.

References

1 Vogel, F. and Motulsky, A. G. (1997) *Human Genetics: Problems and Approaches*. Springer-Verlag

2 Crow, J. F. (1997) The high spontaneous mutation rate: is it a health risk? *Proc. Natl. Acad. Sci. U. S. A.* 94, 8380–8386

3 Behossian, H.B. *et al.* (2000) Unexpectedly similar rates of nucleotide substitution found in male and female hominids. *Nature* 406, 622–625

4 Shimmin, L.C. *et al.* (1993) Potential problems in estimating the male-to-female mutation rate ratio from DNA sequence data. *J. Mol. Evol.* 37, 160–166

5 Richardson, C. *et al.* (1998) Double-strand break repair by interchromosomal recombination: suppression of chromosomal translocation. *Genes Dev.* 12, 3831–3842

6 Engels, W.R. *et al.* (1994) Long-range *cis* preference in DNA homology search over the length of a *Drosophila* chromosome. *Science* 263, 1623–1625

7 Crow, J.F. (2000) The origins, patterns and implications of human spontaneous mutation. *Nat. Rev. Genet.* 1, 40–47

8 Risch, N. *et al.* (1987) Spontaneous mutation and parental age in humans. *Am. J. Hum. Genet.* 41, 218–248

Detecting the form of selection from DNA sequence data

Clues to our evolutionary history lie hidden within DNA sequence data. One of the great challenges facing population geneticists is to identify and accurately interpret these clues. This task is made especially difficult by the fact that many different evolutionary processes can lead to similar observations. For example, low levels of polymorphism within a region can be explained by a low local mutation rate, by selection having eliminated deleterious mutations, or by the recent spread to fixation of a beneficial allele. Theoretical advances improve our ability to distinguish signals left by different evolutionary processes. In particular, a new test might better detect the footprint of selection having favored the spread of a beneficial allele.

Sarah P. Otto
otto@zoology.ubc.ca

Department of Zoology,
University of British
Columbia, Vancouver,
British Columbia,
Canada V6T 1Z4.

As the amount of sequence data has grown, so too has the body of population genetics theory that aims to interpret this data in terms of past evolutionary processes. What makes this task especially difficult is that evolutionary history is complex. At particular times in the past and at particular sites in the genome, selection has favored the spread of new alleles (directional selection; see Box 1), but at other times and sites, selection has prevented the spread of mutations (purifying selection). The nature of selection, the demographic history of a population, and the spectrum of mutations have major effects on the evolution of genetic sequences. This complexity must be addressed directly if we are to isolate the traces of a particular evolutionary event from others that occur simultaneously. Two recent papers[1,2] make important strides in this direction by distinguishing the patterns left in DNA sequences by directional and purifying selection.

From early in the history of population genetics, mathematical models have been used to infer past evolutionary processes from extant population data. In his pioneering work on mutations, J.B.S. Haldane[3,4] demonstrated how the current frequency and fitness effects of deleterious alleles can be used to infer past rates of mutation. Using this method, Haldane obtained one of the first mutation rate estimates for a human disease: 1 mutant every 50 000 generations for hemophilia, an X-linked disorder[4]. Haldane's method has been widely applied to estimate mutation rates for human diseases[5], but it works only in cases where there is extensive data on the survival and fertility of affected individuals. What if the nature of selection is unknown? In cases where selection is difficult or impossible to measure directly, how can we use genetic data to infer the history of selection?

The neutral expectation

To infer that selection has acted within a region of the genome, one must first reject the null hypothesis that no selection has acted on the sequences in question. The neutral theory of molecular evolution, developed largely by Kimura[6], establishes what to expect under the null hypothesis, by describing the pattern of sequence evolution under the forces of mutation and random genetic drift without the action of selection. One particularly important result from the neutral theory is that the rate at which one nucleotide is replaced by another nucleotide throughout a population (a substitution) should equal the mutation rate at that site (ν), regardless of the population size. At first this result is rather surprising because so many more mutations arise in a large population each generation than in a small population. However, the higher number of mutations is counterbalanced by the smaller chance that each mutation will fix within a large population. In addition, the neutral theory makes predictions about the pattern of polymorphism expected among sequences drawn from the same species. For example, the probability that any two randomly sampled sequences differ at a

BOX 1. Glossary

Directional selection

Selection that acts on an allele whose carriers have higher than average fitness. Such alleles are called ‘beneficial’. In the absence of other forces, the frequency of a beneficial allele will increase over time at a rate that is proportional to the strength of directional selection.

