Biased Gene Conversion Skews Allele Frequencies in Human Populations, Increasing the Disease Burden of Recessive Alleles

Joseph Lachance[1,2,*] and Sarah A. Tishkoff[1,*]

Gene conversion results in the nonreciprocal transfer of genetic information between two recombining sequences, and there is evidence that this process is biased toward G and C alleles. However, the strength of GC-biased gene conversion (gBGC) in human populations and its effects on hereditary disease have yet to be assessed on a genomic scale. Using high-coverage whole-genome sequences of African hunter-gatherers, agricultural populations, and primate outgroups, we quantified the effects of GC-biased gene conversion on population genomic data sets. We find that genetic distances (F_{ST} and population branch statistics) are modified by gBGC. In addition, the site frequency spectrum is left-shifted when ancestral alleles are favored by gBGC and right-shifted when derived alleles are favored by gBGC. Allele frequency shifts due to gBGC mimic the effects of natural selection. As expected, these effects are strongest in high-recombination regions of the human genome. By comparing the relative rates of fixation of unbiased and biased sites, the strength of gene conversion was estimated to be on the order of $Nb \approx 0.05$ to 0.09. We also find that derived alleles favored by gBGC are much more likely to be homozygous than derived alleles at unbiased SNPs (+42.2% to 62.8%). This results in a *curse of the converted*, whereby gBGC causes substantial increases in hereditary disease risks. Taken together, our findings reveal that GC-biased gene conversion has important population genetic and public health implications.

Introduction

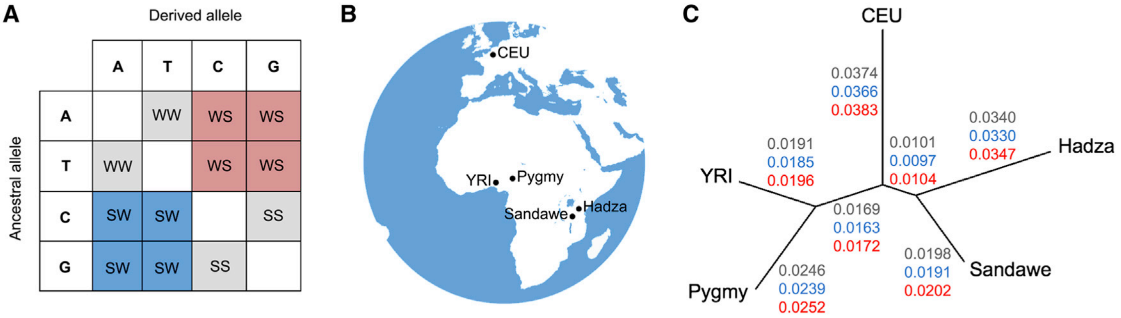
Meiotic recombination results in either crossover or noncrossover events, and gene conversion can occur in either case.[1] In humans the mean tract length of these gene conversion events is approximately 500 base pairs.[2] Gene conversion is defined here as the nonreciprocal exchange of genetic information between homologous sequences, and two kinds of gene conversion exist: conversion between two alleles of the same gene (allelic gene conversion) and conversion between paralogs (interlocus gene conversion).[1,3] In humans there is evidence that allelic gene conversion has affected the fast-evolving *ADCYAP1* gene[4] (MIM 102980), and interlocus gene conversion has shaped the evolution of genes that encode erythrocyte glycoproteins in malaria-endemic African populations.[5] In this paper, we focus on the population genetic and public health implications of allelic gene conversion.

Recombination results in the formation of heteroduplex DNA, and mispairing due to differences in parental alleles is corrected by the mismatch repair machinery.[1] However, mismatch repair preferentially retains guanine (G) and cytosine (C) over adenine (A) and thymine (T) alleles.[6] This causes gene conversion to be biased toward G or C alleles. GC-biased gene conversion (gBGC) likely evolved as a response to high mutation rates caused by the deamination of methylated cytosine.[1,7] Strong (G or C) alleles are represented by the IUPAC code S, and weak (A or T) alleles are represented by the IUPAC code W. Listing the ancestral allele first and derived allele second, pairs of IUPAC codes can be used to describe different types of SNPs (Figure 1). For example, a SNP with an ancestral A allele and derived G allele is labeled WS. One implication of gBGC is that when an individual is heterozygous for a strong (G or C) and weak (A or T) allele, the strong allele is more likely to be passed on to their offspring. Because equal segregation does not occur, gene conversion results in non-Mendelian inheritance. WW and SS SNPs are unbiased, SW SNPs have ancestral alleles that are favored by gBGC, and WS SNPs have derived alleles that are favored by gBGC.

