Research Article

Human Mutation

OFFICIAL JOURNAL
HGVS
HUMAN GENOME VARIATION SOCIETY
www.hgvs.org

Meiotic Recombination Favors the Spreading of Deleterious Mutations in Human Populations

Anamaria Necşulea,[1] Alexandra Popa,[1] David N. Cooper,[2] Peter D. Stenson,[2] Dominique Mouchiroud,[1] Christian Gautier,[1] and Laurent Duret[1]*

[1]Université de Lyon, Université Lyon 1, CNRS, UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, Villeurbanne, France; [2]Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff, United Kingdom

Communicated by Mark H. Paalman

Received 24 June 2010; accepted revised manuscript 28 October 2010.
Published online 30 November 2010 in Wiley Online Library (wileyonlinelibrary.com). DOI 10.1002/humu.21407

ABSTRACT: Although mutations that are detrimental to the fitness of organisms are expected to be rapidly purged from populations by natural selection, some disease-causing mutations are present at high frequencies in human populations. Several nonexclusive hypotheses have been proposed to account for this apparent paradox (high new mutation rate, genetic drift, overdominance, or recent changes in selective pressure). However, the factors ultimately responsible for the presence at high frequency of disease-causing mutations are still contentious. Here we establish the existence of an additional process that contributes to the spreading of deleterious mutations: GC-biased gene conversion (gBGC), a process associated with recombination that tends to favor the transmission of GC-alleles over AT-alleles. We show that the spectrum of amino acid-altering polymorphisms in human populations exhibits the footprints of gBGC. This pattern cannot be explained in terms of selection and is evident with all nonsynonymous mutations, including those predicted to be detrimental to protein structure and function, and those implicated in human genetic disease. We present simulations to illustrate the conditions under which gBGC can extend the persistence time of deleterious mutations in a finite population. These results indicate that gBGC meiotic drive contributes to the spreading of deleterious mutations in human populations.
Hum Mutat 32:198–206, 2011. © 2011 Wiley-Liss, Inc.

KEY WORDS: disease-associated mutations; meiotic recombination; gene conversion; polymorphisms; derived allele frequencies

Introduction

The majority of disease-causing mutations (DMs) detected in human populations are very recent, having only been transmitted over a few generations at most [Slatkin and Rannala, 2000]. A substantial fraction of DMs nevertheless correspond to more ancient mutations that have persisted for a large number of generations. Several nonexclusive hypotheses have been proposed to explain why such detrimental mutations could have escaped negative selection. First, detrimental mutations that have a limited impact on reproductive success (e.g., mutations causing late-onset diseases) can spread simply by genetic drift [Kryukov et al., 2007]. Second, some DMs confer a selective advantage upon heterozygotes (overdominance) [Dean et al., 2002]. Third, some DMs may have attained a high population frequency in the past because they were once advantageous under environmental conditions that no longer pertain [Di Rienzo and Hudson, 2005]. Finally, some DMs may occur at high frequency because of a high de novo mutation rate or a germ-line selective advantage [Choi et al., 2008].

Population genetic models indicate that in addition to genetic drift and natural selection, there is a third process that can contribute to the spreading of mutations within a population: biased gene conversion (BGC). Gene conversion occurs during homologous recombination and involves the nonreciprocal transfer of sequence information between the two recombining DNA molecules. This process is said to be biased if one of the two DNA molecules involved is more likely than the other to be the donor. Gene conversion can affect paralogous sequences duplicated in the genome or different alleles at a given locus [Chen et al., 2007]. In the case of allelic gene conversion, BGC leads to an excess of the "favored" allele in the pool of gametes and hence tends to increase the frequency of this allele in the population. Theoretical analyses have shown that, as with selection, BGC can increase the probability of fixation of the favored allele [Nagylaki, 1983].

Although the theoretical consequences of the BGC process have been known for some time, the potential practical importance of this phenomenon has remained largely unstudied. Recently, the analysis of polymorphism and nucleotide substitution patterns in primates has provided firm evidence for BGC acting genome-wide, favoring GC alleles over AT alleles (for a review, see [Duret and Galtier, 2009a]). Indeed, this process of GC-biased gene conversion (gBGC) appears to be the major determinant of the evolution of base composition at silent sites (noncoding regions, synonymous codon positions) in primate genomes [Duret and Arndt, 2008]. Further, there is now good evidence that gBGC has impacted upon the evolution of functional sequences, both in regulatory noncoding sequences [Duret and Galtier, 2009b; Galtier and Duret, 2007] and in protein-coding exons [Berglund et al., 2009; Galtier et al., 2009]. Importantly, these results indicate that, in our species' evolutionary past, gBGC is likely to have hampered the action of purifying selection and led to the fixation of deleterious mutations.

Additional Supporting Information may be found in the online version of this article.

*Correspondence to: Laurent Duret, Laboratoire de Biométrie et Biologie Evolutive, Université Claude Bernard, Lyon 1, 43 Boulevard du 11 Novembre 1918, Villeurbanne F-69622, France. E-mail: laurent.duret@univ-lyon1.fr

Contract grant sponsors: Centre National de la Recherche Scientifique; The Agence Nationale de la Recherche; Contract grant number: ANR-08-GENO-003-01.

Here we have sought to determine whether gBGC influences the frequency of deleterious nonsynonymous polymorphisms in extant human populations. To this end, we investigated the segregation patterns of AT→GC and GC→AT single nucleotide polymorphisms (SNPs) according to the local recombination rate. We also analyzed different classes of nonsynonymous SNPs, predicted to be deleterious or known to be involved in genetic disease, using synonymous and noncoding SNPs as a neutral control. All classes of SNPs were found to display the hallmarks of the gBGC process. Further, we provide evidence that these segregation patterns cannot be explained by ascertainment bias in SNP detection, artifacts in SNP orientation, or other biological processes such as natural selection. In support of these observations, we present simulations to illustrate the conditions under which gBGC can extend the persistence of deleterious mutations in finite populations. We conclude that gBGC has not only had a substantial impact on human evolution but is also highly relevant to human health and disease.

Materials and Methods

Single Nucleotide Polymorphism Data

To determine the frequency of SNPs in human populations, we used the data gathered in the HapMap Project phase III, release 27 [Frazer et al., 2007]. We analyzed data from four HapMap populations: YRI (Yoruba in Ibadan, Nigeria), JPT (Japanese in Tokyo), CHB (Han Chinese in Beijing), and CEU (Utah residents with ancestry from northern and western Europe) and we grouped the CHB and JPT samples into a single set. We analyzed only SNPs that were polymorphic in the unrelated individuals genotyped in each sample (3,566,377 total, Supp. Table S1). Ensembl annotations [Hubbard et al., 2009] were used to determine the positions of SNPs with respect to transcripts and coding sequences. Four classes of polymorphisms were retained for analysis: intergenic, intronic, protein-coding synonymous and protein-coding nonsynonymous.

