Supplementary data

Supplementary data associated with this article can be found at doi:10.1016/j.tig.2004.11.007

References

1 Enright, A.J. *et al*. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature* 402, 86–90

2 Enright, A.J. and Ouzounis, C.A. (2001) Functional associations of proteins in entire genomes by means of exhaustive detection of fusion events. *Genome Biol.* 2, RESEARCH0034, doi:10.1186/gb-2001-2-9-research0034 (http://genomebiology.com/2001/2/9/research/0034)

3 Marcotte, E.M. *et al*. (1999) Detecting protein function and protein–protein interactions from genome sequences. *Science* 285, 751–753

4 Yanai, I. *et al*. (2001) Genes linked by fusion events are generally of the same function category: a systematic analysis of 30 microbial genomes. *Proc. Natl. Acad. Sci. U. S. A.* 98, 7940–7945

5 Snel, B. *et al*. (2000) Genome evolution: gene fusions versus gene fission. *Trends Genet.* 16, 9–11

6 Snel, B. *et al*. (2002) Genomes in flux: the evolution of archaeal and proteobacterial gene content. *Genome Res.* 12, 17–25

7 Mirkin, B.G. et al. (2003) Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol. Biol.* 3, 2, doi:10.1186/gb-2001-2-9-research0034 (http://www.biomedcentral.com/1471-2148/3/2)

8 Murzin, A.G. *et al*. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* 247, 536–540

9 Gough, J. and Chothia, C. (2002) SUPERFAMILY:HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.* 30, 268–272

10 Apic, G. *et al*. (2001) Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J. Mol. Biol.* 310, 311–325

11 Bashton, M. and Chothia, C. (2002) The geometry of domain combination in proteins. *J. Mol. Biol.* 315, 927–939

12 Wolf, Y.I. *et al*. (2002) Genome trees and the tree of life. *Trends Genet.* 18, 472–479

13 Koonin, E.V. *et al*. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu. Rev. Microbiol.* 55, 709–742

14 Tatusov, R.L. *et al*. (1997) A genomic perspective on protein families. *Science* 278, 631–637

15 Remm, M. *et al*. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *J. Mol. Biol.* 314, 1041–1052

16 Pearson, W.R. and Lipman, D.J. (1988) Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. U. S. A.* 85, 2444–2448

17 Tatusov, R.L. *et al*. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res.* 29, 22–28

18 Andersson, S.G. *et al*. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396, 133–140

19 Cole, S.T. *et al*. (2001) Massive gene decay in the leprosy bacillus. *Nature* 409, 1007–1011

20 Jordan I.K. et al. (2003). Phylogenomic analysis of the *Giardia intestinalis* transcarboxylase reveals multiple instances of domain fusion and fission in the evolution of the biotin-dependent enzymes. *J. Mol. Microbial. Biotech.* 5, 172–189

0168-9525/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.11.007

A limited role for balancing selection

Saurabh Asthana, Steffen Schmidt and Shamil Sunyaev

Genetics Division, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Harvard Medical School New Research Building, 77 Ave Louis Pasteur, Boston, MA 02115, USA

Balancing selection has been shown to act on several genes in short-term evolutionary contexts, but it is not known whether this force is responsible for maintaining a significant number of long-term polymorphisms. We aligned 7628 chimpanzee virtual transcripts and 5524 chimp ESTs to the 4× chimp draft genome assembly and identified polymorphisms in chimpanzee that also occurred in the human single nucleotide polymorphism database (dbSNP). Our analysis suggests that the incidence of ancestral polymorphism is low or absent and that balancing selection on the time-scale of chimpanzee–human divergence has not been a significant force in human evolution.

The debate over the role of balancing selection in maintaining genetic polymorphism has a long history. Balancing selection was in the past a frequently postulated factor used to explain high levels of genomic variation, notably in legendary debates between Dobzhansky and Muller [1,2]. The development of the neutral theory of molecular evolution provided a competing explanation for the high frequency of genetic polymorphism, but it remains unclear how common balancing selection really is. This form of selection has been demonstrated in several contexts: a number of common human genetic diseases are believed to be maintained in the population as a result of balancing selection, for example, sickle-cell anemia [3], glucose-6-phosphate dehydrogenase deficiency [4], thalassemia [5] and cystic fibrosis [6]. In addition, there is evidence that the extremely high rate of polymorphism in mammalian MHC proteins is due to balancing selection [7].