There is increasing evidence that gene conversion is an important evolutionary phenomenon. Gene conversion influences GC content[8–10] and decreases linkage disequilibrium over small scales.[11,12] Haplotypes containing variants that increase recombination rates are more likely to be converted, and this leads to what has been called the "recombination hotspot paradox."[13,14] Alleles favored by gBGC are evolutionarily (and mathematically) equivalent to semidominant mutations under positive selection.[1,15] Because of this, gene conversion results in shifts in site frequency spectra. Low-coverage whole-genome sequences from the 1000 Genomes Project reveal that these allele frequency shifts are stronger in high-recombination regions of the human genome,[16] and gBGC modifies the allele frequencies of nonsynoymous SNPs that are likely to contribute to hereditary disease.[17] Comparisons with other primate genomes have identified human accelerated regions (HARs), and these genomic regions are enriched for WS substitutions, a pattern that is consistent with gene conversion.[18,19] However, substitutions caused by gBGC can be nonadaptive and these substitutions may be "the Achilles' heel of our genome."[20] Indeed, gBGC modifies

[1]Departments of Biology and Genetics, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Present address: School of Biology, Georgia Institute of Technology, Atlanta, GA 30332, USA
*Correspondence: lachance.joseph@gmail.com (J.L.), tishkoff@mail.med.upenn.edu (S.A.T.)

http://dx.doi.org/10.1016/j.ajhg.2014.09.008.



	A	T	C	G
A		WW	WS	WS
T	WW		WS	WS
C	SW	SW		SS
G	SW	SW	SS	

Figure 1. Biased Gene Conversion, Study Populations, and Genetic Distances
(A) Depending on ancestral and derived states, SNPs can be classified as WW or SS (gray), SW (blue), and WS (red). Weak (A or T) alleles are represented by W and strong (G or C) alleles are represented by S.
(B) Study populations: Pygmy, YRI (Yoruba), Sandawe, Hadza, and CEU (Northern and Western European ancestry). Five high-coverage whole genomes were analyzed for each population.
(C) Genetic distances using population branch statistics. Mean values are shown for thee different types of SNPS: WW or SS (gray), SW (blue), and WS (red). Branch lengths depicted are for WW or SS SNPs.

d_N/d_S ratios and has contributed to the fixation of deleterious mutations in primate lineages.[21]

At present there is a lack of studies that analyze the effects of GC-biased gene conversion using high-coverage whole-genome sequence data from diverse global populations. It also is unknown how gBGC affects genetic distances between human populations, and there is a need to estimate the strength of gBGC from genomic data that are free from ascertainment bias. In addition, the effects of gBGC-induced allele frequency shifts on hereditary disease risks are yet to be quantified. GC-biased gene conversion results in the following predictions: (1) increased genetic distances and modified evolutionary rates for variants favored by gene conversion, (2) detectable shifts in allele frequency distributions, (3) greater effects in high-recombination regions of the human genome, and (4) effects of gBGC observable in every population.

In this study we use high-coverage whole-genome sequences from five global populations to test each of the above predictions. We determine how much gBGC perturbs population genetics statistics that are commonly used for demographic inference and scans of selection. We then use relative rates of fixation of biased and unbiased SNPs to infer the strength of gene conversion. Finally, because allele frequency shifts modify allele frequencies and the chance of observing recessive homozygotes, we quantify the effects of gBGC on the risk of hereditary disease.

