

Absence of the TAP2 Human Recombination Hotspot in Chimpanzees

Susan E. Ptak¹, Amy D. Roeder^{1,2}, Matthew Stephens², Yoav Gilad^{1,2}, Svante Pääbo¹, Molly Przeworski^{1,3*}

1 Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, **2** Department of Statistics, University of Washington, Seattle, Washington, United States of America

Recent experiments using sperm typing have demonstrated that, in several regions of the human genome, recombination does not occur uniformly but instead is concentrated in “hotspots” of 1–2 kb. Moreover, the crossover asymmetry observed in a subset of these has led to the suggestion that hotspots may be short-lived on an evolutionary time scale. To test this possibility, we focused on a region known to contain a recombination hotspot in humans, TAP2, and asked whether chimpanzees, the closest living evolutionary relatives of humans, harbor a hotspot in a similar location. Specifically, we used a new statistical approach to estimate recombination rate variation from patterns of linkage disequilibrium in a sample of 24 western chimpanzees (*Pan troglodytes verus*). This method has been shown to produce reliable results on simulated data and on human data from the TAP2 region. Strikingly, however, it finds very little support for recombination rate variation at TAP2 in the western chimpanzee data. Moreover, simulations suggest that there should be stronger support if there were a hotspot similar to the one characterized in humans. Thus, it appears that the human TAP2 recombination hotspot is not shared by western chimpanzees. These findings demonstrate that fine-scale recombination rates can change between very closely related species and raise the possibility that rates differ among human populations, with important implications for linkage-disequilibrium based association studies.

Introduction

Recombination is a fundamental biological feature about which we still know remarkably little, especially in mammals. Understanding recombination is also of practical importance for evolutionary inference and human genetics (Nachman 2002; Arnheim et al. 2003). Unfortunately, the process is difficult to study, because recombination events occur extremely rarely (e.g., with a probability of $\sim 10^{-8}$ per bp per generation in a typical region of the human or *Drosophila melanogaster* genome; Ashburner 1989; Kong et al. 2002). Thus, direct measurements for closely linked sites often require the examination of a prohibitive number of individuals. As a result, our knowledge of recombination rates stems primarily from estimates for markers that are megabases apart, obtained from crosses or, for humans, obtained from pedigrees (e.g., Kong et al. 2002).

One way to learn about finer-scale recombination rates in males is sperm typing (Li et al. 1988; Hubert et al. 1994; Jeffreys et al. 2001). In this approach (reviewed by Arnheim et al. 2003), genetic markers are amplified and typed from a large number of sperm in order to estimate the fraction of recombinant sperm and hence the recombination rate. Fine-scale rates can also be measured indirectly from patterns of allelic associations, or linkage disequilibrium (LD), observed in samples from natural populations (Hudson 1987; Pritchard and Przeworski 2001). In humans, both direct estimates of the recombination rate using sperm typing techniques and indirect approaches based on analyses of LD have suggested the existence of substantial heterogeneity in rates of recombination at small scales (Daly et al. 2001; Jeffreys et al. 2001; Gabriel et al. 2002; Schneider et al. 2002; Wall and Pritchard 2003). In particular, sperm typing experiments have demonstrated that, in several regions of the human genome, crossover resolutions are not uniformly distributed but

instead tend to cluster within narrow regions of 1–2 kb termed “recombination hotspots” (de Massy 2003 and references therein).

While there has been recent progress characterizing the extent of spatial variation in recombination rates, the time scale over which recombination rates change remains an open question. It has been known for decades that natural populations harbor genetic variation for recombination rates (Brooks 1988 and references therein). In humans, in particular, there are significant differences in recombination rates among females (Kong et al. 2002) as well as among males (Cullen et al. 2002). Thus, there is a clear potential for the evolution of recombination rates. However, there are only a couple of demonstrated cases that help to delimit the time scale on which this might occur: at the megabase scale, the best example is probably *D. melanogaster* and *D. simulans*, two

Received December 18, 2003; Accepted March 21, 2004; Published June 15, 2004

DOI: 10.1371/journal.pbio.0020155

Copyright: © 2004 Ptak et al. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abbreviations: LD, linkage disequilibrium; PAC model, Product of Approximate Conditionals model; SNP, single nucleotide polymorphism.

Academic Editor: Andy Clark, Cornell University

*To whom correspondence should be addressed. E-mail: Molly_Przeworski@Brown.edu

©These authors contributed equally to this work.

[†]1Current address: Biodiversity and Ecological Processes Group, Cardiff School of Biosciences, Cardiff, United Kingdom

[†]2Current address: Department of Genetics, Yale University, New Haven, Connecticut, United States of America

[†]3Current address: Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island, United States of America



sibling species that differ in their recombination landscape (True et al. 1996). Among primates, the genetic map of humans is approximately 28% longer than that of an Old World monkey, the baboon (*Papio hamadryas*; Rogers et al. 2000), suggesting that—if physical maps are roughly similar—recombination rates in humans may be higher overall. These instances demonstrate that large-scale recombination rates can change between species that differ on average at roughly 6% to 10% of nucleotide positions (Betancourt and Presgraves 2002; Thomas et al. 2003).

