Letter

Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion

Timothy R. Dreszer,[1] Gregory D. Wall,[2] David Haussler,[1,3,5] and Katherine S. Pollard[2,4,5]

[1]*Department of Biomolecular Engineering, University of California, Santa Cruz, California 95064, USA;* [2]*Department of Statistics, University of California, Davis, California 95616, USA;* [3]*Howard Hughes Medical Institute, University of California, Santa Cruz, California 95064, USA;* [4]*UC Davis Genome Center, University of California, Davis, California 95616, USA*

We examined fixed substitutions in the human lineage since divergence from the common ancestor with the chimpanzee, and determined what fraction are AT to GC (weak-to-strong). Substitutions that are densely clustered on the chromosomes show a remarkable excess of weak-to-strong "biased" substitutions. These unexpected biased clustered substitutions (UBCS) are common near the telomeres of all autosomes but not the sex chromosomes. Regions of extreme bias are enriched for genes. Human and chimp orthologous regions show a striking similarity in the shape and magnitude of their respective UBCS maps, suggesting a relatively stable force leads to clustered bias. The strong and stable signal near telomeres may have participated in the evolution of isochores. One exception to the UBCS pattern found in all autosomes is chromosome 2, which shows a UBCS peak midchromosome, mapping to the fusion site of two ancestral chromosomes. This provides evidence that the fusion occurred as recently as 740,000 years ago and no more than ~3 million years ago. No biased clustering was found in SNPs, suggesting that clusters of biased substitutions are selected from mutations. UBCS is strongly correlated with male (and not female) recombination rates, which explains the lack of UBCS signal on chromosome X. These observations support the hypothesis that biased gene conversion (BGC), specifically in the male germline, played a significant role in the evolution of the human genome.

[Supplemental material is available online at www.genome.org.]

With the sequencing of the chimpanzee genome (Chimpanzee Sequencing and Analysis Consortium 2005), detailed analysis of the genetic determinants of our humanity has begun. In pursuit of the fastest evolving regions of the human genome, it has been observed that single-base substitutions in the top scoring regions were dramatically biased in favor of changes from AT to GC base pairs (Pollard et al. 2006a,b). In the top four fastest evolving regions, there were 33 cases of an AT pair being replaced by a GC pair, but only one case of a GC being replaced by an AT. Thus, bases that pair with two hydrogen bonds ("weak") were replaced by bases that pair with three ("strong"). This substitution bias is particularly surprising given that overall numbers of strong-to-weak and weak-to-strong mutations are roughly equal in the human genome (Eyre-Walker 1999; Maki 2002; Lipatov et al. 2006).

The force that has biased substitutions in recent human evolution may be related to the pressures that have shaped isochores, areas of warm-blooded vertebrate genomes as large as ~300 kb with strikingly greater or lesser proportions of G+C than surrounding areas (Bernardi et al. 1985; Bernardi 2000). Three main theories have been proposed to explain the existence of isochores (Eyre-Walker and Hurst 2001). The first involves variation in mutation rates (Sueoka 1988; Wolfe et al. 1989; Fryxell and Zuckerkandl 2000) in different areas of a genome. Under this model, initial mutations vary in the proportion of G+C at different locations, and this imbalance is carried forward as a substitution bias by neutral evolution. The second theory is that natural selection (Bernardi and Bernardi 1986; Eyre-Walker 1999) for GC alleles has driven the formation and maintenance of isochores. Variability in G+C content might be evolutionarily advantageous, since G+C% is known to be correlated with (Lercher et al. 2003) and positively affect (Kudla et al. 2006) gene expression. G+C% may also be related to thermal stability. A third model that could lead to widespread variation in G+C content involves biased gene conversion (BGC) (Eyre-Walker 1993). BGC is a recombination-driven process that results in the biased fixation of GC alleles due to a biochemical bias in the DNA repair mechanism (Brown and Jiricny 1988; Webster and Smith 2004) acting on short stretches of hetero-duplexed DNA during crossing-over. BGC acts much like a *selection pressure* (Nagylaki 1983), providing a general mechanism for increased G+C, although it remains to be demonstrated (theoretically or experimentally) how BGC could lead to very large increases in substitution rate. Correlation between recombination rates and G+C content in a number of species supports the BGC hypothesis (Brown and Jiricny 1988; Birdsell 2002; Meunier and Duret 2004).

The three models to explain bias in G+C give rise to different predictions regarding patterns of bias in polymorphisms and fixed substitutions (Eyre-Walker and Hurst 2001). If mutation bias is at work, then single nucleotide polymorphisms (SNPs) should show a similar pattern of bias as substitutions (Eyre-Walker and Hurst 2001; Lercher et al. 2003). But if higher G+C is the result of a selection pressure, there should be a more pronounced bias in substitutions than in SNPs (Lercher et al. 2002;

[5]Corresponding authors.
E-mail haussler@soe.ucsc.edu; fax (831) 459-1809.
E-mail kspollard@ucdavis.edu; fax (530) 754-9658.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.6395807. Freely available online through the *Genome Research* Open Access option.

17:1420–1430 ©2007 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/07; www.genome.org

Schmegner et al. 2007). Distinguishing between natural selection and BGC-mediated increases in G+C can be more difficult (Duret 2002; Meunier and Duret 2004). If natural selection is the source, then substitution bias may be different between areas of low and high conservation. However, if BGC is the cause of bias, then genomic G+C would be correlated with historical recombination rates or hot spots (Kong et al. 2002). To the extent that regional recombination rates remain constant (Myers et al. 2005), biased substitutions might be correlated with measures of current recombination.

This study has been undertaken to document the characteristics of biased substitution patterns across the human and chimp genomes. We examine nucleotide bias in human SNPs and fixed chimp–human differences in an effort to characterize the roles of mutation bias, natural selection, and BGC in the observed patterns of substitution bias in the last 6 million years of human evolution.

Results

For every chimp–human nucleotide difference and every human SNP, we determined the ancestral allele using the chimp and macaque genome sequences as described in the Methods. In each substitution where the chimp–human ancestral allele could be determined, we then inferred the lineage (human vs. chimp vs. both) and type of nucleotide change (weak-to-strong vs. strong-to-weak vs. neither). Similarly, we inferred the type of nucleotide change for each human SNP. Henceforth, we use the word "bias" to refer to "weak-to-strong bias" and describe all other types of bias explicitly, e.g., "strong-to-weak bias."

Clusters of substitutions are biased

Genome-wide there is no nucleotide bias in human substitutions. Nearly as many ancestral AT pairs underwent substitution as GC pairs (5,351,332 vs. 5,520,349). Further, weak-to-strong substitutions and strong-to-weak substitutions occur at similar rates, constituting 43.1% and 42.8% of all substitutions, respectively. Clusters of five or more nearby substitutions, however, do show evidence of weak-to-strong bias (Fig. 1). For example, in windows of 100 bp containing seven substitutions, weak-to-strong changes outnumber strong-to-weak changes genome-wide (49.2% vs. 35.8%). This relationship can be seen in every human autosome, although the strength of the relationship varies by chromosome. On GC-rich chromosome 19, strong-to-weak substitutions outnumber all other types (as expected), but clusters of substitutions are nonetheless biased weak-to-strong. This pattern cannot be explained by hypermutability of CpG dinucleotides, since our findings do not change qualitatively when using only sites for which the ancestral state can reliably be inferred to be non-CpG ("class 1 sites," Meunier and Duret 2004) (Supplemental Fig. S1).

Regions near telomeres of autosomes show greatest bias

To locate and quantify this pattern of observed clustered bias within genomic regions, we developed a measure of unexpected biased clustered substitutions (UBCS; see Methods). UBCS measures the excess bias among clustered substitutions in a region and is equal to zero when clustered and nonclustered substitutions have the same pattern of bias. Patterns of UBCS along the chromosomes show a significant increase in bias at the distal ends of all autosomes (Fig. 2). This pattern is not a function of distance from the centromere or length of chromosome, nor does it disappear when only non-CpG class 1 sites are analyzed (Supplemental Fig. S1).