Some patterns of balancing selection develop in unique, short-term contexts. In human evolution, a number of alleles that are pathogenic in the homozygous state can confer significant selective advantages in the heterozygous state (overdominance) (Box 1). For example, several mutations in the cystic fibrosis transmembrane conductance regulator, ATP-binding cassette gene (*CFTR*) confer cystic fibrosis. By far the most common variant is

Corresponding author: Sunyaev, S. (ssunyaev@rics.bwh.harvard.edu).
Available online 11 November 2004

Box 1. Definitions

Balancing selection

Genetic variation typically has a finite lifetime. Even under neutral evolution, in the absence of natural selection, polymorphisms will eventually vanish, as allele frequencies slowly fluctuate because of genetic drift until one allele becomes fixed. On average this will take $4N_e$ generations, where N_e is the effective population size of the organism. Both positive and negative selection will tend to shorten this average lifespan; if an allele is positively selected for, it will increase in frequency more quickly than it would because of genetic drift, whereas an allele that is selected against will decrease in frequency more quickly.

However, some forms of selection will protect genetic variation and increase the average lifespan of a polymorphism, possibly indefinitely. These are collectively called 'balancing selection' and can result from several causes:

Overdominance or heterozygote advantage

The heterozygous form has a selective advantage over either homozygous form. Often this can maintain a deleterious phenotype in the population.

Frequency dependence

An advantage is conferred by a rare feature; for example, mate preference for unique appearance.

Variable environments

If an organism occupies multiple environments, polymorphisms might be maintained in the population if the possession of either allele is advantageous in a different environment.

the *CFTR* ΔF508 allele – although the effects are severe in homozygotes for this mutation, in the heterozygote form it might protect against childhood asthma [6]. Such patterns are unlikely to be stable, because they are highly prone to being replaced with alternatives that are not associated with a strongly deleterious phenotype.

However, balancing selection might have preserved some polymorphisms for a considerably longer period, if there is no negative pressure against any variant but some advantage is conferred by heterozygosity then such ancestral polymorphisms might appear in several closely related species. For example, New World primates are polymorphic in an X-linked opsin allele, which confers color vision; males and female homozygotes have two-color vision, but female heterozygotes have three-color vision, which might confer a selective advantage through increased ability to identify ripe fruit or young leaves [8]. In other instances, polymorphism might be maintained by frequency-dependent selection (e.g. rare alleles that confer a selective advantage). Other examples of genes with ancestral polymorphisms that are believed to be maintained through balancing selection include the ABO blood group genes [9] and the major histocompatibility complex (MHC) class I and II antigen genes, for which trans-species polymorphisms have been identified [10,11].

Identifying trans-species polymorphism

The incidence of trans-species polymorphisms between two species can serve as an indicator of the importance of balancing selection in maintaining polymorphism. The availability of polymorphism information for the chimpanzee genome enables us to make such a comparison between chimps and humans. We have made use of three chimpanzee datasets – 7628 virtual transcripts [12], GenBank chimpanzee ESTs obtained from several large-scale studies [13,14] and the Arachne 4× draft chimp assembly. Because balancing selection should occur at functional positions, our analysis only considered coding sequence; although there is increasing evidence that non-coding sequences might contain many positions of high functional importance, the proportion of functionally significant single nucleotide polymorphisms (SNPs) is much higher in the coding sequence.

We aligned the virtual transcripts and ESTs against the Arachne draft chimp assembly using BLAT [15] and identified SNPs. Because the draft assembly is incomplete in its coverage, we restricted our alignments to those where both chimp sequences aligned with the same locus in the University of California, Santa Cruz (UCSC) human genome assembly (http://genome.ucsc.edu/). The average SNP density was one SNP every 968 bases.

