Quantification of GC-biased gene conversion in the human genome

Sylvain Glémin,[1,2] Peter F. Arndt,[3] Philipp W. Messer,[4] Dmitri Petrov,[5] Nicolas Galtier,[1] and Laurent Duret[6]

[1]*Institut des Sciences de l'Evolution (ISEM - UMR 5554 Université de Montpellier-CNRS-IRD-EPHE), 34095 Montpellier, France;* [2]*Department of Ecology and Genetics, Evolutionary Biology Centre, Uppsala University, SE-752 36 Uppsala, Sweden;* [3]*Department of Computational Molecular Biology, Max Planck Institute for Molecular Genetics, 14195 Berlin, Germany;* [4]*Department of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA;* [5]*Department of Biology, Stanford University, Stanford, California 94305-5020, USA;* [6]*Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, 69622 Villeurbanne, France*

Much evidence indicates that GC-biased gene conversion (gBGC) has a major impact on the evolution of mammalian genomes. However, a detailed quantification of the process is still lacking. The strength of gBGC can be measured from the analysis of derived allele frequency spectra (DAF), but this approach is sensitive to a number of confounding factors. In particular, we show by simulations that the inference is pervasively affected by polymorphism polarization errors and by spatial heterogeneity in gBGC strength. We propose a new general method to quantify gBGC from DAF spectra, incorporating polarization errors, taking spatial heterogeneity into account, and jointly estimating mutation bias. Applying it to human polymorphism data from the 1000 Genomes Project, we show that the strength of gBGC does not differ between hypermutable CpG sites and non-CpG sites, suggesting that in humans gBGC is not caused by the base-excision repair machinery. Genome-wide, the intensity of gBGC is in the nearly neutral area. However, given that recombination occurs primarily within recombination hotspots, 1%–2% of the human genome is subject to strong gBGC. On average, gBGC is stronger in African than in non-African populations, reflecting differences in effective population sizes. However, due to more heterogeneous recombination landscapes, the fraction of the genome affected by strong gBGC is larger in non-African than in African populations. Given that the location of recombination hotspots evolves very rapidly, our analysis predicts that, in the long term, a large fraction of the genome is affected by short episodes of strong gBGC.

[Supplemental material is available for this article.]

The process of GC-biased gene conversion (gBGC) has a major impact on the evolution of mammalian genomes (Duret and Galtier 2009; Romiguier et al. 2010; Katzman et al. 2011) and is known or suspected to a play a role in many other groups of eukaryotes (Webster et al. 2006; Escobar et al. 2011; Pessia et al. 2012; Serres-Giardi et al. 2012). gBGC is a recombination-associated process favoring G:C (S for strong, hereafter) over A:T (W for weak, hereafter) bases during the repair of mismatches that occur within heteroduplex DNA during meiotic recombination (Marais 2003; Lesecque et al. 2013). From a population genetics point of view, gBGC is equivalent to natural selection in favor of S alleles, increasing their frequency and probability of fixation (Nagylaki 1983). gBGC therefore tends to increase GC content and W → S substitution rates in highly recombining regions.

There are at least two reasons why we should be concerned about gBGC. First, as recombination rate is highly heterogeneous across the genome and most recombination events occur in evolutionarily short-lived hotspots (Myers et al. 2005; Ptak et al. 2005; Winckler et al. 2005; Coop and Myers 2007; Auton et al. 2012), gBGC-induced GC-enrichment is expected to occur through short, localized episodic events. Such a sudden locus- and lineage-specific acceleration of substitution rates can easily mimic the signature of positive selection (Galtier and Duret 2007; Berglund et al. 2009; Ratnakumar et al. 2010; Kostka et al. 2012). Accordingly, it was estimated that up to 20% of signatures of positive selection in the human genome could be explained by gBGC (Ratnakumar et al. 2010). Clearly, the effects of gBGC must be taken into account seriously in studies of molecular adaptation in humans, mammals, and other taxa.

Secondly, gBGC can actually oppose natural selection. This occurs when the S allele is less favorable for the fitness than the W allele. In this case, gBGC tends to maintain deleterious alleles at intermediate or high frequency in populations, possibly until fixation, depending on selection and dominance coefficients (Glémin 2010). Accordingly, gBGC tracts are enriched in disease-associated polymorphisms (Capra et al. 2013), and W → S disease-causing mutations segregate at higher frequency than S → W mutations (Necsulea et al. 2011; Lachance and Tishkoff 2014). High rates of fixation of nonsynonymous, likely deleterious, mutations are also associated with gBGC episodes in primates (Galtier et al. 2009).

The magnitude of the above-mentioned effects strongly depends on the intensity of gBGC that can be measured by the population-scaled coefficient $B = 4N_e b$, where N_e is the effective

Corresponding author: sylvain.glemin@univ-montp2.fr

Article published online before print. Article, supplemental material, and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.185488.114.

population size and b is the intensity of the conversion bias (Nagylaki 1983). Similar to selection, gBGC is only considered to be effective, in that it dominates over random genetic drift, if B is substantially greater than one. For example, the magnitude of gBGC-induced deleterious effects depends on the distribution of B values relative to selection: The occurrence of strong gBGC episodes in a few hotspots is a more harmful situation than homogeneous but low gBGC level (Glémin 2010). For a proper assessment of the impact of gBGC on genome evolution, it is therefore essential to accurately quantify the B parameter.

Previous studies have used substitution patterns along phylogenetic lineages to estimate the intensity of gBGC. On average over the whole genome, gBGC was found to be relatively weak $B = 0.2$–0.36 (Lachance and Tishkoff 2014). However, based on the estimated proportion of recombination hotspots, Duret and Arndt (2008) evaluated that an average gBGC intensity of $B = 5$–6.5 in these hotspots is required to explain the patterns of substitution rates in the human lineage. Recently, Lartillot (2013b) developed a Bayesian method that directly estimates B along a phylogeny, incorporating variations both among branches and among genes. Analyzing sets of exons at the scale of the mammalian phylogeny, he showed that B could reach average values of about 5 in small-sized mammalian lineages that have high effective population size, with a small percentage of exons evolving under very strong gBGC ($B > 10$). He also confirmed that gBGC is weaker in the human lineage, and more generally in primates, than in small-sized, short-lived mammals, which can explain the erosion of GC-rich isochores in this group (Duret et al. 2002, 2006). Capra et al. (2013) also developed a phylogenetic method to capture gBGC heterogeneity and detect gBGC tracts, which they applied to the human and chimp genomes. However, these authors did not quantify the intensity of gBGC in these tracts. In fact, their method requires fixing the value of B expected in hotpots (they used $B =$ 3). These two methods were successful in capturing (part of) the heterogeneity of gBGC genome-wide, but they describe and quantify the process over millions of years of evolution. Because recombination hotspots, and hence also gBGC hotspots, have a very short lifespan (Ptak et al. 2005; Winckler et al. 2005; Auton et al. 2012; Lesecque et al. 2014), the intensity of gBGC currently experienced by the human population cannot be properly estimated by the methods described above.

Estimates of gBGC in more recent time periods can, in principle, be obtained from polymorphism data by fitting models of gBGC to the site frequency spectra (SFS) of W → S and S → W mutations (hereafter denoted WS and SW, respectively). Within this framework, Spencer et al. (2006) estimated $B = 1.3$ for the 20% highest recombination fraction of the human genome. However, several methodological issues have not been considered in their approach. First, as demography also affects SFS, it must be taken into account in inference approaches. This can be achieved by incorporating a demographic scenario into the model (usually a simple change in population size is used) (Eyre-Walker et al. 2006; Boyko et al. 2008) or by adding noise parameters to account for the nonselective factors that affect the shape of the SFSs (Eyre-Walker et al. 2006, and see below). Second, errors in the polarization of mutations into ancestral and derived alleles, especially because of homoplasy due to CpG hypermutability, are known to affect the SFS, which can lead to spurious signatures of gBGC (Hernandez et al. 2007). One way to circumvent this problem is to use folded spectra, in which mutations are not polarized. However, gBGC intensity can be estimated from the shape of the folded SFS only under the assumption of mutation/gBGC/drift balance equilibrium (Smith and Eyre-Walker 2001). When this assumption is relaxed, derived allele frequency (DAF) spectra are required to disentangle mutation bias and gBGC. Recently, De Maio et al. (2013) combined polymorphism and divergence data in a global framework to both correct for polarization errors due to CpG and distinguish mutation bias from gBGC (De Maio et al. 2013). However, they assumed a constant population size in their model. Finally, a third issue concerns the dynamics of gBGC episodes. Both Spencer et al. (2006) and De Maio et al. (2013) found rather low values of gBGC (maximum around $B = 1$), but they did not properly take gBGC heterogeneity into account, and it is not clear how this affects B estimates.

Here, we propose a new framework for estimating the intensity of gBGC that solves the issues discussed above. Though mispolarization due to CpG hypermutability has been taken into account previously, we show by simulations that another important effect of mispolarization has been consistently overlooked in SFS-based estimates of gBGC but can be fully corrected within the framework of our method. We also show that strong heterogeneity in B can lead to its underestimation and develop an extension of our approach that accounts for this problem. We apply our inference method to the African (AFR), European (EUR), and East Asian (EAS) populations of the 1000 Genomes data set (The 1000 Genomes Project Consortium 2012) to quantify gBGC and its variation across the human genome and analyze the effect of local recombination rate on these variations.

Results

Genome-wide signature of gBGC in the human genome

To investigate fixation biases affecting WS and SW mutations in the human genome, we first analyzed SNP data from the AFR population of the 1000 Genomes Project (The 1000 Genomes Project Consortium 2012). We selected all SNPs located in noncoding regions (i.e., presumably neutrally evolving SNPs) from autosomes. We excluded sex chromosomes to avoid biases due to their specific features—both in terms of mutation pattern and demography. Mutations were polarized using ancestral state predictions based on four-way multiple alignments (*Homo sapiens*, *Pan troglodytes*, *Pongo pygmaeus*, *Macaca mulatta*) (Paten et al. 2008), which are provided in the original SNP data file (The 1000 Genomes Project Consortium 2012). We excluded SNPs for which information about the ancestral state was reported as being unreliable (see Methods).

We first focused our analyses on non-CpG SNPs. In agreement with previous reports (Katzman et al. 2011), we observed that, on a genome-wide scale, the DAF spectra of WS SNPs are significantly biased toward higher frequencies compared with the DAF spectra of SW SNPs (Fig. 1A). As predicted by the gBGC model, this shift in DAF spectra is much stronger for SNPs located in regions of high recombination (Fig. 1C) compared to SNPs located in regions of low recombination (Fig. 1B), which is in agreement with previous analyses (Katzman et al. 2011; Lachance and Tishkoff 2014). The difference in mean DAF between WS and SW non-CpG mutations increases steadily with increasing recombination rate, from almost 0 (as expected in the absence of gBGC and selection) to ~3.5% (Fig. 1D).

