ARTICLES

Genetic evidence for complex speciation of humans and chimpanzees

Nick Patterson[1], Daniel J. Richter[1], Sante Gnerre[1], Eric S. Lander[1,2] & David Reich[1,3]

The genetic divergence time between two species varies substantially across the genome, conveying important information about the timing and process of speciation. Here we develop a framework for studying this variation and apply it to about 20 million base pairs of aligned sequence from humans, chimpanzees, gorillas and more distantly related primates. Human–chimpanzee genetic divergence varies from less than 84% to more than 147% of the average, a range of more than 4 million years. Our analysis also shows that human–chimpanzee speciation occurred less than 6.3 million years ago and probably more recently, conflicting with some interpretations of ancient fossils. Most strikingly, chromosome X shows an extremely young genetic divergence time, close to the genome minimum along nearly its entire length. These unexpected features would be explained if the human and chimpanzee lineages initially diverged, then later exchanged genes before separating permanently.

The genetic divergence between two species (the proportion of nucleotides differing between representative individuals of the two species) can be converted into a divergence time in terms of millions of years, provided that differences between genomes have accumulated at a constant rate as a result of new mutations[1,2]. The average genetic divergence, τ_{genome}, is sometimes used to estimate the speciation time, τ_{species}. However, $\tau(x)$, the genetic divergence at any position x, fluctuates across the genome and is everywhere larger[3] than τ_{species} (Fig. 1a, and Supplementary Note 1). Thus, its average τ_{genome} necessarily exceeds τ_{species}.

Inferring ancient speciation from genetic data

With the availability of large-scale sequencing, enough data can now be obtained to study not only the average τ_{genome} but the distribution $\tau(x)$. This should allow direct inferences about τ_{species}, which must be less than the minimum time divergence, and about variability in $\tau(x)$, which conveys information about the speciation process. Several issues must be considered in studying $\tau(x)$ for any pair of modern species. First, the genetic divergence should be corrected for local variation in the neutral mutation rates across the genome. This can be done by dividing the local divergence between two species by the divergence of one from an outgroup, for example macaque for the human–chimpanzee comparison. Second, any estimate, $\hat{\tau}(x)$, of local genetic divergence should be corrected for the effects of recurrent mutation; we do this using two independent methods. Third, variability of $\hat{\tau}(x)$ should be assessed by studying large enough subsets of the genome for the resulting estimates to be reliable.

Although chimpanzees are our closest relatives, there are many loci at which humans and gorillas (or chimpanzees and gorillas) are the most closely related[4,5]; we estimate that this is the case over about 18–29% of the genome (Supplementary Note 2). In such places, the genetic divergence time of human and chimpanzee must precede gorilla speciation (Fig. 1b–d). Thus, humans and chimpanzees show large variation in $\tau(x)$, making them an excellent system in which to explore how taking into account the difference between τ_{genome} and τ_{species} can affect inferences about history.

Previous genetic analyses of great apes studied small data sets (the largest was about 25 kilobases (kb))[5] and suggested that the time since average genetic divergence of humans and chimpanzee genes is much greater than the time of speciation[4,6–8]. However, they produced inconsistent estimates of ancient diversity owing to small data sizes, ignoring the effects of recurrent mutation[9], and simplifying assumptions about the demography of ancient populations[10].

Genome comparisons of five primates

To create a much larger data set, we generated shotgun sequence from the gorilla (about 115,000 fragments comprising about 87 megabases (Mb)), and compared it against the human[11] and chimpanzee genomes[12] and the unpublished sequence of orangutan and macaque (Methods). We overlapped the sequences, producing four-species human–chimpanzee–gorilla–macaque (HCGM) alignment (18.3 Mb) and five-species human–chimpanzee–gorilla–orangutan–macaque (HCGOM) alignment (9.3 Mb) (Supplementary Tables 1 and 2). We also studied 1.2 Mb (ref. 13) from contiguous regions of chromosomes 7 and X (Methods). Altogether, these data represent more than 800-fold more aligned bases than the largest data set previously available[5], and enough data to compare chromosome X with the autosomes.

To analyse these data we identified all 'divergent sites', places at which two alternative alleles were observed across the aligned sequences of the species. We eliminated sites in hypermutable CpG dinucleotides[11], and those not flanked by at least one base of completely conserved sequence (our qualitative results are unaffected by these filters; Supplementary Table 3 and Supplementary Fig. 1). This produced 858,941 divergent sites for the HCGM shotgun data, 498,771 for HCGOM shotgun data, and 78,290 for the contiguous data.

We categorized the divergent sites according to how they partitioned the species (Table 1). In the four-species HCGM alignment, there are seven possible partitions: four in which one species differs from the other three (denoted H, C, G and M) and three in which two species differ from the other two (denoted HC, HG and CG). If all divergent sites were due to single historical mutations, the proportion of each class, $n_{\text{H}}{:}n_{\text{C}}{:}n_{\text{G}}{:}n_{\text{HC}}{:}n_{\text{HG}}{:}n_{\text{CG}}{:}n_{\text{M}}$, would be strictly

[1]Broad Institute of Harvard and Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02142, USA. [3]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.

proportional to relative 'branch lengths' of the genealogical tree—that is, the elapsed time on each branch averaged across the genome (assuming that mutations have accumulated at a constant rate over time[1,2]) (Fig. 1b–d). However, the correspondence is not exact because some sites are due to more than one mutation[9]. Recurrent mutation has a particularly distorting effect on short branches. Because HG and CG sites occur rarely and can easily be generated by recurrent mutation, branch lengths t_{HG} and t_{CG} would be overestimated without this correction.

By using five-species alignment data (HCGOM), we were able to estimate the effect of recurrent mutation on branch length estimates, controlling for the biases it introduces. In particular, the HCGOM data revealed six classes of divergent site that could not be due to single historical mutations (Table 1). An example is HO, which clusters together human and orangutan (Supplementary Table 4). We developed two methods that use the rates of these sites to correct the estimates of branch lengths for the impact of recurrent mutations (Supplementary Methods and Supplementary Note 3). Our analysis indicates that past studies that failed to correct for recurrent mutation[5–8] obtained roughly twofold higher estimates of the rates of HG and CG sites (Supplementary Table 5). Thus, these studies were biased towards overestimating the proportion of the genome in which humans and chimpanzees are not most closely related[5–7].