To determine whether any of these polymorphisms were trans-species polymorphisms, we searched within the database of human SNPs (dbSNP, revision 118), and identified those SNPs that occurred at the same position in human and chimpanzee (the position was based on the UCSC Dec 2003 human–chimp genome alignment). We identified 11 SNPs that occurred in the same position in chimp and human for the draft assembly and virtual transcript alignment. Seven of these were the same polymorphism by sequence (e.g. an A to C polymorphism occurred in both genomes); of the six SNPs that occurred in the draft assembly–EST alignment, only one was the same polymorphism by sequence.

Of these eight trans-species polymorphisms, one occurred in untranslated sequence, one was ambiguous for synonymy and three occurred at synonymous coding sites. Only three occurred at non-synonymous sites, and these could a result of balancing selection. However, the surrounding regions of all three sites (one in the myosin IIIA gene (*MYO3A*) [16], one in the *SMARCAD1* gene [17] and the final one in a hypothetical gene of unknown function) lack the high polymorphism density that would certainly be expected of an ancient polymorphism [18]. Four out of eight of these polymorphisms occurred at highly mutable CpG sites.

As a control we examined trans-species polymorphism in a set of MHC Class I genes. We obtained the sequence for 30 alleles from three different chimpanzee MHC Class I genes (*Patr-A*, *Patr-B* and *Patr-C*) from the GenBank database. We aligned these sequences against the chimp draft genome assembly using BLAT and looked for trans-species polymorphisms in human. We identified 12 trans-species polymorphisms that occurred at non-synonymous positions, 11 of which were the same sequence polymorphism.

An absence of ancestral polymorphism

There are three possible sources of trans-species polymorphisms: (i) neutral ancestral polymorphisms that have survived due to random chance; (ii) ancestral polymorphisms that have been maintained as a result of balancing selection; and (iii) coincidental mutations that occurred subsequent to speciation at the same locus.

Given the evolutionary distance between humans and chimpanzees, it is unlikely that a neutral ancestral polymorphism could be maintained by chance in the absence of balancing selection. On average, a neutral polymorphism would persist for 4 N_e generations, where N_e is the effective population size. In humans, this is estimated to be 10 000 [19]. Estimates of the effective population size of chimpanzee have not reached a consensus; as a conservative figure we can use the effective population size of the common ancestor of all chimp variants, estimated at 50 000 [20]. We calculated the probability of detecting a neutral shared polymorphism in six human lineages (the fold coverage for humans in dbSNP) and two individual chimpanzee lineages according to coalescence [12]. According to these estimates of effective population size, a shared polymorphism would survive for 4.6 million years (a lower limit for estimated divergence time between chimps and humans [21]) with a probability of 2×10^{-6}. For 6.7×10^3 sites (the number of SNPs detected in our virtual transcript and draft assembly alignment), we would not expect to see a neutral ancestral polymorphism surviving with any significant probability.

It should be noted that this analysis excludes polymorphisms that might have been balancing for some portion of their history but subsequently became neutral and died out; similarly, long-lived balancing polymorphisms that emerged after speciation might also exist. Such polymorphisms might have been balancing for several million years, and thus might be considered ‘long-term’. Our analysis will not detect such polymorphisms. We should therefore be circumspect in speaking of ‘long-term’ balancing selection, by which we must mean polymorphisms with a lifetime of greater than ~4.6 million years.

A significant proportion of trans-species polymorphisms will be the result of coincidental mutation. We counted 10^4 human database SNPs in the 6.3×10^6 human coding bases that corresponded to our chimpanzee virtual transcript dataset. If we presume a uniform rate of mutation across all coding positions, we would expect to find approximately ten polymorphisms co-occurring by site with our 6.7×10^3 chimpanzee polymorphisms. However, there are significant context dependencies resulting in much higher mutation rates for CpG dinucleotides. Forty two percent of human coding SNPs occur at CpG sites, 67% of these are the result of fast CpG-destroying mutations. If we assume: (i) that the same proportion of the 1.7×10^3 chimpanzee SNPs that occur at CpG sites are the result of CpG-destroying mutations; and (ii) that 7% of coding sites occur in CpG dinucleotides, then we would expect to find ~12 human polymorphisms co-occurring by random chance with our 6.7×10^3 chimpanzee SNPs compared with the 11 SNPs that we actually identified. This number is a conservative estimate, because disparity between SNP density at synonymous and non-synonymous positions, heterogeneous mutation rates and selective effects will tend to increase the likelihood of co-occurrence. A final source of apparent coincidental mutations might be errors in sequencing that produce false-positive SNPs. However, because we see few trans-species polymorphisms, even if our dataset is rife with sequencing errors, it can not have significantly altered our result.