As a complement, we used an independent polymorphism dataset comprising 39,440 autosomal SNPs, found exclusively in coding sequences, at both synonymous and nonsynonymous positions [Lohmueller et al., 2008]. These SNPs were determined by direct exon sequencing in 10,150 transcripts, for two population samples (hereafter termed AFR and CAU): 15 African-American individuals (30,718 SNPs) and 20 European-American individuals (22,514 SNPs, Supp. Table S2).

Inference of Ancestral and Derived Alleles

We determined the ancestral and derived states of human polymorphisms using human–chimpanzee whole-genome alignments, obtained from the UCSC Genome Browser [Rhead et al., 2010] through Galaxy [Giardine et al., 2005].

To infer the most likely ancestral and derived alleles for each SNP, we used a maximum likelihood approach that takes into account the hypermutability of CpG dinucleotides [Duret and Arndt, 2008]. Starting from whole-genome alignments of the human and chimpanzee sequences, we constructed triple alignments that included two sequences for the human population, corresponding to the two alleles observed for each SNP. The allocation of alleles to the two human sequences was performed randomly. We then inferred the ancestral sequence for the human population, thereby obtaining for each genomic position a probability distribution for the identity of the ancestral nucleotide. The ancestral nucleotide was randomly drawn according to these four probabilities. In our analysis, we included only SNPs with a constant 5′-3′ context (i.e., positions with two neighboring SNPs were removed, and we required that the human and chimpanzee nucleotides should be identical).

To confirm that this first approach had not been misled by ancestral "misinference" issues, we also used a second approach, developed by Hernandez et al. [2007a], which corrects the spectrum of derived allele frequencies, obtained by parsimonious reasoning, using a context-dependent model of sequence evolution (software kindly provided by Ryan D. Hernandez). We only considered SNPs found within a constant 5′-3′ context, as defined above. As indicated by the authors, we further restricted our dataset to positions where the chimpanzee nucleotide corresponded to one of the two alleles observed in the human population. The context-dependent site frequency spectrum obtained by maximum parsimony was then corrected using the model proposed by Hernandez et al. [2007a].

As noted previously [Gibbs et al., 2007] for disease-associated mutations, the disease-associated allele sometimes represents the ancestral state; here, we focused exclusively on SNPs for which the derived allele was associated with the disease.

SNP Sampling and Derived Allele Frequency Spectrum

The number of genotyped chromosomes varies widely between individual SNPs. The correction method developed by Hernandez et al. [2007a] requires the derived allele frequency spectrum to be constructed employing the same number of chromosomes for all SNPs. To fulfill this requirement, we applied the following procedure (as proposed by [Hernandez et al., 2007b]): we computed the minimum number of sampled chromosomes ($n_{\min}$) for a given SNP dataset and then estimated the derived allele frequencies for a dataset reduced to $n_{\min}$ chromosomes. For a SNP that was originally present in n out of m sampled chromosomes, the probability that it will be present at a frequency i in the reduced sample is given by the hypergeometric distribution:

$$\frac{C_n^i \times C_{m-n}^{n_{\min}-i}}{C_m^{n_{\min}}},$$

where C_u^v is the number of choices of v elements among u. Using this formula, we can generate the expected derived allele frequency spectrum in a subsample of $n_{\min}$ chromosomes. Note that this procedure was applied independently for each class of SNPs analyzed here (intergenic, intronic, synonymous SNPs, etc.). The $n_{\min}$ values for each SNPs sample and for each region are given in Supp. Table S5.

Recombination Rates and Hot Spots

The positions of 34,136 recombination hotspots were taken from HapMap release 21 [Myers et al., 2005], and converted from hg17 to hg18 assembly coordinates using the *liftover* utility from the UCSC Genome Browser [Rhead et al., 2010]. We also computed the regional recombination rates in 10 kb sliding windows for autosomal mutations using the genetic maps provided by [Frazer et al., 2007], release 36.

Disease-Associated Mutations

We extracted 45,751 disease-associated mutations occurring in protein-coding sequences from HGMD release 2008.3 [Stenson et al., 2009]. Using annotations from the Ensembl database

[Hubbard et al., 2009] release 49, we were able to map unambiguously onto the human genome the positions of 43,953 disease-associated mutations. A total of 193 mutations were synonymous and hence were excluded—here we only analyzed nonsynonymous mutations (34,814 missense and 8,946 nonsense).

HGMD mutations are allocated to four distinct classes with respect to their association with disease: DM, mutations regarded as being a direct cause of disease; DP, polymorphisms exhibiting a significant statistical association with disease but without additional functional evidence supporting their involvement; DFP, disease-associated polymorphisms with additional functional evidence supporting their direct involvement; FP, polymorphisms reported to affect the structure, function or expression of the gene (or gene product), but with no known disease association (Supp. Table S3).

PolyPhen Predictions

To predict which nonsynonymous SNPs present in HapMap are potentially damaging for protein structure and function, we used PolyPhen predictions for dbSNP build 126 [Sunyaev et al., 2001]. For the exon sequencing dataset, we used the PolyPhen predictions provided by the authors [Lohmueller et al., 2008] (Supp. Table S4). We focused on the SNPs predicted to be "probably damaging," for which the derived allele has been shown to be the deleterious allele in 99% of cases [Lohmueller et al., 2008].

Definition of Recombination Classes

To define regions of high and low recombination, we sorted each SNP dataset according to the minimum distance to a recombination hotspot, and then divided the dataset into three equal-sized classes. Only the first and the third classes were compared in order to maximize the crossover rate difference between the high and low recombination regions. This procedure was applied independently for each genomic region (intergenic, intronic, coding synonymous, etc.) and for each HGMD and PolyPhen subset of SNPs.

Statistical Analyses

All statistical analyses were performed with the R environment [R Development Core Team, 2008]. To test the effect of gBGC, we compared the mean derived allele frequencies (DAF) for AT→GC and GC→AT mutations. Given that the distribution of DAF is non-Gaussian, we used a randomization procedure to test the statistical significance of the mean difference [$d = \text{mean}(\text{AT}\rightarrow\text{GC}) - \text{mean}(\text{GC}\rightarrow\text{AT})$]. To do this, we randomized the direction of AT→GC and GC→AT SNPs and compared the observed d value with those obtained from 1,000 randomized datasets. We computed a P-value corresponding to the proportion of simulated datasets for which the d value was higher than that observed in the real dataset; our test was thus one tailed.

We also analyzed the difference in mean DAF between the two mutation classes (d) for regions of high and low recombination. To test if the difference in d (δd) between the two recombination classes was statistically significant, we developed a randomization procedure: we drew randomly two sets of sites (from all possible SNPs in a given genomic region), equal in size to the original low recombination and high recombination classes, and computed δd for the simulated dataset. A one-tailed p-value was computed by comparing the observed δd value with 1,000 simulated datasets.