At a finer scale, the only evidence stems from a recent study of the β -globin gene, where a hotspot had been characterized by sperm typing in humans. Wall et al. (2003) found no evidence of rate variation in LD data collected from the rhesus macaque (*Macaca mulatta*), another Old World monkey. For more closely related species, nothing is known. However, observations in yeast (e.g., Petes 2001; Steiner et al. 2002) and mammals (Jeffreys and Neumann 2002; Yauk et al. 2003) raise the possibility that local recombination rates could change rapidly. Indeed, at the MS32 and DNA2 hotspots in humans (Jeffreys et al. 1998; Jeffreys and Neumann 2002) as well as at the E_{β} hotspot in mice (*Mus* sp.; Yauk et al. 2003), some haplotypes were found to lead to higher rates of initiation of crossover events. Such haplotypes tended to be under-transmitted in crossover products (Jeffreys and Neumann 2002), an asymmetry that favors the loss of recombination hotspots (Boulton et al. 1997). If this is a common phenomenon, it may lead hotspots to be short-lived on an evolutionary time scale (Jeffreys and Neumann 2002).

To evaluate whether fine-scale recombination rates can change rapidly, we were interested in comparing rates in humans with those in their closest evolutionary relative, the chimpanzee (*Pan troglodytes*). The two species are thought to have had a common ancestor five to six million years ago and differ at approximately 1.2% of base pairs on average (Ebersberger et al. 2002). Since it is difficult to use sperm typing techniques in chimpanzees, not least of all because of the need for chimpanzee sperm, we took an indirect approach and estimated the extent of recombination rate variation from patterns of LD in a population sample. To do so, we modified a recently developed statistical approach (Li and Stephens 2003). The method estimates recombination rates by exploiting the fact that patterns of LD reflect the rate and distribution of recombination events in the ancestors of the sample (see Materials and Methods for more details). Although it is based on simplistic assumptions about population demography, it has been shown to produce reliable estimates of recombination rates for data sets simulated under a range of demographic assumptions (Li and Stephens 2003; D. C. Crawford, T. Bhangale, N. Li, G. Hellenthal, M. J. Rieder, et al., unpublished data). We focused on the TAP2 genic region, where a sperm typing study of humans characterized a ~ 1.2 kb recombination hotspot in one of the introns (Jeffreys et al. 2000). Application of the statistical method to polymorphism data collected for this region (Jeffreys et al. 2000) led to estimates similar to those obtained by sperm typing, providing further evidence for its reliability (Li and Stephens 2003).

Samples that include individuals from diverged populations are expected to harbor high levels of LD that may lead to incorrect estimates of recombination rate variation (Pritchard and Przeworski 2001). This is of particular concern

in chimpanzees, for which previous studies have reported high levels of genetic differentiation between subspecies (Morin et al. 1994; Stone et al. 2002; Fischer et al. 2004). In addition, there appears to be a high proportion of less informative, rare alleles in samples from central (*P. t. troglodytes*) but not western (*P. t. verus*) chimpanzees (Gilad et al. 2003; Fischer et al. 2004). We therefore collected polymorphism data from a sample of 24 chimpanzees that were all known to be from the western subspecies. Strikingly, we found no evidence for recombination rate variation at TAP2 in these data.

Results

In humans, LD data for the TAP2 region were previously collected by Jeffreys et al. (2000), who resequenced ~ 9.7 kb in a sample of eight individuals from the United Kingdom (UK) and found 46 single nucleotide polymorphisms (SNPs), excluding insertion-deletions. The SNPs were then typed in a sample of 30 individuals from the UK, in whom haplotypes were determined experimentally (by allele-specific PCR). We collected genotype data for the same region in western chimpanzees by resequencing 24 individuals (see Materials and Methods for details). This led to the discovery of 57 SNPs. When differences in study design are taken into account, diversity levels in western chimpanzees are very similar to those observed in samples of humans from the UK ($\theta_W = 0.145\%$ versus $\theta_W = 0.144\%$ per bp, respectively), consistent with previous findings (e.g., Gilad et al. 2003; Fischer et al. 2004).

The LD data are summarized in Figure 1; overall, there is much less LD in humans than in chimpanzees. In particular, in humans, strong allelic associations are only seen between pairs of sites in close physical proximity, while in chimpanzees, such associations are also found among more distant pairs. Whether this reflects differences in the underlying recombination landscape or chance variation is unclear from visual inspection of these plots alone. We therefore used a statistical approach to assess the evidence for recombination rate variation in the two species. Specifically, we assumed that there is (at most) one hotspot in the region and, as a first step, specified its location according to the results of the sperm typing study in humans. We then applied our modification of the method of Li and Stephens (2003) to estimate a background population recombination rate, ρ , and the relative intensity of recombination in the hotspot segment, λ (see Materials and Methods). Within this model, a λ value of 1 corresponds to an absence of recombination rate variation, while values of λ greater than 1 indicate a hotspot. The approach taken here is Bayesian (see Materials and Methods) so, as a measure of support for a hotspot in the LD data, we report estimates for the probabilities $\Pr(\lambda > 1)$ and $\Pr(\lambda > 10)$; these are the posterior probabilities of a hotspot of any kind and of a hotspot of intensity at least ten times the background rate, respectively.