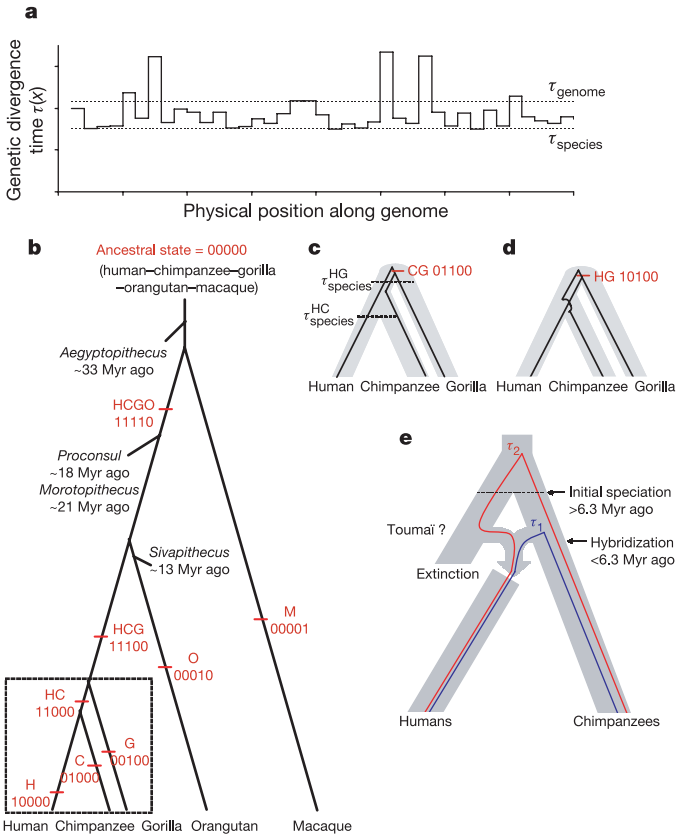

Figure 1 | Genetic relationships differ from species relationships. **a**, The genetic divergence time between two species, $\tau(x)$, varies across the genome and is always greater than or equal to the speciation time, $\tau_{species}$, which is the time of last gene flow between the species' ancestors. The average genetic divergence (τ_{genome}) thus always exceeds $\tau_{species}$. **b**, The historical relationships of the species included in this study, along with relationships to various fossils (adapted from ref. 43). The relative lengths of the branches can be estimated from the data by the number of divergent sites of each type. **c**, **d**, Genealogical relationships are not always the same as the species relationships (grey), because humans and chimpanzees can sometimes share a common ancestor that is older than the gorilla speciation (greater than $\tau^{HG}_{species}$). For example, although humans and chimpanzees are most closely related in most sections of the genome, there are regions in which chimpanzees and gorilla are most closely related[4–6] (producing 'CG' sites, **c**), or in which humans and gorilla are most closely related ('HG' sites, **d**). **e**, A revised model that could explain our data is that the first hominins became isolated from chimpanzee ancestors more than 6.3 Myr ago, but then hybridized back to the chimpanzee lineage. This could explain the great variation in divergence time across the genome, with humans and chimpanzees sharing a common ancestor around the time of hybridization in some regions (blue line) and before the initial speciation in others (red line). A model with chimpanzee ancestors as the hybrids is equally consistent with the data.

Large variation in divergence time across genome

We used a straightforward approach to study variation in $\tau(x)$, avoiding the assumptions about the demography of ancient populations from previous studies[5–8]. We selected subsets of the genome in which we proposed that the divergence would be different from the average. We then calculated the average value of $\hat{\tau}(x)$ across each subset and divided by $\hat{\tau}_{genome}$ to obtain the relative age, A, compared with the autosomal average (Supplementary Table 6).

We began by considering subsets of the genome consisting of the neighbourhoods of HC sites. The subsets $S_{HC}(d)$ were defined as all bases within a distance d of an HC site on an autosome (excluding the site itself, to obtain unbiased estimates of local genetic divergence). We expect that human and chimpanzee would be more closely related near HC sites. Analysis of the HCGOM shotgun data shows that relative age decreases with d (Fig. 2) and reaches a limit of $A \approx 0.862 \pm 0.009$—that is, 86.2% of the average genetic divergence across the autosomes (Supplementary Note 4). Because the human–chimpanzee genome divergence time is thought to be about 7 Myr ago (refs 5, 10), this translates to a roughly 1-Myr reduction. The true difference must be even larger, because the average of $\tau(x)$ near HC sites must be greater than $\tau_{species}$.

Second, we considered the neighbourhoods of HG or CG sites. The subsets $S_{HG/CG}(d)$ were defined as all bases within a distance d of either an HG or a CG site. We would expect that human and chimpanzee would be more distantly related near such sites. The relative age increases with d (Fig. 2) and reaches a limit of $A \approx 1.342 \pm 0.022$ (Supplementary Note 4). Near two HG or CG sites the increase is even greater: $A \approx 1.45 \pm 0.07$. This is primarily because two HG or CG sites close together are less likely to reflect recurrent mutations (comprising about 32% of HG and CG sites in the HCGOM data; Supplementary Table 5). When we use modelling to extrapolate the value of A near HG or CG sites in the absence of recurrent mutation, we infer that the true limit is even greater: $A \approx 1.47$ (Methods).

Table 1 | Main data sets in the study

Class	Pattern		Autosomes		X chromosome	
		Species Bases	HCGOM 8,899,720	HCGM 17,552,410	HCGOM 372,354	HCGM 747,260
n_H	10000		28,504	59,175	936	2,008
n_C	01000		28,495	59,844	944	1,935
n_G	00100		38,677	81,671	1,430	3,138
n_{HC}	11000		8,561	20,408	457	1,035
n_{HG}	10100		1,302	4,809	14	95
n_{CG}	01100		1,430	4,600	12	86
n_{HCG}	11100		41,928		1,493	
n_O	00010		82,670		3,086	
n_M	11110		244,270	596,939	9,621	23,198
n_{HO}	10010		412		9	
n_{CO}	01010		397		11	
n_{GO}	00110		764		22	
n_{HCO}	11010		1,347		56	
n_{HGO}	10110		989		30	
n_{CGO}	01110		872		32	

Each divergent site class is designated by a string of 0s and 1s, the bases seen in human–chimpanzee–gorilla–(orangutan)–macaque. The macaque allele is defined to have state '0'.

Large reduction in divergence time on chromosome X

Third, we considered the divergence along individual chromosomes, and especially chromosome X. The relative divergences for the autosomes are all close to the average (Fig. 3), but divergence is reduced along nearly the entire length of chromosome X (Fig. 3). A slightly reduced age for chromosome X is in fact expected from population genetic theory; the population size of chromosome X should be three-quarters of that of the autosomes, and thus the coalescent time should be three-quarters as large. Calculations would predict $A \approx 0.918$–0.943 (Supplementary Note 5), but the observed value is much younger: $A \approx 0.835 \pm 0.016$ (Supplementary Table 6).

