Gene 443 (2009) 70–75

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

Patterns of autosomal divergence between the human and chimpanzee genomes support an allopatric model of speciation

Matthew T. Webster

Department of Medical Biochemistry and Microbiology, Uppsala University, Box 582, 75123 Uppsala, Sweden

ARTICLE INFO

Article history:
Received 26 November 2008
Received in revised form 13 May 2009
Accepted 13 May 2009
Available online 20 May 2009

Received by C. Feschotte

Keywords:
Speciation
Divergence
Human
Chimpanzee
Incomplete lineage sorting
Nucleotide substitution

ABSTRACT

There is a large variation in divergence times across genomic regions between human and chimpanzee. It has been suggested that this could partly result from selection against ancestral gene flow between incipient species in regions of the genome containing genetic incompatibilities. It is possible that such barriers to gene flow could arise in specific genes or in chromosomal inversions. I analysed patterns of lineage sorting that occur between human, chimpanzee and gorilla genomic sequences by examining divergent site patterns in >18 Mb genomic alignments. I develop a method to normalise site patterns by the mutational spectrum to minimise errors caused by misinference caused by recurrent mutation. Here I show that divergence times appear to be uniform between coding and noncoding sequences and between inverted and non-rearranged portions of chromosomes. I therefore find no evidence to support the large-scale accumulation of genetic incompatibilities at speciation genes or chromosomal inversions in the ancestral population of humans and chimpanzees. In addition, site patterns that are discordant with the species tree occur more frequently in regions with high human recombination rates. This could indicate the action of selective sweeps in the ancestral population, but could also be indicative of increased rates of homoplasy in these regions. I argue that these observations are compatible with a neutral allopatric model of speciation.

1. Introduction

Under the allopatric model of speciation, incipient species are completely separated by geographic barriers leading to reproduction isolation (Mayr, 1942). This has served as a useful null model with which to compare empirical observations. However, in nature gene flow often occurs between nascent species. Under these conditions, genetic incompatibilities could arise at some loci before others (Dobzhansky, 1937). Selection could then prevent such loci from crossing the species barrier, resulting in large divergence times for these loci, whereas gene flow proceeds freely at other parts of the genome (Machado et al., 2002; Wu and Ting, 2004).

The time of divergence between two species at a particular locus is equal to the time of speciation (t_S), which is constant across loci, plus the time to coalescence at that locus in the ancestral population (t_C). Under a standard neutral model with Wright–Fisher demography, t_C should be exponentially distributed with a mean of 2*Ne* generations (Kingman, 1982). The exponential distribution has a long tail, so a small proportion of loci will exhibit an extremely long time to coalescence. Under this neutral model, 1% of loci are expected to coalesce earlier than 9.2*Ne* generations ago (Barton, 2006). The distribution of t_C can also be influenced by natural selection in the ancestral population, which could cause certain loci to have abnormally high or low divergence times. Regardless of the presence of selection, however, it may often be the case that regions of the genome with deep genealogies coalesce earlier than previous bifurcations of the species tree. In this case a phenomenon called incomplete lineage sorting can occur, where gene trees and species trees are not concordant.

Due to the recent availability of large-scale sequencing from closely related species, it is possible to analyse lineage sorting during speciation using genome comparisons. Pollard et al. (2006) used closely related *Drosophila* species to reveal mosaic relationships between genomes, indicating that speciation occurs before lineages from previous speciation events have fully separated. Incomplete lineage sorting is also important in the divergence of humans with our closest ancestors (Takahata et al., 1995; Takahata and Satta, 1997; Satta et al., 2000; Chen and Li, 2001; Patterson et al., 2006). Fig. 1 illustrates the situation for two hypothetical loci in human and chimpanzee. The majority of loci are represented by Fig. 1A, where human and chimpanzee are most closely related. The occurrence of HC sites (where human and chimpanzee share a difference from other species) is expected to be greater in these regions. However, in the rest of the genome, human and gorilla (or chimpanzee and gorilla) are more closely related than human and chimpanzee. These regions are characterised by clusters of HG and CG sites (where human and gorilla or chimpanzee and gorilla are identical). The presence such sites may therefore indicate regions of the genome with deep

Abbreviations: t_C, time to coalescence; t_S, time to speciation; *Ne*, effective population size; WS, weak-to-strong; SW, strong-to-weak.

