Genome Analysis

Single nucleotide polymorphisms and recombination rate in humans

Michael W. Nachman

Levels of heterozygosity for single nucleotide polymorphisms vary by more than one order of magnitude in different regions of the human genome. Regional differences in the rate of recombination explain a substantial fraction of the variation in levels of nucleotide polymorphism, consistent with the widespread action of natural selection at the molecular level.

The last few years have seen an explosion of interest in documenting levels and patterns of nucleotide variability in the human genome. One fruit of the Human Genome Project has been the identification of millions of single nucleotide polymorphisms (SNPs), and a central challenge for population geneticists now is to explain the frequency and distribution of these SNPs throughout the genome. There are at least three reasons for the interest in human nucleotide polymorphisms. First, SNPs hold great promise as markers for mapping polygenic disease loci in natural populations. Knowledge of the frequency and underlying patterns of association among SNPs unrelated to disease is essential for interpreting patterns of linkage disequilibrium between markers and candidate disease genes. Second, nucleotide polymorphisms can shed light on human history, including relationships among ethnic groups, migrations and changes in population size. Third, the distribution of variation can teach us about the relative importance of forces such as selection, mutation, migration, recombination and genetic drift, and thereby can help us understand the nature of the evolutionary process at the molecular level.

Predictions of population genetic models

One of the simplest and most important population genetic models is the neutral theory, which posits that levels of genetic variation are determined solely by the input of new alleles through mutation and the loss or fixation of alleles because of genetic drift[1]. According to this model, natural selection is a negligible force at the molecular level. Without selection, levels of nucleotide heterozygosity are independent of the rate of recombination. Recombination might shuffle nucleotide polymorphisms to create new haplotypes, but it will not alter the average density of SNPs across the genome.

By contrast, selection can alter levels of nucleotide polymorphism at linked sites to a degree determined by the local rate of recombination. In regions of the genome with no recombination, selection will have a pronounced effect on linked neutral sites, whereas in regions of the genome with a high level of recombination, selection will have a negligible impact on linked sites. Two distinct models of selection are expected to reduce levels of heterozygosity at linked sites (Fig. 1). The first, genetic hitchhiking, occurs when an adaptive mutation sweeps through a population to fixation and drags with it linked neutral variants[2]. In the aftermath of a selective sweep, variation will be reduced or eliminated from the region surrounding the site under selection, and the strength of this effect will be inversely proportional to recombination rate. If adaptive mutations are common, genetic hitchhiking is expected to produce an overall positive correlation between rates