Noncoding sequences are not entirely neutral: Overall, ~8% of the human genome is under negative selective pressure (Rands et al. 2014). To test whether selection could affect the shift in DAF spectra that we observed between WS and SW mutations,

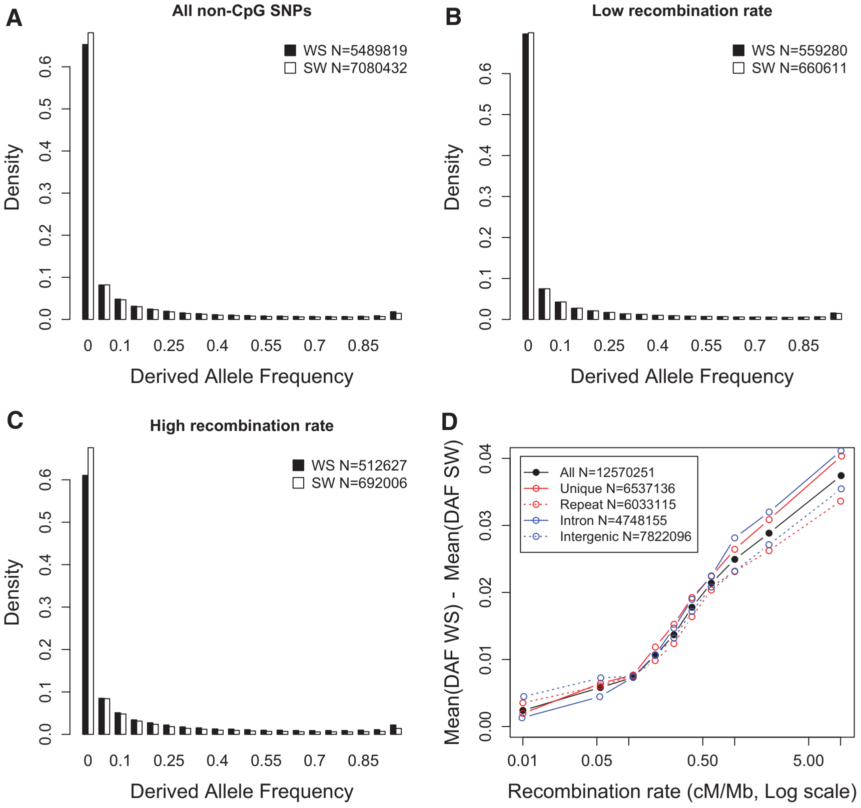

Figure 1. Variations in derived allele frequencies (DAF) according to mutation type (WS or SW) and local recombination rate: non-CpG sites. SNP allele frequencies and polarizations were retrieved from the 1000 Genomes phase 1 data set (population panel: AFR). We selected all non-CpG SNPs located in noncoding regions from autosomes. Recombination rates are measured over 5-kb windows centered on each SNP, using HapMap data (Myers et al. 2005). (*A*) DAF spectra of all SNPs. (*B*) DAF spectra of the subset of SNPs located in regions of low recombination (bottom 10%). (*C*) DAF spectra of the subset of SNPs located in regions of high recombination (top 10%). (*D*) Differences in mean DAF spectra between WS and SW mutations, according to the local recombination rate. These differences are displayed for the entire set of SNPs, for the subsets of SNPs located in introns versus intergenic regions (blue) or in unique versus repeat sequences (red). The values in the *insets* indicate the genome-wide count of SNPs of each category.

we analyzed separately noncoding SNPs located in unique sequences and in repeat sequences (for which there is strong evidence that they evolve essentially neutrally [Lunter et al. 2006]). We also compared SNPs located in introns and intergenic regions. In all cases, we observed very similar patterns (Fig. 1D). This indicates that the shift in DAF spectra between WS and SW mutations is driven by a process that affects all genomic compartments (as predicted by the gBGC model) and that the impact of selection on the observed pattern is (if anything) very limited.

Signatures of gBGC in DAF spectra are obscured by (unexpected) polarization artifacts

The difference in DAF spectra between WS and SW mutations provides information about the intensity of gBGC. We previously developed a generic maximum-likelihood model that allows one to quantify the strength of gBGC from the comparison of the DAF spectra of WS and SW mutations, using the DAF spectrum of WW and SS mutations as a neutral reference (Muyle et al. 2011). For completeness, this method is summarized in Supplemental Text S1. One important difficulty with this approach is that the estimation of DAF spectra remains highly sensitive to polarization errors: Any WS (respectively, SW) mutation observed at frequency $x = i/n$ in the sample that is mispolarized is considered as a SW (WS) mutation at frequency $(n - i)/n$. Given that the majority of derived alleles are rare (i.e., x is generally much smaller than 0.5), polarization errors shift the inferred DAF spectra toward higher frequencies. And, given that the SW mutation rate is higher than the rate of WS mutation, the risk of mispolarization is higher for SW mutations (which are then erroneously counted as WS mutations) (Eyre-Walker 1998). Hence, this polarization artifact leads to overestimating the fixation bias in favor of WS mutations (Supplemental Text S1; Hernandez et al. 2007). This artifact is expected to be particularly strong at hypermutable CpG sites, where the inference of the ancestral state is less reliable, and indeed, CpG sites show very peculiar DAF spectra, with a strong peak of WS SNPs segregating at very high frequency (Fig. 2A). One possible interpretation is that gBGC might be much stronger on CpG than on non-CpG sites. However, this peak is observed regardless of recombination rate (Fig. 2B,C), and the difference in mean DAF between WS and SW mutations is very high (~8%) even in regions of very low recombination (Fig. 2D). All these observations indicate that the strong excess of WS CpG SNPs segregating at very high frequency is not due to gBGC.

To assess the impact of polarization errors on DAF spectra and on estimators of gBGC strength, we performed extensive simulation analyses (see details in Methods). Simulation parameters were set so as to mimic the situation observed in the human genome, where we estimate that the polarization error rate is ~1%–4% when using the polarization provided by the 1000 Genomes data (see below). In the human genome, as in other mammals, the base composition varies strongly along chromosomes and generally does not correspond to the mutational equilibrium (Duret and Arndt 2008). We therefore simulated genomes composed of sequences of different GC-content, subject to the same mutational bias ($\lambda = 2$). We simulated both genomes with gBGC (with stronger gBGC in regions of higher GC-content) and genomes not subject to any gBGC.

Our simulations revealed both expected and unexpected patterns. As expected, and in agreement with previous reports (Hernandez et al. 2007), gBGC is overestimated when the polarization error rate is higher for SW mutations than for WS mutations (typically as for CpG sites) (Supplemental Fig. S1). In principle, it is possible to use more reliable methods that take into account CpG hypermutability to provide unbiased estimates of ancestral states (e.g., Duret and Arndt 2008; De Maio et al. 2013). Such methods do not prevent polarization errors but ensure that error rates are symmetrical (i.e., the rate of polarization error is the same for WS and SW mutations: $e_{WS} = e_{SW} > 0$). However, our simulations

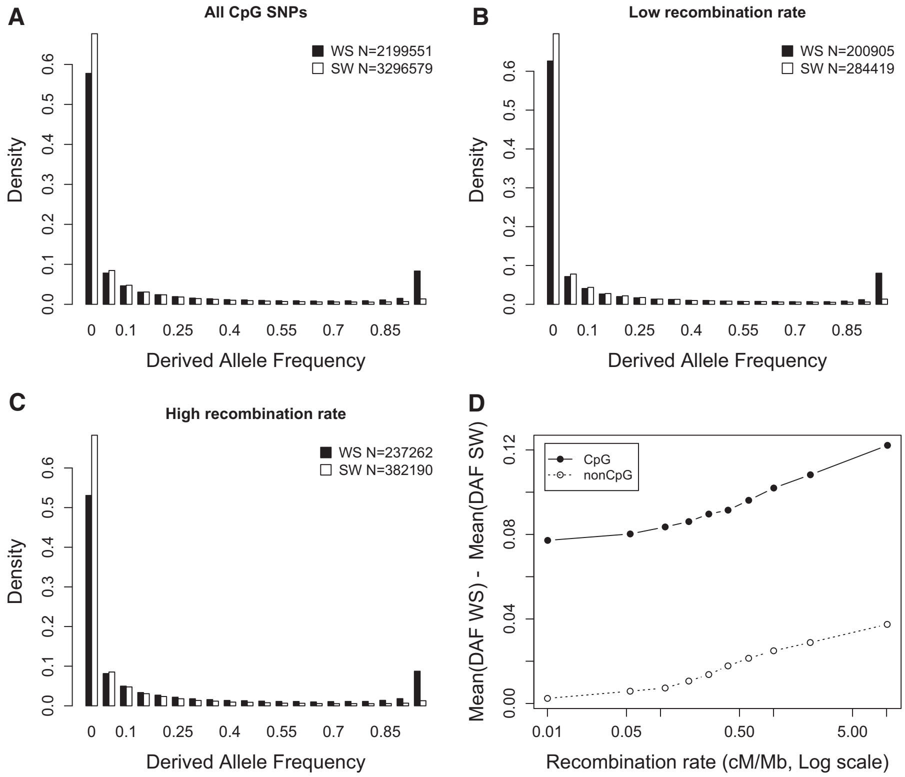

Figure 2. Variations in derived allele frequencies according to mutation type (WS or SW) and local recombination rate: CpG sites. (*A–C*) Information similar to that given in Figure 1, but here for CpG sites only. (*D*) Differences in mean DAF spectra between WS and SW mutations, according to the local recombination rate. For a comparison, both CpG and non-CpG SNPs (same as Fig. 1D) are presented.

show that even when polarization error rates are symmetrical, estimates of *B* are biased. This bias leads to a spurious positive relationship between *B* and the local GC-content and can even lead to the inference of negative average *B* values (Fig. 3A,B). This unexpected result is explained by the fact that, assuming constant mutation rates and probabilities of polarization error, the ratio of the number of WS to SW mutations, increases with GC content, and so does the ratio of the number of WS to SW mispolarized mutations. This is so because we modeled situations departing from the mutational equilibrium, $p_{GC} = 1/(1 + \lambda)$. When GC content is higher (respectively, lower) than the mutation equilibrium, there are more (respectively, less) mispolarized WS than SW mutations and *B* is over- (respectively, under-) estimated. The bias is only suppressed when there is an equal number of WS and SW mutations, i.e., when the base composition closely reflects the mutational equilibrium.

It is therefore crucial to take this bias into account for any method based on DAF spectra that distinguish between WS and SW polymorphisms.