Simulation of the Impact of gBGC on the Derived Allele Frequency Spectrum

We used simulations to determine the expected distribution of derived allele frequencies (DAF) at loci that are subject to mutation, negative selection, and biased gene conversion. The initial population was homozygous and finite following a Fisher-Wright probabilistic model with multinomial sampling, ensuring a constant population size over time. The evolution of the derived allele frequency was simulated independently for each locus. Each simulation was performed for over 20,000 generations, at the end of which the DAF of the derived allele was calculated.

The alleles that can segregate at each locus belong to one of two classes: S(trong) (G or C) or W(eak) (A or T). The fitness of genotypes SS, SW and WW are denoted respectively ω_{SS}, ω_{SW} and ω_{WW}. The mean fitness value in the population is $\bar{\omega}$:

$$\bar{\omega} = z_{SS}\omega_{SS} + z_{SW}\omega_{SW} + z_{WW}\omega_{WW}$$

where z denotes the zygotic frequencies.

For individuals that are heterozygous at a given locus (SW), we termed u the probability of conversion $S \rightarrow W$ and v the probability of conversion $W \rightarrow S$. The gene conversion bias at this site is measured through $\delta = v - u$ and has positive values when gBGC occurs. The frequency of the S allele is denoted p and hence the frequency of allele W is $1-p$. The model describes the transition from one generation, n, to the next, $n+1$, admitting panmixia, with the following equations:

$$\begin{aligned}
&\text{adults } n\text{: } f_{SS};\quad f_{SW};\quad f_{WW};\\
&\text{gametes } n\text{: } g_S = \frac{2f_{SS} + (1+\delta)f_{SW}}{2};\quad g_W = 1 - g_S\\
&\text{zygotes } n+1\text{: } z_{SS} = g_S^2;\quad z_{SW} = 2g_S g_W;\quad z_{WW} = g_W^2\\
&\text{adults } n+1\text{: } f^*_{SS} = \frac{\omega_{SS}}{\omega} z_{SS};\quad f^*_{SW} = \frac{\omega_{SW}}{\omega} z_{SW};\quad f^*_{WW} = \frac{\omega_{WW}}{\omega} z_{WW}\\
&\text{alleles } n+1\text{: } p_s = f^*_{SS} + \tfrac{1}{2} f^*_{SW};\quad p_W = 1 - p_S
\end{aligned}$$

where f represents the frequency of individuals at generation n, g the frequency of gametes at generation n, and f^* the frequency of individuals at generation $n+1$.

Here we only considered mutations that are both deleterious and recessive. We termed s the selection coefficient, so that the fitness of individuals homozygous for the mutant allele is $\omega = 1-s$. Thus, for the simulations of the fate of a newly-arisen $W \rightarrow S$ mutation in a WW population, we have $\omega_{SS} = \omega$ and $\omega_{SW} = \omega_{WW} = 1$, whereas for the simulations of the fate of a newly-arisen $S \rightarrow W$ mutation in an SS population, we have $\omega_{SS} = \omega_{SW} = 1$ and $\omega_{WW} = \omega$.

Simulations were run in populations of size $N_e = 10{,}000$ with a mutation rate of 10^{-8} mutations per base-pair per individual per generation, using different combinations of gBGC coefficient ($\delta = 0$, $\delta = 0.00013$, and $\delta = 0.0013$) and selection coefficient ($s = 0$, $s = 10^{-4}$, $s = 10^{-3}$, and $s = 10^{-2}$).

Supporting Information

The dataset used in this publication is freely available at the following Website: ftp://pbil.univ-lyon1.fr/pub/datasets/Necsulea2010

Results

gBGC Hallmarks are Observed for Deleterious SNPs

To investigate whether gBGC affects the segregation of deleterious mutations in human populations, we studied the spectrum of derived allele frequencies (DAFs) of nonsynonymous SNPs as a function of the local recombination rate across human chromosomes. We first analyzed the HapMap dataset of human SNPs, which provides frequencies of each allele in different human populations [Frazer et al., 2007]. We inferred the ancestral and derived alleles for SNPs by means of a maximum likelihood approach that incorporates CpG hypermutability [Duret and Arndt, 2008], using the chimpanzee genome as an outgroup. Three distinct subsets of nonsynonymous polymorphisms were investigated: (1) all HapMap nonsynonymous SNPs; (2) HapMap nonsynonymous mutations for which the impact on the function of the protein was predicted by PolyPhen [Sunyaev et al., 2001] to be "probably damaging"; and (3) HapMap SNPs corresponding to disease-associated nonsynonymous mutations reported in the HGMD database [Stenson et al., 2009]. We further split the HGMD dataset in order to analyze specifically those inherited mutations that are considered to be a direct cause of disease (DM), thereby excluding those mutations that have only been associated statistically with disease (Supp. Table S3). As a control, we also analyzed SNPs at silent sites, for which evidence of gBGC has already been reported [Galtier et al., 2001; Spencer et al., 2006; Webster and Smith, 2004]. As expected, DAFs were found to be negatively correlated with the strength of purifying selection: SNPs in noncoding regions or at synonymous codon positions exhibited the highest mean DAFs, whereas the lowest mean DAFs were observed for mutations that are known to be involved in genetic disease or that were predicted by PolyPhen to be deleterious (Fig. 1 and Supp. Tables S8–S10).

The gBGC model makes two firm predictions: first, in regions of high recombination, the spectrum of derived allele frequencies (DAFs) for SNPs is expected to be skewed, with higher frequencies for AT→GC than for GC→AT mutations; second, this skewing is expected to be weaker in genomic regions characterized by a lower recombination rate. To test these predictions, we classified SNPs into groups of high and low recombination on the basis of their physical distance to the nearest recombination hotspot [Myers et al., 2005]; similar results were obtained when the recombination classes were computed on the basis of the average crossover rate in fixed-size sliding windows (not shown). We found that in regions of high recombination, AT→GC mutations segregated at higher frequencies than GC→AT mutations (Fig. 2, Supp. Tables S8–S10, and Supp. Figs. S2–S4). This difference was statistically significant in all HapMap samples, both for silent SNPs and for the three sets of nonsynonymous SNPs (Table 1). This pattern was evident even within the DM subset. For this class, the tests remained significant in only one of the HapMap samples. Nevertheless, given that our observations for the more abundant classes of mutations (silent sites, nonsynonymous SNPs) were always in agreement with the gBGC hypothesis, and significantly so, the uncertainty related to the DM class is most likely only a consequence of the reduced sample size. As predicted by the gBGC model, the difference between the mean AT→GC and GC→AT frequencies is much stronger for SNPs located in regions of high recombination rate compared to SNPs located in regions of low recombination rate (Fig. 1, Table 1, and Supp. Tables S13–S15). Thus, all classes of SNPs exhibit the hallmarks of the gBGC process, not only the silent sites but also the three subsets of nonsynonymous sites.