To confirm the low divergence of chromosome X, we performed the same analysis for human–gorilla divergence and found no discrepancy between the expected $A \approx 0.932$–0.958 (Supplementary Note 5) and the observed $A \approx 0.977 \pm 0.028$ (Supplementary Table 6). This excludes the possibility that the reduced divergence in the human–chimpanzee comparison reflects a slowdown of the mutation rate on chromosome X in the apes, because this would affect the gorilla comparison as well. As an independent line of evidence, we note that if human–chimpanzee divergence on chromosome X is recent, we would not expect segments of human–gorilla or chimpanzee–gorilla clustering (Fig. 1c, d). In fact, the rate of HG and CG sites is one-quarter of the autosomal rate. The rate is slightly lower than would be expected if the sites were due entirely to recurrent mutation; the 95% credible interval for the proportion of HG and CG sites not due to recurrent mutation is 0–15% of the autosomal rate (Supplementary Note 6). The data are consistent with a complete absence of regions where humans and gorillas, or chimpanzees and gorillas, are most closely related.

The data point to an enormous decrease in genetic divergence for chromosome X in comparison with the autosomes. On the basis of $A \approx 0.835 \pm 0.016$ for chromosome X (Supplementary Table 6) and calibrating to an estimate of human–chimpanzee genome divergence about 7 Myr ago[2,5], the average age difference between chromosome X and the autosomes must be about 1.2 Myr. The age difference between chromosome X and the autosomes in humans today is an order of magnitude smaller (Supplementary Note 7), again indicating an unusual history in the ancestral population at the time of speciation.

We note that the reduced divergence of humans and chimpanzees on chromosome X also resolves a controversy about mutation rates in males versus females. Comparisons of genomes have shown a lower rate of sequence divergence on chromosome X than the autosomes for many species. On the basis of the hypothesis that this reflects a lower mutation rate in the female germline, it has been used to estimate the ratio α of male:female mutation rates. Studies of human genome repeats[11] and human–rat comparisons[14] have indicated that $\alpha \approx 1.9$–2.1, but comparisons of human and chimpanzee[12,15] have indicated that $\alpha \approx 6$–7. The discrepancy can be resolved if the low human–chimpanzee divergence on chromosome

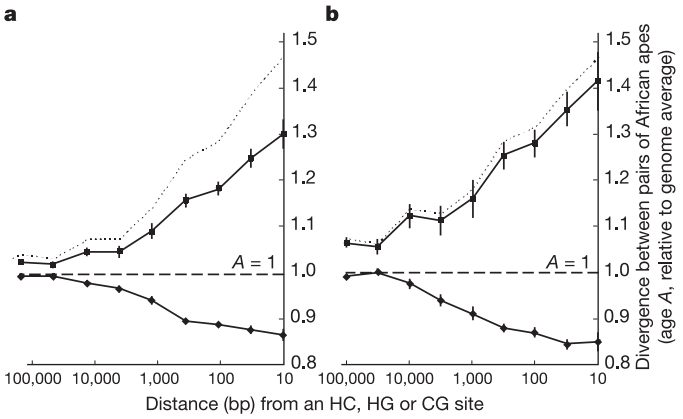

Figure 2 | Near a region of HG, CG or HC clustering, $\tau(x)$ deviates strikingly and significantly from the genome average. **a**, For the HCGM shotgun data, the human–chimpanzee divergence is less than 88% of the genome average near an HC site (upper solid line) and at least 124% of the genome average near an HG or CG site (lower solid line). **b**, For the HCGOM shotgun data, the observed proportions are 86% and 134% (Supplementary Note 4). The true range is certainly larger than shown in either graph because HG and CG sites resulting from recurrent mutation (which are more frequent for HCGM than HCGOM data) dilute the estimated increase in human–chimpanzee divergence near these sites. Correcting for this by using a modelling analysis fitted to the HCGOM shotgun data (Supplementary Table 11) indicates that the divergence near true HG and CG clusters might be about 147% of the genome average. A dotted line corresponding to this extrapolated divergence is shown. (No line is shown for HC because it is very similar to the unextrapolated line.) Each data point is obtained by averaging all bases within a window geometrically centred on distance d ($d/\sqrt{10}$ to $d\sqrt{10}$). Error bars here and in Fig. 3 give ± 1 standard error.

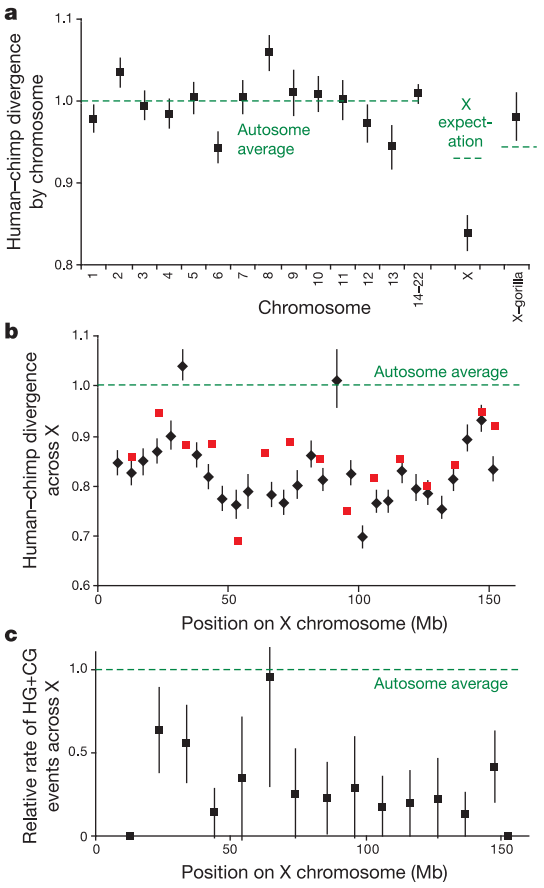

Figure 3 | Reduced human–chimpanzee time divergence across chromosome X. **a**, Human–chimpanzee genetic divergence is lower on chromosome X than on every other chromosome (HCGOM shotgun data, correcting for recurrent mutation), and lower than the theoretical expectation of about 93% for a constant-sized, freely mixing ancestral population (Supplementary Note 5). By contrast, the gorilla chromosome X comparison shows no decrease beyond the expectation of about 95% from theory. We have included a category pooling all chromosomes of less than 100 Mb in size, because the smaller chromosomes do not have much data and thus have large standard errors. **b**, The low time divergence is seen along nearly the entire X chromosome. Black symbols show the ratio of human–chimpanzee to human–macaque divergence plotted in non-overlapping 5-Mb bins in the 15.1 Mb of HCOM alignment. For comparison we also show the ratio of human–chimpanzee to human–gorilla divergence (10-Mb bins) in 372 kb of HCGOM shotgun data (red, ± 1 standard error bars removed for clarity). **c**, The rate of HG and CG sites (HCGOM shotgun data, 10-Mb bins) is also greatly reduced along chromosome X, which is consistent with humans and chimpanzees being most closely related essentially everywhere along chromosome X.