Correcting for polarization error in estimating the intensity of gBGC: a new method

Several methods have been developed to cope with polarization errors, especially to take CpG hypermutability into account (Hernandez et al. 2007; Duret and Arndt 2008; De Maio et al. 2013). However, although these methods suppress the bias in the inference of ancestral states, symmetrical polarization errors remain, and our simulations clearly showed that even unbiased mispolarization is problematic as far as SFS analysis is concerned. One possible solution to circumvent the problem is to remove CpG sites. However, this leads to bias sampling toward SNPs located in GC-poor regions (see Discussion). Here, we propose an alternative approach that incorporates polarization error rates directly into the estimation procedure. This is a priori possible because the impact of polarization errors on the shape of DAF spectra is very different from the impact of gBGC (see Supplemental Text S2). The rationale of the method is the same as for the generic model described in Muyle et al. (2011) (see Supplemental Text S1), except that, here, the probability of observing k_i SNPs having *i* derived alleles out of *n* follows a Poisson distribution, $P(\mu, k_i)$, with mean:

$$\mu^{\mathrm{obs}}_{\mathrm{neutral}}(i) = (1 - e_{\mathrm{neutral}})\mu_{\mathrm{neutral}}(i) + e_{\mathrm{neutral}}\mu_{\mathrm{neutral}}(n - i) \quad \text{for neutral SNPs,} \tag{1a}$$

$$\mu^{\mathrm{obs}}_{W\to S}(i) = (1 - e_{\mathrm{WS}})\mu_{W\to S}(i) + e_{\mathrm{SW}}\mu_{S\to W}(n - i) \quad \text{for WS SNPs,} \tag{1b}$$

$$\mu^{\mathrm{obs}}_{S\to W}(i) = (1 - e_{\mathrm{SW}})\mu_{S\to W}(i) + e_{\mathrm{WS}}\mu_{W\to S}(n - i) \quad \text{for WS SNPs,} \tag{1c}$$

where the "true" μ's are given by equation (S1.1) (see Supplemental Text S1) and e_{neutral}, e_{WS}, and e_{SW} are polarization error probabilities, which are estimated jointly with the other parameters of the model. We thus have four possible models: $B = 0$ (M0) and $B \neq 0$ (M1), without error correction, and the same with error correction (M0* and M1*). The four models can be compared by a likelihood ratio test (LRT) with the appropriate degrees of freedom (see Supplemental Table S1). The goodness of fit of these

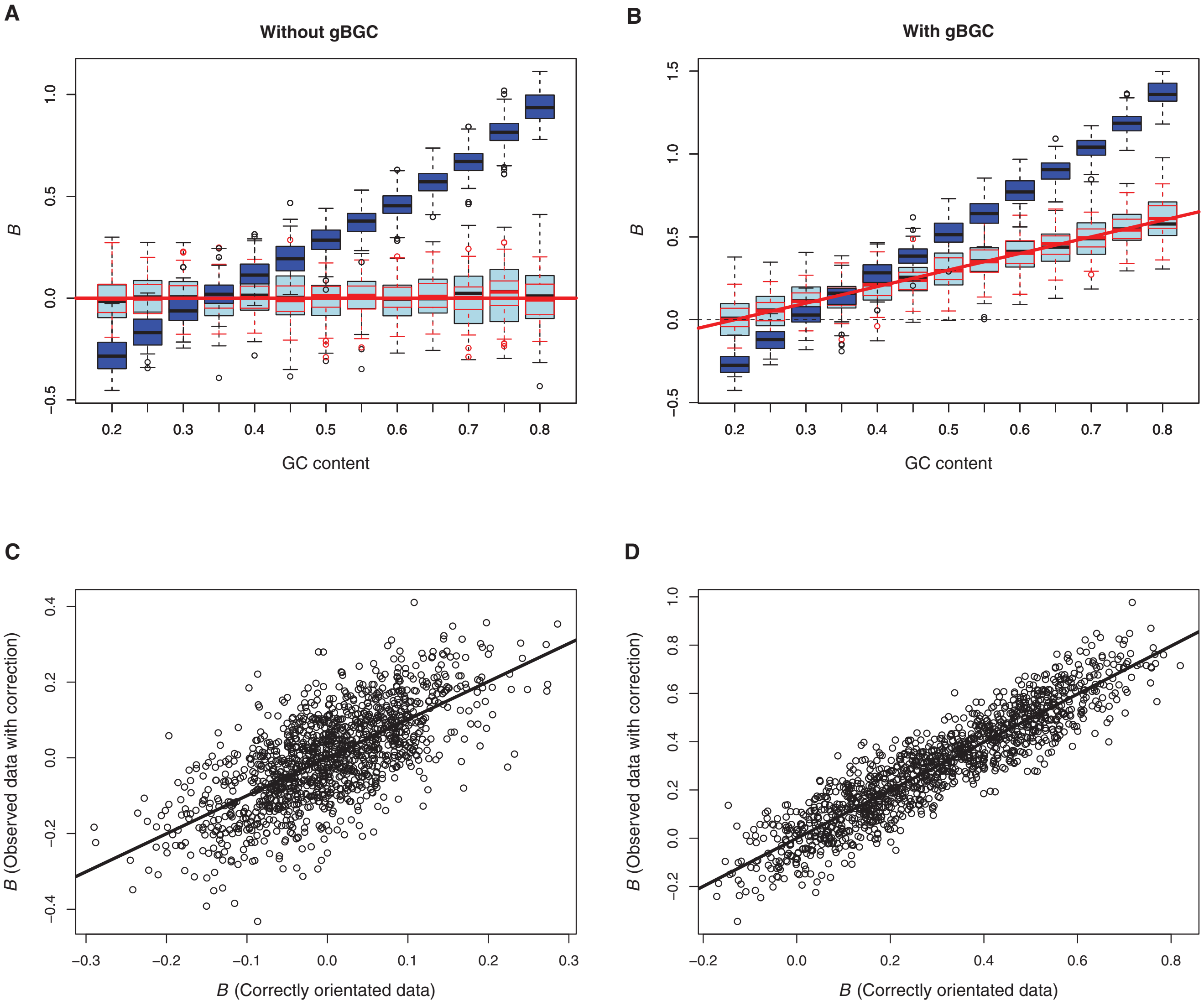

Figure 3. Effect of polarization errors on *B* estimates and accuracy of the correction method. Estimation of *B* as a function of GC content in simulated data sets: $B = 0$ for any GC content (*A*,*C*), and *B* linearly increases with GC content (*B*,*D*) (red lines in *A*,*B*). For a given simulated data set (= correctly orientated data, red lines), polarization errors ($e_{\text{neutral}} = e_{WS} = e_{SW} = 0.03$) were secondarily added (= observed data). *B* values were estimated using the model without error correction (M1) (see main text and Supplemental Table S1) and with error correction (M1*). Box plots correspond to the *B* estimates of 100 correctly orientated data sets using the M1 model (red), 100 observed data sets using the M1 model (dark blue), 100 observed data sets using the M1* model (light blue). Figure parts *C* and *D* show the correlation between *B* values estimated from correctly orientated data with the M1 model and *B* values estimated from observed data with the M1* model. The regression line is indistinguishable from the diagonal $y = x$.

models can then be assessed by comparison with the likelihood of the saturated model, in which every class of each SFS has its own parameter.

We evaluated our new method under different conditions (symmetrical versus asymmetrical error rates, stable versus non-equilibrium populations). Simulations show that our method performs well in all tested conditions and accurately recovers the true simulated value of *B* (Fig. 3; Supplemental Figs. S1, S2). We compared the M1* model applied to data sets with polarization errors to the M1 model applied to the same data sets without errors. The two estimates of *B* were very well correlated with no bias (the regression line was indistinguishable from the $y = x$ line) (Fig. 3C, D). We also checked the accuracy of the estimation of polarization error rates. These estimates suffer from a large variance. Because error rates are low and bounded to zero, large variance tends to increase error rate estimates on average. As a consequence, the mean estimate of e_{WS} (respectively, e_{SW}) tended to slightly increase (respectively, decrease) with GC content. Once again, this is explained by the fact that the number of WS (respectively, SW) mutations, and hence the power to estimate e_{WS} (respectively, e_{SW}), decreases (respectively, increases) with GC-content (see Supplemental Fig. S3). However, this bias did not affect the estimation of *B*, as shown above.

A moderate genome-average gBGC intensity in humans

We applied our method on the AFR population of the 1000 Genomes data set (The 1000 Genomes Project Consortium

2012), using all noncoding SNPs (whole data set), only non-CpG SNPs (non-CpG data set), and only CpG SNPs (CpG data set). Several parameters of the model (mutation rates, gBGC strength, polarization error rates) are susceptible to variations along chromosomes. We therefore performed parameter estimations individually on 1-Mb-long windows (nonoverlapping) across the genome. For a few windows, one or more models did not converge, and these windows were excluded from the analyses. The final numbers of windows for the three data sets are 2669, 2665, and 2644, respectively. Each window was characterized by its average GC-content and recombination rate. Because local GC-content and recombination rates can be different between CpG and non-CpG sites within a given window, for each data set we computed GC-content at 100 bp and a recombination rate at 5 kb around each SNP and averaged these values over the SNPs of each window.

Over the whole data set, estimates of B obtained with model M1* ranged from −0.70 to 2.06 with a median of 0.35 and a mean of 0.38 (Fig. 4). A negative B was estimated for only 232 out of 2669 windows, and only 11 (5%) were significantly different from 0. In contrast, over the 2437 windows with a positive B, 1458 (60%) were significantly different from 0 (Fig. 4). As expected, B was strongly correlated with the recombination rate in the 1-Mb window ($R^2 = 0.32$). We also observed a correlation between B and GC content ($R^2 = 0.14$). Multiple regression analysis showed that this correlation is essentially due to the known correlation between recombination rate and GC-content ($R^2 = 0.18$): The variance of B explained by GC-content and recombination together ($R^2 = 0.33$) was only slightly greater than that explained by recombination alone ($R^2 = 0.32$). Given that the density in CpG sites increases with GC-content, CpG SNPs are enriched in GC-rich genomic regions and, hence, in regions of high recombination. Consistently, on average, B was higher for CpG sites compared to non-CpG sites (Table 1). However, for a given recombination rate and local GC content there was virtually no difference in strength of gBGC between CpG and non-CpG sites (Fig. 5). Although the effect of the category of sites was still significant after error correction (because of the size of the data set), it only explained 2% of the variance in B when polarization error was correctly accounted for (but 39%, otherwise) (see Fig. 5).