Concluding remarks

The total incidence of trans-species polymorphism in our analysis is low, and can be parsimoniously attributed to coincidental mutation rather than to surviving ancestral polymorphism. It is plausible that individual cases of balancing selection might be found in the future, because we screened only a fraction of the polymorphisms. However, our analysis suggests that statistically the effect of balancing selection is limited.

Acknowledgements

The authors are grateful for the financial support of the Genome Canada Foundation and the William F. Milton Fund.

References

1 Crow, J.F. (1987) Muller, dobzhansky and overdominance. *J. Hist. Biol.* 20, 351–380
2 Dobzhansky, T. (1955) A review of some fundamental concepts and problems of population genetics. *Cold Spring Harb. Symp. Quant. Biol.* 20, 1–15
3 Aidoo, M. *et al*. (2002) Protective effects of the sickle cell gene against malaria morbidity and mortality. *Lancet* 359, 1311–1312
4 Verrelli, B.C. *et al*. (2002) Evidence for balancing selection from nucleotide sequence analyses of human g6pd. *Am. J. Hum. Genet.* 71, 1112–1128
5 Weatherall, D.J. (1997) Thalassemia and malaria, revisited. *Ann. Trop. Med. Parasitol.* 91, 885–890
6 Schroeder, S.A. *et al*. (1995) Protection against bronchial asthma by cftr delta f508 mutation: a heterozygote advantage in cystic fibrosis. *Nat. Med.* 1, 703–705
7 Hedrick, P.W. and Thomson, G. (1983) Evidence for balancing selection at HLA. *Genetics* 104, 449–456
8 Surridge, A.K. and Mundy, N.I. (2002) Trans-specific evolution of opsin alleles and the maintenance of trichromatic colour vision in callitrichine primates. *Mol. Ecol.* 11, 2157–2169
9 Saitou, N. and Yamamoto, F. (1997) Evolution of primate abo blood group genes and their homologous genes. *Mol. Biol. Evol.* 14, 399–411
10 Hughes, A.L. and Nei, M. (1998) Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. *Nature* 335, 167–170
11 Hughes, A.L. and Nei, M. (1989) Nucleotide substitution at major histocompatibility complex class II loci: evidence for overdominant selection. *Proc. Natl. Acad. Sci. U. S. A.* 86, 958–962
12 Clark, A.G. *et al*. (2003) Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302, 1960–1963
13 Sakate, R. *et al*. (2003) Analysis of 5′-end sequences of chimpanzee cDNAs. *Genome Res.* 13, 1022–1026
14 Hellmann, I. *et al*. (2003) Selection on human genes as revealed by comparisons to chimpanzee cdna. *Genome Res.* 13, 831–837
15 Kent, W.J. (2002) Blat – the blast-like alignment tool. *Genome Res.* 4, 656–664
16 Walsh, T. *et al*. (2002) From flies’ eyes to our ears: mutations in a human class III myosin cause progressive nonsyndromic hearing loss dfnb30. *Proc. Natl. Acad. Sci. U. S. A.* 99, 7518–7523
17 Adra, C.N. *et al*. (2000) Smarcad1, a novel human helicase family-defining member associated with genetic instability: cloning, expression, and mapping to 4q22–q23, a band rich in breakpoints and deletion mutants involved in several human diseases. *Genomics* 69, 162–173
18 Richman, A. (2000) Evolution of balanced genetic polymorphism. *Mol. Ecol.* 9, 1953–1963
19 Takahata, N. (1993) Allelic genealogy and human evolution. *Mol. Biol. Evol.* 10, 2–22
20 Fischer, A. *et al.* Evidence for a complex demographic history of chimpanzees. *Mol. Biol. Evol.* (in press)
21 Chen, F.C. and Li, W.H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68, 444–456

doi:10.1016/j.tig.2004.11.001