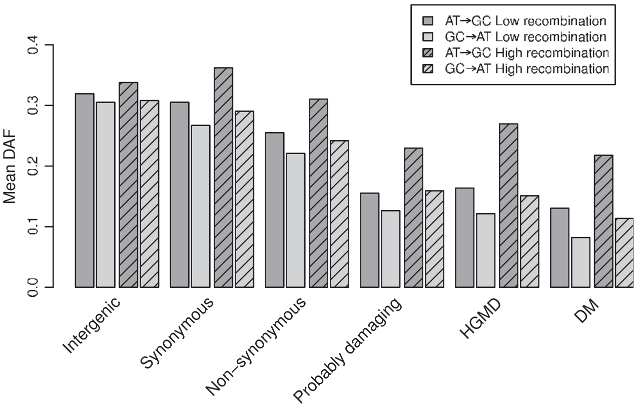

Figure 1. Mean derived allele frequencies for AT→GC and GC→AT alleles in regions of high and low recombination, for the HapMap YRI sample, for different genomic regions and classes of nonsynonymous SNPs. Dark gray: AT→GC, light gray: GC→AT mutations. Solid bars: low recombination, hatched bars: high recombination. Probably damaging: HapMap nonsynonymous SNPs predicted by Polyphen to be probably damaging. HGMD: entire HGMD dataset. DM: inherited mutations known to be a direct cause of disease (HGMD mutations minus those that have only been associated statistically with disease).

Control for Variations in Selective Pressure on Nonsynonymous Mutations

We observed that at nonsynonymous sites, GC→AT mutations segregate at lower frequency than AT→GC mutations. One potential explanation for this observation is that AT→GC nonsynonymous mutations might be, on average, less deleterious than GC→AT nonsynonymous mutations. To test this hypothesis, we compared AT→GC and GC→AT SNPs that lead to the same amino acid replacement, and hence are expected to have the exact same fitness impact. In total, there are 10 amino acid changes that can be caused both by AT→GC and GC→AT mutations. For each of the three populations, we performed pairwise comparisons of the mean DAF of AT→GC and GC→AT SNPs causing the same amino acid changes: in 23 out of 30 comparisons, the AT→GC SNP had the highest mean DAF (Supp. Table S19). For example, the mean DAF of Q→H nonsynonymous SNPs in the CEU population is 0.19 when it results from an AT→GC mutation, compared to 0.16 when it results from a GC→AT mutation. Conversely, the mean DAF of the reverse amino acid change (H→Q) is 0.35 when it results from an AT→GC mutation, compared to 0.23 when it results from a GC→AT mutation. Thus, the mean DAF varies according to the direction of the GC-content change (AT→GC vs. GC→AT), independently of the nature of the amino acid change. Hence, the observed differences in mean DAF between AT→GC and GC→AT nonsynonymous SNPs cannot be attributed to differences in selective pressure on the corresponding amino acid changes.

Control for SNP Ascertainment Bias and Ancestral Misidentification

The HapMap dataset is known to be biased toward high-frequency polymorphisms, and this representation bias can confound some population genetic analyses [Clark et al., 2005]. There is, however, no a priori reason why this ascertainment bias should differentially affect AT→GC- and GC→AT-derived allele frequencies. This notwithstanding, to ensure that our observations

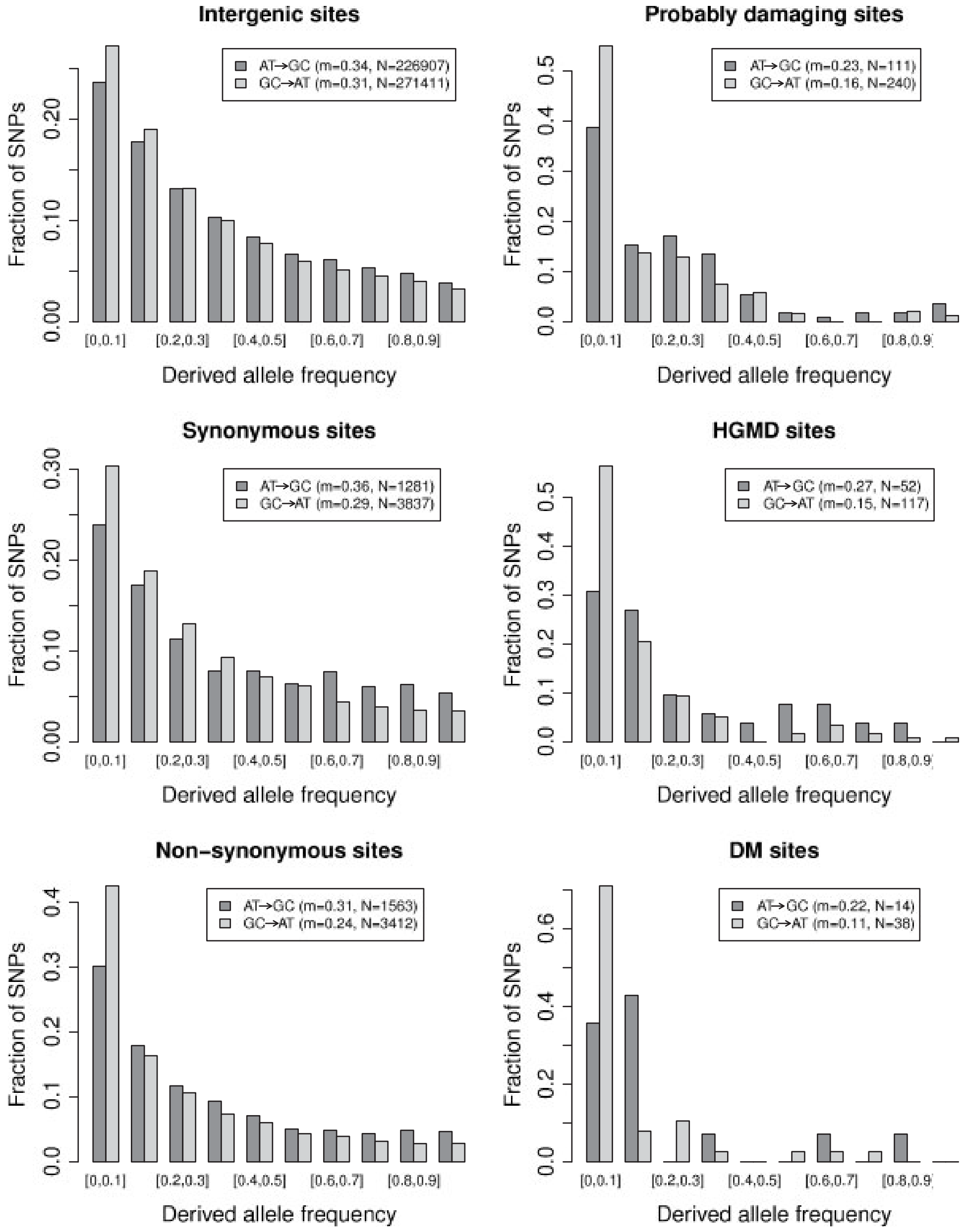

Figure 2. Derived allele frequency spectra for the HapMap YRI sample, for different genomic regions and classes of nonsynonymous SNPs. The data presented here relate only to the high recombination class. Dark gray: AT→GC mutations, light gray: GC→AT mutations.