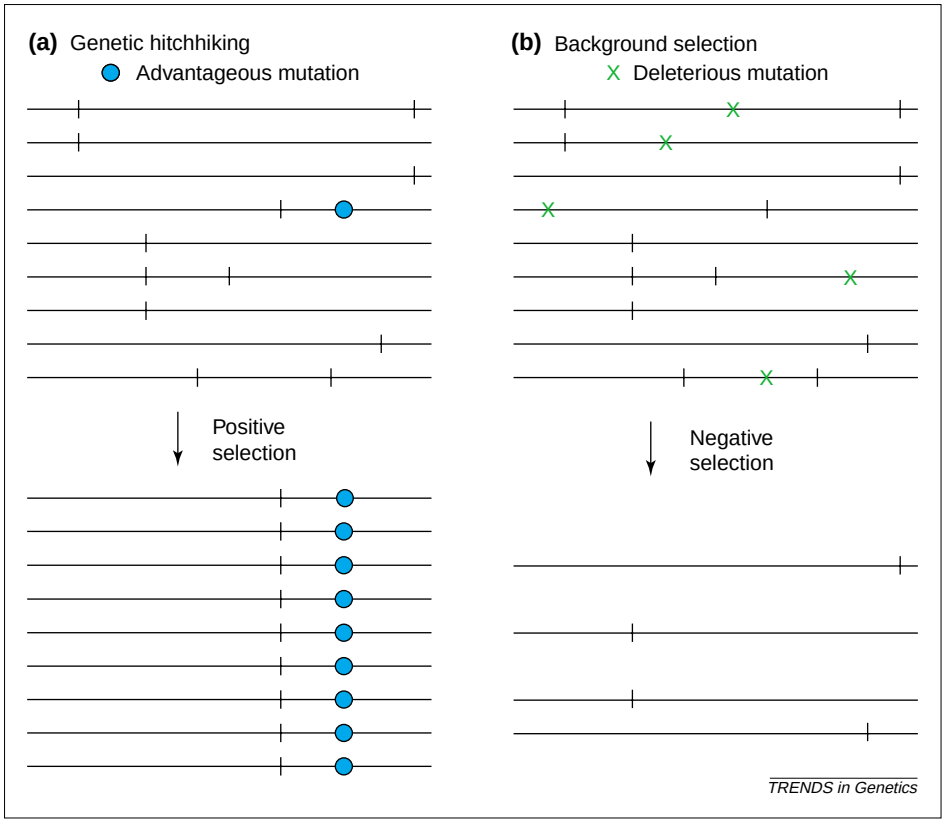

Fig. 1. Schematic model of genetic hitchhiking and background selection without recombination. Horizontal lines depict haplotypes in a population, and vertical marks depict neutral mutations. Under genetic hitchhiking (a), an advantageous mutation arises and is fixed by positive selection, dragging linked neutral variants with it. In the aftermath of a complete selective sweep without recombination, all individuals possess the same haplotype. If recombination occurs during the selective sweep (not shown), some variation might remain in the population. Under background selection (b), deleterious mutations arise and are eliminated by selection, eliminating linked neutral variants with them. In the presence of recombination (not shown), neutral variants might escape elimination. Formally, background selection is equivalent to a reduction in the effective population size by a fraction f_0, the equilibrium frequency of chromosomes free of deleterious mutations.

  PII: S0168-9525(01)02409-X

Table 1. Nucleotide variability at autosomal and X-linked genes in humans[a]

Locus	Chromosome	N[b]	L (bp)	π(%)[c]	θ(%)[d]	Tajima's D[e]	Divergence (%)[f]	GC (%)	cM Mb^{-1}[g]	Refs
β-globin	11	349	2670	0.157	0.110	1.06	1.34	39.9	1.84	29
Lpl	8	142	9734	0.169	0.147	0.48	1.48	40.8	2.26	30
Hox B6	17	210	1000	0.057	0.068	–0.28	0.59	50.8	0.97	31
Mc1r	16	242	951	0.096	0.104	–0.15	1.58	63.3	1.92	32
Ace	17	22	24070	0.093	0.089	0.11	–	58.7	0.79	33
Apoe	19	192	5491	0.053	0.069	–0.62	1.18	59.5	1.37	34
OR ψ	17	66	4535	0.110	0.097	0.40	2.34	53.3	1.59	35
Duffy[h]	1	34	2931	0.131	0.108	0.68	–	54.6	1.39	36
22q11.2	22	128	9901	0.088	0.132	–1.03	1.35	46.0	1.04	37
1q24	1	122	8991	0.058	0.060	–1.21	0.62	31.5	1.54	38
Plp	X	10	769	0.095	0.092	0.12	0.65	38.5	3.44	7
Gk	X	10	1861	0.019	0.019	0.02	0.64	33.9	3.25	7
Il2rg	X	10	1147	0.000	0.000	–	0.78	48.0	0.45	7
Ids	X	10	1909	0.000	0.000	–	0.26	49.3	3.92	7
Hprt	X	10	2485	0.038	0.057	–1.25	0.97	40.5	3.46	7
Pdha1	X	35	4153	0.189	0.146	1.03	1.10	45.9	4.42	39
Xq13.3	X	69	10163	0.036	0.068	–1.61	0.92	37.8	0.67	40
Zfx	X	336	1089	0.082	0.144	–0.95	1.17	39.1	3.57	41
Dmd I44	X	41	3000	0.141	0.148	–0.16	0.90	37.4	3.39	42
Dmd I7	X	41	2389	0.034	0.088	–1.79	1.63	33.1	3.39	42
Msn	X	41	4622	0.035	0.045	–0.67	0.80	42.3	0.57	[i]
Alas	X	41	5125	0.014	0.032	–1.53	0.63	42.5	0.64	[i]