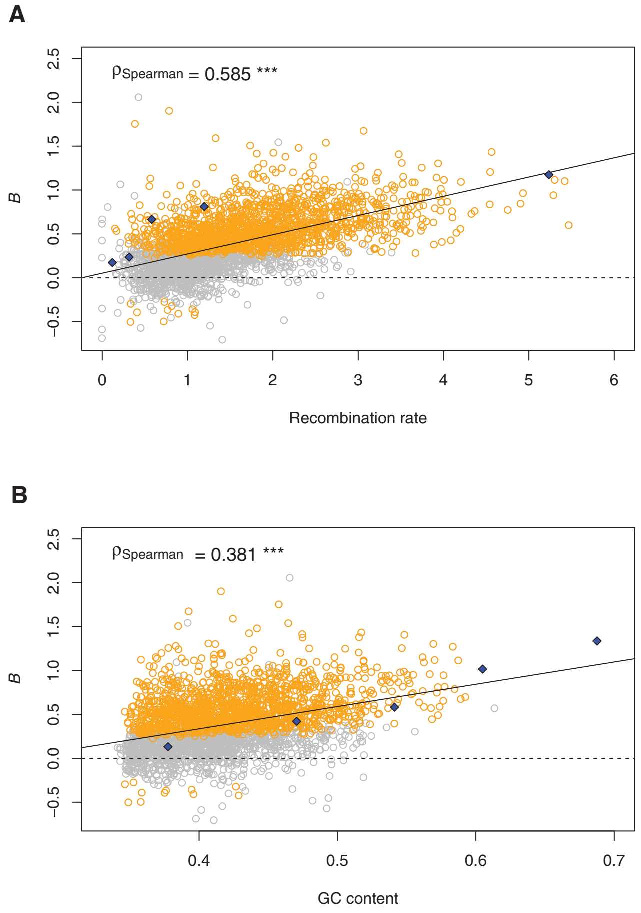

Figure 4. B estimates for 1-Mb windows as a function of recombination rate and GC content. Values of B were estimated on autosomes with the M1* model. Gray (respectively, orange) dots correspond to B values nonsignificantly (respectively, significantly) different from 0. The regression lines and the Spearman correlation coefficients are given in the plots. (***) P-values < 10^{-15}. P-values were computed by a likelihood ratio test with 1 degree of freedom between models M1* and M0*. The blue diamonds correspond to estimates of B on synonymous sites grouped into five recombination rate (A) or GC content (B) quintiles. To be congruent with Figure 5, GC-content was measured over 100 bp and recombination rate over 5 kb around each SNP and then averaged over each window.

Table 1. Estimates of average polarization error rates and gBGC strength (B) for models M1* and M1

				B	
Sites	e_{SW}	e_{WS}	e_{neutral}	M1*	M1
All	1.8%	1.0%	0.8%	0.38	0.55
Non-CpG	0.7%	1.0%	0.9%	0.33	0.32
CpG	4.1%	0.7%	0.6%	0.50	1.00

Estimates were obtained with model M1* on 1-Mb-long genomic windows and compared with estimates of B obtained without correction for polarization errors (model M1). SNP data set: AFR. Polarization error rates: (e_{SW}) SW mutations, (e_{WS}) WS mutations, (e_{neutral}) SS + WW mutations.

Quantification of polarization errors in human SNP data sets

Our method estimated an average polarization error rate of ~4% for SW mutations at CpG sites and 0.6%–1% for other categories of mutations and sites (Table 2; Supplemental Figs. S4–S6). As a negative control, we also grouped the 0.7% of SNPs at CpG sites located in CpG islands (which are generally unmethylated and less mutable) and applied model M1*. We found a lower polarization error rate for SW mutations (1%), similar to that estimated for other mutations. All these rates are consistent with the expected rate of homoplasy along the chimpanzee branch, given the branch length between human and chimp. This suggests that the method accurately estimates error probabilities, on average, despite the fact that no prior information was included in the model. As predicted by simulations, there was a slight effect of GC content on error estimates: The variance and the mean of e_{WS} increased with GC content, the variance of e_{SW} decreased with GC content, while there was no effect of GC content on e (Supplemental Fig. S4). Although these error rates are relatively low, they have a strong impact on the quantification of gBGC: On the whole data set, the estimate of B is 49% higher when ignoring polarization errors than when these errors are modeled, and this overestimate reached 96% for CpG sites (Table 2). Importantly, the difference in estimates of B between CpG and non-CpG sites disappeared when polarization errors were accounted for (Fig. 5). As predicted by

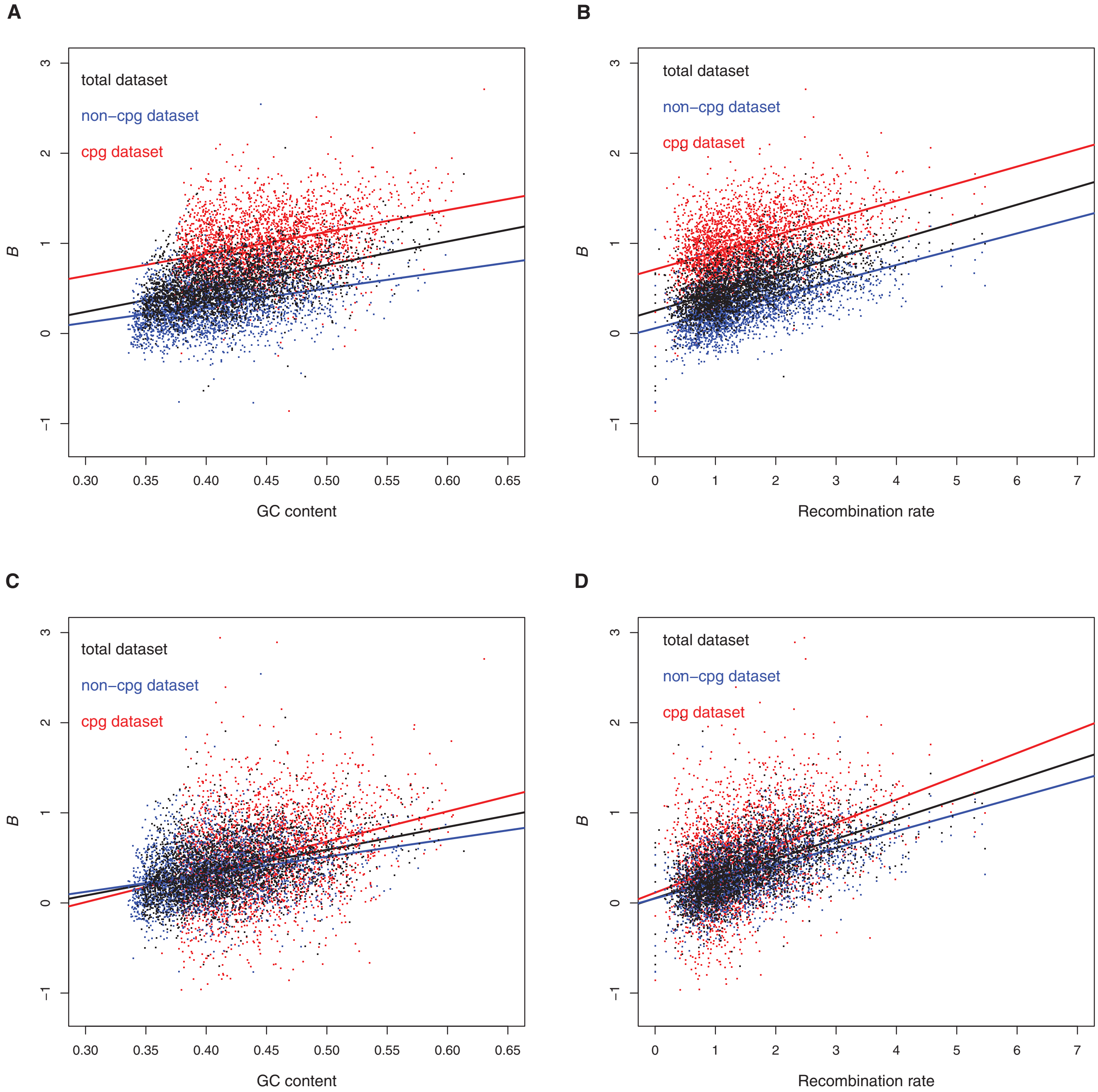

Figure 5. Comparison of the *B* estimated with and without error correction. Values of *B* estimated without error correction (M1 model, *A,B*) and with error correction (M1* model, *C,D*) as a function of GC content (*A,C*) and recombination rate (*B,D*) for the whole data set (black), the non-CpG data set (blue), and the CpG data set (red). To take into account differences in local GC-content and recombination rates between CpG and non-CpG sites in the same window, we measured GC-content over 100 bp and recombination rate over 5 kb around each SNP and then averaged them over each window. When the non-CpG and the CpG data sets are analyzed jointly, recombination rate, GC-content, and the category of sites explain, respectively, 11%, 17%, and 38% of the variance in *B* without error correction, and 16%, 4%, and 1% with error correction.

simulations, the correlation between *B* and GC content was lower when error correction was applied ($R^2 = 0.14$) than without correction ($R^2 = 0.21$). On the contrary, error correction did not affect the correlation between *B* and recombination rate ($R^2 = 0.27$ with correction versus $R^2 = 0.29$ without correction). Our method thus appears efficient to correct for biases induced by GC-content dependent polarization errors at both CpG and non-CpG sites.

In what follows, all results are presented for the whole data set (CpG + non-CpG) with correction for misorientation of SNPs.

gBGC is underestimated when its strength varies along a chromosome

In agreement with previous studies (Spencer et al. 2006; De Maio et al. 2013), our genome-wide estimates of *B* are relatively low, in

Table 2. Characteristics of the hotspot maps in the three populations

	AFR	EUR	EAS
Number of hotspots	36,571	30,621	28,442
Average length (bp)	5076.0	5219.1	5607.7
f	0.074	0.062	0.061
Average ρ	14.4	27.6	28.0
Average r_0 (cM/Mb)	0.73	0.55	0.54
Average r_1 (cM/Mb)	9.5	13.6	13.7

(*f*) Fraction of hotpsot, (r_0) and (r_1) recombination rate in coldspots and hotspots, respectively; $\rho = r_1/r_0$.

the nearly neutral area. At first sight, this appears to be in contradiction with other analyses reporting episodes of very strong gBGC (Galtier and Duret 2007; Ratnakumar et al. 2010). However, the model we used above assumes that all sites in a given window evolve under the same gBGC regime. We thus performed additional simulations to test the robustness of our approach to spatially heterogeneous levels of gBGC. We modeled recombination/gBGC hotspots by considering two categories of SNPs: A fraction, *f*, of SNPs was affected by recombination hotspots with mean gBGC B_1, whereas the other fraction, $1 - f$, was affected by a basal gBGC level B_0, with $0 \leq B_0 < B_1$. We fixed B_0, and we let B_1 vary to simulate variation in hotspot intensities. For simplicity reasons, here, we did not include polarization errors in the simulation, nor in estimations. Under this model, the average *B* is equal to $(1 - f)B_0 + f B_1$ and increases linearly with B_1. Contrary to this expectation, we observed that the estimated *B* quickly saturated as B_1 increased (Fig. 6A). gBGC is thus underestimated by model M1 when its strength is highly heterogeneous along the chromosome.