Application of this method to the human haplotype data led to extremely strong support for rate variation: we estimated $\Pr(\lambda > 1) = 1$ and $\Pr(\lambda > 10) = 0.982$. When the same method was applied to the human genotype data (i.e., ignoring the information about the phase of multiple heterozygotes), we estimated $\Pr(\lambda > 1) = 1$ and $\Pr(\lambda > 10) = 0.992$. The results are almost identical, suggesting minimal

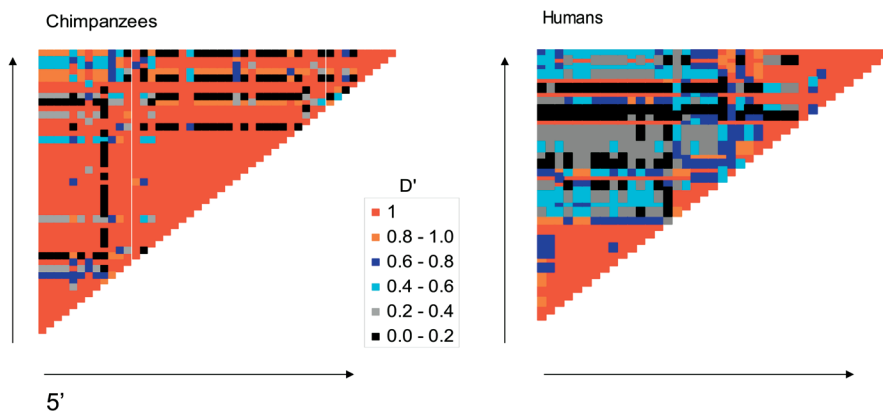


Figure 1. Patterns of Pairwise LD in Humans and Chimpanzees

Only SNPs with minor allele frequencies above 0.1 are included. The rows correspond to the consecutive SNPs in the region, as do the columns. Each cell indicates the extent of LD between a pair of sites, as measured by $|D'|$ (estimated using the Expectation Maximization algorithm, as implemented by Arlequin: <http://lgb.unige.ch/arlequin/>). DOI: 10.1371/journal.pbio.0020155.g001

loss of information with the use of genotypes. Interestingly, the point estimate of λ using either haplotypes (28.4) or genotypes (32.1) is higher than the corresponding estimate from sperm typing (11). This difference may reflect error in the estimates; alternatively, it may point to a more intense hotspot in females than in males (Jeffreys et al. 2000).

Next, we applied the same method to the genotype data collected from western chimpanzees. The estimate of the background rate of recombination, $\hat{\rho}$, is 5.0×10^{-4} per base pair, which is very similar to the estimate from the human genotype data (Figure 2). However, in contrast to what is found in humans, there is no evidence for recombination rate variation: our estimate of λ is 1, suggesting a uniform rate of recombination throughout the region, and our estimates of $\Pr(\lambda > 1) = 0.200$ and $\Pr(\lambda > 10) = 0.006$, reflecting tepid support for a hotspot of any kind and almost no support for a hotspot similar to the one observed in humans. Indeed, the latter figure represents very strong evidence *against* a hotspot of moderate intensity and rules out the possibility that the chimpanzee polymorphism data are simply uninformative, because of, for example, insufficient sample size or diversity.

To assess how likely we would be to obtain such weak support if there were in fact a hotspot in western

chimpanzees similar to the one in humans, we generated 200 simulated genotype data sets under a model with a hotspot of intensity $\lambda = 11$ and then tabulated the proportion with posterior probability estimates as low or lower than that observed (see Materials and Methods). We took the λ value estimated from sperm typing because it is the lowest of the various estimates for humans and hence its use was conservative for our purposes. With the ρ value estimated from the data (5.0×10^{-4} per bp), the probability of obtaining $\Pr(\lambda > 1) \leq 0.200$ is $p = 0.010$ and the probability of obtaining $\Pr(\lambda > 10) \leq 0.006$ is $p = 0.005$. With a lower ρ value (2.7×10^{-4} per bp; see Materials and Methods), the probability of obtaining $\Pr(\lambda > 1) \leq 0.200$ is $p = 0.020$. In other words, we can reject the null hypothesis that there is a hotspot in western chimpanzees similar to the one in humans, because we would expect to see more support for a hotspot in these data if one were there. It appears that western chimpanzees do not harbor a hotspot in the same location as humans.

The possibility remains, however, that there is a hotspot in a slightly different position in chimpanzees. To evaluate this, we used a more general model in which there is at most one hotspot in the region, but where the location is unknown and estimated together with ρ and λ (see Materials and Methods). Again, we found very little evidence for recombination rate variation: across all pairs of consecutive segregating sites, the largest posterior probability of elevated recombination is estimated to be < 0.060 (Figure 3). Thus, the hotspot appears to be entirely absent from the ~ 9.4 kb surveyed in western chimpanzees.

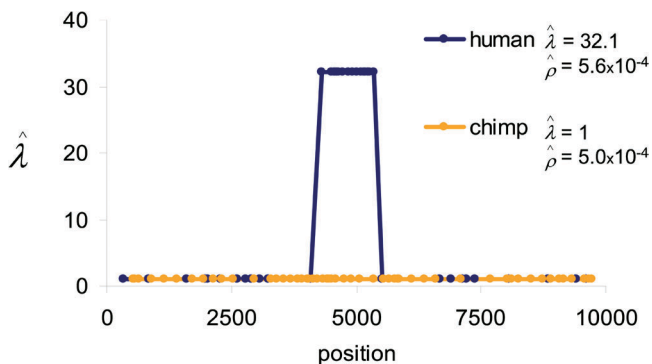


Figure 2. Estimates of the Recombination Hotspot Intensity, λ , Based on Genotype Data

We assumed that, if the hotspot is present, it is in the same location as estimated by sperm typing in humans (see Materials and Methods). A λ value of one corresponds to the absence of recombination rate variation, while values of λ greater than one indicate a hotspot. The estimates for humans from the UK are shown in blue and those for western chimpanzees in orange.