Table 1. Summary Table for the BGC Hallmarks for the HapMap and Resequencing SNP Datasets

		Intergenic		Introns		Synonymous		Nonsynonymous		HGMD		DM		Probably damaging	
Dataset	Population	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd	d_H	Δd
HapMap	**CEU**	**0.03**	**0.013**	**0.03**	**0.022**	**0.08**	**0.043**	**0.09**	*0.016*	**0.09**	*0.052*	*0.07*	*0.031*	**0.08**	*0.055*
	CHB+JPT	**0.03**	**0.013**	**0.03**	**0.02**	**0.08**	**0.052**	**0.11**	*0.015*	**0.09**	*0.026*	**0.11**	*0.062*	**0.14**	**0.086**
	YRI	**0.03**	**0.016**	**0.03**	**0.028**	**0.07**	**0.033**	**0.07**	**0.034**	**0.12**	*0.076*	*0.1*	*0.056*	**0.07**	*0.042*
Resequencing	**AFR**					**0.1**	**0.073**	**0.06**	**0.05**	**0.1**	*0.09*	*0.05*	*0.052*	*0.02*	*0.046*
	CAU					**0.1**	**0.092**	**0.05**	**0.035**	**0.07**	*0.098*	−0.01	*0.039*	*0.04*	*0.026*

The difference in mean derived allele frequencies between AT→GC and GC→AT SNPs is denoted by d. d_H is the value of d in regions of high recombination. Δd represents the difference in d between the high and low recombination regions. Bold font: values are positive and significantly different from zero, with a P-value <0.05. Italic font: values are positive but not significantly different from zero. Normal font: values are negative but not significantly different from zero. No cases were found where d_H or Δd were significantly lower than zero.

were not affected by this intrinsic bias in HapMap data, we repeated our analysis on an independent polymorphism dataset that was acquired through direct exon resequencing in two human populations [Lohmueller et al., 2008], and which should therefore be free of ascertainment bias. Our conclusions remained unchanged with the resequencing dataset: in regions of high recombination, AT→GC mutations segregated at higher frequencies than GC→AT mutations, and this excess was higher than in regions of low recombination. This pattern was observed in both populations, not only for the synonymous sites but also for the three datasets of nonsynonymous sites (Table 1, Supp. Tables S11–S12, S16–S17, and Supp. Figs. S5–S6). We may therefore conclude that the observed skewing of derived allele frequencies was not simply a consequence of ascertainment bias. It may be noted that the pattern appears to be stronger with the HapMap dataset compared to the resequencing dataset (Table 1). By means of simulations, we showed that this difference is due to the fact that the HapMap SNP sampling strategy provides greater power to detect gBGC (see Supporting Information).

One other potential artifact that had to be considered and assessed was the possibility that the observed gBGC-like pattern stemmed from ancestral "misinference" [Hernandez et al., 2007a]: when the mutational pattern is biased toward AT, and most notably in the case of strong context dependence (such as CpG dinucleotide mutational hotspots in mammalian genomes), maximum parsimony tends to incorrectly ascribe directionality for GC→AT mutations, yielding an apparent excess of high-frequency AT→GC SNPs [Hernandez et al., 2007a]. Nevertheless, we are confident that this artifact has not influenced our results for the following reasons. First, instead of using parsimony-based reasoning, we determined SNP directionality using a maximum-likelihood approach that takes CpG hypermutability into account [Duret and Arndt, 2008]. Second, our conclusions were unchanged when CpG sites were excluded (Supp. Table S7). Third, we repeated our analyses using the context-dependent model proposed by Hernandez and colleagues [2007a] to correct for potential ancestral allele misidentification. With this method, the results remained in agreement with our previous observations (Supp. Table S6). Finally, it should be highlighted that the difference between the mean DAFs of AT→GC and GC→AT mutation was found to be much stronger in regions of high recombination (Fig. 1). This observation, which is consistent with the gBGC model, cannot be explained by an ancestral misinference artifact. Indeed, the pattern of substitution is more biased toward AT in regions of low recombination compared to regions of high recombination [Duret and Arndt, 2008]. Thus, an artifactual increase in AT→GC DAFs caused by ancestral misinference would be expected to be stronger in regions of low recombination, in contradiction to our own observations (Fig. 1).

Simulation of the Impact of gBGC in a Finite Population

To investigate the impact of gBGC on the fate of deleterious mutations (AT→GC or GC→AT), we performed simulations in a finite population (effective population size $N_e = 10{,}000$), considering recessive mutations subject to different selection coefficients (s) and gBGC coefficients (δ; see Materials and Methods section). The population-scale gBGC coefficient ($N_e\delta$) in the human genome was estimated by Spencer et al. [2006] by analyzing the DAF spectra of noncoding SNPs. In genomic regions of high recombination (defined as the top 20% of the genome with the highest recombination rate; average crossover rate = 2.5 cM/Mb) their estimate was $N_e\delta = 0.325$. Given that, in the human genome, recombination is essentially confined to hotspots (typically less than 2 kb long) with an average crossover rate of about 40 cM/Mb [Myers et al., 2006], it is expected that the gBGC coefficient should be about 16 times higher in these hotspots. Recombination hotspots vary in intensity [Myers et al., 2006]. We therefore considered two values for the population-scale gBGC coefficient: $N_e\delta = 1.3$ (for a moderate recombination hotspot) and $N_e\delta = 13$ (for a more intense recombination hotspot).

With gBGC parameters corresponding to those of a moderate human recombination hotspot, the impact of gBGC on the DAF spectrum was clearly detectable for both nearly neutral ($|N_es| = 1$) and mildly deleterious mutations ($|N_es| = 10$): compared to a situation without gBGC ($N_e\delta = 0$), AT→GC segregate at higher frequency, whereas GC→AT segregate at lower frequency (Fig. 3). For the more intense recombination hotspots, the impact of gBGC on the DAF spectrum was detectable even for highly deleterious mutations ($|N_es| = 100$). Recombination hotspots occupy only a small fraction of the genome: among the nonsynonymous SNPs that we analyzed, 6% were located within 2 kb of the center of a recombination hotspot. Thus, only a limited fraction of SNPs is expected to be affected by gBGC. This explains why the skewing observed in real data (Fig. 2) is intermediate between the patterns obtained in simulations corresponding to moderate hotspots ($N_e\delta = 1.3$) or to the absence of gBGC ($N_e\delta = 0$) (Fig. 3). Thus, the pattern observed with real data appears to be compatible with the hypothesis that the skew in the DAF spectrum is due to gBGC affecting deleterious mutations in recombination hotspots. It should be noted that the location of recombination hotspots is extremely dynamic [Baudat et al., 2010; Myers et al. 2010], which suggests that the fraction of SNPs that are at some time affected by gBGC, might be larger than that estimated above. To obtain a more realistic estimation of the expected DAF spectra, it would be necessary to take into account not only the intensity recombination hotspots but also their dynamics.