Patterns of bias nearly identical in chimpanzee

Due to sequencing and assembly errors in the less-complete chimpanzee genome, we expect to overestimate chimp substitu-

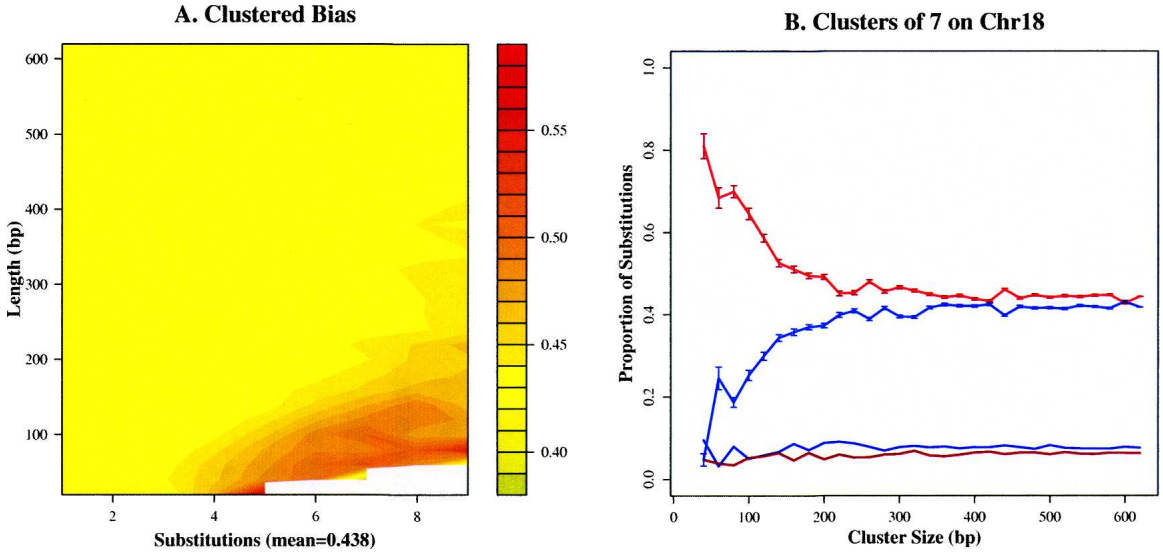

Figure 1. Clusters of substitutions are biased weak (AT) to strong (GC). (*A*) This heat map shows the proportion of substitutions that are weak-to-strong (color scale) as a function of the local substitution rate for a range of sizes and densities of substitutions. The overall proportion of substitutions that are weak-to-strong is 0.438 genome-wide (yellow). As substitution density approaches 4 or more within 100 bp, the proportion rises to 0.58 (red). (*B*) Chromosome 18 shows a particularly strong relationship between substitution density and bias, as can be seen for clusters of seven substitutions in the "zipper" plot. A similar pattern was observed on all autosomes. Additional graphics are available at http://www.soe.ucsc.edu/research/compbio/ubcs/.

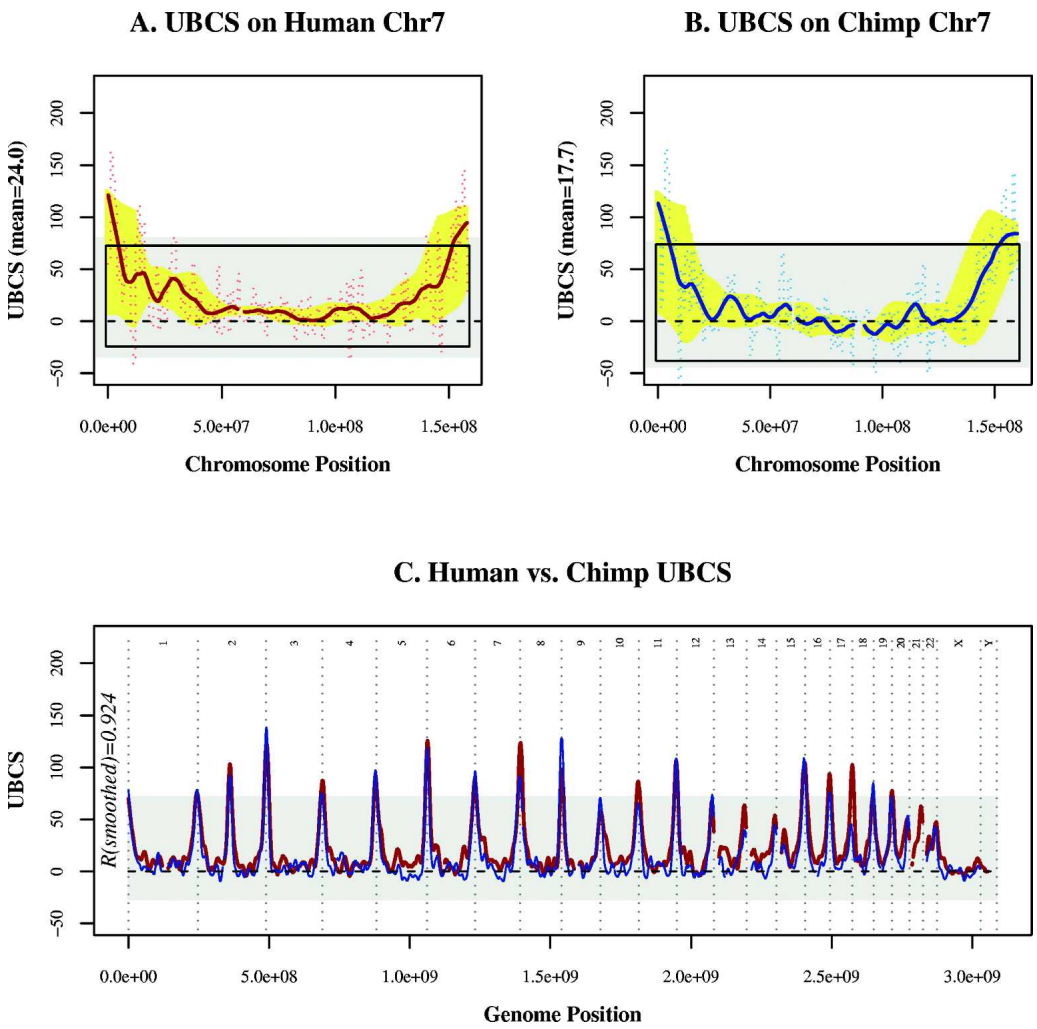

Figure 2. Patterns of substitution bias are nearly identical in human and chimp. (*A*) Unexpected biased clustered substitutions (UBCS) (faint line) for human chromosome 7 is above zero, indicating GC bias, along most (91.1%) of the chromosome and rises significantly at the distal ends. Smoothed UBCS (dark line) and 95% confidence band (yellow) are shown. The 95% confidence region is above the null expectation (zero) for more than half of the chromosome (61.8%). UBCS only exceeds the genome-wide 95% confidence interval (gray) near the telomeres. (*B*) Chimpanzee chromosome 7 has a remarkably similar profile (Spearman correlation $\rho = 0.87$). (*C*) The pattern of bias on chromosome 7 is mirrored on all autosomes (ordered sequentially; red, human; blue, chimp). Elevated UBCS near telomeres exceeds the human genome-wide 95% confidence interval (gray) on almost all autosomes. Here the chimpanzee sequence has been aligned to the human genome.

tion rates of all types. Our use of only high-quality chimp base calls should reduce this effect, but some residual overestimation may exist in our data. Nonetheless, we observe in the substitutions specific to the chimpanzee lineage the same pattern of UCBS as found in the human-specific substitutions in the human genome, indicating that our findings are not specific to the human lineage. In fact, there is a striking similarity between the human and chimp UBCS profiles along each chromosome, despite the fact that (by definition) no substitutions are shared between the two genomes (Fig. 2). The human and chimp smoothed UBCS values are highly correlated genome-wide (Spearman correlation: $\rho = 0.70$, permutation $P \approx 0$) and for each chromosome (Supplemental Table S3). This correlation is much lower ($\rho = 0.44$ genome-wide), though still significant, if the distal 16 Mb of each chromosome arm are dropped from the analysis. Analyzing only every 100th window along the genome in order to reduce possible effects of autocorrelation between adjacent regions does not change the correlation between human and chimp UBCS nor its statistical significance. Rough (unsmoothed) UBCS values are less highly correlated but are still significantly similar between the two species (Supplemental Table S3). This suggests that the correlation between chimp and human UBCS is regional rather than fine-scale. Thus, the force that is responsible for creating biased clusters of substitutions is clearly location-dependent and evolutionarily conserved.