X reflects a low time divergence. Correcting for this, we estimate that $\alpha \approx 1.9$ (Supplementary Note 8), giving no evidence for an increase in α on the primate lineage[16].

Implications for current models of human–chimp speciation

The inference that human–chimpanzee genetic divergence varies over more than 4 Myr, and that genetic divergence is about 1.2 Myr less on chromosome X than the autosomes, raises two issues about human–chimpanzee speciation.

First, these results place an upper bound on the age of human–chimpanzee speciation that poses conflict with some inferences from the fossil record. The Toumaï fossil (*Sahelanthropus tchadensis*), with its bipedalism and hominin dental features, is usually interpreted as being on the hominin line and setting a minimum date for human–chimpanzee speciation[17,18]. The fossil was originally dated to 6–7 Myr ago (refs 17, 18), and a more recent study estimates 6.5–7.4 Myr ago (refs 19). We compared inferences for the human–chimpanzee speciation date based on the Toumaï fossil with those that would be obtained by extrapolating from older species divergences. We first used an extreme upper bound of 20 Myr ago for human–orangutan genetic divergence[20] (Supplementary Table 7). On the basis of the relative genetic divergence of human and chimpanzee (Supplementary Tables 7 and 8; Supplementary Note 9), we infer $\tau_{\text{genome}} < 7.6$ Myr ago and, from the bound $\tau_{\text{species}} < 0.835\tau_{\text{genome}}$, we can infer that $\tau_{\text{species}} < 6.3$ Myr ago. Using a more realistic estimate of human–orangutan genome divergence of less than 17 Myr ago, we obtain a younger bound of $\tau_{\text{species}} < 5.4$ Myr ago. The first bound is not compatible with the older range for Toumaï. The second bound is difficult to reconcile not only with interpretations of Toumaï but also with other fossils recognized as early hominins: *Orrorin tugenensis* at about 5.8 Myr ago (ref. 21) and *Ardipithecus kadabba* at about 5.6–5.8 Myr ago (ref. 22). We emphasize that these calibrations to the older fossil record are not likely to be compromised by molecular clock errors: a 'rate test' shows evidence of only slight lineage-specific changes in the mutation rate since the divergence of the great apes[23] (Supplementary Tables 8 and 9). Similar bounds on human–chimpanzee speciation time are obtained by calibration to macaque fossil divergence (Supplementary Note 9).

Second, the properties of chromosome X indicate an unusual evolutionary history around the time of human–chimpanzee ancestral speciation—proving that the structure of the population around the time of speciation was unlike that in any modern human or apes. In a freely mixing population under neutral drift, the ratio R of the genetic divergence on chromosome X and the autosomes should be about 0.75 (the effective population size of chromosome X relative to the autosomes). Such values are, in fact, observed in humans ($R \approx 0.59$–0.87) and chimpanzees ($R \approx 0.56$–0.76) and inferred for the population ancestral to human and gorilla ($R \approx 0.68$–1.00) and chimpanzee and bonobo ($R \approx 0.75$) (Supplementary Note 10). By contrast, the inferred value of R in the population ancestral to human and chimpanzee is 0–0.29 (Supplementary Table 10). We were not able to devise a demographic history consistent with such a low R, even with models of asymmetry between the sexes (Supplementary Note 11). However, strong selection across chromosome X could produce this effect.

The apparent conflict with interpretations of the fossil record could be explained if Toumaï were somewhat younger than previously reported[19], or if there was a problem with the molecular clock used for the calibrations to older fossil divergences. These factors, however, would not explain the more than 4 Myr spread of genetic divergence times across the genome, or the evidence for intense natural selection on chromosome X. We note that, on general grounds, we might expect to see greater evidence for natural selection on chromosome X than the autosomes, because recessive genetic variants are subject to selection in hemizygous males. However, we see no evidence for an unusual X chromosome divergence in the human–gorilla comparison.

Possible hybridization in the human–chimp lineage

We suggest a provocative explanation for multiple features of these data: that the hominin and chimpanzee lineages initially separated but then exchanged genes before finally separating less than 6.3 Myr ago (Fig. 1e). First, this could explain how Toumaï could have dates older than hominin speciation and yet still have hominin features[17–19]. Second, it could explain the wide range of divergence times (more than 4 Myr): at some loci human and chimpanzee lineages share ancestry around initial separation, whereas at others the genetic ancestry is more recent at the time of hybridization. Third, it could explain the low divergence of human and chimpanzee on chromosome X. An empirically observed pattern, documented in *Drosophila*, tsetse flies, mosquitoes, butterflies and guinea-pigs[24], is that "the genes having the greatest effect on hybrid sterility and inviability are X-linked"[24]. The reasons for this 'second rule of speciation'[24] are not fully understood[25–27], although they are thought to be related to Haldane's rule about hybrid sterility affecting the heterogametic sex more than the homogametic sex[28]. A corollary—not previously suggested—is that if gene flow between two diverged populations occurs, chromosome X should be subject to strong and rapid selection to eliminate alleles, from one parental population or the other, that contribute to reduced fitness. The presence of multiple hybrid incompatibility loci could lead to selection across much or all of chromosome X, as in our data (Fig. 3). As a specific example, if human and chimpanzee ancestors initially speciated and then interbred, hybrid males might have been infertile, consistent with Haldane's rule. A viable population could then only have arisen if the fertile females mated back to one of the ancestral populations (for example, chimpanzee ancestors), producing fertile male hybrids when they transmitted X chromosomes derived almost entirely from that ancestral population. This could explain why humans and chimpanzees are most closely related throughout chromosome X. We note that in wild mice in the European *Mus musculus/domesticus* hybrid zone there has been reported to be a gradient of genetic variants on the autosomes, but a sharp geographic transition for chromosome X (ref. 29). This indicates that wild mouse hybrid populations might have difficulty carrying X chromosomes from multiple ancestral populations, which is consistent with what would be expected from our model and proposed corollary to the second rule of speciation.

Speciation in animals is generally believed to occur by allopatry—that is, by the formation of an isolation barrier with no subsequent gene flow. When subsequent hybridization does occur, it is generally believed that the resulting population dies out[30,31]. However, there are known examples of adapted hybrid populations in nature[32–34], and hybridization could be advantageous, allowing nascent species to derive traits from several ancestral populations, combining them to adapt to new environments[35]. The failure to observe more instances of successful hybridizations in field studies so far[30,31] might simply be due to ascertainment bias—the fact that hybridizations occur too episodically to be observed practically. With comparative genomic methods, one can project backwards in time to make inferences about what happened at the time at which speciation occurred. Allopatric speciation without subsequent gene flow predicts that population genetic structure before and after speciation should be similar[30,31]. By contrast, hybridization predicts a wide range of divergence times and different coalescence times in parts of the genome, such as chromosome X. By comparing the genomes of modern species, one could systematically test whether hybridization is a widespread process in evolution.