E-mail address: matthew.webster@imbim.uu.se.

0378-1119/$ – see front matter © 2009 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2009.05.006

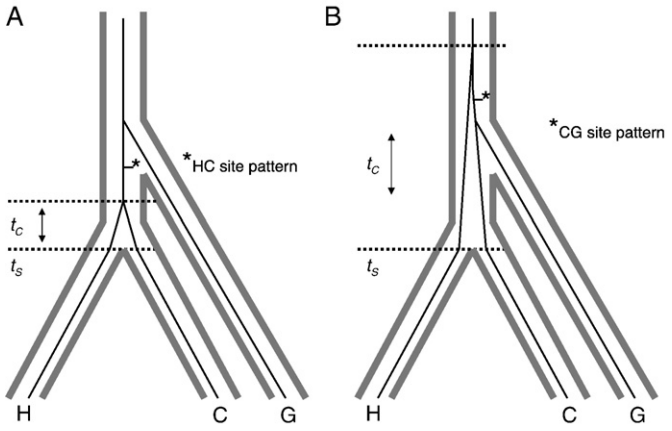

Fig. 1. The divergence time between human and chimpanzee varies across the genome and is equal to the time of speciation (t_S) plus the time to coalescence in the ancestral population (t_C). The majority of the genome supports genealogy A, which is indicated by the presence of HC sites. However, a significant proportion of the genome has a deeper genealogy B, which is indicated by the presence of HG or CG sites. Mutations on the genealogy that would lead to these different site patterns in alignments are indicated by asterisks.

divergence. Recent studies of several megabases of genomic alignments suggested that this portion comprises 18–29% of the genome (Patterson et al., 2006; Ebersberger et al., 2007; Hobolth et al., 2007). However, many incidences of HG and CG site patterns could also be explained by recurrent mutations in genomic regions where the gene tree and species tree are identical.

Two main methods have been used to estimate ancestral population parameters using orthologous sequences from humans and other primates, which both analyse variation in t_C. One approach is to estimate the proportion of loci where the inferred gene tree does not match the species tree. This can be used to estimate *Ne* in the human–chimpanzee ancestral population, assuming an allopatric model of speciation (Ruvolo, 1997; Chen and Li, 2001). A second approach is to analyse variation in divergence time between human and chimpanzee, and to use this data to jointly estimate t_S and the ancestral *Ne* using maximum-likelihood (Takahata et al., 1995; Takahata and Satta, 1997; Yang, 1997). Studies using both of these methods indicate that *Ne* in the human-chimpanzee ancestor was ~5–10 times greater than current *Ne*, due to a large variation in estimates of t_C across the genome.

One reason for the large estimates of ancestral *Ne* could be that variation in t_C between humans and chimpanzees is too large to be accommodated by an allopatric model of speciation with no selection. If selective sweeps were common in the ancestral population of human and chimpanzees, then this would result in more loci with low t_C than predicted. Conversely, if the ancestral population was subdivided, then more loci with high t_C would be observed. Genomic incompatibilities within the ancestral population would also tend to generate regions with high t_C. On the basis of a large variation in t_C across the genome, it has been argued that speciation was characterised by a split and subsequent rehybridisation between lineages (Patterson et al., 2006). However, other authors have argued that this variation can be explained by an allopatric model of speciation with no selection if *Ne* in the ancestral population was large (Barton, 2006; Innan and Watanabe, 2006; Wakeley, 2008). Patterson et al. (2006) also observe a significant reduction in divergence in the X chromosome between human and chimpanzee relative to other chromosomes. This pattern is difficult to reconcile with an allopatric model and suggests the action of a selective sweep in the ancestral human–chimpanzee population. A possible cause of this sweep could be the removal of genetic incompatibility loci on the X chromosome upon rehybridisation.