Purifying selection

Selection that acts against an allele whose carriers have lower than average fitness. Such alleles are called ‘deleterious’. In the absence of other forces, the frequency of a deleterious allele will decrease over time at a rate that is proportional to the strength of purifying selection.

Random genetic drift

Changes in the frequency of a gene that occur over time because the number of offspring born to parents carrying the gene is subject to chance variation. Even if the expected number of offspring per parent is one, an individual might be lucky and leave several offspring (or unlucky and leave none). Consequently, the number of copies of a gene transmitted to the next generation might be greater or less than expected.

Effective population size (N_e)

The effective population size is defined by the amount of random genetic drift that a population exhibits. When gene frequencies drift substantially from generation to generation, the population has a low N_e and vice versa. The effective population size equals the census size for the ‘idealized’ population used in mathematical models. For real populations, the effective size can be calculated given sufficient demographic information.

Hitchhiking

An allele on a chromosome that carries a new beneficial allele remains coupled to the beneficial allele until separated by recombination. During this time, the neighboring allele will increase in frequency along with the beneficial allele, even if it does not directly affect fitness.

Background selection

An allele is less likely to persist within a population and might be eliminated if it occurs on a chromosome (a ‘genetic background’) with deleterious alleles. Note that purifying selection acts on the deleterious mutations themselves, whereas background selection acts on the neighboring alleles.

BOX 2. Allele genealogies with and without selection

Both background selection and hitchhiking reduce the average level of genetic variation relative to the neutral expectation. The mathematics behind this result are novel and interesting. The proofs rely on looking backwards over evolutionary time to see when two lineages last shared a common ancestor at a marker site. This event is called a ‘coalescent.’ Consider the genealogy illustrated in Fig. Ia for a simple population of only four alleles (circles), which can represent either four haploid or two diploid individuals. Going forwards in time (top to bottom), parent alleles are connected to their offspring by lines, and every individual in the population must be tracked. In the absence of selection (Fig. Ia), an offspring allele is equally likely to be descended from any of the alleles in the previous generation. Of course, this does not mean that every allele leaves offspring; in particular, the alleles marked with an X do not reproduce. Going backwards in time (bottom to top), offspring alleles are connected to their parents, but only alleles that are ancestors of present-day alleles must be examined. It is the fact that alleles not contributing to the current population can be ignored (shaded circles connected by shaded lines) that makes the coalescent approach so powerful, especially when analysing large populations.