[a]One recently surveyed[44] region flanking the hypervariable minisatellite MS205 from chromosome 16p13.3 is not included in the table. This region lies very near the p telomere of chromosome 16 and reliable estimates of recombination rate are not available. Watterson's θ for this region is 0.46%, more than four times the average value for autosomes. This locus lies adjacent to a recombinational hotspot with rates estimated to be more than tenfold above the genomic average recombination rate.
[b]Sample size is number of chromosomes.
[c]Nucleotide heterozygosity, π, is equivalent to the average pairwise difference among all alleles in the sample, expressed per site[16].
[d]Watterson's θ is a measure of nucleotide variability based on the number of polymorphic sites[17].
[e]Tajima's D is a measure of the skew in the frequency distribution of polymorphic nucleotides relative to an equilibrium neutral model. A negative value indicates an excess of low-frequency polymorphisms, while a positive value indicates a deficiency of low-frequency polymorphisms[26].
[f]Divergence is based on a randomly chosen allele from humans and a randomly chosen allele from chimpanzees.
[g]Recombination rates taken from Ref. 11.
[h]Based on European sample; African alleles, believed to be under selection, are excluded.
[i]M.W. Nachman *et al.*, unpublished.

of recombination and levels of nucleotide heterozygosity at neutral sites[3,4]. A second model, background selection, is based on selection removing deleterious mutations and linked neutral variants from a population[5]. If the deleterious mutation rate is sufficiently large and selection is weak, background selection is also expected to result in a positive correlation between nucleotide heterozygosity and recombination rate[6].

Are levels of nucleotide variation positively correlated with recombination rate in humans, as predicted by models of selection? Answering this question requires accurate estimates of recombination rate and accurate measures of nucleotide variability for different regions of the genome. Previous studies based on a few loci suggested that recombination rate and nucleotide variability might be correlated[7,8], but recent data from the Human Genome Project now allow us to address this question in greater detail.

Theory meets data

Variation in recombination rate for different genomic regions can be estimated by comparing genetic and physical distances between markers, and is expressed in cM Mb^{-1}. Several large-scale physical maps of the human genome have been assembled and integrated with genetic maps[9,10]. Comparison of genetic and radiation-hybrid-based physical maps has shown that there is substantial variation in recombination rate in different regions of the genome, with high values close to 8 cM Mb^{-1} and low values below 0.20 cM Mb^{-1} (Ref. 11). In general, lower recombination rates are observed near the centromeres of metacentric chromosomes. The ultimate physical map comes from the complete sequence of a genome. The human genome sequence[12,13] provides a new opportunity to evaluate variation in recombination rate. Weber and colleagues[14] have provided the first detailed look at recombination rate variation from a sequence-based physical map, and they found megabase-sized chromosomal regions with particularly low and particularly high recombination rates, ranging from 0–9 cM Mb^{-1}, which agrees well with the data based on radiation hybrid maps[11]. Nonetheless, estimates of recombination rate are necessarily still imprecise owing to the relatively small number of meioses used in the construction of genetic maps.

An accurate measure of nucleotide variability requires a population sample. The variance in estimates of

heterozygosity for a single panmictic population decreases substantially when the sample size is greater than ten individuals[15]. For species in which there is some geographic structure to the genetic variation, as is the case for humans, global samples of much more than ten individuals are likely to be necessary for an accurate estimate of species-wide levels of nucleotide variability. There have been several studies aimed at measuring nucleotide variability in humans in the last four years (summarized in Table 1). All of these datasets are based on complete sequences, and most include primarily introns and other noncoding regions. Thus, most sites surveyed are themselves unlikely to be targets of selection, although they might be linked to sites under selection. Two different measures of nucleotide variability, π and θ, are given in Table 1. Nucleotide heterozygosity (π; Ref. 16) is the average number of nucleotide differences between two sequences drawn at random from a population, and thus depends not only on the number of polymorphic sites but also on their frequency. The proportion of polymorphic sites in a sample, corrected for sample size (θ; Ref. 17), does not depend on the frequency of segregating nucleotides. Two observations are noteworthy from Table 1. First, the average level of heterozygosity in humans is very low ($\pi = 0.10\%$). Only one difference per thousand bases is seen between two alleles drawn at random from humans. Second, levels of nucleotide heterozygosity vary by more than one order of magnitude among loci, with low values of zero and high values nearly 0.2%.

Scatterplots of π and θ versus recombination rate are shown in Fig. 2. In these comparisons, only studies with sample sizes greater than ten are included. In both cases, there is a strong positive correlation between nucleotide variability and recombination rate (π, $R^2 = 0.54$, $P < 0.001$; θ, $R^2 = 0.63$, $P < 0.001$). Thus, it appears that variation in recombination rate explains approximately 60% of the variation in levels of polymorphism among loci. (Inclusion of all samples, including those with small sizes, also reveals a significant positive correlation, although with more scatter; π, $R^2 = 0.24$, $P < 0.05$; θ, $R^2 = 0.22$, $P < 0.05$).