Material and Methods

Whole-Genome Sequences

A total of 25 high-coverage (~60×) genomes sequenced by Complete Genomics[22] were analyzed in this study. Error rates for these genomes are on the order of 1 per 100,000 base pairs.[22–24] The standard Complete Genomics bioinformatics pipeline was used for sequence alignment, read mapping, assembly, and SNP calling (Assembly Pipeline v.1.10 and CGA Tools 1.4). Five genomes were analyzed per population, and the geographic locations of study populations are shown in Figure 1B. Populations sampled include Pygmies from Cameroon (Baka, Bakola, and Bedzan), Yoruba from Nigeria (YRI), Sandawe from Tanzania, Hadza from Tanzania, and individuals with Northern and Western European ancestry (CEU). Pygmy, Sandawe, and Hadza genomes were previously analyzed in a recent study of African hunter-gatherers,[23] and YRI and CEU genomes were obtained from the Complete Genomics public data release. Prior to collection of Pygmy, Hadza, and Sandawe samples, informed consent was obtained from all research participants. Permits were received from the Ministry of Health and National Committee of Ethics in Cameroon and from COSTECH and NIMR in Dar es Salaam, Tanzania. In addition, appropriate IRB approval was obtained from both the University of Maryland and the University of Pennsylvania.

Data Processing

After merging genomes using CGA Tools, we selected sites that were polymorphic in at least one study population. We then filtered out sex-linked and mitochondrial variants and required that sites be fully called in all 25 genomes. For each autosomal locus, this gives a sample size of $n = 10$ per population. As per a previous study, derived and ancestral states of SNPs were found via maximum likelihood using chimpanzee, orangutan, and rhesus macaque genomes as outgroups.[23] A total of 10,770,084 SNPs remained after obtaining derived allele frequencies of fully called autosomal SNPs.

Hypermutable CpG dinucleotides can cause derived allele frequencies to be misestimated.[25] To correct for this, base pairs flanking each SNP were identified and variants were flagged if they belong to a CpG dinucleotide in either humans or chimpanzees. Flagged variants were then excluded from subsequent polymorphism analyses, resulting in a total of 7,539,623 non-CpG SNPs. GC-biased gene conversion favors strong alleles (G or C, denoted by the IUPAC code S) over weak alleles (A or T, denoted by the IUPAC code W). Because of this, we binned SNPs into three broad categories (Figure 1A): SNPs unaffected by gBGC (WW or SS SNPs), SNPs where the ancestral allele is favored by gBGC (SW), and SNPs where the derived allele is favored by gBGC (WS).

To generate 95% confidence intervals of genetic distances and population genetics statistics (see below), we bootstrapped whole-genome data sets for five different populations and four different types of data: WW or SS, SW, WS, and all non-CpG

SNPs. For each combination of population and SNP type, we bootstrapped whole-genome data sets 1,000 times, generating 10,000 unlinked SNPs per bootstrap run. Mann-Whitney U tests were used to compare bootstrapped values of population genetics statistics and to generate two-tailed p values. For recombination rate tests, bootstrapped statistics for SNPs in the lowest quintile (0%–20%) were compared to SNPs with recombination rates in the highest quintile (80%–100%) using Mann-Whitney U tests. Because 10,000 SNPs were analyzed for each bootstrap run, even small effect sizes resulted in low p values.

Calculating Genetic Distances

Pairwise F_{ST} statistics were used to estimate genetic distances between populations and population branch statistics (PBS) were used to estimate relative rates of evolution. To correct for small sample size, F_{ST} was calculated using Weir and Cockerham's method.[26] PBS statistics measure the amount of sequence change along branches of a population tree.[27,28] These statistics have also been called locus-specific branch length[29] and relative rate statistics.[30] Whenever four or more populations are analyzed, such as in this present study, PBS statistics require a known topology, and we used a neighbor joining tree generated from whole-genome sequencing data:[23] (Pygmy, (YRI, ((Hadza, Sandawe), CEU))). An unrooted version of this tree is shown in Figure 1C. PBS for all internal and external branches of the population tree were calculated using pairwise F_{ST} statistics and Equations A1–A7 in Appendix A. Negative values of F_{ST} and PBS statistics were treated as 0. To assess the effects of gBGC on genetic distances, mean values of PBS statistics were calculated for WW or SS SNPs, SW SNPs, and WS SNPs using the R programming language.[31]

Allele Frequency Distributions and Summary Statistics

Normalized site frequency spectra (SFS) were obtained for each population and type of SNP (WW or SS, SW, and WS). For each pooled set of SNPs (i.e., population and type of SNP), we calculated Tajima's D,[32] normalized Fay and Wu's H,[33,34] and mean derived allele frequency (DAF). We also calculated a summary statistic that compares WS and SW DAF spectra (W→S DAF skew). This statistic involves performing a Mann-Whitney U test and then normalizing by the maximum possible value of the test.[16] To test whether the effects of gBGC are stronger in regions of high recombination, we obtained recombination rates from the deCODE 2010 data set,[35] averaged over 100 kb intervals, and annotated each SNP. SNPs were then binned into five different recombination rate fractions, and population genetics statistics (Tajima's D, Fay and Wu's H, mean DAF, and W→S DAF skew) were calculated for sets of pooled SNPs for each recombination rate quintile (20% bin).