Discussion

We have shown that all functional classes of SNPs, including nonsynoynmous SNPs known to be implicated in human disease, and nonsynonymous SNPs predicted to be damaging for protein structure and function, exhibit the hallmarks of gBGC: the derived allele frequency of AT→GC mutations is higher than that of GC→AT mutations, and this is more pronounced in regions characterized by high recombination rates. Importantly, we demonstrated that the observed excess of high-frequency SNPs in regions of high recombination does not result from sampling biases nor from artifacts of SNP directionality determination.

Is gBGC the only possible explanation for these observations? One alternative hypothesis to explain the fact that nonsynonymous GC→AT mutations segregate at lower frequency than AT→GC mutations is that GC→AT mutations could be more deleterious that the AT→GC mutations. For instance, it has been recently shown that GC→AT mutations at hypermutable CpG sites within coding regions are under stronger purifying selection than other nonsynonymous mutations [Schmidt et al., 2008]. Several observations however argue against this hypothesis. First, we note that our conclusions remained unchanged when SNPs occurring within a CpG context were excluded (Supp. Table S7). Second, comparison of GC→AT and AT→GC mutations causing the same amino acid changes confirmed that the higher mean DAF of the latter cannot be attributed to a weaker impact on the encoded protein. Moreover, this hypothesis that AT→GC mutations are relatively less deleterious cannot explain why their

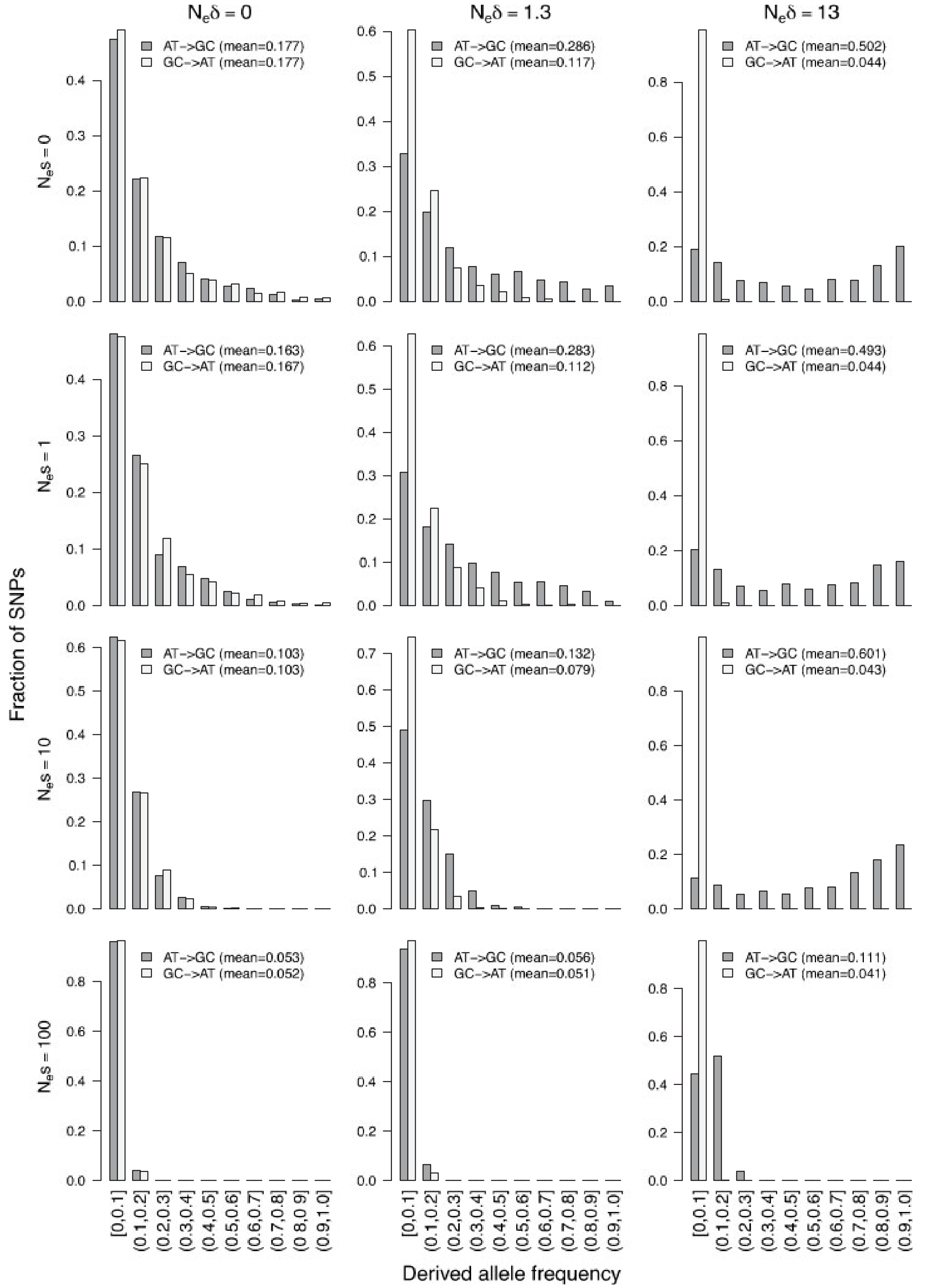

Figure 3. Derived allele frequency spectrum obtained through simulations with different parameter sets. Represented in light gray are the distributions of derived allele frequencies for GC→AT alleles, and in dark gray, those of AT→GC alleles. The population-scaled selection coefficient (N_es) and the population-scaled biased gene conversion parameter ($N_e\delta$) is indicated for each graph.

mean DAF increases with the recombination rate. Finally, we have shown that the DAF pattern is consistent over all classes of SNP, including those located in intergenic and intronic regions, which may be presumed to be largely free of selective pressure. It has been previously demonstrated that the relationship between recombination and the evolution of GC-content in noncoding regions is the consequence of gBGC and not selection [Duret and Arndt, 2008]. Hence, the most parsimonious explanation for our findings is that both silent sites and nonsynonymous sites are subject to gBGC.