We have shown that human and chimpanzee speciation was complex; furthermore, our model makes predictions that can be tested with larger data sets[36]. First, it predicts that evidence of natural selection should be seen not only on chromosome X but also at some autosomal loci. Such ancient selective sweeps might be detected as long regions with unusually low (or high) rates of human–chimpanzee divergence and HG and CG sites. (It might even be

possible to identify autosomal genes under selection). Second, if a hybridization involving a single episode of gene flow occurred, it might result in a bimodal distribution of $\tau(x)$. Third, speciation involving hybridization could give rise to distinctive patterns of 'frozen linkage disequilibrium': differences in the lengths and distribution of HC, HG or CG clustering (Fig. 2). All these hypotheses can be tested once the gorilla genome is complete and aligned to the genomes of humans, chimpanzees and more distant primates.

METHODS

DNA sequence data. We sequenced 115,152 fragments of DNA ('shotgun reads') from a western lowland gorilla and 2,710 from a black-handed spider monkey (*Ateles geoffryi*) (these data are publicly available in the NCBI trace archive, http://www.ncbi.nlm.nih.gov/Traces). We combined our data with whole-genome shotgun sequence data from orangutan and macaque from the Washington University Genome Sequencing Center, the Baylor College of Medicine and the Venter Institute (http://www.ncbi.nlm.nih.gov/Traces/trace.cgi; we thank our colleagues for making these data publicly available) (Supplementary Table 1). For a supporting data set, we analysed finished contiguous sequence from bacterial artificial chromosome (BAC) sequencing of sections of chromosomes 7 and X, generated by the NIH Intramural Sequencing Center[13].

Genome alignments. All shotgun reads were aligned to the NCBI Build 34 human genome assembly using the Arachne[37] or BLASTZ[38] programs (Supplementary Table 1). At loci with at least 100 base pairs of DNA aligned across all species of interest, we used the Multiple Alignment Program[39] to obtain optimized local alignments (Supplementary Methods). We then applied four filters (Supplementary Table 2 and Supplementary Methods) to eliminate the following: first, alignments with an extremely high rate of intraspecific polymorphism; second, alignments with an extreme rate of reads from any one species; third, alignments with a very high rate of divergent sites from some species; or fourth, alignments that mapped to known segmental duplications in humans or chimpanzees[40]. Application of these filters minimized misalignment in our data but did not qualitatively change our main inferences (Supplementary Table 3). The Threaded Block Set Aligner[41] was used for alignment of the BAC data. Aligned data sets are available online and at our website (http://genepath.med.harvard.edu/~reich).

Identification of divergent sites for analysis. For the shotgun data, we used divergent sites only if they had a sequencing quality score of at least 30 and 5 bases on each side with quality 25 or more. We additionally required sites to be single nucleotide substitutions, to show exactly two alternate alleles across the species, to be outside hypermutable CpG dinucleotides, to be at least three base pairs from a repeat identified by Tandem Repeat Finder[42], and to be flanked on either side by at least one base of perfect alignment across species. When multiple reads were available, we used data from the read with the highest sequence quality. Application of these filters did not qualitatively affect the main conclusions from our analysis (Supplementary Table 3). The filtered data sets are available online and at our laboratory website (genepath.med.harvard.edu/~reich).

Genetic divergence estimates. The main measurement in this paper is genetic divergence between two species. To calculate this over a stretch of sequence, we always counted the number of differences per base pair between the two species and normalized by the difference between human and macaque (or another outgroup) over the same stretch. This corrects for variability in mutation rate from locus to locus. In particular, it corrects for a high local mutation rate. As an example for the HCGOM data, the normalized divergence is proportional to $\tau(x) = (n_{\mathrm{H}} + n_{\mathrm{HG}} + n_{\mathrm{C}} + n_{\mathrm{CG}})/(n_{\mathrm{H}} + n_{\mathrm{HC}} + n_{\mathrm{HG}} + n_{\mathrm{HCG}} + n_{\mathrm{M}})$. Using human–macaque divergence for normalization is slightly conservative for X–autosome comparisons. Like all species pairs, humans and macaques are slightly less time-diverged on chromosome X; if we could correct for this, human–macaque divergence on chromosome X could be up to a few per cent less than shown in Supplementary Table 6.

Received 5 November 2005; accepted 7 April 2006.
Published online 17 May 2006.