Estimating the Strength of GC Bias

Relative rates of fixation were used to estimate the strength of GC-biased gene conversion (b). We define r as the relative rate of fixation of biased WS or SW substitutions compared to unbiased WW or SS substitutions. Accelerated evolution yields $r > 1$ and decelerated evolution yields $r < 1$. Alleles favored by gBGC are evolutionarily equivalent to semidominant mutations under selection.[1,15] Because of this, the mathematics of natural selection can be repurposed to estimate the strength of GC bias. Here, gBGC coefficients (b) are used instead of selection coefficients. Note that the frequency of G or C alleles among gametes produced by WS or SW heterozygotes is equal to $(1+b)/2$.[1] The probability of fixation of selected (or biased) substitutions follows from Kimura's diffusion approximation[36] and the probability of fixation of neutral (or unbiased) substitutions is equal to 1/2N, where N is the population size. The ratio of these two expressions yields equations for the relative rate of fixation at biased sites compared to unbiased sites. For small values of b:

$$r_{WS} = \frac{4Nb}{1 - e^{-4Nb}} \qquad \text{(Equation 1)}$$

$$r_{SW} = \frac{4Nb}{e^{4Nb} - 1} \qquad \text{(Equation 2)}$$

Empirical estimates of r can be calculated from the number of substitutions per site (D) and mutation rates (μ). Ancestral states were inferred at a total of 2.58×10^9 autosomal sites. The ancestral allele was A or T at 1.52×10^9 sites and G or C at 1.06×10^9 sites. We observed a total of 6.85×10^6 WS substitutions, 7.65×10^6 SW substitutions, and 2.68×10^6 WW or SS substitutions. Mutation rates were obtained from the Kong et al.[35] ($\mu_{WS} = 6.89 \times 10^{-9}$, $\mu_{SW} = 1.48 \times 10^{-8}$, and $\mu_{WWorSS} = 1.91 \times 10^{-9}$). For WS and SW sites:

$$r_{WS} = \frac{D_{WS}/\mu_{WS}}{D_{WWorSS}/\mu_{WWorSS}} \qquad \text{(Equation 3)}$$

$$r_{SW} = \frac{D_{SW}/\mu_{SW}}{D_{WWorSS}/\mu_{WWorSS}} \qquad \text{(Equation 4)}$$

By combining Equations 1, 2, 3, and 4, we were able to numerically estimate the population-scaled strength of gBGC (Nb) from accelerated evolutionary rates at WS sites and decelerated evolutionary rates at SW sites. These estimates were made for the full set of autosomal sites and for each recombination rate quintile.

$$\frac{4Nb}{1 - e^{-4Nb}} = \frac{D_{WS}/\mu_{WS}}{D_{WWorSS}/\mu_{WWorSS}} \qquad \text{(Equation 5)}$$

$$\frac{4Nb}{e^{4Nb} - 1} = \frac{D_{SW}/\mu_{SW}}{D_{WWorSS}/\mu_{WWorSS}} \qquad \text{(Equation 6)}$$

Estimating Disease Burden from Site Frequency Spectra

We used SFS shifts to determine the extent that gBGC influences the burden of hereditary disease. Predicted disease burden (β) is influenced by penetrance (π), derived allele frequency (p), the probability that disease alleles are derived alleles (d), the inbreeding coefficient (F), and the dominance coefficient (h). For a set of j SNPs, the mean disease burden can be found by averaging across allele frequencies and weighting by the probability that pathogenic alleles are derived as opposed to ancestral (see Appendix A for additional equations).

$$\overline{\beta} = \sum_{i=1}^{j} \pi_i \Big[\Big(d_i p_i^2 + (1 - d_i)(1 - p_i)^2 + p_i(1 - p_i)F \Big) + (2p_i(1 - p_i)(1 - F))h_i \Big] \Big/ j \qquad \text{(Equation 7)}$$

Because small sample sizes bias SFS toward intermediate frequency alleles,[37] we corrected the empirical SFS from whole-genome sequencing using *trueFS*.[38] Sites included in this analysis were required to be polymorphic in at least one population.