A simple explanation for the positive correlation might be that recombination is mutagenic; i.e. that regions with more recombination have higher levels of variability because they have a higher input of new mutations. This hypothesis can be tested by comparing recombination rate with levels of divergence between humans and chimpanzees for this same set of genes. If genes with higher recombination rates also have higher mutation rates, this should be reflected in higher levels of divergence. However, for the genes in Table 1, there is no correlation between divergence and recombination rate ($R^2 = 0.02$, $P = 0.57$). Another way to ask whether variation in mutation rate might underlie some of the variation in levels of heterozygosity is to ask whether GC content is correlated with levels of polymorphism. There is some evidence that GC content is positively correlated with recombination rate[18], and it is well known that in mammalian genomes CpG dinucleotides are hotspots for mutation[19]. However, there is no correlation between GC content and either π or θ for the genes in Table 1 (π, $R^2 < 0.01$, $P = 0.99$; θ, $R^2 = 0.04$, $P = 0.42$). Thus, the relationship shown in Fig. 2 does not appear to result from underlying differences in mutation rate among loci and therefore is not consistent with a neutral model of molecular evolution. Instead, this result suggests that selection is having a strong effect on patterns of variation at linked sites throughout the genome.

The relationship between levels of genetic variation and recombination rate can also be addressed by looking at the density of SNPs throughout the genome. As part of the Human Genome Project, over one million SNPs have now been identified and mapped in a concerted public effort[20]. Several strategies have been used for SNP detection, including: (1) shotgun sequencing of genomic fragments drawn from a publicly available panel of ethnically diverse individuals, (2) comparison of randomly sequenced clones to finished genome sequence, and (3) sequence comparisons in regions of overlap between large-insert clones. Thus SNP density depends, in part, on the method of ascertainment and the availability of genome sequence, and on the true level of nucleotide diversity. None of the samples used for SNP discovery is a true population sample, and in most cases, the number of chromosomes surveyed is two or three. Thus, many low frequency variants are likely to be missed. To analyze nucleotide diversity in a more homogeneous sample, each chromosome was divided into bins of 200 kb, and π was calculated for a sample of size two[20]. In spite of this small sample, 95% of the bins displayed nucleotide diversities between 0.02% and 0.16%, in general agreement with the studies in Table 1. A recent analysis shows that the density of SNPs is positively correlated with recombination rate ($R^2 = 0.55$, $P = 0.01$)[21], using recombination rate estimates derived from the sequence-based physical map[14]. Thus, the new data from the Human Genome Project corroborate the result in Fig. 2.

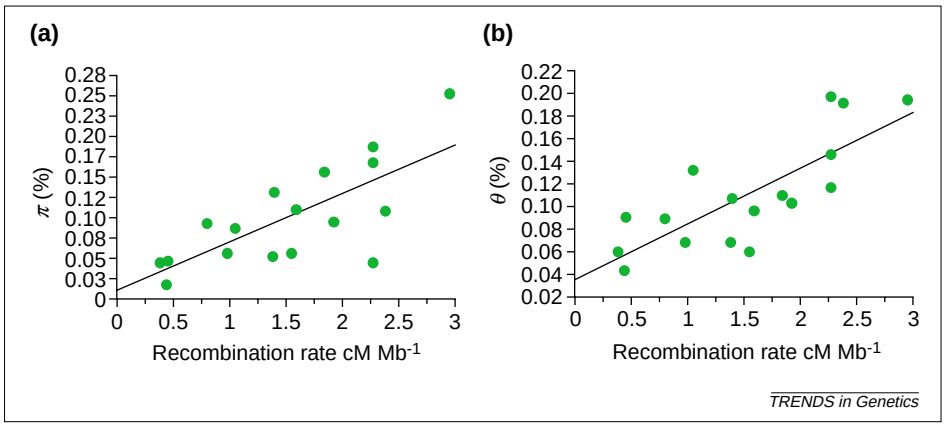

Fig. 2. (a) Scatterplot of nucleotide diversity versus recombination rate for the 17 loci in Table 1 for which sample size is greater than ten. Linear regression $R^2 = 0.54$, $P < 0.001$. (b) Scatterplot of proportion of segregating sites versus recombination rate for the seventeen loci in Table 1 for which sample size is greater than ten. Linear regression $R^2 = 0.63$, $P < 0.001$. Recombination rate estimates are from Ref. 11. Recombination rates for X-linked genes have been multiplied by two-thirds to account for the fact that the X chromosome spends two-thirds of its time in females where it recombines, and one-third of its time in males where no recombination occurs. Both π and θ for X-linked genes have been multiplied by four-thirds, to account for the fact that the effective population size of the X chromosome is three-quarters that of the autosomes.