Bias absent from chromosomes X and Y

The human sex chromosomes do not share the consistent pattern of bias found on all the autosomes (Fig. 3). On human chromosome Y, UBCS varies randomly and does not reach significant levels. This observation is consistent with the BGC model, which predicts no bias in the absence of recombination. Data for chimpanzee chromosome Y was insufficient for analysis of UBCS. Surprisingly, despite the presence of substantial recombination (al-

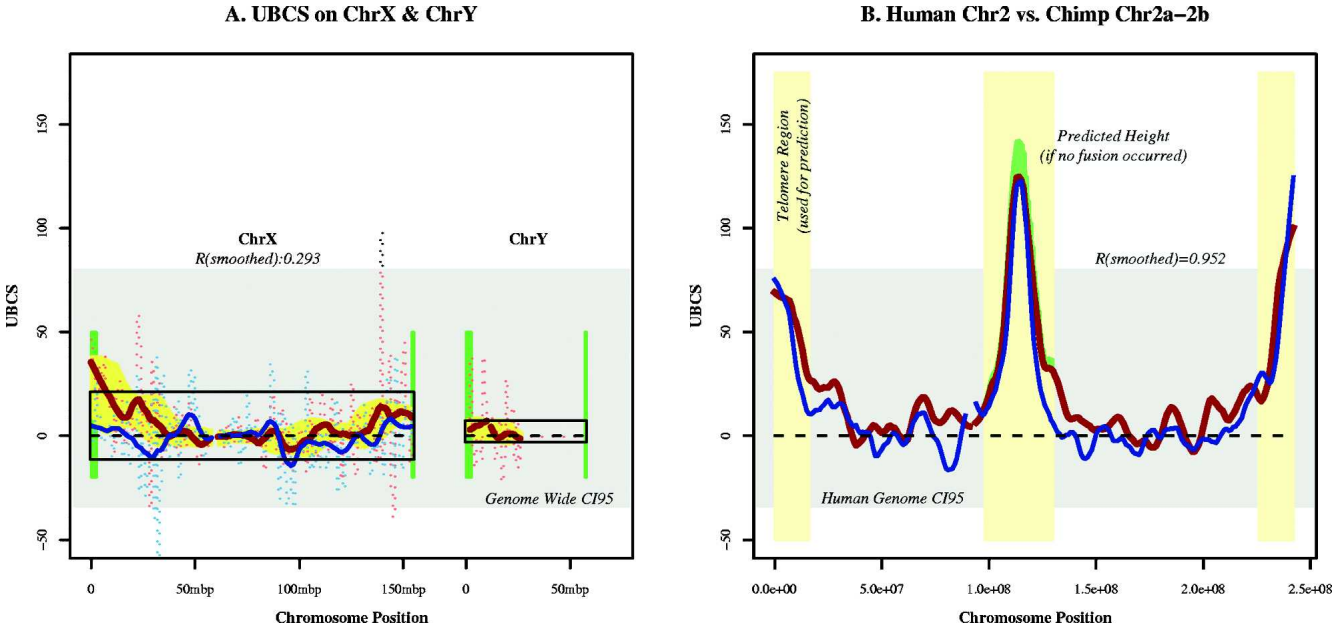

Figure 3. Exceptions to genome-wide patterns of bias. (*A*) There is little evidence of UBCS on either human or chimp sex chromosomes. There is also poor agreement between the human and chimp smoothed UBCS curves ($\rho = 0.25$ on chrX). The pseudo-autosomal regions (PAR) are shown in green. (*B*) Human chromosome 2 shows an atypical central peak, which is likely due to the fusion of two ancestral chromosomes (red, human chr2; blue, chimp chr2a and chr2b). The two smoothed curves are in remarkable agreement ($\rho = 0.84$). The UBCS signal for hypothetical telomeres (if no fusion had occurred) was predicted (green) by using 16 Mb of distal sequence (yellow).

beit reduced relative to the autosomes) both the human and chimpanzee X chromosomes show almost no UBCS signal (Fig. 3). Overall, clusters of substitutions on chromosome X are in fact biased strong-to-weak, as would be expected in the absence of selective pressures, due to AT mutation bias. In fact, correlation between the smoothed UBCS curves in human and chimp is much weaker on chromosome X ($\rho = 0.24$) compared to the autosomes ($\rho = 0.39$ to 0.96; Supplemental Table S3).

Since the X and Y chromosomes recombine with each other in the pseudo-autosomal regions (PAR), we specifically examined these regions for evidence of UBCS. In humans there are two PAR regions. One may be too small (~320 kb) to detect any significant UBCS, and we have found none. The only significant level of UBCS on chromosome X is found along the distal end of the p arm peaking in the larger (2.6 Mb) PAR, but extending ~5 Mb beyond its current boundary (Fig. 3). If UBCS is a product of recombination, we might expect a dramatic peak in the PAR region. We detect a mild elevation, but only in comparison with the levels of UCBS in other regions of X, and not from a genome-wide perspective. This milder elevation may be due to the same processes that render X and Y much less polymorphic (International SNP Map Working Group 2001) and diverse (Hellborg and Ellegren 2004; Baines and Harr 2007) than autosomes. Further investigation is needed to determine why the UCBS effect is not as dramatic in the PAR regions as it is in autosomal telomeres.

Bias correlated with male recombination rate

In order to further investigate genome-wide bias trends and to better understand this "X exception," we computed correlations between human UBCS and a variety of genome characteristics using a window size of 1 Mb. Male recombination rate ($\rho = 0.38$), recombination hot spots ($\rho = 0.37$), and G+C% ($\rho = 0.33$) show the strongest associations with UBCS, all of which are statistically significant ($P \approx 0$). This level of correlation is high, especially given that UBCS has accumulated for 6 million years under changing patterns of recombination, which should be only partially reflected in current rates (Kong et al. 2002). UBCS, G+C%, and recombination rates all rise substantially near the telomeres of each autosome. Computing the correlations without the last 16 Mb of sequence from each chromosome arm reduces the associations by ~50%. Compared to the male rate, the female recombination rate has a much weaker, though statistically significant, correlation with UBCS ($\rho = 0.25$). Both transcription density ($\rho = 0.009$) and conservation ($\rho = -0.069$) show little relation to UBCS. We also examined genome-wide correlations over a range of window sizes from 10 kb to 1 Mb. Results at finer scales were similar to those for 1 Mb windows, and we observed that the stronger correlations all increase with window size (Fig. 4). Repeating this analysis with smoothed UBCS values increases the magnitude of the observed correlations but does not qualitatively alter our findings.

To further investigate these correlations, we fit linear regression models on the data for 1-Mb windows. A multiple linear regression analysis that adjusts for the effects of other variables indicates that recombination hot spots, G+C%, and male recombination rate all have significant positive associations with UBCS. These associations are reduced in magnitude, but remain significant, when distal 16-Mb regions are dropped from the analysis. Conservation and female recombination rate show weak negative associations with UBCS in the model adjusting for other genomic variables. Furthermore, the female recombination rate is not correlated with the residuals from a simple linear regression of the male recombination rate on UBCS, suggesting that the female rate offers no additional explanatory power. If clusters of biased substitutions are a consequence of some process associated primarily with recombination in the male germline,

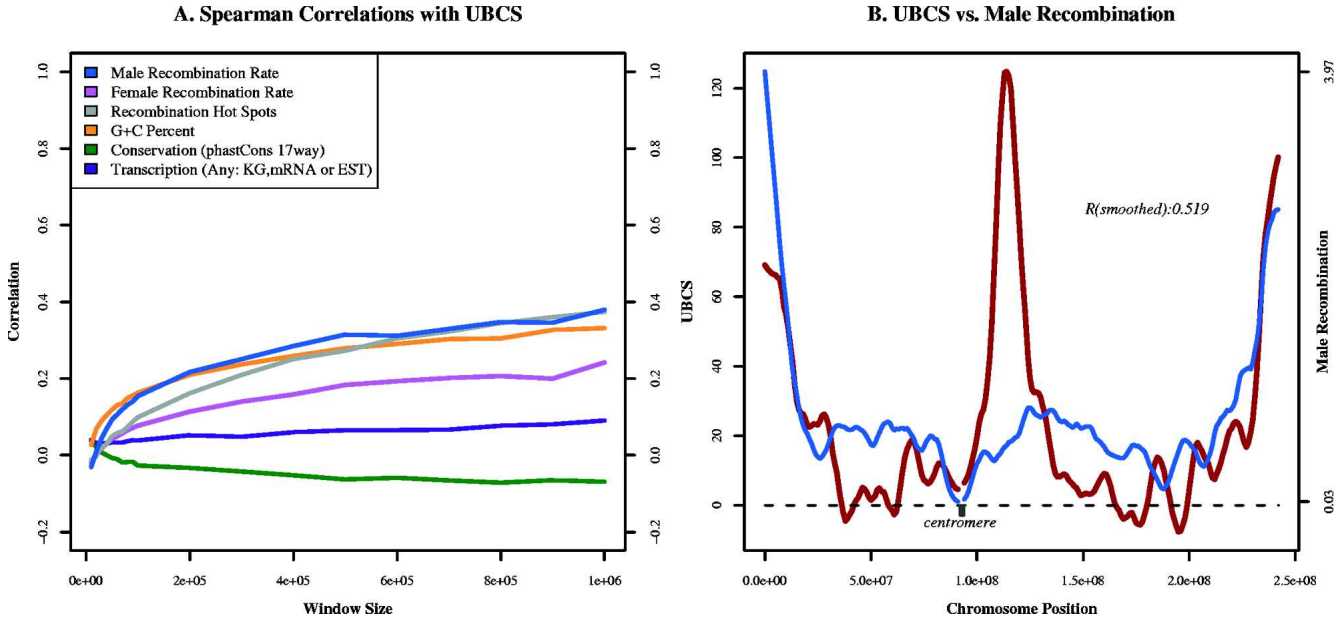

Figure 4. UBCS is strongly correlated with male recombination rates. (*A*) Spearman correlation between unsmoothed UBCS and different genomic variables for windows of size 10 kb to 1 Mb over the whole human genome. Male recombination rate (blue), recombination hot spots (gray), and G+C% (orange) have the highest correlations with UBCS. These correlations rise with increasing window size, particularly in the case of recombination hot spots. Female recombination rate (pink) is less highly correlated with UBCS. Transcription (purple) and conservation (green) are not significantly correlated with UBCS. (*B*) UBCS (red) and male recombination rate (blue) vary together along all the autosomes, with distal regions showing steep rises in both variables. Chromosome 2 (shown) has the weakest correlation, due mostly to the low male recombination rate in the zone of fusion.

then this would explain their absence on the X chromosome. We observed a positive regression of the residuals with G+C content. This association may reflect an historical relationship between male recombination, G+C, and BGC that cannot be measured with present-day male recombination rate estimates.