DOI: 10.1371/journal.pbio.0020155.g002

Discussion

These estimates of recombination rate parameters are based on assumptions of neutrality, constant population size, and random mating, raising the concern that the hotspot is not absent but instead masked by departures from model assumptions. However, we chose to focus on western chimpanzees precisely because previous studies reported allele frequencies in rough accordance with the assumptions of our model. Consistent with these studies (Gilad et al. 2003; Fischer et al. 2004), the allele frequencies at TAP2 are not significantly different from the expectations of the standard neutral model (as assessed by Tajima's $D = 0.848$, $p = 0.237$; see Materials and Methods). Moreover, simulations suggest that the power to detect a hotspot is not strongly affected by

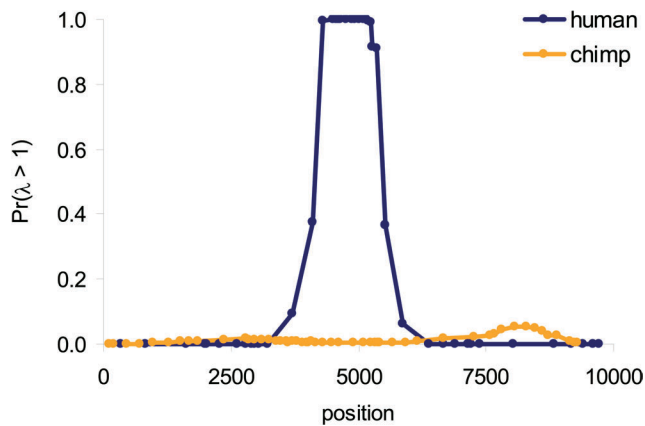


Figure 3. Estimates of Recombination Rate Variation in Humans and Western Chimpanzees

In this model, there is at most one hotspot in the region, the location and width of which are unknown and estimated along with λ and ρ . On the y-axis is an estimate of the posterior probability of elevated recombination, $\Pr(\lambda > 1)$, between each pair of consecutive SNPs (plotted at the midpoint position).
DOI: 10.1371/journal.pbio.0020155.g003

population history (Li and Stephens 2003). To some extent, this is expected, as population history tends to affect LD in the entire region, not only in the hotspot, so that estimates of the relative rates of recombination are unlikely to be substantially altered. In summary, there is no evidence for a marked departure from model assumptions in the allele frequencies, and the method is expected to be robust to small departures. Consistent with this, in humans, the approach yields similar results to sperm typing experiments that do not rely on the same assumptions. On this basis, it seems that the hotspot is truly absent from the homologous region in western chimpanzees.

This finding implies that the hotspot was lost in chimpanzees or gained in humans, or that it moved in one of the species (over a larger distance than we surveyed). This in turn raises a number of more general questions. Are hotspots frequently born de novo or do they tend to migrate within circumscribed regions of the genome? Are particular sequence motifs sufficient to produce recombination hotspots, or are larger-scale requirements, such as chromatin accessibility, required for their formation (Petes 2001)? The systematic comparison between closely related species with different recombination landscapes may be helpful in addressing these problems. As an illustration, in these data, we found two motifs that were previously implicated in the formation of recombination hotspots (Smith et al. 1998; Badge et al. 2000 and references therein) and that varied between the two species: a Pur binding motif that is present in humans but absent in chimpanzees (because of a single base pair difference) and two scaffold attachment sites that are in different positions in the two species. The significance of these differences cannot be determined on the basis of a single example; however, once a larger sample of hotspot regions has been surveyed, one can begin to test for an association between particular sequence motifs or features and the presence of hotspots.

Comparative studies of hotspot regions will also increase our understanding of the determinants of mutation rates. As

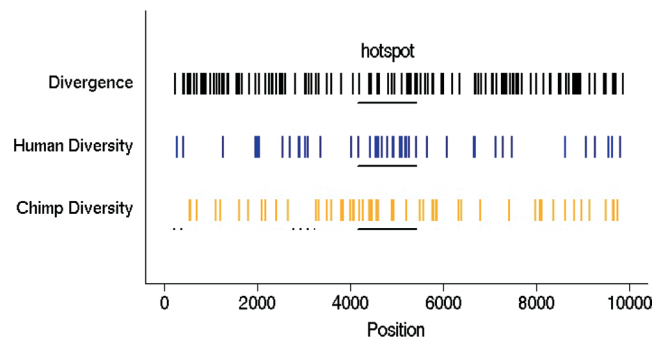


Figure 4. Distribution of Variable Sites in the Genomic Region

The positions of sites that differ between humans and chimpanzees are shown on the first line, while the positions of sites polymorphic in humans from the UK or in western chimpanzees are shown on the next two lines. The human hotspot region is underlined. The dashed lines indicate regions not surveyed for variation in western chimpanzees (see Materials and Methods).
DOI: 10.1371/journal.pbio.0020155.g004

noted by Jeffreys et al. (2000), there is a significant excess of diversity within the hotspot region in humans from the UK (Figure 4): when the hotspot region is compared to the 8,735 other windows of the same size, only 0.3% have as many or more SNPs. In contrast, in western chimpanzees, levels of diversity are not higher than elsewhere in the region (Figure 4): 17.0% of comparable windows harbor at least as many SNPs as the hotspot. Nor are levels of human–chimpanzee divergence unusual in the hotspot region: 67.3% of windows show the same or higher numbers of fixed differences between species (Figure 4). Given the evidence for a recombination hotspot in humans but not in chimpanzees, these observations are consistent with an association between recombination and mutation in primates (Hellmann et al. 2003) and, in particular, with a mutagenic effect of recombination (Ratray et al. 2002). If indeed recombination events introduce mutations, the lack of a peak of human–chimpanzee divergence in the hotspot region (Jeffreys et al. 2000; Figure 4) would suggest that the hotspot arose fairly recently in human evolution.