Positive or negative selection?

Although it appears that selection is influencing patterns of variation at linked neutral sites across much of the genome, we would like to know whether the pattern in Fig. 2 is caused mainly by positive selection driving the fixation of adaptive mutations, by negative selection eliminating deleterious mutations, or by some combination of both processes. Models of background selection and genetic hitchhiking make several predictions that could help us distinguish between these alternatives. First, for loci with very high mutation rates, such as microsatellites, background selection is expected to produce a correlation between heterozygosity and recombination rate, whereas genetic hitchhiking is not expected to produce such a correlation, unless selective sweeps are very frequent[22]. In humans, microsatellite variability is not correlated with recombination rate[11,14], arguing against background selection. Second, under many conditions, positive selection will result in lower levels of variability on the X chromosome relative to the autosomes, whereas negative selection will result in higher levels of variability on the X chromosome relative to the autosomes[23]. Both the data in Table 1 and the density of mapped SNPs[20] show a slightly lower level of variability on the X chromosome than is seen on the autosomes, consistent with genetic hitchhiking. Third, simple models of genetic hitchhiking are expected to produce a skew in the frequency distribution of segregating variants, with a large number of low-frequency polymorphisms[24], whereas background selection is generally not expected to produce this pattern[25]. Tajima's *D* (Ref. 26; Table 1) is a statistic that summarizes the frequency distribution and is negative when there is an excess of low-frequency polymorphisms, and positive when there is an excess of intermediate-frequency polymorphisms. There is a weak, non-significant positive association between Tajima's *D* and recombination rate for the loci in Table 1 with sample sizes greater than ten ($R^2 = 0.17$, $P = 0.10$); this trend is in the direction expected under genetic hitchhiking.

Although there is some evidence suggesting that genetic hitchhiking might be more important than background selection in producing the observed correlation, determining the relative contribution of each process will require theoretical models that combine both processes, in addition to further empirical studies. It is important to bear in mind that these two processes are not mutually exclusive, and it is likely that both contribute somewhat to the observed pattern[27].

If we suppose for a moment that genetic hitchhiking is primarily responsible for the correlation between recombination rate and nucleotide diversity, we can ask how much adaptive evolution is consistent with the observed patterns. Using the model of Wiehe and Stephan[28] and the data in Table 1, it is possible to estimate the parameter $\alpha = 2N_e sv$, where N_e is the effective population size, s is the average selection coefficient for beneficial mutations, and v is the rate of adaptive evolution. In this case, $\alpha = 10^{-7}$. If we assume that $N_e = 10^4$ and $s = 10^{-2}$, then $v = 5 \times 10^{-10}$. This is approximately 2% of the neutral substitution rate, suggesting that roughly one out of every 50 substitutions are fixed by positive selection. If we assume that 5% of the three billion sites in the genome are functional, and if we take the average human–chimpanzee sequence divergence of 1% (Table 1), then humans and chimpanzees differ at approximately 1.5×10^6 functional sites. If one out of every 50 of these differences are driven by positive selection, this suggests that humans and chimpanzees differ by 30 000 adaptive substitutions, or approximately one every 300 years. Of course, this calculation is crude and makes many assumptions. To the extent that background selection is operating, for example, we might view this as an overestimate. Refinements of theory and additional data could eventually help us to estimate this quantity more precisely.

Regardless of the relative impact of positive and negative selection, it now appears likely that variation in recombination rate explains a substantial fraction of the variability in levels of nucleotide polymorphism across the human genome.

Acknowledgements

I thank H.E. Hoekstra, G. Marth, B.A. Payseur, J.L. Weber and two anonymous reviewers for comments on the manuscript, and G. Marth for sharing his unpublished results. This work was supported by the NSF.