Human chromosome 2 fusion retains bias

The UBCS plot for human chromosome 2 (Fig. 3) has an internal peak, which maps to the region of fusion of two ancestral chromosomes (Ijdo et al. 1991). It is possible that the force causing biased clusters of substitutions is still acting in the middle of chromosome 2 even though it is no longer distal. However, the male recombination rate is not currently elevated in this region (Fig. 4). A simpler explanation is that the peak was created by the accumulation of biased substitutions prior to fusion. If so, then the size of the peak (only slightly less biased than the corresponding chimpanzee regions) suggests that the fusion occurred recently in human evolution. Using the assumptions of a relatively constant rate of UBCS accumulation before fusion and cessation of UBCS accumulation upon fusion, it is possible to estimate the date of the fusion itself based on the reduction of bias on the fused region of human chromosome 2 compared to the orthologous regions of chimpanzee chromosomes 2a and 2b (see Methods). Based on a speciation event 6 million yr ago (Mya), we estimate the fusion date at 0.74 Mya with a 95% confidence interval 0–2.81 Mya. This finding argues against the hypothesis that this fusion was the speciation event that separated the human and chimp lines (Navarro and Barton 2003).

Many extremely biased regions are transcribed

We identified the 200 most-biased regions of the human genome using a method agnostic to the conservation level of the regions (see Methods). Surprisingly, despite a lack of correlation between UBCS and transcription density genome-wide, the most extremely biased regions of the genome contain a disproportionate number of genes. Of the top 200 most-biased regions, 108 (54%) are in segments spanned by the pre-mRNAs of known genes (Apweiler et al. 2004). This compares to an expected 33.5% for similar sized and located regions (see Methods). A total of 161 (80.5%) lie in an mRNA or EST (National Center for Biotechnology Information 2002; Benson et al. 2005) transcribed in humans (Table 1). The 200 most-biased regions are also disproportionately found in the exons of known genes (15.5%, expected 6%). Among the known genes that overlap the top 10 most-biased regions (Table 2) are *VLDLR* (Webb et al. 1994) (associated with autism, bipolar disorder, and schizophrenia; may interact with Reelin), *HAR1* (Pollard et al. 2006b) (implicated in fetal brain development), *CHL1* (Wei et al. 1998) (neural adhesion molecule; active in fetal brain), and a *HTR5B* pseudogene (Matthes et al. 1993) (serotonin receptor; may have been destroyed by weak-to-strong substitutions separately in humans and chimps [Dreszer 2006]). A full list of biased regions is available at http://www.soe.ucsc.edu/research/compbio/ubcs/sub18_d32.html. The enrichment for genes in the most extremely biased regions suggests the possible involvement of non-neutral evolutionary forces.

Patterns of substitution bias are not reflected in SNPs

In contrast to the distinct and consistent patterns of bias observed in human and chimp substitutions, there is little evidence of genome-wide bias in human SNPs. This is the case overall (only 39.97% of SNPs are weak-to-strong) and among clusters of nearby SNPs. Weak-to-strong SNPs are less prevalent in regions with high G+C%, reflecting ancestral base

Table 1. Evidence of transcription in biased regions

Transcription evidence	Top 200 weak-to-strong biased regions	Top 200 strong-to-weak biased regions	200 Control regions
Known genes			
Pre-mRNA	108 (54%)	69 (34.5%)	67 (33.5%)
Exon	31 (15.5%)	12 (6%)	12 (6%)
Coding exon	18 (9%)	7 (3.5%)	9 (4.5%)
Transcription evidence in humans	161 (80.5%)	131 (65.5%)	123 (61.5%)

The number of regions that overlap various gene annotations varies dramatically between the most weak-to-strong (column 1) and strong-to-weak (column 2) biased regions of substitutions in the human genome. Transcription density in the strong-to-weak biased regions resembles that of a comparable set of 200 control regions (see Methods). In contrast, the weak-to-strong biased regions are unusually likely to be transcribed in humans and are five times more likely to be found in mature mRNAs than control regions. Known genes are from the UCSC "Known Genes" track (Apweiler et al. 2004; Benson et al. 2004, 2005).

composition. Human substitutions generally mirror genome-wide patterns of mutational bias. In distal regions, however, weak-to-strong substitutions are much more common than expected given the distribution of polymorphisms (Supplemental Table S4). These findings are similar to those reported by Webster et al. (2003). Thus, it appears that the patterns of UBCS that we observe along each autosomal chromosome cannot be explained by a higher rate of mutations toward GC unless historical mutational patterns were very different from those observed today.

Although there is no genome-wide polymorphism bias, some extremely biased clusters of SNPs do exist. The region with the most biased SNPs (8 within 148 bp) falls in the intron of a gene required for pain perception (Kim et al. 1996), and two of the top five regions occur in genes associated with cancer in humans (e.g., *SERINC1* [Zhang et al. 2003] and *CSMD1* [Scholnick and Richter 2003]). These data are consistent with the assumption that the force leading to the fixation of clusters of biased changes is still currently operating and remind us of the possibility that this force may occasionally compete with purifying selection.

Fixation bias is mostly weak to strong

By comparing levels of bias among SNPs and substitutions at different loci in the human genome, we can identify regions where the pattern of fixation in the last 6 million years deviates the most from that expected given current patterns of bias among polymorphic sites. We performed a modified McDonald–Kreitman test (see Methods) to compare counts of polymorphisms and fixed differences of different types (weak-to-strong vs. strong-to-weak vs. neither). In order to use a SNP data set free from ascertainment bias, we analyzed only the 828 genome regions (19.72 Mb) sequenced by the Seattle SNPs project (http://pga.gs.washington.edu; http://egp.gs.washington.edu). We found 34 loci with significant differences in patterns of bias between fixed and polymorphic changes (FDR < 20%). Consistent with the results above, 30 of these regions (88.2%) had many more weak-to-strong substitutions than expected given the pattern of bias in currently segregating sites. This finding is true for exons, UTRs, introns, and intergenic sequences, although most observed changes (fixed and polymorphic) fell in introns due to the design of the sequencing project. Only three (8.8%) of the 34 significant loci showed a strong-to-weak bias among fixed differences. These findings further support the argument that the regional weak-to-strong substitution bias we observe is associated with areas of increased probability of fixation and not with variation in mutation rate.

In order to evaluate whether these results might have been biased by underestimation of strong-to-weak substitutions due to homoplasy at CpG sites, we repeated the analysis using only sites for which the ancestral state can reliably be inferred to be non-CpG ("class 1 sites"; Meunier and Duret 2004). We have lower power to detect differences in the ratio of fixed to polymorphic differences in this case, due to the smaller number of sites (both fixed and polymorphic). Nonetheless, we found five loci with significantly different fixed-to-polymorphic ratios among the three types of changes. All five significant tests have an excess of weak-to-strong fixations, consistent with our conclusions based on all sites.