To check this prediction, we analyzed the human AFR data set in a distinct way: Rather than using genomic windows, we grouped SNPs into centiles of local recombination rate (measured on 5-kb windows centered on SNPs), thus maximizing the range of expected gBGC intensities among groups of SNPs. As predicted by simulations, the estimated *B* did not increase linearly but roughly log-linearly with recombination rate (Fig. 6B). We thus did not estimate very high *B* values, even for the highest recombination rate centiles: The maximum was only $B = 1.47$. This suggests that gBGC is too heterogeneous to be accurately estimated by the simple constant gBGC model (M1*), even when SNPs are grouped by similar recombination rates.

In order to capture this heterogeneity, we introduced additional models (called M2) with two fractions of sites experiencing different gBGC levels, a low but nonnull basal intensity of gBGC (B_0) for a fraction, $1 - f$, of sites, and higher gBGC intensity ($B_1 > B_0$) for the remaining fraction, *f* (see Supplemental Text S3). However, simulations showed that it is difficult to jointly estimate *f* and the two gBGC levels. To circumvent this difficulty, we used external information to constrain the model by fixing either *f*, or the ratio $\rho = B_1/B_0$, or both. Simulations show that the most constrained model, noted M3*, is the most robust (see Supplemental Text S3). We thus applied it by setting *f* to the fraction of sites in recombination hotspots measured in each window (using the AFR recombination map from the 1000 Genomes Project), and B_1 to ρB_0, where ρ is the ratio of recombination rates measured within and outside hotspots in each 1-Mb window (see Methods). This modeling of the dependence between B_1 and B_0 is justified because the mechanism of gBGC implies a linear relationship between recombination rate and *B* (see Supplemental Text S4). M3* therefore includes a single free gBGC parameter but still allows taking gBGC heterogeneity into account.

Applying this model to the human AFR data set, we estimated the distribution of *B* outside (B_0) and within hotspots (B_1) across 2620 1-Mb windows (Fig. 7). Excluding the 1% most extreme values, gBGC intensities ranged from −0.45 to 1.28 with a median of 0.21 and a mean of 0.23 outside hotspots, and from −4.38 to 17.93 with a median of 2.67 and a mean of 2.97 within hotspots. Averaging over hotspots and coldspots, the mean *B* equaled 0.43, which is 13% higher than the mean estimated with model M1* (mean $B = 0.38$) (Table 2). The more negative extreme values within rather than outside hotspots are simply explained by the constraint that $B_1 = \rho B_0$. Overall, 437 of 2620 windows exhibited values of B_1 higher than five. Given that hotspots cover, on average, 7.4% of each window, this indicates that ~1% of the genome experiences a gBGC intensity $>B = 5$.

gBGC intensity varies among human populations

In previous analyses, we focused on the AFR population. However, patterns of gBGC are expected to vary among populations because

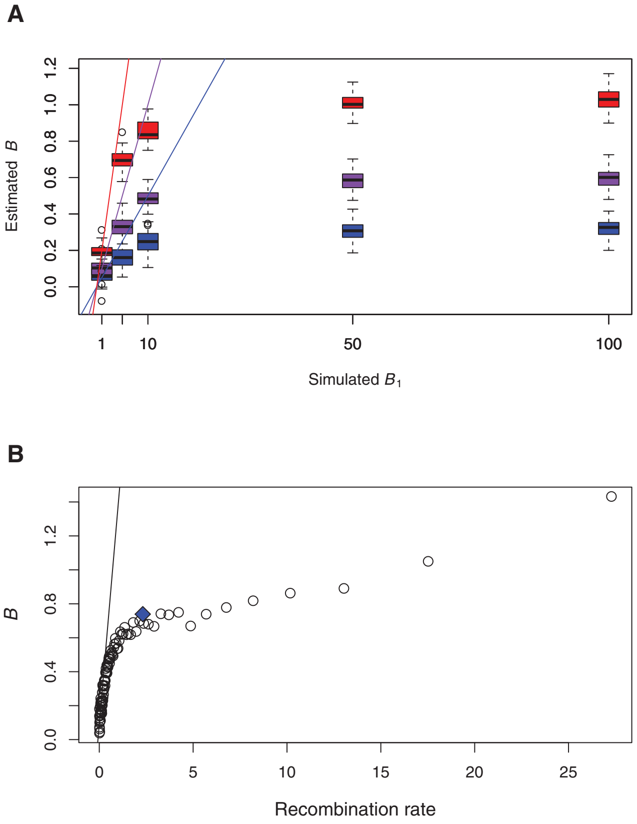

Figure 6. Effect of hotspots on *B* estimates. (*A*) Simulations were performed under model M2b with $B_0 = 0$, $f = 0.05$ (blue), 0.1 (purple), and 0.2 (red). For each B_1 value (*x*-axis), 100 simulations were performed and the M1 model was applied to estimate *B*. The lines correspond to the expectation $B = (1 - f)B_0 + fB_1$. Very similar results were obtained for $B_0 = 0.25$, and $B_0 = 0.5$ (not shown). (*B*) The whole data set was divided into centiles of recombination rates computed over 5 kb around each SNP. The line corresponds to the regression performed on centiles for which the recombination rate was lower than 0.1 cM/Mb. Dots correspond to *B* estimates under the M1* model. The blue diamond corresponds to *B* estimated using the SNPs belonging to the gBGC tracts detected by Capra et al. (2013).

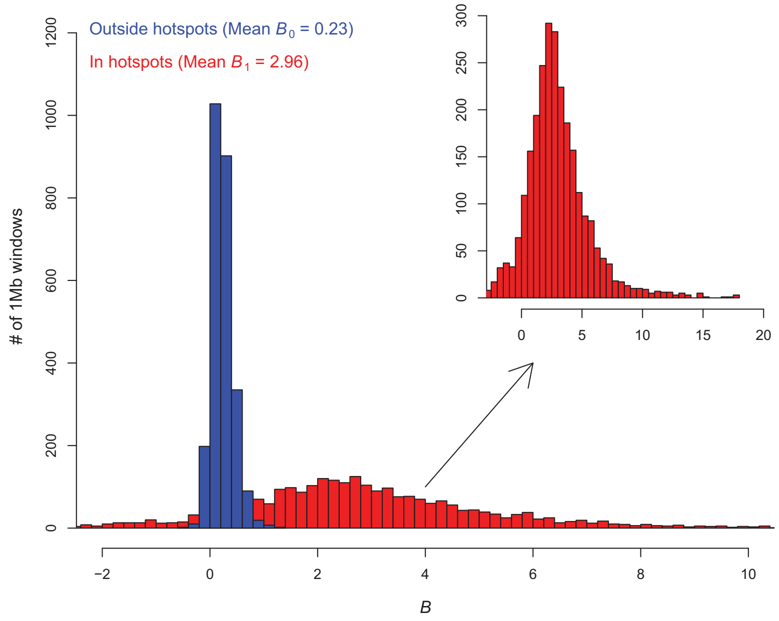

Figure 7. Distribution of B within and outside of recombination hotspots for 1-Mb windows. Distribution of the estimate of B_0 (= outside hotspots, blue) and B_1 (= within hotspots, red) for each 1-Mb window. The M3* model was used, and f, the fraction of sites within recombination hotspots, was fixed to the observed fraction. The ratio $\rho = B_1/B_0$ was also constrained to be proportional to the ratio of recombination rates within and outside hotspots. In the *inset*, the distribution of B_1 is magnified and extended to the whole range of values.

of both different recombination patterns and effective population sizes. We thus performed the same analyses (using M1* and M3* models) on the European and Asian populations. For the M3* model, we used hotspot data sets inferred from population-specific recombination maps of the 1000 Genomes Project. Genome-wide, gBGC intensity is the highest in the African, intermediate in the European, and the lowest in the Asian population (see B, M1* model and mean B, M3* model in Table 3; Supplemental Fig. S7). However, because the recombination landscape is more heterogeneous in the EUR population (Table 3; see The 1000 Genomes Project Consortium 2010), the highest B_1 was estimated in this population (Table 3; Supplemental Fig. S7). Accordingly, this population shows a higher proportion of sites experiencing strong gBGC ($B > 5$) (Table 3).

In the EUR and EAS populations, B also correlates with recombination rates and GC content (Table 4). It should be noted that at this genomic scale (1 Mb), recombination landscapes are strongly conserved across populations, despite fine scale variations in recombination hotspot locations (Pearson correlation coefficient between African and European recombination rates ≥ 0.94 [Hinch et al. 2011]). This explains why the correlations observed between B and recombination rates are quite similar for the different population-specific genetic maps, both LD-based or pedigree-based (Table 4). However, unexpectedly, estimates of B in EUR and EAS correlate more strongly with the AFR than with their population-specific recombination rates (Table 4). It is known that recombination maps are noisy (Kong et al. 2010; Hinch et al. 2011). Hence, the stronger correlation observed with AFR could be due to the fact that this recombination map is more precise than the others (possibly because the AFR population is more polymorphic). Another possible explanation is that the measures of B integrate the impact of gBGC over a long period, predating the out-of-Africa migration, and that this historical contribution to the current B estimate in EUR and EAS might be better represented by the African recombination map.

Comparison with direct measures of gBGC in noncrossovers

Direct observations of gBGC have been first provided by the analysis of a few specific recombination hotspots (Odenthal-Hesse et al. 2014; Arbeithuber et al. 2015). Recently, Williams and colleagues published a large-scale study of gene conversion tracts associated with noncrossover (NCO) recombination events in humans (Williams et al. 2015). Their analysis identified 103 NCOs and provided the first genome-wide quantification of GC-biased transmission distortion in humans: Among the AT/GC heterozygous SNPs that were involved in a NCO gene conversion, the GC allele was transmitted at a frequency $F_S = 68\%$ (95% confidence interval: 58%–78%). The gBGC coefficient (conditional to the fact that the SNP is affected by a NCO event) is $b_i = 2 \times F_S - 1$. Given the rate of NCO gene conversion ($r_{\mathrm{NCO}} = 5.7 \times 10^{-6}$/bp/generation; CI: 4.5×10^{-6}–7.3×10^{-6}) (Williams et al. 2015), and assuming an effective population size of 10,000–20,000, this implies that the population-scaled gBGC coefficient associated with NCOs ($B_{\mathrm{NCO}} = 4\ N_e \times r_{NCO} \times b_i$) is ∼0.08–0.16. In humans and mice, crossover recombination events (COs) are ∼5–15 times less frequent than NCOs, but their conversion tracts are ∼2–8 times longer (Jeffreys and May 2004; Cole et al. 2014). Hence, under the assumption that the transmission bias (F_S) is the same for COs and NCOs, the population-scaled gBGC coefficient associated with COs (B_{CO}) should be of the same order as B_{NCO}. Thus, our estimates of the total genome-wide gBGC coefficient ($B = B_{\mathrm{NCO}} + B_{\mathrm{CO}} = 0.27$–$0.43$) (Table 3) are compatible with the direct measurements of gBGC in NCO events.