Under strictly allopatric speciation, the distribution of divergence times should be equal across genomic regions. Osada and Wu (2005) analysed alignments of 76 protein coding and 53 intergenic regions from human, chimpanzee, gorilla and orangutan, and found evidence that coding regions had significantly greater divergence times than noncoding regions. A greater proportion of the protein coding sequences supported deep genealogies, indicated by the greater occurrence of HG and CG sites in these alignments (see Fig. 1). These results support the hypothesis that 'speciation genes' were involved in driving reproductive isolation between humans and chimpanzees. In contrast, Ebersberger et al. (2007) used ~23 Mb of genomic alignments and did not find evidence for discordance in divergence times between coding and noncoding regions.

It has also long been believed that chromosomal inversions could be the basis for reproductive isolation between incipient species. Recent theories stress the importance that inversions have in reducing recombination in heterozygotes in leading to their fixation in diverging populations (Noor et al., 2001; Rieseberg, 2001; Navarro and Barton, 2003). Fixation of inversions can then create genetic barriers between species (Barton and Bengtsson, 1986). For example, there are barriers to gene flow between *Drosphila pseudoobscura* and closely related species at loci associated with inversions (Wang et al., 1997; Machado et al., 2002). One study reported faster rates of protein evolution at genes on chromosomes that were rearranged between human and chimpanzee, indicating that population-specific selective sweeps may have driven reproductive isolation in these regions (Navarro and Barton, 2003; Marques-Bonet et al., 2007). However, this is not supported by other large-scale analyses (Zhang et al., 2004; Chimpanzee Sequencing and Analysis Consortium, 2005). If chromosomal inversions were involved in generating reproductive isolation in the ancestral human/chimpanzee population then we would expect them to have a higher average divergence times, although this hypothesis has not been tested for noncoding regions, which are assumed to evolve neutrally.

It is important to distinguish between sites caused by single mutations on alternative genealogies and sites caused by recurrent mutations (O'HUigin et al., 2002). Using maximum-likelihood, Patterson et al. (2006) estimate the proportion of HG and CG sites caused by recurrent mutations in HCGM shotgun alignments to be >55%. For single base H, C or G sites, these figures are 1.2–3.1%, with HC sites intermediate at 6.7–19%. Different mutational and evolutionary forces affect genes and noncoding regions and this could cause biases in the distribution of HG and CG sites due to recurrent mutations rather than differing genealogies. The most common form of mutation is CpG mutations, which are likely to occur at different frequencies in genes and intergenic regions, because these regions differ in the frequency of occurrence of CpG sites. Furthermore, it is known that the GC content of the genome is not at equilibrium, with a general trend towards the decay of GC-rich regions (Duret et al., 2002), and that the relative frequency of substitutions from G:C to A:T (strong-to-weak, or SW) and A:T to G:C (weak-to-strong, or WS) is believed to vary across the genome (Meunier and Duret, 2004). However, spatially clustered sites that support alternative genealogies (HG or CG clusters) are more likely to result from single mutations, whereas single occurrences are more likely to be due to recurrent mutation on the standard genealogy. It has also been suggested that recombination is mutagenic (Lercher and Hurst, 2002; Hellmann et al., 2003), which could increase the rate of homoplasy. However, how rates of homoplasy vary between loci is not well understood.

The main aim of this paper is to test the null hypothesis of equal divergence times across genomic regions using megabase-scale sequences data now available from multiple primate species. In particular, I examine whether divergence times are uniform between coding and noncoding regions and between regions involved in chromosomal inversions and non-rearranged regions. In contrast to previous studies, I implemented a normalisation procedure based on