Discussion

Motivated by the observation that patterns of extremely accelerated nucleotide substitution are associated with a disproportionate number of weak-to-strong changes in the human genome, we

Table 2. Top 10 regions of biased clustered substitutions in humans

	Location in hg18	Length (bp)	Total subs.	W-to-S subs.	*P* value	Annotation
1	chr2: 113977236–113978604	1369	74	61	2.00×10^{-6}	*NT_022135.55*
2	chr9: 2612396–2613708	1313	48	43	9.85×10^{-6}	*VLDLR* (intron 1)
3	chr20: 61203595–61204231	637	37	34	2.54×10^{-4}	*HAR1* (RNA gene)
4	chr2: 117529420–117530267	848	43	38	3.68×10^{-4}	*NT_022135.102*
5	chr2: 115136243–115136977	735	39	35	6.74×10^{-4}	*DPP10* (intron 1)
6	chr3: 213300–214258	959	42	37	7.54×10^{-4}	*CHL1* (UTR exon 1)
7	chr2: 118333477–118334224	748	32	30	8.01×10^{-4}	*HTR5B* (exon 1)
8	chr8: 1752398–1753394	997	38	34	1.40×10^{-3}	mRNA *AF123758*
9	chr2: 113630962–113631930	969	43	37	3.13×10^{-3}	EST *BM926122*
10	chr13: 112033076–112033732	657	26	25	4.77×10^{-3}	EST *BG722997*

A surprising number of genes appear in the list of regions showing the most substitution bias (8 out of 10 are transcribed). VLDLR, HAR1, CHL1, and HTR5B are all involved in brain development or function. DPP10 is associated with asthma. AF123758 is a putative transmembrane protein that may be implicated in a neurodegenerative disorder. Regions 1 and 4 are predicted genes transcribed in mammals. W-to-S = weak-to-strong (Matthes et al. 1993; Webb et al. 1994; Burge 1998; Wei et al. 1998; Chen et al. 2003; Benson et al. 2004; Pollard et al. 2006b).

undertook a whole-genome analysis of substitution bias. We found that while there is no evidence of wide spread bias, genomic regions with a high density of substitutions since divergence from the chimp–human ancestor (≥5 in 100 bp) are enriched for weak-to-strong changes. We also found the accumulation of such biased clusters of substitutions is highly localized and dramatically increases near the telomeres of all autosomes. A strikingly similar pattern is seen in the chimpanzee genome. To understand the evolutionary forces generating clusters of biased substitutions at different but nearby locations in both primate genomes, we examined the correlation between bias and various genomic variables.

A detailed comparison of bias in human SNPs and fixed substitutions found no evidence of bias in polymorphic sites. In particular, human SNPs do not show the same pattern of increasing bias with higher density that we observe in human substitutions. Our findings are completely consistent with and expand the results of a previous study of compositional bias among SNPs and fixed differences in 1.8 Mb of the human genome (Webster et al. 2003). Furthermore, data from the Seattle SNPs sequencing project suggests that the vast majority (~90%) of genomic regions with a different pattern of bias in polymorphic sites compared to fixed differences have a disproportionate number of weak-to-strong substitutions. These findings indicate the presence of a fixation bias. It is nonetheless possible that the observed patterns of bias were created by a mutation bias that has changed over time. Since recombination is mutagenic, a mutation bias would be consistent with the association we find between bias and current large-scale recombination rates. In order to evaluate the possibility of a local, directed mutation bias, an analysis of the site-frequency spectrum in a large resequencing study is needed. Such investigation will be the focus of future work.

A fixation bias could result from either directional selection or a neutral process, such as BGC. We found that substitutions and clusters of biased substitutions are not correlated with conservation or transcription density, and bias is not limited to G+C rich regions. Thus, natural selection for increased G+C in either genes or at the level of isochores is not a likely explanation for the patterns of bias seen here. On the other hand, the easily recognizable clustering we have documented might be expected as the calling card of a BGC process. Supporting this, we observe that UBCS is correlated with male recombination rates and is absent from the sex chromosomes. These observations agree with previous studies that showed (1) elevated male recombination rates at the telomeres (Yu et al. 2001) and (2) a correlation between BGC and male recombination in human *Alu* repeats (Webster et al. 2005). Our results suggest that recombination is an integral component of the generation of UBCS.

It should be noted that not all variation in human–chimp substitution patterns can be attributed to variation in mutation and/or fixation. Recent studies have shown that variation in coalescent times along the genome can also result in differences in the number of substitutions (Patterson et al. 2006). It is not clear, however, how this phenomenon could lead to the pattern of UBCS that we describe.

The degree to which BGC pressure remains stable over time may prove revealing. We observed a moderate correlation of UBCS and current recombination hot spots, but a strong correlation with current male recombination rates, which is consistent with a model of hot spots moving or "burning out" through recombination events (Jeffreys and Neumann 2002; Pineda-Krch and Redfield 2005; Winckler et al. 2005), while regional recombination rates remain relatively constant (Myers et al. 2005). Further, our observation that human and chimp UBCS profiles are nearly identical attests to a relatively stable force across 12 million years of genetic divergence. By using this assumption of stability, we were able to calculate a date of fusion for human chromosome 2. Additionally, the highly localized profile of UBCS accumulation suggests a possible model for the evolution of isochores. Duret et al. (2002, 2006) have proposed a *GC factory* model, where high recombination rates at specific genomic locations may allow the build up of isochores. We found dramatic accumulation of UBCS near the telomeres of almost all human and chimp autosomes, suggesting that these may be current GC factories. Regions of high GC content within 10 Mb of autosomal telomeres have gained in GC in the last 6 million years, even while declining genome wide as part of the overall decline of mammalian isochores (Belle et al. 2004).

Perhaps the most intriguing aspect of a BCG force in genome evolution is that it acts like non-Darwinian selection. According to the theory of neutral evolution (Takahata 1996), most genome change is fueled by neutral stochastic forces acting upon random mutations, and the remainder is due to adaptive selection. As described by Nagylaki (1983), gene conversion is a neutral evolutionary force that can lead to increased genetic drift and faster evolution, mimicking directional selection for GC base pairs. When BGC acts upon a cluster of biased SNPs, the result will often be the creation of a novel cluster not seen in either parent chromosome. This new allele is more likely to be harmful than beneficial, which may account for the presence of some biased SNP clusters in association with human disease genes. The *neutral* fixation pressure of BGC may in some cases directly compete against, and slow, purifying selection.

The asymmetric association of BGC with male as opposed to female meiosis is provocative. It is well known that mutation rates are higher in males (Crow 1993; Li et al. 2002; Goetting-Minesky and Makova 2006), leading to a process dubbed "male-driven evolution" (Li et al. 2002). The higher mutation rate is most frequently ascribed to the larger number of cell divisions occurring in the germ-cell lines of males. However, that may not explain all of the increase in male mutation rates (Lercher et al. 2001; Filatov and Charlesworth 2002; Gaffney and Keightley 2005). BGC in males would also drive genetic variation and could be a significant additional part of male-driven evolution. However, it is not obvious why BGC would be tolerated in males while avoided in females. It has been proposed that transcription-associated recombination (TAR) (Prado et al. 1997; Bell et al. 1998; Nickoloff and Reynolds 1990; Aguilera 2002) might be driving BGC (Vinogradov 2003). If this were true, then the demands of gamete generation would make BGC hard to avoid in males, at the same time that "sperm selection" (Holt and Van Look 2004) might mitigate the dangers. This model would explain our finding that the most-biased regions are disproportionately transcribed. Additionally, if subtelomeric regions are GC factories building isochores, then an association of BGC with TAR might explain why (1) genes appear to lead the accumulation of GC in GC-rich isochores (Press and Robins 2006) and (2) widely expressed housekeeping genes are more often found in GC-rich isochores than are tissue-specific genes (Vinogradov 2003).

In conclusion, we have found a strong link between male recombination rates and GC bias among clusters of substitutions that have occurred in human and chimp genomes since divergence, suggesting a significant presence of BGC during male meiosis. By studying biased clusters, which are the footprints of

this process, chromosome fusions can be dated and perhaps a greater understanding of the evolution of isochores can be obtained. These observations suggest that biased gene conversion could have been an important contributor to the evolution of our species.

Methods

All sequence, assembly, and alignment data are freely available from the UCSC Genome Browser (http://www.genome.ucsc.edu/; Karolchik and Kent 2002; Kent et al. 2002; Karolchik et al. 2003). Recombination hot-spot data were downloaded from the HapMap website (http://www.hapmap.org/downloads/). Recombination rate data were obtained from deCODE Genetics and are available through the UCSC Table Browser (Karolchik et al. 2003). Data analysis was performed using custom programs written in the C and R languages (R Development Core Team 2004). A more complete explanation of methods can be found in Supplemental Text S1.