Discussion

Many lines of evidence show that gBGC is a major determinant of the evolution of GC content in mammalian genomes. Quantifying its intensity throughout the genome is necessary to appreciate its

Table 3. **Average estimates of B in the M1* and M3* models for the three populations**

		AFR	EUR	EAS
Model M1*	B	0.38	0.32	0.21
Model M3*	B_0	0.23	0.17	0.11
	B_1	2.97	4.20	2.74
	Mean B	0.43	0.41	0.27
	% $B_1 > 5$	0.84	1.78	1.13
	B_1/0.3%	9.03	15.21	13.61

Mean B is computed as $(1 - f)B_0 + fB_1$. B_1/0.3% corresponds to the mean B_1 of the 0.3% of the genome with the highest gBGC (see Discussion).

Table 4. **Pearson correlation coefficients between recombination rate or GC content and *B* in the three populations**

		Recombination map					GC content
		AFR	EUR	EAS	HapMap	deCODE	
B (M1*)	AFR	0.56	0.53	0.51	0.52	0.50	0.37
	EUR	0.41	0.38	0.37	0.38	0.35	0.25
	EAS	0.40	0.38	0.34	0.35	0.34	0.22
Mean B (M3*)	AFR	0.56	0.53	0.51	0.52	0.49	0.37
	EUR	0.38	0.36	0.34	0.35	0.34	0.24
	EAS	0.36	0.35	0.31	0.35	0.31	0.20

In the M3* model, mean B is computed as $(1-f)B_0 + fB_1$. The recombination maps of the 1000 Genomes Project for the three populations and of the HapMap and deCODE projects were used. The correlation is always stronger with the AFR recombination map.

evolutionary and functional impact. As gBGC is driven largely by recombination, which is highly heterogeneous along the genome and episodic in time (Myers et al. 2005; Ptak et al. 2005; Winckler et al. 2005; Coop and Myers 2007; Auton et al. 2012), it is especially important to obtain estimates over short genomic scales and short time scales. So far, such quantifications were still lacking. To achieve this goal, we used sequence polymorphism data and tackled several issues associated with the use of such kinds of data. We proposed a new efficient method and provided a fine description of the heterogeneity of the gBGC process along the human genome.

Methodological issues

DAF spectra potentially contain information about the gBGC process and, more generally, about selection-like processes. However, to correctly infer the intensity of gBGC, two issues need to be addressed: the effect of demography and/or sampling on spectra and the problem of polarization errors. Two alternatives have been proposed to correct for demographic effects. Demographic parameters can be imposed on the estimation model (Boyko et al. 2008) or jointly inferred with selection/gBGC parameters (Keightley and Eyre-Walker 2007). Eyre-Walker et al. (2006) proposed to correct for demography by adding correction parameters for each frequency category. This latter approach is more general because it is valid for any scenario, including specific sampling schemes, which cannot be easily modeled by a simple change in population size. However, it assumes that distortions from the equilibrium expectation are the same for neutral and selected spectra, which should be accurate for weak selection but not for strong selection. Because gBGC is relatively weak globally, it is fully justified to use the second approach, which makes our method quite general and practical for many conditions.

The most serious issue is the spurious signature of gBGC created by polarization errors (Hernandez et al. 2007). Contrary to previous approaches that seek to get accurate reconstruction of ancestral states before applying an inference model, we proposed to include polarization errors directly in the inference model and to estimate them jointly with the other parameters of interest. The advantage of this approach is that it is blind to the underlying process creating polarization errors. It therefore does not require a priori information about processes of sequence evolution, such as context-dependent mutation rates that take CpG hypermutability into account (Hernandez et al. 2007; Duret and Arndt 2008). Moreover, we showed by simulations that simply correcting the polarization bias between WS and SW mutations is not sufficient because even symmetrical error rates can be problematic (Fig. 3).

Overall, we showed by simulations that our joint-inference method performed well under various scenarios. Practically, we also showed that the method corrected well for CpG effects: We observed a clear difference between CpG and non-CpG sites with the basic model without polarization errors, whereas this difference disappeared when we used the model with error correction (Fig. 4). For non-CpG sites, the correction for polarization errors did not affect the estimate of B (Table 2). One might therefore argue that the simplest option to avoid biases due to polarization errors consists of excluding CpG sites from the analysis. However, an important drawback of this option is that CpG sites are not uniformly distributed along the genome: The exclusion of CpG sites, therefore, leads to biases in the sampling toward SNPs located in GC-poor regions, where the recombination rate is, on average, lower, and thus gBGC is weaker. Hence, to obtain an unbiased estimate of gBGC strength across the entire genome, it is necessary to analyze all categories of SNPs. Moreover, the quantification of gBGC at CpG sites is also interesting in itself for understanding the molecular mechanisms causing gBGC (see below).

Our method also allows estimating the mutational bias toward AT bases and provides insights into the mutational process genome-wide (see details in Supplemental Text S4). Using the M3* model on the whole data set, we obtained the mean mutational bias across the genome to be $\lambda = 2.08$, 2.10, and 2.02 for the AFR, EUR, and EAS populations, respectively. This is very close to the direct estimate obtained by Kong et al. (2012) and suggests that accurate inference of mutation bias can also be obtained with our method. Moreover, the comparison of observed and expected GC content under mutational equilibrium (given by $p_{\mathrm{GC}} = 1/[1 + \lambda]$) indicates that most of the genome has a higher GC content than expected under mutational equilibrium, highlighting the genome-wide effect of gBGC (see Supplemental Text S4 for a quantification of this disequilibrium).

Finally, we showed that the strong heterogeneity of the gBGC process made its accurate quantification difficult. On average, the signature of gBGC is weakened by heterogeneity. We thus extended the constant gBGC model to take recombination/gBGC hotspots into account, taking advantage of the detailed knowledge of the recombination landscape in humans that we used to constrain the model and limit the variance on estimates. To evaluate how sensitive our results are to the definition of hotpots, we reran the M3* model with two other sets of hotspots based on HapMap data (see Methods; Supplemental Table S1). Despite moderate quantitative variations, the different B estimates are highly correlated, which suggests that the results of the M3* model are not very sensitive to hotspot definition (see Supplemental Tables S2, S3).

It is important to note that the location of recombination hotspots evolves very rapidly. Notably, we have shown that human recombination hotspots are, at most, 0.7–1.3 Myr old (Lesecque et al. 2014). It is therefore likely that DAF spectra at sites that correspond to previous recombination hotspots that are no longer active still retain the hallmarks of past gBGC activity. Conversely, DAF spectra at human recombination hotspots are probably not yet at mutation/drift/gBGC equilibrium. This is why the strength of gBGC cannot be estimated simply by analyzing DAF spectra at presently active recombination hotspots. Here, we modeled hotspot dynamics by considering DAF spectra as a mixture of two categories of sites, supposed to evolve under a stationary regime, which is mathematically convenient. This is clearly an oversimplification, and we suspect that the signature of gBGC is also weakened because gBGC is episodic. In the future, a challenging perspective to better quantify the heterogeneity of the gBGC process would be to develop nonstationary models taking into account both heterogeneity between sites and short-lived episodes.

Despite the limitations mentioned above, we suggest that our method can be applied to a broad set of organisms and data sets because a specific knowledge of the demographic history is not required and the effect of polarization errors can be easily corrected for. In addition, we suggest that including polarization errors should also improve other inference methods based on the analysis of DAF spectra.

No difference in gBGC strength between CpG and non-CpG sites

Our analyses of DAF spectra indicate that the strength of gBGC is very similar at CpG and non-CpG sites (Fig. 5). This result corroborates a recent study of NCO recombination events in human pedigrees, which revealed the same segregation bias in favor of GC-allele at CpG and non-CpG sites (Williams et al. 2015). These observations provide insights about the molecular mechanisms causing gBGC in humans. It is known that the methylation of cytosines at CpG sites is responsible for their hypermutability: The spontaneous deamination of 5-methylcytosine causes the formation of G/T mismatches in DNA that, if not repaired, lead to G:C → A:T mutations in the next round of DNA replication. The base excision repair system (BER) plays a major role in the repair of such mismatches. This pathway is initiated by the activity of DNA glycosylases that recognize the G/T mismatch and specifically excise thymines. The resulting gap is ultimately repaired into a G:C base pair (for review, see Sjolund et al. 2013). Mammalian cells possess four enzymes with thymine glycosylase activity (Sjolund et al. 2013). Two of these thymine glycosylases act preferentially at CpG dinucleotides, presumably to limit the hypermutability of these sites: Methyl-CpG Domain Protein 4 (MBD4) and Thymine DNA Glycosylase (TDG) (Sjolund et al. 2013).

Given that the repair of G/T mismatches by BER is systematically directed toward G:C base pairs, it has been hypothesized that this process might be responsible for gBGC in mammals (Brown and Jiricny 1987; Birdsell 2002; Duret et al. 2002; Marais 2003). If this were indeed the case, given the preferential activity of BER at CpG sites, one would then expect a stronger gBGC on CpG than on non-CpG sites. The fact that we do not observe such a pattern strongly argues against this hypothesis. This observation is in accordance with recent results demonstrating that in yeast, gBGC is not caused by BER (Lesecque et al. 2013). The prominent repair pathway during recombination is the mismatch repair (MMR) system (Surtees et al. 2004). In yeast, the analysis of gene conversion tracts indicates that gBGC is most probably caused by MMR (Lesecque et al. 2013). Our observations suggest that this might also be the case in humans.

Intensity and dynamics of gBGC across the human genome and across populations

In agreement with previous studies (Spencer et al. 2006; Capra et al. 2013; Lartillot 2013b; Lachance and Tishkoff 2014), we found that gBGC is weak on average, but widespread along the human genome, which is sufficient to explain that GC content is higher than the expected mutational equilibrium in most regions of the genome (Supplemental Fig. S4.1). The genome-wide estimates of B (obtained by averaging M3* estimates over hotspots and coldspots) are in the nearly neutral area (0.27–0.43) (Table 3). However, average values mask the strong heterogeneity we detected. In highly recombining hotspots, gBGC values can reach high values ($B > 10$) (Fig. 7; Supplemental Fig. S7), and we evaluated that ∼1%–2% of the genome experience gBGC higher than $B = 5$ (Table 3). Given that the location of hotspots evolves continually (Myers et al. 2005; Ptak et al. 2005; Winckler et al. 2005; Coop and Myers 2007; Auton et al. 2012), this implies that over the long term this process affects a large fraction of the genome.

Genome-wide gBGC intensity varies across populations (B = 0.43, 0.41, and 0.27 in the AFR, EUR, and EAS populations, respectively). Though demographic effects and variation in recombination patterns on gBGC remain to be established in details, these variations are consistent with differences in effective population sizes (N_e): AFR population has the highest N_e and EAS the lowest, with both EUR and EAS populations having experienced a demographic bottleneck (Tenesa et al. 2007; Gutenkunst et al. 2009). However the impact of larger N_e in AFR is mitigated by the fact that, due to higher *PRDM9* allelic diversity, the distribution of recombination events is more uniform (i.e., hotspots are weaker, at the population scale) in AFR than in EUR or EAS (Table 3; The 1000 Genomes Project Consortium 2010; Berg et al. 2011). This explains why the fraction of the genome affected by strong gBGC ($B > 5$) is higher in EUR and EAS than in AFR populations (Table 3).