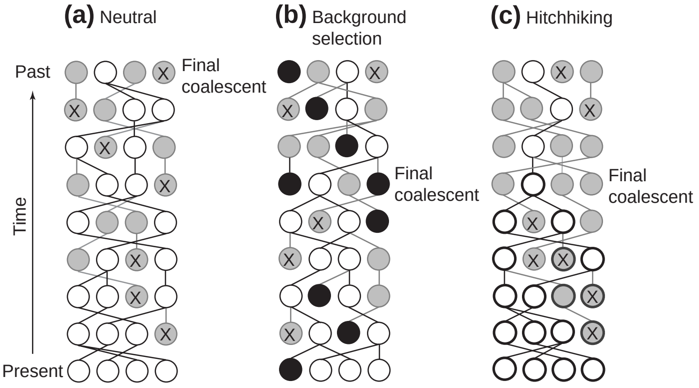


FIGURE I

trends in Genetics

In Fig. Ia, the first coalescent event occurs in one generation (the third and fourth alleles descend from the same parent), and all present-day alleles share a common ancestor eight generations ago (the final coalescent event, marked at the top). With background selection (Fig. Ib), some parent alleles are especially unlikely to leave offspring because they are located on chromosomes with many deleterious mutations (dark circles). This makes it more likely that two alleles will descend from the same parent allele (i.e. one that carries few mutations), which speeds up the coalescent process[15]. In the example illustrated, all present-day alleles share a common ancestor five generations ago. With hitchhiking (Fig. Ic), some parent alleles are especially likely to leave offspring because they are located on the same chromosome as a beneficial allele (denoted by a thick perimeter). Again, this makes it more likely that two alleles will descend from the same parent (i.e. one that carries the beneficial allele), which speeds up the coalescent process[16]. Indeed, mathematical analyses demonstrate that the coalescence time is shorter, on average, with either background selection[15] or hitchhiking[16]. (How much shorter depends on several factors, including how tightly linked the marker gene is to the genes subject to selection[15,16].) When the coalescence time is shorter, there is less time for mutations to accumulate while two lineages are distinct. Therefore, the average number of genetic differences between two sequences and, consequently, the level of genetic variation are reduced when either background selection or hitchhiking act in a region of the genome.

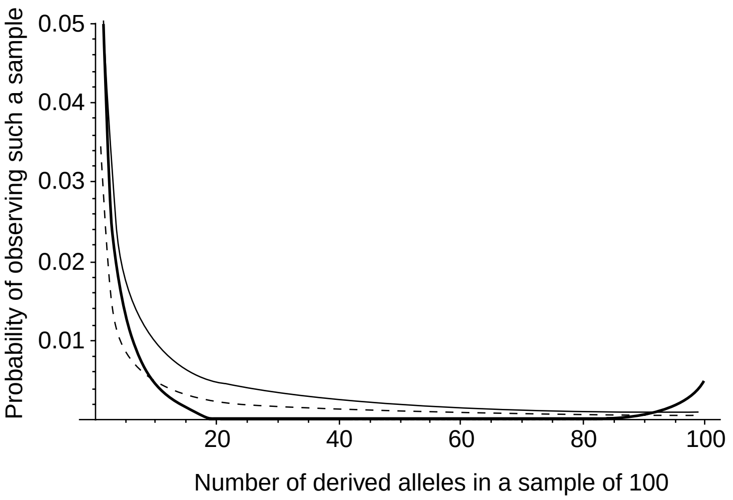
particular position in the genome is given by the equation $\theta = 4N_e\nu$, provided that θ is small, for a diploid species with an effective population size of N_e. This makes sense; two individuals are more likely to differ if the population is large or the mutation rate is high. It is the wealth of such testable predictions that has made the neutral theory so central to modern evolutionary biology.

Tests of selection

To test for selection, population geneticists identify some quantitative feature of DNA sequences and then determine whether the feature observed in data differs significantly from the neutral expectation. For example, one simple yet ingenious test for selection is the McDonald–Kreitman test[7]. As mentioned above, the number of substitutions between two species and the level of polymorphism within a species are both proportional to the mutation rate ν, under the neutral theory. This prediction holds true whether one counts nucleotide differences that alter the amino acid sequence (replacement changes, with a mutation rate ν_r) or those that do not (silent changes, with a mutation rate ν_s). This prediction can be tested using a 2 × 2 contingency table that examines whether the ratio of replacement to silent changes within a species differs significantly from the ratio of replacement to silent changes between species. If a significant difference is found, then at least one category of sites is not evolving as expected under the neutral theory, and the null hypothesis that selection is absent can be rejected. Despite the elegance and simplicity of this test, it does have its shortcomings. In particular, if both purifying and directional selection have taken place, then any result is possible, including failure to reject a false null hypothesis.

Building upon previous studies, two recent papers[1,2] improve our ability to distinguish the effects of directional

FIGURE 1. The distribution of new marker alleles



trends in Genetics

Consider a sample of 100 typed alleles at a polymorphic marker site where the mutation rate is $\nu = 10^{-6}$ in a population of effective size $N_e = 25\,000$ ($\theta = 4N_e\nu = 0.1$). We say that a marker allele is 'new' or 'derived' if it arose by mutation from a previous ('old' or 'ancestral') marker allele. As noted by Fay and Wu[1], closely related species can be used to identify which marker alleles are derived and which are ancestral. Under the neutral theory, the probability of observing a new marker allele m times in a sample of alleles is approximately θ/m (thin curve)[17]. Notice that it is very unusual for a derived marker allele to constitute a majority of the sample (right). Background selection essentially lowers the effective population size[15], reducing θ, and making it even less likely to observe a derived marker at high frequency (dashed curve). Following a hitchhiking event, however, derived alleles will occasionally be observed at high frequency if they occur on the chromosome where the beneficial allele appeared (thick curve; data from Ref. 1). Here we illustrate the pattern observed when the recombination rate between the marker gene and the selected gene is 10^{-3}. With much lower recombination rates, one of the marker alleles will almost certainly fix, and polymorphism will be absent following the hitchhiking event. With much higher recombination events, the marker alleles do not remain in association with the beneficial allele for long, and the frequency distribution will be little affected by hitchhiking.