Data sets

We used the following data sets:

1. Unless otherwise stated, all analyses were based upon the sequence and locations found in the May 2004 assembly of the human genome (hg17). Some work used the March 2006 assembly (hg18).
2. All analyses on human substitutions and SNPs were based upon the reciprocal best alignment to hg17 of the November 2003 assembly of the chimpanzee (*Pan troglodytes*) genome (Chimpanzee Sequencing and Analysis Consortium 2005) (pt1). Chimp substitution work was based upon the January 2006 assembly (pt2) aligned to hg18.
3. The January 2005 pre-release assembly of the Rhesus macaque (*Macaca mulatta*) genome (Rhesus Macaque Genome Sequencing Consortium 2007) (rh0) was used as an out-group for the hg17–pt1 comparisons. The January 2006 assembly of macaque (rh1) was used for the same purpose with the hg18–pt2 alignment.
4. Analysis of bias in SNPs was undertaken using the International HapMap Project's October 2005 release of haplotype map for humans (http://www.hapmap.org/downloads/index.html.en; International HapMap Consortium 2005).
5. Recombination hot spots were located using the September 2005 release of HapMap Phase I data from the International HapMap Project (http://www.hapmap.org/downloads/recombination/latest/hotspots/; International HapMap Consortium 2005).
6. Recombination rates were provided by the deCODE genetic map (Kong et al. 2002). Rates were available for males and females separately as well as the sex-averaged rate, in the form of a mean rate across 1 Mb segments of the genome.
7. Gene locations were taken from the "Known Genes" track of the UCSC browser, which was compiled using protein data from UniProt (Apweiler et al. 2004) and mRNA data from NCBI (National Center for Biotechnology Information 2002; Benson et al. 2005). mRNA locations are derived from the UCSC browser "Human mRNA" track. Locations of expressed sequence tags (ESTs) are from the UCSC "Human EST" track. Sources for both tracks were international public sequence databases (Benson et al. 2004). A descending hierarchy of transcription evidence was as follows: known exons, known genes, human mRNAs, human ESTs.
8. Human sequencing data from the Seattle SNPs Project was downloaded from the PGA and EGP websites (http://pga.gs.washington.edu; http://egp.gs.washington.edu) in July 2006.

Identification of human–chimp fixed differences and human SNPs

Genome sequences were parsed into three distinct data sets: fixed substitutions in humans, fixed substitutions in chimps, and human SNPs. Reciprocal best human–chimp alignments were used to identify fixed differences and to polarize human polymorphisms. Human–chimp differences were mapped to the macaque assemblies using single-coverage lift over chains.

Preparation of the fixed substitutions data sets involved the creation of a set of single-nucleotide differences between human and chimp in regions of high-quality chimp and macaque sequence. Human–chimp differences were only included if the following criteria for chimp base quality were met: (1) chimp base of quality $\geq$30, (2) all bases in an 11-base window of quality $\geq$25, (3) no more than 2 base differences in the 11-base window, and (4) no insertions or deletions in the 11-base window. Each single fixed difference between humans and chimps was considered "derived" in humans if the corresponding macaque base matched chimp and was deemed "biased" if the human derived pair was CG or GC, while the "ancestral" pair was AT or TA. All substitutions with indeterminate ancestry were excluded. The same method was used to create a data set of chimp-derived substitutions, using the more recent hg18, pt2, and rh1 assemblies. When examining fixed substitution data for chimp, sequencing and assembly errors may exist despite base-quality filtering. If a base pair is AT in human and macaque, but GC in chimp, then the difference might be due to either a fixed substitution or an error in chimp. For this reason, the majority of work concentrates upon characterizing patterns found in the human genome.

The human SNP data set was generated by similar methods, except that the "ancestral" allele was the one that matched chimp, macaque, or both when both were present. All SNPs without at least one out-group or with indeterminate ancestry were excluded from the data sets. Final processing of the substitutions data set involved subtracting any locations found in both the substitution and SNP data sets.

Of 28,937,901 high-quality single-base differences between humans and chimps found in hg17, 24,795,278 remained after aligning with macaque and 23,916,284 remained after subtracting SNPs. Of these, 22,784,742, showed unambiguous ancestry with 10,871,714 derived in humans, and 4,685,510 (43.1%) of those were biased. Another 12,502,294 were derived in the chimp line (pt2, hg18 and rh1), and 5,342,016 (42.7%) were biased. Of 3,874,080 SNPs in the hg17 data set, 3,424,895 had an ancestry that could be determined and of those, 1,368,922 (40.0%) were biased.

Window method

A windowing method was used to characterize bias across the genome. Windows are capable of detecting clusters of substitutions, but the method does not assume that bias is related to clustering. Using windows of 100 bp sliding 50 bp results in 61,535,590 possible windows of the human genome (hg17). For the fixed substitutions data set, 16,633,481 windows were discovered with at least one substitution in the human line. Analysis of SNPs using 300-bp windows sliding 150-bp resulted in 20,511,865 possible windows, 1,900,453 of which contained at least one SNP of clear ancestry.

Windows were characterized based on a variety of genomic variables. Conservation scoring was done using the phastCons

methods (Siepel et al. 2005) available in the "conservation" track of the UCSC Genome Browser (hg17). Of 16,633,481 fixed substitution windows, 15,245,997 received a non-zero conservation score. A window was considered distal (or "telomeric") if it overlapped the first or last chromosome band (unless otherwise stated). Additionally, distance to a telomere was considered as a proportional distance along each chromosome. Windows overlapping a recombination hot spot in the "SNP Recombination Hots" track of the UCSC Genome Browser were considered "hot." The distance of a window to its nearest hot spot was also analyzed.

Unexpected biased clustered substitutions

A second method of viewing biased substitutions was developed in order to fully characterize bias as a function of clustering. We defined a novel statistic, called UBCS, which measures excess bias among clustered substitutions in a genomic region. UBCS is computed as follows. First, a substitution is defined as a *clustered substitution* (CS) if it is one of at least five substitutions within 300 bp. A CS is further characterized as a *biased clustered substitution* (BCS) if it belongs to a cluster with at least 80% weak-to-strong substitutions. Then, for each genomic region, we calculate the expected number of BCS under the null hypothesis of no association between bias and clustering. This calculation is complicated by the fact that a CS can belong to more than one cluster, any one of which can be biased (and hence make the CS a BCS). For a CS that falls in only one cluster, the null probability that it is a BCS can be calculated using the Bernoulli distribution with the probability of bias estimated by the proportion of all substitutions (CS and non-CS) in the region that are biased. For any CS that falls in more than one cluster, the null probability that it is a BCS can be derived through a combinatorial calculation based on the number, size, and overlap of clusters (Supplemental Text S1). The expected number of BCS in a region is then computed based on the observed number of CS and the null probability that each is a BCS. Finally, the region's *unexpected biased clustered substitutions* (UBCS) statistic is defined as the actual number of BCS minus the expected number.

UBCS is a measure of the excess bias in a region and can be positive or negative. UBCS is zero under a null model with no relationship between substitution density and bias. In this case, the pattern of bias in clustered substitutions is identical to that of nonclustered substitutions in the region. When UBCS is greater than zero, there is more bias among densely clustered substitutions than expected based on the overall bias of the region. An individual substitution can be a CS or a BCS, but not an unexpected biased clustered substitution. UBCS is a characteristic of a genomic region. It should be noted that this study is not attempting to characterize the forces that create clusters of substitutions (or SNPs) but only the forces that *bias* clusters, however they may be created.

UBCS was computed for 1-Mb regions along all human and all but two chimp chromosomes (Figs. 2, 3). Additionally, the UBCS signal along each chromosome was smoothed using the loess function in *R* with a Gaussian kernel and least-squares fitting. Loess smoothing estimates a local polynomial surface. The degree of smoothing is determined by the span, which we selected to be either 15 Mb (individual chromosomes) or 25 Mb (genome-wide). A 95% confidence interval for smoothed UBCS was generated, either genome-wide, by chromosome, or for a sliding window of 25 Mb, using ± 1.96 SD of the smoothed UBCS score. When both human and chimp maps were superimposed, a small distortion in the chimp maps was introduced by plotting it using human coordinates. However, the chimp UBCS signal was at all times generated using chimp locations. A similar analysis was performed for SNPs.

Correlation with genomic variables

Spearman's rank correlation coefficients (ρ) were calculated between rough (nonsmoothed) and smoothed UBCS and a number of genomic variables: chimp UBCS, G+C content, recombination hot spots, recombination rates (male, female, and sex-averaged), conservation score, and transcription density. Calculations were performed for 1-Mb regions, since recombination data were measured on this scale. Because X and Y recombination has a special character, genome-wide correlations between recombination rates and UBCS were computed using only the autosomes. Inclusion of chromosome X for female recombination rate does not change the correlation. Excluding both sex chromosomes in the calculation of other correlations (hot spots, G+C%, transcription, conservation) does not significantly alter any of the reported values.

We also looked at correlations for a variety of scales from 10 kb to 1 Mb (Fig. 4). We fit linear regression models for each genomic variable and examined plots of residuals versus the remaining factors to determine if a second variable might explain any of the unexplained variance in UBCS. Finally, a multiple linear regression model was fit using stepwise selection. This model allowed us to compute associations of each genomic variable with UBCS while adjusting for the effects of the other variables. Analyses were performed with both unsmoothed and smoothed UBCS values. All correlations reported are for unsmoothed UBCS except for the comparison of human and chimp UBCS, where we discuss both analyses.