References

1 Kimura, M. (1983) *The Neutral Theory of Molecular Evolution*, Cambridge University Press
2 Maynard Smith, J. and Haigh, J. (1974) The hitch-hiking effect of a favourable gene. *Genet. Res.* 23, 23–35
3 Kaplan, N.L. *et al.* (1989) The 'hitchhiking effect' revisited. *Genetics* 123, 887–899
4 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
5 Charlesworth, B. *et al.* (1993) The effect of deleterious mutations on neutral molecular variation. *Genetics* 134, 1289–1303
6 Hudson, R.R. and Kaplan, N.L. (1995) Deleterious background selection with recombination. *Genetics* 141, 1605–1617
7 Nachman, M.W. *et al.* (1998) DNA variability and recombination rates at *X*-linked loci in humans. *Genetics* 150, 1133–1141
8 Przeworski, M. *et al.* (2000) Adjusting the focus on human variation. *Trends Genet.* 16, 296–302
9 Gyapay, G. *et al.* (1996) A radiation hybrid map of the human genome. *Hum. Mol. Genet.* 5, 339–346
10 Olivier, M. *et al.* (2001) A high-resolution radiation hybrid map of the human genome draft sequence. *Science* 291, 1298–1302
11 Payseur, B.A. and Nachman, M.W. (2000) Microsatellite variation and recombination rate in the human genome. *Genetics* 156, 1285–1298
12 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921
13 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351
14 Yu, A. *et al.* (2001) Comparison of human genetic and sequence-based physical maps. *Nature* 409, 951–953
15 Pluzhnikov, A. and Donnelly, P. (1996) Optimal sequencing strategies for surveying molecular genetic diversity. *Genetics* 144, 1247–1262
16 Nei, M. and Li, W.-H. (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5269–5273
17 Watterson, G.A. (1975) On the number of segregating sites in genetical models without recombination. *Theor. Popul. Biol.* 7, 256–276
18 Fullerton, S.M. *et al.* (2001) Local rates of recombination are positively correlated with GC content in the human genome. *Mol. Biol. Evol.* 18, 1139–1142
19 Cooper, D.N. and Krawczak, M. (1993) *Human Gene Mutation*, Bios Scientific Publishers
20 The International SNP Map Working Group (2001) A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409, 928–933
21 Marth, G. *et al.* The structure of single-nucleotide variation in overlapping regions of human genome sequence. *Proc. Natl. Acad. Sci. U. S. A.* (in press).
22 Wiehe, T. (1998) The effect of selective sweeps on the variance of the allele distribution of a linked multiallele locus: hitchhiking of microsatellites. *Theor. Popul. Biol.* 53, 272–283
23 Begun, D.J. and Whitley, P. (2000) Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. *Proc. Natl. Acad. Sci. U. S. A.* 97, 5960–5965

24 Braverman, J.M. *et al.* (1995) The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *Genetics* 140, 783–796
25 Charlesworth, D. *et al.* (1995) The pattern of neutral molecular variation under the background selection model. *Genetics* 141, 1619–1632
26 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595
27 Kim, Y. and Stephan, W. (2000) Joint effects of genetic hitchhiking and background selection on neutral variation. *Genetics* 155, 1415–1427
28 Wiehe, T.H.E. and Stephan, W. (1993) Analysis of a genetic hitchhiking model, and its application to DNA polymorphism data from *Drosophila melanogaster*. *Mol. Biol. Evol.* 10, 842–854
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 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
31 Dienard, A. and Kidd, K. (1999) Evolution of a HOXB6 intergenic region within the great apes and humans. *J. Hum. Evol.* 36, 687–703
32 Rana, B.K. *et al.* (1999) High polymorphism at the human melanocortin 1 receptor locus. *Genetics* 151, 1547–1557
33 Rieder, M.J. *et al.* (1999) Sequence variation in the human angiotensin converting enzyme. *Nat. Genet.* 22, 59–62
34 Fullerton, S.M. *et al.* (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am. J. Hum. Genet.* 67, 881–900
35 Gilad, Y. *et al.* (2000) Dichotomy of single-nucleotide polymorphism haplotypes in olfactory receptor genes and pseudogenes. *Nat. Genet.* 26, 221–224
36 Hamblin, M.T. and Di Rienzo, A. (2000) Detection of the signature of natural selection in humans: evidence from the Duffy blood group locus. *Am. J. Hum. Genet.* 66, 1669–1679
37 Zhao, Z. *et al.* (2000) Worldwide DNA sequence variation in a 10 kb noncoding region on human chromosome 22. *Proc. Natl. Acad. Sci. U. S. A.* 97, 11354–11358
38 Yu, N. *et al.* (2001) Global patterns of human DNA sequence variation in a 10 kb region on chromosome 1. *Mol. Biol. Evol.* 18, 214–222
39 Harris, E.E. and Hey, J. (1999) X chromosome evidence for ancient human histories. *Proc. Nat. Acad. Sci. U. S. A.* 96, 3320–3324
40 Kaessmann, H. *et al.* (1999) DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.* 22, 78–81
41 Jaruzelska, J. *et al.* (1999) Spatial and temporal distribution of the neutral polymorphisms in the last Zfx intron: analysis of haplotype structure and genealogy. *Genetics* 152, 1091–1101
42 Nachman, M.W. and Crowell, S.L. (2000) Contrasting evolutionary histories of two introns of the Duchenne muscular dystrophy locus, *Dmd*, in humans. *Genetics* 155, 1855–1864
43 Alonso, S. and Armour, J.A.L. (2001) A highly variable segment of human subterminal 16p reveals a history of population growth for modern humans outside Africa. *Proc. Natl. Acad. Sci. U. S. A.* 98, 864–869