Taken together, the data presented are consistent with the hypothesis that biased gene conversion is responsible for the excess of AT→GC SNPs segregating at high frequency in regions of high recombination. This result has important implications for human health because it indicates that recombination, via gBGC, leads to an increase in the frequency of disease-causing AT→GC mutations in human populations. It should be stressed that the impact of gBGC on deleterious mutations is not always negative. Indeed, a majority (58.7%) of known DMs correspond to GC→AT mutations. Thus, for a majority of DMs, gBGC acts in such a way as to limit their probability of spreading. However, the price to pay for this positive influence of gBGC is that it can lead to an increase in the frequency of disease-causing AT→GC mutations in human populations. We speculate that the genes most likely to be influenced by this effect will be those that are AT-rich (i.e., for which there are more opportunities for AT→GC mutations) and which coincide with recombination hotspots: an additional argument for these hotspots being an Achilles' heel of the human genome [Duret and Galtier, 2009b; Galtier and Duret, 2007].

Acknowledgments

All computations were performed at the IN2P3 Computing Center. We thank Nicolas Galtier, Sylvain Glémin, Alex Kondrashov, and one anonymous referee for their helpful comments; Rasmus Nielsen and Dara Torgersson for providing us with SNP data; and Ryan D. Hernandez for providing the software used to correct the DAF spectrum.

References

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B. 2010. *PRDM9* is a major determinant of meiotic recombination hotspots in humans and mice. Science 327:836–840.

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. PLoS Biol 7:e26.

Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. Nat Rev Genet 8:762–775.

Choi SK, Yoon SR, Calabrese P, Arnheim N. 2008. A germ-line-selective advantage rather than an increased mutation rate can explain some unexpectedly common human disease mutations. Proc Natl Acad Sci USA 105:10143–10148.

Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R. 2005. Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15:1496–1502.

Dean M, Carrington M, O'Brien SJ. 2002. Balanced polymorphism selected by genetic versus infectious human disease. Annu Rev Genomics Hum Genet 3:263–292.

Di Rienzo A, Hudson RR. 2005. An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21:596–601.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet 4:e1000071.

Duret L, Galtier N. 2009a. Biased gene conversion and the evolution of mammalian genomic landscapes. Annu Rev Genomics Hum Genet 10:285–311.

Duret L, Galtier N. 2009b. Comment on "Human-specific gain of function in a developmental enhancer." Science 323:714; author reply 714.

Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, Gibbs RA, Belmont JW, Boudreau A, Hardenbol P, Leal SM, Pasternak S, Wheeler DA, Willis TD, Yu F, Yang H, Zeng C, Gao Y, Hu H, Hu W, Li C, Lin W, Liu S, Pan H, Tang X, Wang J, Wang W, Yu J, Zhang B, Zhang Q, Zhao H, Zhao H, Zhou J, Gabriel SB, Barry R, Blumenstiel B, Camargo A, Defelice M, Faggart M, Goyette M, Gupta S, Moore J, Nguyen H, Onofrio RC, Parkin M, Roy J, Stahl E, Winchester E, Ziaugra L, Altshuler D, Shen Y, Yao Z, Huang W, Chu X, He Y, Jin L, Liu Y, Shen Y, Sun W, Wang H, Wang Y, Wang Y, Xiong X, Xu L, Waye MM, Tsui SK, Xue H, Wong JT, Galver LM, Fan JB, Gunderson K, Murray SS, Oliphant AR, Chee MS, Montpetit A, Chagnon F, Ferretti V, Leboeuf M, Olivier JF, Phillips MS, Roumy S, Sallée C, Verner A, Hudson TJ, Kwok PY, Cai D, Koboldt DC, Miller RD, Pawlikowska L, Taillon-Miller P, Xiao M, Tsui LC, Mak W, Song YQ, Tam PK, Nakamura Y, Kawaguchi T, Kitamoto T, Morizono T, Nagashima A, Ohnishi Y, Sekine A, Tanaka T, Tsunoda T, Deloukas P, Bird CP, Delgado M, Dermitzakis ET, Gwilliam R, Hunt S, Morrison J, Powell D, Stranger BE, Whittaker P, Bentley DR, Daly MJ, de Bakker PI, Barrett J, Chretien YR, Maller J, McCarroll S, Patterson N, Pe'er I, Price A, Purcell S, Richter DJ, Sabeti P, Saxena R, Schaffner SF, Sham PC, Varilly P, Altshuler D, Stein LD, Krishnan L, Smith AV, Tello-Ruiz MK, Thorisson GA, Chakravarti A, Chen PE, Cutler DJ, Kashuk CS, Lin S, Abecasis GR, Guan W, Li Y, Munro HM, Qin ZS, Thomas DJ, McVean G, Auton A, Bottolo L, Cardin N, Eyheramendy S, Freeman C, Marchini J, Myers S, Spencer C, Stephens M, Donnelly P, Cardon LR, Clarke G, Evans DM, Morris AP, Weir BS, Tsunoda T, Mullikin JC, Sherry ST, Feolo M, Skol A, Zhang H, Zeng C, Zhao H, Matsuda I, Fukushima Y, Macer DR, Suda E, Rotimi CN, Adebamowo CA, Ajayi I, Aniagwu T, Marshall PA, Nkwodimmah C, Royal CD, Leppert MF, Dixon M, Peiffer A, Qiu R, Kent A, Kato K, Niikawa N, Adewole IF, Knoppers BM, Foster MW, Clayton EW, Watkin J, Gibbs RA, Belmont JW, Muzny D, Nazareth L, Sodergren E, Weinstock GM, Wheeler DA, Yakub I, Gabriel SB, Onofrio RC, Richter DJ, Ziaugra L, Birren BW, Daly MJ, Altshuler D, Wilson RK, Fulton LL, Rogers J, Burton J, Carter NP, Clee CM, Griffiths M, Jones MC, McLay K, Plumb RW, Ross MT, Sims SK, Willey DL, Chen Z, Han H, Kang L, Godbout M, Wallenburg JC, L'Archevêque P, Bellemare G, Saeki K, Wang H, An D, Fu H, Li Q, Wang Z, Wang R, Holden AL, Brooks LD, McEwen JE, Guyer MS, Wang VO, Peterson JL, Shi M, Spiegel J, Sung LM, Zacharia LF, Collins FS, Kennedy K, Jamieson R, Stewart J. 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet 23:273–277.

Galtier N, Duret L, Glemin S, Ranwez V. 2009. GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet 25:1–5.

Galtier N, Piganeau G, Mouchiroud D, Duret L. 2001. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. Genetics 159:907–911.

Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, Zhang Y, Blankenberg D, Albert I, Taylor J, Miller W, Kent WJ, Nekrutenko A. 2005. Galaxy: a platform for interactive large-scale genome analysis. Genome Res 15:1451–1455.