Regions of high density of bias

A list of the longest clusters of substitutions of a given minimum density was generated and then ranked according to degree of bias. This analysis was conducted on the hg18, pt2, rh1 human substitutions data set and the hg17, pt1, rh0 human SNP data set. We looked for clusters of at least six substitutions (or SNPs) with a density of no less than 1 in 32 bases. These initial clusters were extended out as long as the region maintained a density of 1 difference per 32 bases with no barren stretch longer than 96 bases. The regions were then carved down to maximize a score of bias for each cluster in the list. The *P* value used to rank these regions was the binomial probability of the observed number of biased substitutions given the total number of substitutions within the region. It should be noted that this list was generated agnostic to any other factor beyond density of substitutions (or SNPs) and degree of bias. The top 200 regions were compared to control regions of the same size and distance from a chromosome end on other autosomal chromosome arms. Control regions for top-scoring regions near the fusion point on chromosome 2 were identified using distance from the fusion point in place of distance from a chromosome end.

Dating the fusion of chromosome 2

Because of the remarkably similar shape and amplitude of UBCS peaks between human and chimp chromosomes, we were able to use the relative heights of the telomeric regions seen on a chimp chromosome to predict the relative heights of the telomeric regions on the corresponding human chromosome. For human chromosome 2, we compared the UBCS peak of the central fusion region to the predicted peak using the orthologous regions of chimpanzee chromosomes 2a.q and 2b.p. The ratio of actual vs. predicted UBCS for this fusion region provided an estimate of the proportion of the last 6 million years when UBCS accumulated

prior fusion (Supplemental Text S1). One important part of this calculation involves the size of a region used to measure the UBCS signal near a telomere. The size of 16 Mb was used because it is the average region of heightened UBCS at telomeres in humans. Since our calculation relies upon the assumption of cessation of UBCS accumulation upon fusion, the estimated date should be considered as the most recent bound for the fusion event. Using chromosomes with complete human and chimp substitution data over both distal regions (chromosomes 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 16, 17, and 20; Supplemental Table S2), we estimated the standard error and normal 95% confidence interval for the predicted fusion date (Supplemental Text S1).

Modified McDonald–Kreitman tests

For each gene in the Seattle SNPs resequencing project (http://pga.gs.washington.edu; http://egp.gs.washington.edu), we polarized polymorphic sites using the method described above for the whole-genome analysis. Fixed differences were obtained as described, using the hg18, pt2, and rh1 assemblies. Only SNPs and substitutions in "scanned regions" (those covered by sequencing reads) were included in the analysis. Finally, changes (fixed or polymorphic) were classified as weak-to-strong, strong-to-weak, or neither based on the inferred ancestral base. For this analysis, we conservatively classified a substitution as "neither" if both chimp and human differed from macaque. Similarly, we required both of the chimp and macaque bases to agree with one of the two human alleles in order to classify a SNP as weak-to-strong or strong-to-weak. From this data, we constructed a 2-by-3 table for each gene comparing the proportion of each type of change among SNPs vs. fixed substitutions. A Fisher's exact test for association was performed for each table, and *P* values were adjusted for multiple comparisons using Benjamini and Hochberg's FDR controlling procedure (Benjamini and Hochberg 1995). The analysis was performed for each gene locus as a whole and then separately for the exons, introns, UTRs, and intergenic regions.

Acknowledgments

We thank Daryl Thomas for data set preparation, Andy Kern for the suggestion of the modified McDonald–Kreitman test, Laurent Duret for helpful discussions, and Svitlana Tyekucheva for suggesting an alternative to the UBCS calculation.

References

Aguilera, A. 2002. The connection between transcription and genomic instability. *EMBO J.* **21:** 195–201.

Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. 2004. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **32:** D115–D119. http://www.expasy.uniprot.org/about/publications.shtml

Baines, J.F. and Harr, B. 2007. Reduced X-linked diversity in derived populations of house mice. *Genetics* **175:** 1911–1921.

Bell, S.J., Chow, Y.C., Ho, J.Y., and Forsdyke, D.R. 1998. Correlation of chi orientation with transcription indicates a fundamental relationship between recombination and transcription. *Gene* **216:** 285–292.

Belle, E.M.S., Duret, L., Galtier, N., and Eyre-Walker, A. 2004. The decline of isochores in mammals: An assessment of the GC content variation along the mammalian phylogeny. *J. Mol. Evol.* **58:** 653–660.

Benjamini, Y. and Hochberg, Y. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Statist. Soc. Ser. B (Methodol.)* **57:** 289–300.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2004. GenBank: Update. *Nucleic Acids Res.* **32:** D23–D26. doi: 10.1093/nar/gkh045.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. 2005. GenBank. *Nucleic Acids Res.* **33:** D34–D38. doi: 10.1093/nar/gki063.

Bernardi, G. 2000. Isochores and the evolutionary genomics of vertebrates. *Gene* **241:** 3–17.

Bernardi, G. and Bernardi, G. 1986. Compositional constraints and genome evolution. *J. Mol. Evol.* **24:** 1–11.

Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., and Rodier, F. 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228:** 953–958.

Birdsell, J.A. 2002. Integrating genomics, bioinformatics, and classical genetics to study the effects of recombination on genome evolution. *Mol. Biol. Evol.* **19:** 1181–1197.

Brown, T.C. and Jiricny, J. 1988. Different base/base mispairs are corrected with different efficiencies and specificities in monkey kidney cells. *Cell* **54:** 705–711.

Burge, C. 1998. Modeling dependencies in pre-mRNA splicing signals. In *Computational methods in molecular biology* (eds. S. Salzberg et al.), pp. 127–163. Elsevier Science, Amsterdam. http://www.cs.jhu.edu/%7Esalzberg/compbio-book.html

Chen, T., Ajami, K., McCaughan, G.W., Gorrell, M.D., and Abbott, C.A. 2003. Dipeptidyl peptidase IV gene family. The DPIV family. *Adv. Exp. Med. Biol.* **524:** 79–86.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437:** 69–87.

Crow, J.F. 1993. How much do we know about spontaneous human mutation rates? *Environ. Mol. Mutagen.* **21:** 122–129.

Dreszer, T. 2006. "Biased clustered substitutions in the human genome: Sex, gambling and non-Darwinian evolution." M.S. thesis, University of California, Santa Cruz. http://www.soe.ucsc.edu/research/compbio/ubcs/Thesis.pdf

Duret, L. 2002. Evolution of synonymous codon usage in metazoans. *Curr. Opin. Genet. Dev.* **12:** 640–649.

Duret, L., Eyre-Walker, A., and Galtier, N. 2006. A new perspective on isochore evolution. *Gene* **385:** 71–74.

Duret, L., Semon, M., Piganeau, G., Mouchiroud, D., and Galtier, N. 2002. Vanishing GC-rich isochores in mammalian genomes. *Genetics* **162:** 1837–1847.

Eyre-Walker, A. 1993. Recombination and mammalian genome evolution. *Proc. Biol. Sci.* **252:** 237–243.

Eyre-Walker, A. 1999. Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152:** 675–683.

Eyre-Walker, A. and Hurst, L.D. 2001. The evolution of isochores. *Nat. Rev. Genet.* **2:** 549–555.

Filatov, D.A. and Charlesworth, D. 2002. Substitution rates in the X- and Y-linked genes of the plants, *Silene latifolia* and *S. dioica*. *Mol. Biol. Evol.* **19:** 898–907.

Fryxell, K.J. and Zuckerkandl, E. 2000. Cytosine deamination plays a primary role in the evolution of mammalian isochores. *Mol. Biol. Evol.* **17:** 1371–1383.

Gaffney, D.J. and Keightley, P.D. 2005. The scale of mutational variation in the murid genome. *Genome Res.* **15:** 1086–1094.

Goetting-Minesky, M.P. and Makova, K.D. 2006. Mammalian male mutation bias: Impacts of generation time and regional variation in substitution rates. *J. Mol. Evol.* **63:** 537–544.

Hellborg, L. and Ellegren, H. 2004. Low levels of nucleotide diversity in mammalian Y chromosomes. *Mol. Biol. Evol.* **21:** 158–163.

Holt, W.V. and Van Look, K.J. 2004. Concepts in sperm heterogeneity, sperm selection and sperm competition as biological foundations for laboratory tests of semen quality. *Reproduction* **127:** 527–535.

Ijdo, J., Baldini, A., Ward, D.C., Reeders, S.T., and Wells, R.A. 1991. Origin of human chromosome 2: An ancestral telomere–telomere fusion. *Proc. Natl. Acad. Sci.* **88:** 9051–9055.

International HapMap Consortium. 2005. A haplotype map of the human genome. *Nature* **437:** 1299–1320.

International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.

Jeffreys, A.J. and Neumann, R. 2002. Reciprocal crossover asymmetry and meiotic drive in a human recombination hot spot. *Nat. Genet.* **31:** 267–271.