Michael W. Nachman
Dept of Ecology and Evolutionary Biology, University of Arizona, Tuscon, AZ 85721, USA.
e-mail: nachmann@email.arizona.edu

Modularity in the gain and loss of genes: applications for function prediction

Thijs Ettema, John van der Oost and Martijn Huynen

Genes that are clustered on multiple genomes and are likely to functionally interact tend to be gained or lost together during genome evolution. Here, we demonstrate that exceptions to this pattern indicate relatively distant functional interactions between the encoded proteins. Hence, this can be used to divide predicted clusters of functionally interacting proteins into sub-clusters, and as such, to refine the prediction of their function and functional interactions.

The increasing availability of sequenced genomes allows for the analysis of differences between genomes in terms of the gain and loss of genes. This enables the study of modularity in genome evolution: do genes that are functionally linked, for example because they encode enzymes from the same pathway, tend to be gained and lost simultaneously? The modularity of genome evolution bears relevance for the prediction of gene function, as the co-occurrence of genes in genomes has been proposed[1] and demonstrated[2] to indicate functional relations between their protein products. Recently, modularity has been observed in the loss of genes in one species: *Saccharomyces cerevisiae*[3]. Here, we present the first systematic analysis of the gain and loss of genes and their modularity within the first genus of which three genomes are available, *Pyrococcus furiosus*, *Pyrococcus abyssi* and *Pyrococcus horikoshii*.

Functional patterns in the gain and loss of genes

In the three pyrococcal genomes, a total of 1071 genes are present in at least one *Pyrococcus* species that have no orthologs in the other sequenced Archaea. Thus, this subset of genes has most likely been gained in the evolution of *Pyrococcus*. Conversely, 325 genes are present in both the Euryarchaea and the Crenarchaea, but absent in at least one *Pyrococcus* species, and so have probably been lost. A functional classification of the gained and lost genes along the cluster of orthologous genes (COG) scheme[4] revealed the dominance of a few classes. Genes involved in 'amino acid transport and metabolism' and 'energy production and conversion' have both been gained (12% and 15%, respectively) and lost more frequently (22% for both) than other functional classes. Interestingly, the gained genes that are involved in 'amino acid transport and metabolism' and those of a third functional class of genes that showed significant gain, that of 'carbohydrate transport and metabolism' (13%), are dominated by heterotrophic functions, indicating the evolution of *Pyrococcus* towards a heterotrophic lifestyle. The corresponding proteins are involved in the import and catabolism of amino acids and carbohydrates, including a putative galactoside utilization cluster (Fig. 1b), α-amylases, cellulases and several other hydrolases capable of degrading α- and β-linked carbohydrate substrates (http://www.dove.embl-heidelberg.de/Pyrococcus).

Functional coupling within the gain and loss of genes

The fact that the genes from a few functional classes are most frequently gained or lost suggests that their functions are coupled. To obtain an increased level of resolution of the functional coupling between genes that are gained or lost, we examined how often

 0168–9525/01/$ – see front matter PII: S0168-9525(01)02384-8