Gibbs RA, Rogers J, Katze MG, Bumgarner R, Weinstock GM, Mardis ER, Remington KA, Strausberg RL, Venter JC, Wilson RK, Batzer MA, Bustamante CD, Eichler EE, Hahn MW, Hardison RC, Makova KD, Miller W, Milosavljevic A, Palermo RE, Siepel A, Sikela JM, Attaway T, Bell S, Bernard KE, Buhay CJ, Chandrabose MN, Dao M, Davis C, Delehaunty KD, Ding Y, Dinh HH, Dugan-Rocha S, Fulton LA, Gabisi RA, Garner TT, Godfrey J, Hawes AC, Hernandez J, Hines S, Holder M, Hume J, Jhangiani SN, Joshi V, Khan ZM, Kirkness EF, Cree A, Fowler RG, Lee S, Lewis LR, Li Z, Liu YS, Moore SM, Muzny D, Nazareth LV, Ngo DN, Okwuonu GO, Pai G, Parker D, Paul HA, Pfannkoch C, Pohl CS, Rogers YH, Ruiz SJ, Sabo A, Santibanez J, Schneider BW, Smith SM, Sodergren E, Svatek AF, Utterback TR, Vattathil S, Warren W, White CS, Chinwalla AT, Feng Y, Halpern AL, Hillier LW, Huang X, Minx P, Nelson JO, Pepin KH, Qin X, Sutton GG, Venter E, Walenz BP, Wallis JW, Worley KC, Yang SP, Jones SM, Marra MA, Rocchi M, Schein JE, Baertsch R, Clarke L, Csürös M, Glasscock J, Harris RA, Havlak P, Jackson AR, Jiang H, Liu Y, Messina DN, Shen Y, Song HX, Wylie T, Zhang L, Birney E, Han K, Konkel MK, Lee J, Smit AF, Ullmer B, Wang H, Xing J, Burhans R, Cheng Z, Karro JE, Ma J, Raney B, She X, Cox MJ, Demuth JP, Dumas LJ, Han SG, Hopkins J, Karimpour-Fard A, Kim YH, Pollack JR, Vinar T, Addo-Quaye C, Degenhardt J, Denby A, Hubisz MJ, Indap A, Kosiol C, Lahn BT, Lawson HA, Marklein A, Nielsen R, Vallender EJ, Clark AG, Ferguson B, Hernandez RD, Hirani K, Kehrer-Sawatzki H, Kolb J, Patil S, Pu LL, Ren Y, Smith DG, Wheeler DA, Schenck I, Ball EV, Chen R, Cooper DN, Giardine B, Hsu F, Kent WJ, Lesk A, Nelson DL, O'brien WE, Prüfer K, Stenson PD, Wallace JC, Ke H, Liu XM, Wang P, Xiang AP, Yang F, Barber GP, Haussler D, Karolchik D, Kern AD, Kuhn RM, Smith KE, Zwieg AS. 2007. Evolutionary and biomedical insights from the rhesus macaque genome. Science 316:222–234.

Hernandez RD, Williamson SH, Bustamante CD. 2007a. Context dependence, ancestral misidentification, and spurious signatures of natural selection. Mol Biol Evol 24:1792–1800.

Hernandez RD, Williamson SH, Zhu L, Bustamante CD. 2007b. Context-dependent mutation rates may cause spurious signatures of a fixation bias favoring higher GC-content in humans. Mol Biol Evol 24:2196–2202.

Hubbard TJ, Aken BL, Ayling S, Ballester B, Beal K, Bragin E, Brent S, Chen Y, Clapham P, Clarke L, Coates G, Fairley S, Fitzgerald S, Fernandez-Banet J, Gordon L, Graf S, Haider S, Hammond M, Holland R, Howe K, Jenkinson A, Johnson N, Kahari A, Keefe D, Keenan S, Kinsella R, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Rios D, Schuster M, Slater G, Smedley D, Spooner W, Spudich G, Trevanion S, Vilella A, Vogel J, White S, Wilder S, Zadissa A, Birney E, Cunningham F, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Kasprzyk A, Proctor G, Smith J, Searle S, Flicek P. 2009. Ensembl 2009. Nucleic Acids Res 37: D690–D697.

Kryukov GV, Pennacchio LA, Sunyaev SR. 2007. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. Am J Hum Genet 80:727–739.

Lohmueller KE, Indap AR, Schmidt S, Boyko AR, Hernandez RD, Hubisz MJ, Sninsky JJ, White TJ, Sunyaev SR, Nielsen R, Clark AG, Bustamante CD. 2008. Proportionally more deleterious genetic variation in European than in African populations. Nature 451:994–997.

Myers S, Bottolo L, Freeman C, McVean G, Donnelly P. 2005. A fine-scale map of recombination rates and hotspots across the human genome. Science 310: 321–324.

Myers S, Bowden R, Tumian A, Bontrop RE, Freeman C, MacFie TS, McVean G, Donnelly P. 2010. Drive against hotspot motifs in primates implicates the *PRDM9* gene in meiotic recombination. Science 327:876–879.

Myers S, Spencer CC, Auton A, Bottolo L, Freeman C, Donnelly P, McVean G. 2006. The distribution and causes of meiotic recombination in the human genome. Biochem Soc Trans 34:526–530.

Nagylaki T. 1983. Evolution of a finite population under gene conversion. Proc Natl Acad Sci USA 80:6278–6281.

Rhead B, Karolchik D, Kuhn RM, Hinrichs AS, Zweig AS, Fujita PA, Diekhans M, Smith KE, Rosenbloom KR, Raney BJ, Pohl A, Pheasant M, Meyer LR, Learned K, Hsu F, Hillman-Jackson J, Harte RA, Giardine B, Dreszer TR, Clawson H, Barber GP, Haussler D, Kent WJ. 2010. The UCSC Genome Browser database: update 2010. Nucleic Acids Res 38:D613–D619.

R Development Core Team. 2008. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Schmidt S, Gerasimova A, Kondrashov FA, Adzhubei IA, Kondrashov AS, Sunyaev S. 2008. Hypermutable non-synonymous sites are under stronger negative selection. PLoS Genet 4:e1000281.

Slatkin M, Rannala B. 2000. Estimating allele age. Annu Rev Genomics Hum Genet 1:225–249.

Spencer CC, Deloukas P, Hunt S, Mullikin J, Myers S, Silverman B, Donnelly P, Bentley D, McVean G. 2006. The influence of recombination on human genetic diversity. PLoS Genet 2:e148.

Stenson PD, Mort M, Ball EV, Howells K, Phillips AD, Thomas NS, Cooper DN. 2009. The Human Gene Mutation Database: 2008 update. Genome Med 1:13.

Sunyaev S, Ramensky V, Koch I, Lathe3rd W, Kondrashov AS, Bork P. 2001. Prediction of deleterious human alleles. Hum Mol Genet 10:591–597.

Webster MT, Smith NG. 2004. Fixation biases affecting human SNPs. Trends Genet 20:122–126.