Karolchik, D. and Kent, W.J. 2002. The UCSC Genome Browser. In *Current protocols in bioinformatics* (ed. A.D. Baxevanis). Wiley, New York.

Karolchik, D., Baertsch, R., Diekhans, M., Furey, T.S., Hinrichs, A., Lu, Y.T., Roskin, K.M., Schwartz, M., Sugnet, C.W., Thomas, D.J., et al. 2003. The UCSC genome browser database. *Nucleic Acids Res.* **31:** 51–54.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. 2002. The human genome browser at UCSC.

Genome Res. **12:** 996–1006.
Kim, E., Cho, K.O., Rothschild, A., and Sheng, M. 1996. Heteromultimerization and NMDA receptor-clustering activity of Chapsyn-110, a member of the PSD-95 family of proteins. *Neuron* **17:** 103–113.
Kong, A., Gudbjartsson, D.F., Sainz, J., Jonsdottir, G.M., Gudjonsson, S.A., Richardsson, B., Sigurdardottir, S., Barnard, J., Hallbeck, B., Masson, G., et al. 2002. A high-resolution recombination map of the human genome. *Nat. Genet.* **31:** 241–247.
Kudla, G., Lipinski, L., Caffin, F., Helwak, A., and Zylicz, M. 2006. High guanine and cytosine content increases mRNA levels in mammalian cells. *PLoS Biol.* **4:** e18. doi 10.1371/journal.pbio.0040180.
Lercher, M.J., Williams, E.J., and Hurst, L.D. 2001. Local similarity in evolutionary rates extends over whole chromosomes in human–rodent and mouse–rat comparisons: Implications for understanding the mechanistic basis of the male mutation bias. *Mol. Biol. Evol.* **18:** 2032–2039.
Lercher, M.J., Smith, N.G.C., Eyre-Walker, A., and Hurst, L.D. 2002. The evolution of isochores: evidence from SNP frequency distributions. *Genetics* **162:** 1805–1810.
Lercher, M.J., Urrutia, A.O., Pavlícek, A., and Hurst, L.D. 2003. A unification of mosaic structures in the human genome. *Hum. Mol. Genet.* **12:** 2411–2415.
Li, W.H., Yi, S., and Makova, K. 2002. Male-driven evolution. *Curr. Opin. Genet. Dev.* **12:** 650–656.
Lipatov, M., Arndt, P.F., Hwa, T., and Petrov, D.A. 2006. A novel method distinguishes between mutation rates and fixation biases in patterns of single-nucleotide substitution. *J. Mol. Evol.* **62:** 168–175.
Maki, H. 2002. Origins of spontaneous mutations: Specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annu. Rev. Genet.* **36:** 279–303.
Matthes, H., Boschert, U., Amlaiky, N., Grailhe, R., Plassat, J.L., Muscatelli, F., Mattei, M.G., and Hen, R. 1993. Mouse 5-hydroxytryptamine5A and 5-hydroxytryptamine5B receptors define a new family of serotonin receptors: Cloning, functional expression, and chromosomal localization. *Mol. Pharmacol.* **43:** 313–319.
Meunier, J. and Duret, L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol. Biol. Evol.* **21:** 984–990.
Myers, S., Bottolo, L., Freeman, C., McVean, G., and Donnelly, P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. *Science* **310:** 321–324.
Nagylaki, T. 1983. Evolution of a finite population under gene conversion. *Proc. Natl. Acad. Sci.* **80:** 6278–6281.
National Center for Biotechnology Information. 2002. *NCBI handbook*. National Library of Medicine (U.S.), National Center for Biotechnology Information, Bethesda, MD. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Books.
Navarro, A. and Barton, N.H. 2003. Chromosomal speciation and molecular divergence—Accelerated evolution in rearranged chromosomes. *Science* **300:** 321–324.
Nickoloff, J.A. and Reynolds, R.J. 1990. Transcription stimulates homologous recombination in mammalian cells. *Mol. Cell. Biol.* **10:** 4837–4845.
Patterson, N., Richter, D.J., Gnerre, S., Lander, E.S., and Reich, D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* **441:** 1103–1108.
Pineda-Krch, M. and Redfield, R.J. 2005. Persistence and loss of meiotic recombination hotspots. *Genetics* **169:** 2319–2333.
Pollard, K.S., Salama, S.R., King, B., Kern, A., Dreszer, T., Katzman, S., Siepel, A., Pedersen, J., Bejerano, G., Baertsch, R., et al. 2006a. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet.* **2:** e168. doi: 10.1371/journal.pgen.0020168.
Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. 2006b. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443:** 167–172.
Prado, F., Piruat, J.I., and Aguilera, A. 1997. Recombination between DNA repeats in yeast *hpr1*$_{\Delta}$ cells is linked to transcription elongation. *EMBO J.* **16:** 2826–2835.
Press, W.H. and Robins, H. 2006. Isochores exhibit evidence of genes interacting with the large-scale genomic environment. *Genetics* **174:** 1029–1040.
R Development Core Team. 2004. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 3-900051-07-0. http://www.r-project.org/index.html.
Rhesus Macaque Genome Sequencing Consortium. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* **316:** 222–234.
Schmegner, C., Hoegel, J., Vogel, W., and Assum, G. 2007. The rate, not the spectrum, of base pair substitutions changes at a GC-content transition in the human *NF1* gene region: Implications for the evolution of the mammalian genome structure. *Genetics* **175:** 421–428.
Scholnick, S.B. and Richter, T.M. 2003. The role of CSMD1 in head and neck carcinogenesis. *Genes Chromosomes Cancer* **38:** 281–283.
Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.* **15:** 1034–1050.
Sueoka, N. 1988. Directional mutation pressure and neutral molecular evolution. *Proc. Natl. Acad. Sci.* **85:** 2653–2657.
Takahata, N. 1996. Neutral theory of molecular evolution. *Curr. Opin. Genet. Dev.* **6:** 767–772.
Vinogradov, A.E. 2003. Isochores and tissue-specificity. *Nucleic Acids Res.* **31:** 5212–5220. doi: 10.1093/nar/gkg699.
Webb, J.C., Patel, D.D., Jones, M.D., Knight, B.L., and Soutar, A.K. 1994. Characterization and tissue-specific expression of the human 'very low density lipoprotein (VLDL) receptor' mRNA. *Hum. Mol. Genet.* **3:** 531–537.
Webster, M.T. and Smith, N.G. 2004. Fixation biases affecting human SNPs. *Trends Genet.* **20:** 122–126.
Webster, M.T., Smith, N.G.C., and Ellegren, H. 2003. Compositional evolution of non-coding DNA in the human and chimpanzee genomes. *Mol. Biol. Evol.* **20:** 278–286.
Webster, M.T., Smith, N.G.C., Hultin-Rosenberg, L., Arndt, P.F., and Ellegren, H. 2005. Male-driven biased gene conversion governs the evolution of base composition in human Alu repeats. *Mol. Biol. Evol.* **22:** 1468–1474.
Wei, M.-H., Karavanova, I., Ivanov, S.V., Popescu, N.C., Keck, C.L., Pack, S., Eisen, J.A., and Lerman, M.I. 1998. In silico-initiated cloning and molecular characterization of a novel human member of the L1 gene family of neural cell adhesion molecules. *Hum. Genet.* **103:** 355–364.
Winckler, W., Myers, S.R., Richter, D.J., Onofrio, R.C., McDonald, G.J., Bontrop, R.E., McVean, G.A.T., Gabriel, S.B., Reich, D., Donnelly, P., et al. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308:** 107–111.
Wolfe, K.H., Sharp, P.M., and Li, W. 1989. Mutation rates differ among regions of the mammalian genome. *Nature* **337:** 283–285.
Yu, A., Zhao, C., Fan, Y., Jang, W., Mungall, A.J., Deloukas, P., Olsen, A., Doggett, A., Ghebranious, N., Broman, K.W., et al. 2001. Comparison of human genetic and sequence-based physical maps. *Nature* **409:** 951–953.
Zhang M., Yu L., Wu Q., Zheng L.H., Wei Y.H., Wan B., and Zhao S.Y. 2003. Identification and characterization of TDE2, a plasma-membrane protein with 11 transmembrane helices, and its variable expression in human lung cancer and liver cancer tissues. Submitted (JUL-2003) to the EMBL/GenBank/DDBJ databases. http://www.expasy.org/uniprot/Q9NRX5.

Received February 14, 2007; accepted in revised form June 28, 2007.

Biased clustered substitutions in the human genome: The footprints of male-driven biased gene conversion

Timothy R. Dreszer, Gregory D. Wall, David Haussler, et al.

Genome Res. 2007 17: 1420-1430 originally published online September 4, 2007
Access the most recent version at doi:10.1101/gr.6395807