1. Zuckerkandl, E. & Pauling, L. Molecules as documents of evolutionary history. *J. Theor. Biol.* **8,** 357–366 (1965).
2. Glazko, G. V. & Nei, M. Estimation of divergence times for major lineages of primate species. *Mol. Biol. Evol.* **20,** 424–434 (2003).
3. Pamilo, P. & Nei, M. Relationships between gene trees and species trees. *Mol. Biol. Evol.* **5,** 568–583 (1988).
4. Ruvolo, M. Molecular phylogeny of the hominoids: inferences from multiple independent DNA sequence data sets. *Mol. Biol. Evol.* **14,** 248–265 (1997).
5. Chen, F. C. & Li, W. H. 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 (2001).
6. Takahata, N. & Satta, Y. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from DNA sequences. *Proc. Natl Acad. Sci. USA* **94,** 4811–4815 (1997).
7. Wall, J. D. Estimating ancestral population sizes and divergence times. *Genetics* **163,** 395–404 (2003).
8. Rannala, B. & Yang, Z. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164,** 1645–1656 (2003).
9. O'hUigin, C., Satta, Y., Takahata, N. & Klein, J. Contribution of homoplasy and of ancestral polymorphism to the evolution of genes in anthropoid primates. *Mol. Biol. Evol.* **19,** 1501–1513 (2002).
10. Pilbeam, D. & Young, N. Hominoid evolution: synthesizing disparate data. *C. R. Palevol.* **3,** 305–321 (2004).
11. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409,** 860–921 (2001).
12. Mikkelsen, T. S. *et al.* Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
13. Hwang, D. G. & Green, P. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. *Proc. Natl Acad. Sci. USA* **101,** 13994–14001 (2004).
14. Rat Genome Sequencing Project Consortium, Genome sequence of the Brown Norway rat yields insights into mammalian evolution. *Nature* **428,** 493–521 (2004).
15. Taylor, J., Tyekucheva, S., Zody, M., Chiaromonte, F. & Makova, K. D. Strong and weak male mutation bias at different sites in the primate genomes: Insights from the human–chimpanzee comparison. *Mol. Biol. Evol.* **23,** 565–573 (2006).
16. Makova, K. D. & Li, W. H. Strong male-driven evolution of DNA sequences in humans and apes. *Nature* **416,** 624–626 (2002).
17. Brunet, M. *et al.* A new hominid from the Upper Miocene of Chad. *Nature* **418,** 145–151 (2002).
18. Vignaud, P. *et al.* Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* **418,** 152–155 (2002).
19. Brunet, M. *et al.* New material of the earliest hominid from the Upper Miocene of Chad. *Nature* **434,** 752–755 (2005).
20. MacLatchy, L., Gebo, D., Kityo, R. & Pilbeam, D. Postcranial functional morphology of *Morotopithecus bishopi*, with implications for the evolution of modern ape locomotion. *J. Hum. Evol.* **38,** 1–25 (2000).
21. Senut, B. *et al.* First hominid from the Miocene (Lukeino Formation, Kenya). *C. R. Acad. Sci. IIA* **332,** 137–144 (2001).
22. WoldeGabriel, G. *et al.* Geology and palaeontology of the Late Miocene Middle Awash valley, Afar rift, Ethiopia. *Nature* **412,** 175–178 (2001).
23. Sarich, V. M. in *New Interpretations of Ape and Human Ancestry* (eds Ciochon, R. L. & Corruccini, R. S.) 137–150 (Plenum, New York, 1983).
24. Coyne, J. A. & Orr, H. A. in *Speciation and its Consequences* (eds Otte, D. & Endler, J. A.) 180–207 (Sinauer, Sunderland, Massachusetts, 1989).
25. Tao, Y., Chen, S., Hartl, D. L. & Laurie, C. C. Genetic dissection of hybrid incompatibilities between *Drosophila simulans* and *D. mauritiana*. I. Differential accumulation of hybrid male sterility effects on the X and autosomes. *Genetics* **164,** 1383–1397 (2003).
26. True, J. R., Weir, B. S. & Laurie, C. C. A genome-wide survey of hybrid incompatibility factors by the introgression of marked segments of *Drosophila* mauritiana chromosomes into *Drosophila* simulans. *Genetics* **142,** 819–837 (1996).
27. Orr, H. A. & Irving, S. Segregation distortion in hybrids between the Bogota and USA subspecies of *Drosophila pseudoobscura*. *Genetics* **169,** 671–682 (2005).
28. Haldane, J. B. S. Sex ratio and unidirectional sterility in hybrid animals. *J. Genet.* **58,** 237–242 (1922).
29. Tucker, P. K., Sage, R. D., Wilson, A. C. & Eichler, E. M. Abrupt cline for sex chromosomes in a hybrid zone between two species of mice. *Evolution Int. J. Org. Evolution* **46,** 1146–1163 (1992).
30. Mayr, E. *Systematics and the Origin of Species* (Columbia Univ. Press, New York, 1942).
31. Coyne, J. A. & Orr, H. A. *Speciation* (Sinauer, Sunderland, Massachusetts, 2004).
32. Rieseberg, L. H. Hybrid origin of plant species. *Annu. Rev. Ecol. Syst.* **28,** 359–389 (1997).
33. Boag, P. T. & Grant, P. R. The classical case of character release: Darwin's finches (Geospiza) on Isla Daphne Major, Galápagos. *Biol. J. Linn. Soc.* **22,** 243–287 (1984).
34. Schwarz, D., Matta, B. M., Shakir-Botteri, N. L. & McPheron, B. A. Host shift to an invasive plant triggers rapid animal hybrid speciation. *Nature* **436,** 546–549 (2005).
35. Barton, N. H. The role of hybridization in evolution. *Mol. Ecol.* **10,** 551–568 (2001).
36. Osada, N. & Wu, C. I. Inferring the mode of speciation from genomic data: a study of the great apes. *Genetics* **169,** 259–264 (2005).
37. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).

38. Schwartz, S. *et al.* Human–mouse alignments with BLASTZ. *Genome Res.* **13,** 103–107 (2003).
39. Huang, X. On global sequence alignment. *Comput. Appl. Biosci.* **10,** 227–235 (1994).
40. Cheng, Z. *et al.* A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* **437,** 88–93 (2005).
41. Blanchette, M. *et al.* Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14,** 708–715 (2004).
42. Benson, G. Tandem repeats finder: a program to analyse DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).
43. Steiper, M. E., Young, N. M. & Sukarna, T. Y. Genomic data support the hominoid slowdown and an Early Oligocene estimate for the hominoid-cercopithecoid divergence. *Proc. Natl Acad. Sci. USA* **101,** 17021–17026 (2004).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements We thank B. Bodamer, J. Caswell, M. Clamp, J. Coyne, J. Cuff, E. Green, G. McDonald, J. Mullikin, H. A. Orr, D. Page, D. Pilbeam, N. Stange-Thomann and M. Zody for discussions, comments and assistance with stages of this study, and E. Green, R. Gibbs and R. Wilson for producing and making publicly available data from large-scale sequencing projects (contiguous data from chromosomes 7 and X, and the orangutan and macaque shotgun data). N.P. was supported by a career transition award from the National Institutes of Health. E.S.L. was supported in part by funds from the National Human Genome Research Institute and the Broad Institute of Harvard and the Massachusetts Institute of Technology. D.R. was supported in part by a Burroughs–Wellcome Career Development Award in the Biomedical Sciences.

Author Contributions The authors all played significant roles in the conception, execution, interpretation and presentation of the study.

Author Information To obtain sequencing reads from the NCBI trace archive (http://www.ncbi.nlm.nih.gov/Traces), use the following queries:
(1) Gorilla data (*Gorilla gorilla*):
CENTER_NAME = 'WIBR' and CENTER_PROJECT = 'G611'
CENTER_NAME = 'WIBR' and CENTER_PROJECT = 'G612'
CENTER_NAME = 'WIBR' and CENTER_PROJECT = 'G618'
CENTER_NAME = 'WIBR' and CENTER_PROJECT = 'G619'
CENTER_NAME = 'WIBR' and CENTER_PROJECT = 'G744';
(2) New world monkey data (*Ateles geoffroyi*):
CENTER_NAME = 'WIBR' and CENTER_PROJECT = 'G820'.
 The authors declare no competing financial interests.
Correspondence and requests for materials should be addressed to D.R. (reich@genetics.med.harvard.edu).

Complex speciation of humans and chimpanzees

Arising from: N. Patterson, D. J. Richter, S. Gnerre, E. Lander & D. Reich *Nature* **441**, 1103–1108 (2006)

Genetic data from two or more species provide information about the process of speciation. In their analysis of DNA from humans, chimpanzees, gorillas, orangutans and macaques (HCGOM), Patterson *et al.*[1] suggest that the apparently short divergence time between humans and chimpanzees on the X chromosome is explained by a massive interspecific hybridization event in the ancestry of these two species. However, Patterson *et al.*[1] do not statistically test their own null model of simple speciation before concluding that speciation was complex, and—even if the null model could be rejected—they do not consider other explanations of a short divergence time on the X chromosome. These include natural selection on the X chromosome in the common ancestor of humans and chimpanzees, changes in the ratio of male-to-female mutation rates over time, and less extreme versions of divergence with gene flow (see ref. 2, for example). I therefore believe that their claim of hybridization is unwarranted.