The strength of gBGC depends both on recombination rate and on N_e. There is evidence that because of selection at linked sites (selective sweeps or background selection), N_e varies along chromosomes (Gossmann et al. 2011). To test whether such variations have a substantial impact on gBGC intensity, we measured (with model M1*) the strength of gBGC at synonymous codon positions, which are tightly linked to sites under negative selective pressure and hence should have reduced N_e. Thus, all else being equal, one would expect gBGC to be weaker at these sites than in the rest of the genome. In contradiction to this prediction, we observed that, on average, gBGC is stronger at synonymous positions ($B = 0.54$ by grouping all synonymous SNPs) than in the rest of the genome. This is probably because protein-coding genes tend to be enriched in GC-rich regions (average local GC content around synonymous SNPs = 0.53), where the recombination rate (and hence, gBGC) is higher (average local recombination rate = 1.49 cM/Mb). However, even when controlling for recombination and GC-content, we observed that the strength of gBGC is not reduced at synonymous sites compared to the rest of the genome (Fig. 4). This suggests that variations in gBGC along chromosomes are mainly driven by the heterogeneity of recombination rate and that the impact of selection at linked sites on the intensity of gBGC (via N_e) is relatively limited.

Temporal dynamics of gBGC episodes

Previous attempts to quantify the impact of gBGC were based on the analysis of substitution patterns along the phylogeny (Capra et al. 2013; Lartillot 2013b). Capra et al. (2013) estimated that ∼0.3% of the human genome has been subject to strong gBGC episodes since the divergence from chimpanzee, whereas Lartillot (2013b) did not detect any signature of strong gBGC episodes in primates. This contrasts with our results, which indicate that 1%–2% of our genome is currently subject to strong gBGC ($B > 5$). The discrepancy is probably due to the fact that these phylogenetic approaches tend to effectively average processes over periods of time (divergence between species) that are much longer than the lifespan of recombination hotspots. Hence, only extremely strong or long-lasting gBGC episodes can be detected by such methods. For a comparison, the distribution of B values obtained under the M3* model (Fig. 7; Supplemental Fig. S7) indicates that the 0.3% of the human genome with the strongest gBGC experience average B values between 9 (AFR) and 15 (EUR) (Table 3).

Our results also allow us to elucidate the dynamics of gBGC hotspots. If the gBGC tracts detected along the genome by Capra et al. (2013) were still active gBGC hotspots, we should observe high B values in these tracts. To test this, we retrieved all SNPs belonging to these tracts from the web site, http://genome-mirror.bscb. cornell.edu, and applied the M1* model. The value we obtained, $B = 0.74$, is higher than the mean computed over the 1-Mb windows ($B = 0.38$ with M1* and 0.43 with M3*), but still rather low. Accordingly, the current average recombination rate around these tracts (2.32 cM/Mb) is higher than the genomic mean (1.42 cM/Mb) but does not reach the most extreme values (Fig. 6B). These observations suggest that, on average, gBGC is currently not extremely active in these tracts. Thus, most of these tracts probably correspond to ancient recombination hotspots that are no longer active. This is in agreement with the recent findings that current human hotspots are <0.7–1.3 Myr old (Lesecque et al. 2014), i.e., much younger than the human-chimpanzee divergence time (7–13 Myr) (Langergraber et al. 2012).

Consequences of transient strong gBGC episodes

As already suspected, our results show that strong gBGC episodes transiently occur along the genome. The consequences of a highly heterogeneous versus a homogeneous gBGC process are strikingly different even when the mean effect in both scenarios is the same. First, strong gBGC episodes are required to explain substitution hotspots (Dreszer et al. 2007; Kostka et al. 2012; Clement and Arndt 2013) and spurious signatures of positive selection (Galtier and Duret 2007; Ratnakumar et al. 2010), but previous studies so far only provided rather low average estimates with maximum B values slightly higher than one (Spencer et al. 2006; De Maio et al. 2013; Lartillot 2013b). Here, we directly show that the intensity of gBGC can locally reach values higher than $B = 5$ and even of the order of $B = 15$, which is largely sufficient to explain substitution hotspots. Beyond these technical consequences for the interpretation of genomic patterns, gBGC can counteract selection (Galtier et al. 2009; Lartillot 2013a) and have deleterious consequences (Necsulea et al. 2011; Capra et al. 2013; Lachance and Tishkoff 2014), and it has been shown that strong gBGC episodes in few hotspots have worse consequences than low gBGC levels that are homogeneous along a chromosome (Glémin 2010). gBGC can thus contribute significantly to the genetic load experienced by human populations. The EUR population, and to a lesser extent, the EAS one, would be the more affected by this gBGC load because they show the highest values of B_1, though the average B is lower. Even though the load a population can tolerate can be high under soft and/or stabilizing selection (Lesecque et al. 2012; Charlesworth 2013), our estimates are quantitatively compatible with potential pathological implications of gBGC as previously proposed (Galtier et al. 2009; Necsulea et al. 2011; Capra et al. 2013; Lachance and Tishkoff 2014).

Methods

Data set

The preparation of data sets is fully detailed in Supplemental Text S5 and summarized here.

We downloaded the 1000 Genomes Project polymorphism data set (phase 1) (The 1000 Genomes Project Consortium 2012) from the EBI web site, ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20120316_phase1_integrated_release_version2/, and filtered out SNPs without ancestral state predictions or with low reliability. Genome annotations (location of coding exons, introns, etc.) were retrieved from Ensembl release 54 (Flicek et al. 2014). The location of repeats (transposable elements or low-complexity sequences identified by RepeatMasker) and CpG islands were retrieved from the UCSC server (Karolchik et al. 2014).

To measure recombination rates, we used a genetic map from the HapMap Phase 2 project (International HapMap Consortium 2007) and population-specific genetic maps from the 1000 Genomes Project (CEU, YRI, CHBJPT) (The 1000 Genomes Project Consortium 2010). These four maps are all based on the analysis of LD within populations. As a control, we also used the pedigree-based genetic map from deCODE (Kong et al. 2010).

We used the list of recombination hotspots published by HapMap (International HapMap Consortium 2007), which we referred to as HM hotspots. We also defined hotspots for the three AFR, EUR, and EAS populations from the 1000 Genomes (1KG) pilot project LD recombination maps specific to these three populations, as regions of length > 2 kb with a recombination rate > 6 cM/Mb. For a comparison, we also used the same approach to identify hotspots in the HapMap LD map. These sets of hotspots are referred to as 1KG hotspots and HMt6 hotspots (t6 stand for threshold at 6 cM/Mb). In each 1-Mb window, we computed, for each recombination map, the total recombination rate, the fraction of the window occupied by hotspots (f), the average recombination rate within hotspots (r_1), the average recombination rate outside hotspots (r_0), and the ratio $\rho = r_1/r_0$.

Maximum-likelihood framework to estimate the intensity of gBGC from site frequency spectra

We fitted population genetics models to the derived allele frequency spectra to estimate B using a maximum-likelihood framework similar to Muyle et al. (2011). The generic model is given by equation (1) in the main text. In equation (1), the first term within the integral corresponds to the binomial sampling of i alleles in a sample of size n given true population-frequency x. When n is high, we can use the continuous approximation that gives very similar results and speeds up numerical computations:

$$\int_0^1 C_n^i x^i (1-x)^{n-i} H(x)dx \approx \frac{1}{n} H(i/n). \tag{2}$$

For each subpopulation of the 1000 genomes data set, the frequencies are given in 1/100 so that we set $n = 100$.

We used the following nested models:

M0: no gBGC:

$$H(x) = \frac{2}{x} \tag{3}$$

M1: constant gBGC of intensity $B = 4N_e b$:

$$H_{\mathrm{WS}}(x) = H(B, x) = 2\frac{1 - e^{-B(1-x)}}{x(1-x)(1-e^{-B})} \tag{4}$$

and $H_{\mathrm{SW}}(x) = H(-B,x)$. B can be either positive or negative.

M2a: gBGC hotspots of intensity $B = 4N_e b$ in frequency f:

$$H_{\mathrm{WS}}(x) = H(B, f, x) = 2\left(f\frac{1 - e^{-B(1-x)}}{x(1-x)(1-e^{-B})} + \frac{(1-f)}{x}\right) \tag{5}$$

and $H_{\mathrm{SW}}(x) = H(-B, f, x)$. B can be either positive or negative.

M2b: gBGC hotspots of intensity $B_1 = 4N_e b_1$ in frequency f and basal gBGC of intensity $B_0 = 4N_e b_0$:

$$H_{\mathrm{WS}}(x) = H(B_0, B_1, f, x)$$

$$= 2\left((1-f)\frac{1-e^{-B_0(1-x)}}{x(1-x)(1-e^{-B_0})} + f\frac{1-e^{-B_1(1-x)}}{x(1-x)(1-e^{-B_1})}\right) \quad (6)$$

and $H_{\mathrm{SW}}(x) = H(-B_0, -B_1, f, x)$. B_0 and B_1 can be either positive or negative.

Polarization errors were included in the four models according to equation (1).

M3: constrained model of gBGC hotspots:

This model is equivalent to the M2b model, except that f is fixed according to the fraction of sites within recombination hotspots detected by HapMap and $B_1 = \rho B_0$, where ρ is the ratio of recombination rates measured in and outside hotspots.

Assuming independence between SNPs, the likelihood of the model can thus be written as:

$$\Gamma = \prod_{i=1}^{n} P(\mu_{\mathrm{neutral}}(i), k_i^{\mathrm{WW,SS}}) P(\mu_{\mathrm{WS}}(i), k_i^{\mathrm{WS}}) P(\mu_{\mathrm{SW}}(i), k_i^{\mathrm{SW}}). \quad (7)$$

Parameters estimates were obtained by maximization of the log-likelihood function using the FindMaximum function of Mathematica v8 (Wolfram 1996) (see Supplemental Text S6 for details of the implementation). Likelihood-ratio tests with one degree of freedom can be performed to compare the different nested gBGC models (M1 versus M0, M2a versus M1, M2b versus M2a, with or without polarization errors). Similarly, the equivalent models with and without polarization errors can be compared. Note that because of possible nonindependence between SNPs, LRT are anti-conservative and must be viewed with caution. However, maximum-likelihood estimates should not be affected by such nonindependence.

Estimated parameters of the different models are given for all 1-Mb windows in Supplemental files.