In conclusion, these analyses demonstrate that fine-scale recombination rates can change between closely related species. Together with the observations that crossover frequencies can depend on specific haplotypes (Jeffreys and Neumann 2002) and that large-scale recombination rates differ among individuals (Cullen et al. 2002; Kong et al. 2002), this finding raises the possibility that local rates can vary among human groups that differ in their allele frequencies. Unfortunately, demonstrating compelling evidence for variation among human populations on the basis of LD data alone promises to be substantially harder than demonstrating such differences between chimpanzees and humans. In particular, human populations share most of their evolutionary history, making differences between extant populations, if they exist, more difficult to detect. Nevertheless, LD studies should be helpful in identifying interesting regions for further study via sperm typing.

The extent to which local recombination rates vary among human populations influences the degree of similarity of LD patterns among them, with important consequences for the design of efficient LD-based association studies (including, for example, the choice of appropriate “haplotype tagging

SNPs” [Johnson et al. 2001] in different human populations) and for the relevance of data generated by the current human HapMap project to populations not currently represented in that study (International Hapmap Consortium 2003). Perhaps most importantly, if local recombination rates do vary among groups, then the study of regions with the most pronounced differences should lead to further insights into the underlying biological processes that cause fine-scale variation in recombination rates.

Materials and Methods

Samples. We used DNA from 24 western chimpanzees (*Pan troglodytes verus*) that were wild caught or known to be unrelated based on recent pedigrees. Twelve samples (Annaclara, Frits, Hilko, Liesbeth, Louise, Marco, Oscar, Regina, Socrates, Sonja, Yoran, and Yvonne) are from the collection stored at the Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany, while 12 other samples (NDH0311G1, NDH0312G1, NDH0313G1, NDH0314G1, NDH0317G1, NDH0320G1, NDH0321G1, NDH0322G1, NDH0325G1, NDH0326G1, NDH0328G1, and NDH0329G1) were kindly provided by P. Morin and the Primate Foundation of Arizona.

Primer design. We amplified 9,491 bp from the TAP2 region, corresponding to base pairs 113102–122585 of the sequence from Beck et al. (1996) (see Supporting Information); the slight discrepancy in the number of base pairs is due to indels. To minimize the chance of allelic dropout, we designed the PCR primers such that most of the sequence would be amplified by two independent sets of primers. The 20 overlapping primer sequences are listed in Protocol S1.

PCR and DNA sequencing. DNA amplification reactions contained 250 μ M of each dNTP, 1–2 mM $MgCl_2$, PCR buffer (10 mM Tris-HCl, 50 mM KCl; pH 8.3), 0.5 U of Taq DNA polymerase (all reagents from Roche, Basel, Switzerland), and 10 pmol of each primer. We used 50–100 ng of DNA in each 30 μ l PCR. Amplification conditions for all regions were the following: incubation for 3 min at 94 $^{\circ}C$, 35 cycles (45 s at 94 $^{\circ}C$, 1 min at 45–62 $^{\circ}C$, and 1 min at 72 $^{\circ}C$) and a final elongation of 5 min at 72 $^{\circ}C$. A nested PCR was performed to obtain regions 6 and 7 by using the product of the primers Tap2 5' and Tap2 8 3' as a template. PCR products were separated from primers and unincorporated dNTPs by treatment with a solution of 10% PEG 8000/1.25 M NaCl followed by centrifugation. PCR products were then air dried and resuspended in 10–15 μ l of H_2O .

Sequencing reactions consisted of 1 μ l of ABI Prism BigDyeTM Terminators version 2.0 (Perkin Elmer Biosystems, Torrance, California, United States), 8–10 ng of purified PCR product, and 1 μ l of 2.5 μ M primer (the same primers used for PCR) in a volume of 7 μ l. Cycling conditions were 96 $^{\circ}C$ for 2 min and then 35 cycles of 96 $^{\circ}C$ (20 s), annealing temperature (30 s), and 60 $^{\circ}C$ (4 min). Isopropanol-precipitated cycle sequencing products were run on an ABI 3730 DNA analyzer. Base calling was done with ABI Prism DNA Sequencing Analysis version 5.0 and ABI Basecaller. BioEdit version 5.0.6 was used for sequence analysis and alignment.

In total, 2-fold coverage of a 9,370 bp sequence was obtained for each individual; these are available from GenBank (see Supporting Information). Most of the region was sequenced from both DNA strands. However, due to the presence of insertions, deletions, and T or A stretches, this was not possible for a subset of segments; for these, 2-fold coverage was achieved by sequencing the same strand. For segment 6, we did not obtain reliable sequence data for all individuals (due to suspected allelic dropout); we therefore excluded this region of 487 bp. Otherwise, there are no missing data. SNPs were identified by visualization of the chromatograms using BioEdit version 5.0.6. The polymorphism data used for the analyses are available in Protocol S1.