Patterson *et al.*[1] estimate the divergence time between humans and chimpanzees on the X chromosome to be 0.835 times the average autosomal divergence time; they note that this is less than the value of 0.94 predicted by their model of simple speciation (see Methods). They also computed a 'genome minimum' divergence time of 0.86 by measuring divergence close to sites where humans and chimpanzees share a derived base ('HC' sites[1]), and cited the similarity between this and the X-chromosome divergence time of 0.835 as support of hybridization. However, the HC speciation time estimated using the data and methods of Patterson *et al.*[1] is only 0.76 (see Methods): the X-chromosome divergence and the genome minimum divergence therefore occurred in the common ancestor of humans and chimpanzees, rather than after an initial speciation event (Fig. 1). The apparently low estimated divergence time on the X chromosome thus provides only indirect evidence for complex speciation.

It has been suggested that the autosomal data may be consistent with the null model of simple speciation and that a test is lacking[3]. With reference just to the apparent reduction in divergence time on the X chromosome, a proper statistical test would compute the probability under the null model of observing a difference between the X chromosome and the autosomes as large as, or larger than, the difference between 0.94 and 0.835. This test would require all relevant sources of variation to be accounted for, including the stochastic variation in coalescence times under the null model and the variation in α, the ratio of male-to-female mutation rates. The jack-knife procedure described in the Supplementary Information of ref. 1 does not do this.

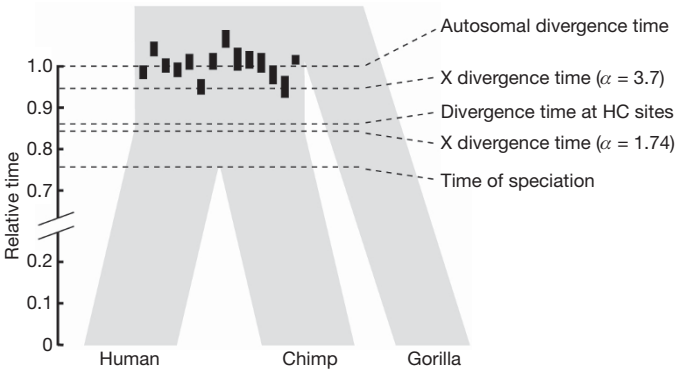

Figure 1 | Simple Speciation of humans and chimpanzees. Black bars show the ranges of autosomal human–chimpanzee divergence by chromosome, redrawn (in the same order) from Fig. 3a of Patterson *et al.*[1]. Dashed lines display quantities discussed in the text. The drawing is to scale, with all times given relative to the average autosomal divergence time between humans and chimpanzees.

Assuming that the null model can be rejected, the suggestion of hybridization needs to be supported by rejecting other kinds of complex speciation, which Patterson *et al.*[1] do not. Their Supplementary Note 11 considers whether modifications of the null model could produce a large reduction of HC divergence on the X chromosome, and they conclude that natural selection must explain the reduction. Note that, because fossil dates were included in the analysis, Supplementary Note 11 seeks to explain a more extreme reduction ($R < 0.29$[1]) than implied by the genetic data alone. Different scenarios that include natural selection and models of complex speciation other than hybridization are not considered.

An alternative to hybridization is that α has changed during the course of primate evolution. Another study[4] finds $\alpha = 3.88$ (95% CI = 2.90–6.07) in primates and $\alpha = 3.79$ (95% CI = 2.71–5.99) in perisodactyls. In contrast, Patterson *et al.*[1] use $\alpha < 2$, based on the observed divergences to macaque in their data. Relative genetic divergences on the X chromosome versus the autosomes are converted into the relative divergence times above by using the factor $3(1+\alpha)/(4+2\alpha)$, which accounts for the different fractions of time the X and the autosomes spend in males versus females. For example, summing the counts of patterns in which human and chimpanzee differ and dividing by the total number of bases in Table 1 of ref. 1, gives genetic divergences of $\pi^{(X)}_{HC} = 0.0053$ and $\pi^{(A)}_{HC} = 0.0070$. The relative genetic divergence is $\pi^{(X)}_{HC}/\pi^{(A)}_{HC} \approx 0.76$. A similar value ($0.0094/0.0123 \approx 0.764$) is obtained from the whole genomes of humans and chimpanzees[5]. If $\alpha = 1.74$, then $0.76 \times 3(1+\alpha)/(4+2\alpha) = 0.835$ (ref. 1). Instead, if $\alpha = 3.7$, then $0.76 \times 3(1+\alpha)/(4+2\alpha) = 0.94$, as predicted by the fitted null model[1]. A similar value ($\alpha = 3.55$) makes the X-autosome relative genetic divergence 0.61 within humans[6], consistent with the simple prediction of 3/4 for the relative divergence time.

METHODS

The values 0.76 and 0.94 are obtained from the HCGOM autosomal data in Supplementary Table 4 of ref. 1 by using the model and method in Supplementary Note 2 of ref. 1, and assuming that the effective population sizes of HC and HCG ancestors are the same. The three model parameters are in units of expected numbers of mutations in the sampled portions of the autosomal genome: γ is the time back to the HC speciation event, β is the time between the HCG and HC speciation events, and $\theta/2$ is the pairwise coalescence time within a species. The expected values in Supplementary Note 2 of ref. 1 become $E[n_H + n_{HG}] = \gamma + \theta/2$; $E[n_H + n_{HC}] = \gamma + \beta + \theta/2$; $E[n_{HG}] = \theta e^{-2\beta/\theta}/6$, where n_H, n_{HC} and n_{HG} are the numbers of mutations unique to humans, shared uniquely by humans and chimpanzees, and shared uniquely by humans and gorillas. After correcting for multiple mutations at single sites[1], the observed values become $n_H = 28{,}173$, $n_{HC} = 7{,}990$ and $n_{HG} = 866$. By equating the expected and observed values, one finds $\gamma \sim 21{,}946$, $\beta \sim 7{,}124$ and $\theta \sim 14{,}186$. Thus, relative to the average autosomal divergence, the estimated HC speciation time is $2\gamma/(2\gamma+\theta) \approx 0.76$. With the standard assumption of a 3/4 ratio of effective population sizes on the X chromosome versus the autosomes[1], the predicted value for the relative HC divergence time on the X chromosome is $(2\gamma + 0.75\theta)/(2\gamma+\theta) \approx 0.94$.