selection from purifying selection. To place their results in context, however, we must first consider how each type of selection acting alone affects neighboring sites not under selection, which, for clarity, will be called 'marker' sites. When a beneficial mutation occurs and spreads as a result of directional selection, marker alleles that were lucky enough to be on the same chromosome as the beneficial mutation will tend to increase in frequency also, a phenomenon known as hitchhiking[8]. Similarly, under purifying selection, a marker allele that happens to be on the same chromosome as a deleterious mutant allele will decrease in frequency as a result of this association. This second phenomenon is referred to as background selection, to emphasize the fact that the persistence of a marker allele within a population depends on the number of deleterious alleles in its genetic background[9]. Interestingly, either form of selection reduces the average amount of genetic variation in a region surrounding the selected sites (see Box 2).

A low level of genetic variation within a region of the genome could, however, result from a locally low mutation rate. To avoid this problem, Tajima[10] developed a test for selection that is insensitive to the level of mutation within a region. His test statistic (D) measures whether the level of heterozygosity is lower than expected under the neutral theory based on the number of polymorphic sites. The amount of heterozygosity observed at a polymorphic site depends on the frequency of each allele – heterozygosity is low if only one allele is common and is high when more than one allele is common. However, observing a significantly low level of heterozygosity for the number of polymorphisms is consistent with several possible processes, including hitchhiking, background selection and a recent expansion in population size[9,11].

How can we identify which process best explains a local reduction in genetic variation? Kim and Stephan[2] analyzed the joint effects of directional and purifying selection on levels of genetic variation. They found that the two forms of selection can interact strongly. For example, if many mutations are deleterious and are eliminated by purifying selection, the rare beneficial mutation that occurs might arise on a genetic background of low fitness and fail to spread. Hence, when purifying selection is acting also, fewer hitchhiking events will result from directional selection, and the level of genetic variation will be less affected. Furthermore, they confirmed that there is a distinct signal left by hitchhiking but not by background selection: with hitchhiking, it is not unusual for marker alleles that have recently arisen by mutation to be found at high frequency because they increased along with the favorable allele (Fig. 1). However, with no selection or with background selection newly arisen marker alleles are almost always less common than marker alleles that have existed within the population for a longer period of time.

Fay and Wu[1] took advantage of this predicted difference to devise a test for past hitchhiking events. Based on theoretical results of Fu[12], they developed a new statistic (H) that is sensitive to an excess of new marker alleles at high frequency, which is consistent with hitchhiking but not background selection. What is important about this new test statistic is that it can be used in combination with other test statistics (such as Tajima's D) to reveal more about the pattern of selection. A significantly negative D is consistent with hitchhiking, but it is consistent with many other processes as well. If H is significant also, then background selection or a recent population expansion can be ruled out. Conversely, if D is significant but not H, other explanations besides hitchhiking gain credibility for the reduced level of genetic variation (as long as the power of the H statistic to detect selection can be shown to be high under the conditions of the empirical study).

Fay and Wu[1] demonstrate the application of their test using data from *Drosophila*, finding evidence for hitchhiking around the genes encoding achaete and accessory gland protein (*Acp26Aa*). In the *Acp26Aa* region, the authors pinpoint the site of directi onal selection to within a region of ~350 base pairs. Within this region, nearly all genetic variation is eliminated, consistent with the prediction that tightly linked alleles should hitchhike all the way to fixation. Around this region, there is a large excess of high-frequency derived alleles, where recombination arrested the hitchhiking process before ancestral alleles were entirely lost.