Data analysis. We estimated the population mutation rate, $\theta = 4N_e\mu$ (N_e is the diploid effective population size and μ is the mutation rate per generation), using Watterson's estimator, θ_w (Watterson 1975), based on the number of segregating sites in the sample. We also calculated a commonly used summary of the allele frequency spectrum, Tajima's D (Tajima 1989); both D and θ_w were calculated with DNAsp (Rozas and Rozas 1999). We used the D statistic to test the fit of the standard neutral model (of a random mating population of constant size) to allele frequencies in western chimpanzees. Specifically, we ran 10^4 coalescent simulations of the standard neutral model with the same number of chromosomes and base pairs as in the actual data, with θ equal to θ_w , and with the population recombi-

nation rate equal to the estimated value (see below). We then tabulated the proportion of simulated runs with a Tajima's D value as or more extreme than that observed.

We calculated the GC content of the region and searched for sequence motifs previously associated with recombination hotspots (Badge et al. 2000; Petes 2001; Wall et al. 2003) using the program “scan_for_matches” available from http://bioweb.pasteur.fr/seqanal/interfaces/scan_for_matches.html. The list of motifs found in the human and chimpanzee sequences is given in Protocol S1.

Analyses of LD. To assess the support in the polymorphism data for a recombination hotspot, we used the Product of Approximate Conditionals (PAC) model of Li and Stephens (2003). Assuming haplotypes are known, the method considers each one in turn and attempts to represent it as a mosaic of the previously considered haplotypes. Qualitatively, the larger the regions over which haplotypes tend to resemble one another, the fewer the pieces required in each mosaic, and the lower the estimates of the recombination rates. The method uses simplistic assumptions about population demography to quantify this qualitative relationship and hence to estimate recombination rates across the region.

More formally, the model of Li and Stephens (2003) defines the probability of observing haplotypes H given the underlying recombination parameters α (which in our case may include the background recombination rate and the hotspot location and intensity; see below). This can be used directly to estimate α from H in situations where haplotypes have been experimentally determined (e.g., Li and Stephens 2003). However, in our case the chimpanzee haplotypes are not known. Rather, we have genotype data G and we wish to estimate α from G . A simple approximate solution to this would be first to use a statistical method (e.g., that of Stephens et al. 2001) to obtain an estimate \hat{H} for the haplotypes H from the genotypes G , and then to estimate α from \hat{H} . However, a risk of this approach is that overconfident conclusions will be drawn by ignoring uncertainty in the estimated haplotypes. A better solution, and the approach we take here, is to jointly estimate H and α from G , or, more specifically, to obtain a sample from the joint posterior distribution, $\Pr(H, \alpha | G)$. To do so, we start with an initial guess for the haplotypes, and iterate the following steps: (i) estimate a new value for α , using the current estimate of H and (ii) estimate a new value for H , using the genotypes G and the current value for α . Step (i) is performed using the PAC-B model of Li and Stephens (2003) and the priors on α described below. Step (ii) is performed by using the method for haplotype inference described in Stephens and Donnelly (2003), but replacing the conditional distribution that they use (which ignores recombination) with the conditional distribution of Fearnhead and Donnelly (2001) (which takes into account recombination) computed using two quadrature points. (Actually, we modified the Fearnhead and Donnelly conditional distribution slightly, replacing the equation $q_i = z_i p / (j + z_i p)$ in their Appendix A with $q_i = 1 - \exp(-z_i p / j)$.) Both the PAC-B model and the Fearnhead and Donnelly conditional require the specification of a mutation parameter, θ , and a mutation process. In each case, we used the value of θ given in Li and Stephens (2003) and a mutation process whereby each mutation event at a biallelic site results in a change from one allele to the other.

This iterative scheme defines a Markov chain whose stationary distribution is the distribution $\Pr(H, \alpha | G)$ from which we wish to sample. Provided that the algorithm is run for sufficiently long, the estimates of α obtained each iteration provide a sample from the distribution $\Pr(\alpha | G)$, and thus allow α (i.e., the underlying recombination process) to be estimated directly from G , taking full account of the fact that the actual underlying haplotypes are not known. The algorithm is implemented within the software package PHASE version 2.1, which is available online at <http://www.stat.washington.edu/stephens/software.html>.

We considered two versions of the simple hotspot model of Li and Stephens (2003). In this model, there is a single hotspot of constant intensity λ . Crossovers occur as a Poisson process (i.e., there is no interference) of constant rate r (per base pair) outside the hotspot and of constant rate λr inside the hotspot; gene conversion is not explicitly modeled. In the first version, we assumed that, if present, the hotspot is at the same location as estimated by sperm typing in humans (4180–5417). (This location is not precisely the same as the one used by Li and Stephens [2003], which is why our estimates differ from theirs.) There are two parameters to be estimated: the background population recombination parameter $\rho = 4N_e r$, where N_e is the effective population size) and λ . We assumed a priori that a hotspot exists with probability 0.5 and that, if the hotspot exists, λ is between one and 100. Specifically, we assumed that $\lambda = 1$ with probability 0.5 and otherwise that $\log_{10}(\lambda)$ is uniformly distributed on

(0, 2). The prior on ρ is uniform on $\log_{10}(\rho)$ in the range $(-8, 3)$, which covers all plausible values.

In the second version, we assumed that the location and width of the hotspot are unknown and to be estimated along with λ and ρ . In this case, we assumed a priori that the hotspot exists with probability 0.18 (corresponding to an assumption that a hotspot occurs roughly once per 50 kb of sequence), that the center of the hotspot is equally likely to be anywhere along the length of the sequence, and that the width of the hotspot is between 200 and $\sim 4,000$ bp (specifically, we assumed that the width had a normal distribution, with a mean of 0 bp and a standard deviation of 2,000 bp, truncated to lie above 200 bp). Priors on ρ and on λ (conditional on there being a hotspot) are as in the first version.