Simulations

We simulated data sets by drawing SNPs from Poisson distributions with expectation values given by the population genetics models M0 to M2b. These are the "true" correctly orientated data sets. Then, from these data sets, we built data sets with a given proportion of polarization errors: e_{neutral}, e_{WS}, and e_{SW}. For these "observed" data sets with polarization errors, the observed numbers of SNPs in frequency classes i/n are thus:

$$f_{\mathrm{obs}}(i) = (1 - e_{\mathrm{neutral}}) f_{\mathrm{true}}(i) + e_{\mathrm{neutral}} f_{\mathrm{true}}(n-i) \quad \text{for neutral SNPs,}$$

$$f_{\mathrm{obs}}(i) = (1 - e_{\mathrm{WS}}) f_{\mathrm{true}}(i) + e_{\mathrm{SW}} f_{\mathrm{true}}(n-i) \quad \text{for WS SNPs,}$$

$$f_{\mathrm{obs}}(i) = (1 - e_{\mathrm{SW}}) f_{\mathrm{true}}(i) + e_{\mathrm{WS}} f_{\mathrm{true}}(n-i) \quad \text{for SW SNPs.}$$

Note that the observed numbers of WS SNPs are proportional to $(1 - p_{\mathrm{GC}})$ θWS and the observed number of SW SNPs to p_{GC} λ θWS. We then applied the different models, without and with error corrections, to the two kinds of data sets. The following parameters are common to all simulations: θneutral = 1000, θWS = 2000, λ = 2, $n = 20$.

Software availability

The Mathematica script used for the analyses with an example file and R scripts are also provided in the Supplemental Material.

Acknowledgments

This work was supported by the Centre National de la Recherche Scientifique and by the Agence Nationale de la Recherche (ABS4NGS: ANR-11-BINF-0001-06). S.G. is supported jointly by the French CNRS and the Marie Curie IEF Grant "SELFADAPT" 623486. We thank the Montpellier Bioinformatics and Biodiversity platform for computing facilities. We thank three anonymous referees for their constructive comments.

References

The 1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061–1073.

The 1000 Genomes Project Consortium. 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* **491:** 56–65.

Arbeithuber B, Betancourt AJ, Ebner T, Tiemann-Boege I. 2015. Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proc Natl Acad Sci* **112:** 2109–2114.

Auton A, Fledel-Alon A, Pfeifer S, Venn O, Segurel L, Street T, Leffler EM, Bowden R, Aneas I, Broxholme J, et al. 2012. A fine-scale chimpanzee genetic map from population sequencing. *Science* **336:** 193–198.

Berg IL, Neumann R, Sarbajna S, Odenthal-Hesse L, Butler NJ, Jeffreys AJ. 2011. Variants of the protein PRDM9 differentially regulate a set of human meiotic recombination hotspots highly active in African populations. *Proc Natl Acad Sci* **108:** 12378–12383.

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

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

Boyko AR, Williamson SH, Indap AR, Degenhardt JD, Hernandez RD, Lohmueller KE, Adams MD, Schmidt S, Sninsky JJ, Sunyaev SR, et al. 2008. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLoS Genet* **4:** e1000083.

Brown TC, Jiricny J. 1987. A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. *Cell* **50:** 945–950.

Capra JA, Hubisz MJ, Kostka D, Pollard KS, Siepel A. 2013. A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. *PLoS Genet* **9:** e1003684.

Charlesworth B. 2013. Why we are not dead one hundred times over. *Evolution* **67:** 3354–3361.

Clement Y, Arndt PF. 2013. Meiotic recombination strongly influences GC-content evolution in short regions in the mouse genome. *Mol Biol Evol* **30:** 2612–2618.

Cole F, Baudat F, Grey C, Keeney S, de Massy B, Jasin M. 2014. Mouse tetrad analysis provides insights into recombination mechanisms and hotspot evolutionary dynamics. *Nat Genet* **46:** 1072–1080.

Coop G, Myers SR. 2007. Live hot, die young: transmission distortion in recombination hotspots. *PLoS Genet* **3:** e35.

De Maio N, Schlotterer C, Kosiol C. 2013. Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. *Mol Biol Evol* **30:** 2249–2262.

Dreszer TR, Wall GD, Haussler D, Pollard KS. 2007. Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. *Genome Res* **17:** 1420–1430.

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

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

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

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

Escobar JS, Glémin S, Galtier N. 2011. GC-biased gene conversion impacts ribosomal DNA evolution in vertebrates, angiosperms, and other eukaryotes. *Mol Biol Evol* **28:** 2561–2575.

Eyre-Walker A. 1998. Problems with parsimony in sequences of biased base composition. *J Mol Evol* **47:** 686–690.

Eyre-Walker A, Woolfit M, Phelps T. 2006. The distribution of fitness effects of new deleterious amino acid mutations in humans. *Genetics* **173:** 891–900.

Flicek P, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S, et al. 2014. Ensembl 2014. *Nucleic Acids Res* **42:** D749–D755.

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

Glémin S. 2010. Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. *Genetics* **185:** 939–959.

Gossmann TI, Woolfit M, Eyre-Walker A. 2011. Quantifying the variation in the effective population size within a genome. *Genetics* **189:** 1389–1402.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet* **5:** e1000695.

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

Hinch AG, Tandon A, Patterson N, Song Y, Rohland N, Palmer CD, Chen GK, Wang K, Buxbaum SG, Akylbekova EL, et al. 2011. The landscape of recombination in African Americans. *Nature* **476:** 170–175.

International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449:** 851–861.

Jeffreys AJ, May CA. 2004. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nat Genet* **36:** 151–156.

Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Guruvadoo L, Haeussler M, et al. 2014. The UCSC Genome Browser database: 2014 update. *Nucleic Acids Res* **42:** D764–D770.

Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hotspots. *Genome Biol Evol* **3:** 614–626.

Keightley PD, Eyre-Walker A. 2007. Joint inference of the distribution of fitness effects of deleterious mutations and population demography based on nucleotide polymorphism frequencies. *Genetics* **177:** 2251–2261.

Kong A, Thorleifsson G, Gudbjartsson DF, Masson G, Sigurdsson A, Jonasdottir A, Walters GB, Jonasdottir A, Gylfason A, Kristinsson KT, et al. 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* **467:** 1099–1103.

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, et al. 2012. Rate of *de novo* mutations and the importance of father's age to disease risk. *Nature* **488:** 471–475.

Kostka D, Hubisz MJ, Siepel A, Pollard KS. 2012. The role of GC-biased gene conversion in shaping the fastest evolving regions of the human genome. *Mol Biol Evol* **29:** 1047–1057.

Lachance J, Tishkoff SA. 2014. Biased gene conversion skews allele frequencies in human populations, increasing the disease burden of recessive alleles. *Am J Hum Genet* **95:** 408–420.

Langergraber KE, Prufer K, Rowney C, Boesch C, Crockford C, Fawcett K, Inoue E, Inoue-Muruyama M, Mitani JC, Muller MN, et al. 2012. Generation times in wild chimpanzees and gorillas suggest earlier divergence times in great ape and human evolution. *Proc Natl Acad Sci* **109:** 15716–15721.

Lartillot N. 2013a. Interaction between selection and biased gene conversion in mammalian protein-coding sequence evolution revealed by a phylogenetic covariance analysis. *Mol Biol Evol* **30:** 356–368.

Lartillot N. 2013b. Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. *Mol Biol Evol* **30:** 489–502.

Lesecque Y, Keightley PD, Eyre-Walker A. 2012. A resolution of the mutation load paradox in humans. *Genetics* **191:** 1321–1330.

Lesecque Y, Mouchiroud D, Duret L. 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Mol Biol Evol* **30:** 1409–1419.

Lesecque Y, Glemin S, Lartillot N, Mouchiroud D, Duret L. 2014. The red queen model of recombination hotspots evolution in the light of archaic and modern human genomes. *PLoS Genet* **10:** e1004790.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2:** e5.

Marais G. 2003. Biased gene conversion: implications for genome and sex evolution. *Trends Genet* **19:** 330–338.

Muyle A, Serres-Giardi L, Ressayre A, Escobar J, Glémin S. 2011. GC-biased gene conversion and selection affect GC content in the *Oryza* genus (rice). *Mol Biol Evol* **28:** 2695–2706.

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.

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

Necsulea A, Popa A, Cooper DN, Stenson PD, Mouchiroud D, Gautier C, Duret L. 2011. Meiotic recombination favors the spreading of deleterious mutations in human populations. *Hum Mutat* **32:** 198–206.

Odenthal-Hesse L, Berg IL, Veselis A, Jeffreys AJ, May CA. 2014. Transmission distortion affecting human noncrossover but not crossover recombination: a hidden source of meiotic drive. *PLoS Genet* **10:** e1004106.

Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, Birney E. 2008. Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Res* **18:** 1829–1843.

Pessia E, Popa A, Mousset S, Rezvoy C, Duret L, Marais GA. 2012. Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol* **4:** 675–682.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S. 2005. Fine-scale recombination patterns differ between chimpanzees and humans. *Nat Genet* **37:** 429–434.

Rands CM, Meader S, Ponting CP, Lunter G. 2014. 8.2% of the human genome is constrained: variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* **10:** e1004525.

Ratnakumar A, Mousset S, Glémin S, Berglund J, Galtier N, Duret L, Webster MT. 2010. Detecting positive selection within genomes: the problem of biased gene conversion. *Phil Trans R Soc Lond B* **365:** 2571–2580.

Romiguier J, Ranwez V, Douzery EJ, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res* **20:** 1001–1009.

Serres-Giardi L, Belkhir K, David J, Glémin S. 2012. Patterns and evolution of nucleotide landscapes in seed plants. *Plant Cell* **24:** 1379–1397.

Sjolund AB, Senejani AG, Sweasy JB. 2013. MBD4 and TDG: multifaceted DNA glycosylases with ever expanding biological roles. *Mutat Res* **743–744:** 12–25.

Smith NG, Eyre-Walker A. 2001. Synonymous codon bias is not caused by mutation bias in G+C-rich genes in humans. *Mol Biol Evol* **18:** 982–986.

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.

Surtees JA, Argueso JL, Alani E. 2004. Mismatch repair proteins: key regulators of genetic recombination. *Cytogenet Genome Res* **107:** 146–159.

Tenesa A, Navarro P, Hayes BJ, Duffy DL, Clarke GM, Goddard ME, Visscher PM. 2007. Recent human effective population size estimated from linkage disequilibrium. *Genome Res* **17:** 520–526.

Webster MT, Axelsson E, Ellegren H. 2006. Strong regional biases in nucleotide substitution in the chicken genome. *Mol Biol Evol* **23:** 1203–1216.

Williams AL, Genovese G, Truax T, Jun G. 2015. Non-crossover gene conversions show strong GC bias and unexpected clustering in humans. *Elife* doi: 10.7554/eLife.04637.

Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, Bontrop RE, McVean GA, Gabriel SB, Reich D, Donnelly P, et al. 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* **308:** 107–111.

Wolfram S. 1996. *The Mathematica book*. Cambridge University Press, Cambridge.

Received October 8, 2014; accepted in revised form May 18, 2015.