John Wakeley[1]

[1]Department of Organismic and Evolutionary Biology, Harvard University, 16 Divinity Avenue, Cambridge, Massachusetts 02138, USA.
e-mail: wakeley@fas.harvard.edu

Received 22 February 2007; accepted 17 January 2008.

1. Patterson, N., Richter, D. J., Gnerre, S., Lander, E. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441**, 1103–1108 (2006).
2. Innan, H. & Watanabe, H. The effect of gene flow on the coalescent time in the human–chimpanzee ancestral population. *Mol. Biol. Evol.* **23**, 1040–1047 (2006).
3. Barton, N. H. Evolutionary biology: how did the human species form? *Curr. Biol.* **16**, R647–R650 (2006).

4. Goetting-Minetsky, M. P. & Makova, K. D. Mammalian male mutational bias: impacts of generation time and regional variation in substitution rates. *J. Mol. Evol.* **63,** 537–544 (2006).
5. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437,** 69–87 (2005).
6. The International SNP Map Working Group. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409,** 928–933 (2001).

doi:10.1038/nature06805

Patterson et al. *reply*

Replying to: J. Wakeley *Nature* 452, doi:10.1038/nature06805 (2008)

In his communication[1], Wakeley does not find any flaw in our argument for complex speciation of humans and chimpanzees[2], nor has he identified a simple demographic model that can explain the notable differences in genetic divergence that we observe between humans and chimpanzees when comparing chromosome X and the autosomes.

Our argument[2] for complex speciation rests on the difference in genetic divergence time that we observe between chromosome X and the autosomes, and not on the wide range of genetic divergence times observed within the autosomes (which can indeed be explained by a large ancestral population size[3,4]). To reiterate our argument for complex speciation, we began with a null model of simple speciation in which the ancestral populations of humans and chimpanzees were separated by a barrier with no subsequent gene flow. Fitting this model to our data, we obtain an expectation for the genetic divergence on chromosome X. However, the observed chromosome X data differ from this, and we could not explain the difference even by using more elaborate demographic histories. Wakeley's model of demographic history can explain only the autosomal data, and does not reconcile the autosomal and chromosome X data, even though comparing these two parts of the genome provides the key signal for complex speciation. Table 1 shows three statistics that are each sensitive to human–chimpanzee genetic divergence on chromosome X: all are significantly reduced (4.4–8.3 s.d.) compared with that predicted from Wakeley's simple model fitted to our autosomal data.

Wakeley[1] suggests that the ratio of male-to-female mutation rate in primates (α) might be higher than we estimated. This would not explain the data, especially in light of our human–gorilla comparison. Using Wakeley's method[1] for calculating α with raw genetic divergence data (rather than by comparison with an outgroup[2]), we estimate $\alpha = 3.19$ from the human–chimpanzee comparison, similar to the value Wakeley reports (the slight difference is owing to our correction for recurrent mutation). But this disagrees with our $\alpha = 1.57$ obtained from the same calculation using the human–gorilla comparison.

If mutation-rate differences alone could explain the observed data, we would expect a consistent value for α from the human–chimpanzee and human–gorilla divergence data, but estimates of α are significantly different ($P = 0.001$). A high value of α also cannot explain other important features in Table 1: the near-absence of sites on chromosome X that cluster humans and gorillas or chimpanzees and gorillas; or why human–gorilla divergence should not be reduced on chromosome X (such a reduction would be expected if high male mutation rate were responsible for low human–chimpanzee genetic divergence on chromosome X).

What could explain the evidence for reduced chromosome-X time divergence? We suggested hybridization[2]. In hybrids, genetic barriers to gene flow often map to chromosome X (ref. 5). A corollary is that if a hybrid population overcomes these barriers, it may experience intense selection to eliminate most or all the chromosome X contribution from one of the ancestral populations—the population cannot tolerate the sequence of both X chromosomes and one is selected away. When hybridization between two populations establishes a third population, the divergence on chromosome X will be large relative to one ancestral population and small relative to the other, depending on which ancestral chromosome X is selected away. Depending on which populations survive to the present day, the divergence on chromosome X will be very high or very low relative to that on the autosomes (but not intermediate), which would explain the low observed chromosome X divergence.

We therefore reject the simple model of speciation that Wakeley proposes, having also investigated other simple speciation models and found none to explain the data[2]. In contrast, hybridization followed by natural selection across the entire chromosome X to eliminate hybrid sterility or inviability loci does explain the data[2]. To argue against the evidence for complex speciation, an alternative simple model is needed that explains the reduced chromosome X divergence in humans and chimpanzees with no similar reduction for humans and gorillas. No one has yet succeeded in identifying such a model.

Nick Patterson[1], Daniel J. Richter[1], Sante Gnerre[1], Eric S. Lander[1,2] & David Reich[1,3]
[1]Broad Institute of Harvard and MIT, Cambridge, Massachusetts 02142, USA.
[2]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, USA.
[3]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.
e-mail: reich@genetics.med.harvard.edu

1. Wakeley, J. Complex speciation of humans and chimpanzees. *Nature* **452,** doi:10.1038/nature06805 (2008).
2. Patterson, N., Richter, D. J., Gnerre, S., Lander, E. S. & Reich, D. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441,** 1103–1108 (2006).
3. Barton, N. H. Evolutionary biology: how did the human species form? *Curr. Biol.* **16,** 647–650 (2006).
4. Innan, H. & Watanabe, H. The effect of gene flow on the coalescent time in the human–chimpanzee ancestral population. *Mol. Biol. Evol.* **23,** 1040–1047 (2006).
5. Coyne, J. A. & Orr, H. A. in *Speciation and its consequences* (eds Otte, D. & Endler, J. A.) 180–207 (Sinauer Associates, Sunderland, Massachusetts, 1989).

doi:10.1038/nature06806

Table 1 | Statistics showing extreme reduction in genetic divergence time between humans and chimps on chromosome X

Statistic of interest*	Expectation of X-to-autosome ratio from Wakeley's model†	Observation of X-to-autosome ratio in our actual data	Significance of difference between observed and expected (s.d.)
Human–chimpanzee genetic divergence averaged across the genome and normalized by human–macaque divergence	0.937 ± 0.002	0.839 ± 0.022	−4.4
Sum of divergent sites clustering humans and gorillas (HG) or chimpanzees and gorillas (CG) and normalized by human–macaque divergence	0.548 ± 0.023	0.022 ± 0.059	−8.3
Ratio of human–chimpanzee to human–gorilla divergence averaged across the genome and calculated on the human side of the tree	0.987 ± 0.001	0.862 ± 0.022	−5.8

* Corrected for recurrent mutation.
† Parameters fitted to autosomal data, then extrapolated to chromosome X. The extrapolation to chromosome X assumes a ratio of chromosome X to autosomal population size of three-quarters throughout history.