To allow for potential problems with convergence of this Markov chain Monte Carlo algorithm, we ran the algorithm ten times for each analysis, using different seeds for the pseudorandom number generator. For each run, we obtained a point estimate of the parameters (using sample posterior medians) and posterior probabilities. The reported estimates are the median of the estimates obtained from the ten runs.

To test how likely we would be to obtain such weak support for a hotspot in the LD data if there were in fact a hotspot similar to the one in humans, we ran 200 coalescent simulations of the standard neutral model (Hudson 1990) with the same number of base pairs and sample size as the actual data (48 chromosomes), a hotspot of intensity $\lambda = 11$, and $\theta = \theta_w$. Haplotypes were randomly paired to form genotypes and phase information was ignored. The data were masked to mimic the actual data structure, i.e., they included a gap of 487 bp in the same position. We then counted the proportion of simulated data sets for which our estimate of the posterior probability was as low as observed or lower (using the first version of the Li and Stephens [2003] model). Since we obtained estimates for the simulated data in the same way as for the actual data, significance values obtained from this analysis are valid independent of the convergence, or even the correctness, of the Markov chain Monte Carlo scheme. In the first set of 200 simulations, we used $\rho = \hat{\rho}$, the background rate that we estimated from the western chimpanzee data. In the second set of simulations, we used $\rho = 4\hat{N}_e\hat{r} = 2.7 \times 10^{-4}$

per bp, where $\hat{N}_e = 17,100$ is an estimate of the effective population size of western chimpanzees (based on Fischer et al. 2004) and $\hat{r} = 0.4$ cM/Mb is the rough estimate of the background recombination rate reported in Jeffreys et al. (2000).

Supporting Information

Protocol S1. Supplementary Materials

Information on the primers used in this study, the polymorphism data used for the analyses, and the search for sequence motifs previously associated with recombination hotspots.

Found at DOI: 10.1371/journal.pbio.0020155.sd001 (91 KB DOC).

Accession Numbers

The GenBank (<http://www.ncbi.nlm.nih.gov/>) accession number for the human TAP2 region of Beck et al. (1996) is X87344. The numbers for the 9,370-bp sequences obtained from the 24 western chimpanzees are AY559252–AY559299.

Acknowledgments

We are grateful to Phil Morin and the Arizona primate facility for providing DNA samples from 12 chimpanzees, to Alec Jeffreys and Rita Neumann for sending us a list of TAP2 primers and conditions, to Jeff Wall for providing us with a computer program, to Anna Di Rienzo and Jeff Wall for helpful discussions, and to the Max Planck sequencing unit. Support for this work was provided by the Max Planck Society and Deutsche Forschungsgemeinschaft grant BIZ6-1/1. MS was supported by National Institutes of Health grant number 1R01HG/LM02585-01.

Conflicts of interest. The authors have declared that no conflicts of interest exist.

Author contributions. MP conceived and designed the experiments. ADR and YG performed the experiments. SEP, MS, and MP analyzed the data. MS, SP, and MP contributed reagents/materials/analysis tools. All authors contributed to the writing of the paper. ■