Future directions

DNA sequences are the modern hieroglyphics, challenging us to decode the signals of our past that lie embedded within them. The Rosetta stone, with the same text written in ancient Greek, Demotic and Egyptian hieroglyphics, allowed Champollion to decipher hieroglyphics in 1822. To translate the signals of evolution left in the genetic code, we must construct our own Rosetta stone based on models of evolutionary processes. Two recent papers[1,2] provide us with tools to extract more information about the form of selection from genetic data. The test developed by Fay and Wu[1], for example, could be used to infer whether genetic differences between humans and chimps represent chance substitutions of neutral mutations, or whether they might be the products of directional selection.

The interpretation of our evolutionary past will not, however, be simple. We need to investigate the robustness of our conclusions to changes in assumptions. For example, do the results of the Fay and Wu test[1] depend on the assumption that each mutation generates a unique allele? Furthermore, we need to account explicitly for the various sources of uncertainty in our conclusions. For example, how certain are we that an allele is ancestral or derived? Finally, we need to investigate the statistical power of the various tests. How can we best estimate a number of parameters, including the strength of background selection, the strength of directional selection and the timing of hitchhiking events? How can we distinguish signals left by selection from those left by demographic events, such as the population expansion and widespread migration that have characterized human history[13,14]? Over the next few years we will learn much about the strengths and weaknesses of different approaches to infer evolutionary events from DNA sequence data. It will be fascinating to see what clues to our evolutionary past are revealed.

Acknowledgements

I gratefully acknowledge helpful comments on the manuscript from P. Awadalla, A. Poon, M. Whitlock and two anonymous reviewers. Funding was provided by the Natural Science and Engineering Research Council (Canada).

References

1 Fay, J.C. and Wu, C.I. (2000) Hitchhiking under positive Darwinian selection. *Genetics* 155, 1405–1413
2 Kim, Y. and Stephan, W. (2000) Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* 155, 1415–1427
3 Haldane, J.B.S. (1927) A mathematical theory of natural and artificial selection, Part V: Selection and Mutation. *Proc. Cambridge Philos. Soc.* 23, 838–844
4 Haldane, J.B.S. (1935) The rate of spontaneous mutation of a human gene. *J. Genetics* 31, 317–326
5 Cavalli-Sforza, L.L. and Bodmer, W.F. (1971) *The Genetics of Human Populations*, W.H. Freeman
6 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press
7 McDonald, J.H. and Kreitman, M. (1991) Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351, 652–654
8 Maynard-Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favorable gene. *Genet. Res. Camb.* 23, 23–35
9 Charlesworth, B. *et al.* (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303
10 Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 437–460
11 Fu, Y.X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. *Genetics* 147, 915–925
12 Fu, Y.X. (1995) Statistical properties of segregating sites. *Theor. Popul. Biol.* 48, 172–197
13 Excoffier, L. and Schneider, S. (1999) Why hunter-gatherer populations do not show signs of pleistocene demographic expansions. *Proc. Natl. Acad. Sci. U. S. A.* 96, 10597–10602
14 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
15 Hudson, R.R. and Kaplan, N.L. (1995) The coalescent process and background selection. *Philos. Trans. R. Soc. London Ser. B* 349, 19–23
16 Kaplan, N.L. *et al.* (1989) The ‘hitchhiking effect’ revisited. *Genetics* 123, 887–899
17 Watterson, G.A. (1975) On the number of segregating sites. *Theor. Popul. Biol.* 7, 256–276

The evolutionary history of ribosomal protein RpS14:

horizontal gene transfer at the heart of the ribosome

The comparative study of the flow of complete genome sequences has helped to reveal the important role of horizontal gene transfer (HGT) in prokaryotic evolution[1–5]. Complete genome sequences not only allow the identification of the set of transferred genes present in an organism, but also help to reconstruct the fate of these genes once incorporated into the receiving genome. In fact, if homologues were already present in the host, transferred genes can replace the host genes via recombination, or they can simply be integrated at another site in the host genome leading to the presence of multiple copies. This is especially interesting in the case of genes belonging to conserved operons, in order to know whether the selective advantage that the transferred genes provide to the host can compensate the disadvantage intrinsic to operon disruption (which hinder the coordinated expression of all genes).

As a model, we have analysed the spectinomycin (*spc*) operon, a conserved operon that groups the genes coding for the ribosomal proteins RpL14, RpL24, RpL5, RpS14,