References

- Arnheim N, Calabrese P, Nordborg M (2003) Hot and cold spots of recombination in the human genome: The reason we should find them and how this can be achieved. *Am J Hum Genet* 73: 5–16.
- Ashburner M (1989) *Drosophila: A laboratory handbook*. Cold Spring Harbor (New York): Cold Spring Harbor Laboratory Press. 1331 p.
- Badge RM, Yardley J, Jeffreys AJ, Armour JA (2000) Crossover breakpoint mapping identifies a subtelomeric hotspot for male meiotic recombination. *Hum Mol Genet* 9: 1239–1244.
- Beck S, Abdulla S, Alderton RP, Glynn RJ, Gut IG, et al. (1996) Evolutionary dynamics of non-coding sequences within the class II region of the human MHC. *J Mol Biol* 255: 1–13.
- Betancourt AJ, Presgraves DC (2002) Linkage limits the power of natural selection in *Drosophila*. *Proc Natl Acad Sci U S A* 99: 13616–13620.
- Boulton A, Myers RS, Redfield RJ (1997) The hotspot conversion paradox and the evolution of meiotic recombination. *Proc Natl Acad Sci U S A* 94: 8058–8063.
- Brooks LD (1988) The evolution of recombination rates. In: Michod RE, Levin BR, editors. *The evolution of sex: An examination of current ideas*. Sunderland, Massachusetts: Sinauer Associates. pp. 87–105.
- Cullen M, Perfetto SP, Klitz W, Nelson G, Carrington M (2002) High-resolution patterns of meiotic recombination across the human major histocompatibility complex. *Am J Hum Genet* 71: 759–776.
- Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. *Nat Genet* 29: 229–232.
- de Massy B (2003) Distribution of meiotic recombination sites. *Trends Genet* 19: 514–522.
- Ebersberger I, Metzler D, Schwarz C, Paabo S (2002) Genomewide comparison of DNA sequences between humans and chimpanzees. *Am J Hum Genet* 70: 1490–1497.
- Fearnhead P, Donnelly P (2001) Estimating recombination rates from population genetic data. *Genetics* 159: 1299–1318.
- Fischer A, Wiebe V, Paabo S, Przeworski M (2004) Evidence for a complex demographic history of chimpanzees. *Mol Biol Evol*. In press.
- Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, et al. (2002) The structure of haplotype blocks in the human genome. *Science* 296: 2225–2229.
- Gilad Y, Bustamante C, Lander D, Paabo S (2003) Natural selection on the olfactory receptor gene family in humans and chimpanzees. *Am J Hum Genet* 73: 489–501.
- Hellmann I, Ebersberger I, Ptak S, Paabo S, Przeworski M (2003) A neutral explanation for the correlation of diversity with recombination in humans. *Am J Hum Genet* 72: 1527–1535.
- Hubert R, MacDonald M, Gusella J, Arnheim N (1994) High resolution localization of recombination hot spots using sperm typing. *Nat Genet* 7: 420–424.
- Hudson RR (1987) Estimating the recombination parameter of a finite population model without selection. *Genet Res* 50: 245–250.
- Hudson RR (1990) Gene genealogies and the coalescent process. In: Futuyma D, Antonovics J, editors. *Oxford surveys in evolutionary biology*, Volume 1. Oxford: Oxford University Press. pp. 1–44.
- International Hapmap Consortium (2003) The International Hapmap project. *Nature* 426: 789–796.
- Jeffreys AJ, Neumann R (2002) Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat Genet* 31: 267–271.
- Jeffreys AJ, Murray J, Neumann R (1998) High-resolution mapping of crossovers in human sperm defines a minisatellite-associated recombination hotspot. *Mol Cell* 2: 267–273.
- Jeffreys AJ, Ritchie A, Neumann R (2000) High resolution analysis of haplotype diversity and meiotic crossover in the human TAP2 recombination hotspot. *Hum Mol Genet* 9: 725–733.
- Jeffreys AJ, Kauppi L, Neumann R (2001) Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nat Genet* 29: 217–222.
- Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, et al. (2001) Haplotype tagging for the identification of common disease genes. *Nat Genet* 29: 233–237.
- Kong A, Gudbjartsson DF, Sainz J, Jonsson GM, Gudjonsson SA, et al. (2002) A high-resolution recombination map of the human genome. *Nat Genet* 31: 241–247.
- Li HH, Gyllenstein UB, Cui XF, Saiki RK, Erlich HA, et al. (1988) Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* 335: 414–417.
- Li N, Stephens M (2003) Modelling linkage disequilibrium, and identifying recombination hotspots using SNP data. *Genetics* 165: 2213–2233.
- Morin PA, Moore JJ, Chakraborty R, Jin L, Goodall J, et al. (1994) Kin selection, social structure, gene flow, and the evolution of chimpanzees. *Science* 265: 1193–1201.
- Nachman MW (2002) Variation in recombination rate across the genome: Evidence and implications. *Curr Opin Genet Dev* 12: 657–663.
- Petes TD (2001) Meiotic recombination hot spots and cold spots. *Nat Rev Genet* 2: 360–369.



- Pritchard JK, Przeworski M (2001) Linkage disequilibrium in humans: Models and data. *Am J Hum Genet* 69: 1–14.
- Ratray AJ, Shafer BK, McGill CB, Strathern JN (2002) The roles of Rev3 and Rad57 in double-strand break repair induced mutagenesis in *Saccharomyces cerevisiae*. *Genetics* 162: 1063–1077.
- Rogers J, Mahaney MC, Witte SM, Nair S, Newman D, et al. (2000) A genetic linkage map of the baboon (*Papio hamadryas*) genome based on human microsatellite polymorphisms. *Genomics* 67: 237–247.
- Rozas J, Rozas R (1999) DnaSP version 3: An integrated program for molecular population genetics and molecular evolution analysis. *Bioinformatics* 15: 174–175.
- Schneider JA, Peto TEA, Boone RA, Boyce AJ, Clegg JB (2002) Direct measurement of the male recombination fraction in the human β -globin hot spot. *Hum Mol Genet* 11: 207–215.
- Smith RA, Joy Ho P, Clegg JB, Kidd JR, Thein SL (1998) Recombination breakpoints in the human β -globin gene cluster. *Blood* 11: 4415–4421.
- Steiner WW, Schreckhise RW, Smith GR (2002) Meiotic DNA breaks at the *S. pombe* recombination hot spot M26. *Mol Cell* 9: 847–855.
- Stephens M, Smith NJ, Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68: 978–989.
- Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73: 1162–1169.
- Stone AC, Griffiths RC, Zegura SL, Hammer MF (2002) High levels of Y-chromosome nucleotide diversity in the genus *Pan*. *Proc Natl Acad Sci U S A* 99: 43–48.
- Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.
- Thomas JW, Touchman JW, Blakesley RW, Bouffard GC, Beckstrom-Sternberg SM, et al. (2003) Comparative analyses of multi-species sequences from targeted genomic regions. *Nature* 424: 788–793.
- True JR, Mercer JM, Laurie CC (1996) Differences in crossover frequency and distribution among three sibling species of *Drosophila*. *Genetics* 142: 507–523.
- Wall JD, Pritchard JK (2003) Haplotype blocks and the structure of linkage disequilibrium in the human genome. *Nat Rev Genet* 4: 587–597.
- Wall JD, Frisse LA, Hudson RR, Rienzo AD (2003) Comparative linkage disequilibrium analysis of the β -globin hotspot in primates. *Am J Hum Genet* 73: 1330–1340.
- Watterson GA (1975) On the number of segregating sites in genetic models without recombination. *Theor Popul Biol* 7: 256–276.
- Yauk CL, Bois PR, Jeffreys AJ (2003) High-resolution sperm typing of meiotic recombination in the mouse MHC E_{β} gene. *EMBO J* 22: 1389–1397.