To determine whether β statistics are reasonable proxies of actual disease risk, we examined whether the effects of inbreeding on Equation 7 are comparable to increases in hereditary disease risk from clinical data. For each population, we used the corrected SFS of all non-CpG SNPs and assumed that disease alleles were derived and recessive. By setting $F = 0$ we inferred the predicted disease burden under random mating. We then simulated the effects of first-cousin mating by setting $F = 0.0625$. The ratio of $\beta_{first\text{-}cousin} / \beta_{random}$ was then calculated for each population. These values were then compared to clinical estimates of increased risks of hereditary disease due to inbreeding obtained from a global panel of 69 populations.[39]

Using corrected SFS and Equation 7, the predicted disease burden was obtained for unbiased WW or SS SNPs, SW SNPs, WS SNPs, and all non-CpG SNPs. We then compared the relative disease burden of each type of SNP (Equations A15, A16, and A17) and used Student's t tests to determine statistical significance. We also assessed whether the effects of gBGC are stronger for recessive alleles, for alleles with intermediate dominance, or for dominant alleles (Equations A10, A11, and A12), and how the proportion of disease alleles that are derived or ancestral influences disease burden (Equation A13). The relative disease burden of different populations was compared using Equations A18, A19, A20, and A21.

Because the SFS of disease alleles can be affected by natural selection, we used theoretical population genetics to examine genetic systems where there is a balance between mutation, selection, and gene conversion. Here, wild-type alleles can mutate to disease alleles. These recessive disease alleles are favored by gBGC when heterozygous and selected against when homozygous. Equilibrium allele frequency ($\hat{p}$) is affected by the strength of biased gene conversion (b), the strength of selection against recessive alleles (s), and the per generation rate of mutation (μ). Using Equation 4d from Glémin:[40]

$$\hat{p} = \frac{b + \sqrt{b^2 + 4\mu(b+s)}}{2(b+s)} \qquad \text{(Equation 8)}$$

Values of μ were obtained from the from the Kong et al., 2012 deCODE data set,[35] and values of b were estimated from relative rates of fixation of biased and unbiased sites.

Results

Genetic Distances and Rates of Evolution Are Modified by gBGC

GC-biased gene conversion results in slower rates of evolution for sites where ancestral alleles are favored (SW SNPs) and faster rates of evolution for sites where derived alleles are favored (WS SNPs). We calculated genetic distances for all ten population pairs using F_{ST} statistics, finding that WS SNPs result in mean values of F_{ST} that are 0.1% to 0.8% greater than WW or SS SNPs and 0.9% to 1.7% greater than SW SNPs. Bootstrapping 10,000 SNPs of each type from each population 1,000 times reveals that although the genome-wide effect of gBGC on F_{ST} is small, differences between types of SNPs are highly significant (p value $< 10^{-15}$ for all comparisons using Mann-Whitney U tests).

Population branch statistics (PBS) indicate that GC-biased gene conversion modifies evolutionary rates across all branches of a human evolutionary tree (Figure 1). When comparisons are made between different types of SNPs, we find the same rank order for each internal and external branch: SW SNPs have the smallest PBS, WW or SS SNPs have intermediate PBS, and WS SNPs have the largest PBS (p value $< 10^{-15}$ for all comparisons using Mann-Whitney U tests). PBS differences between different types of SNP are modest for each branch (WS SNPs have PBS statistics that are 1.9% to 3.1% greater than WW or SS SNPs and 4.5% to 6.6% greater than SW SNPs). However, values shown in Figure 1C are genome-wide estimates, and they include SNPs in recombination hotspots and coldspots. Furthermore, we find that PBS outliers (defined as the top 1% sites) are enriched for WS SNPs (p value < 0.0001 for all branches, two proportion Z-test). Comparing PBS statistics for different branches of the tree in Figure 1, we find that the largest branch lengths are for CEU and Hadza populations and WS SNPs. Note that PBS statistics reflect both divergence times and the amount of genetic drift that occurs along each branch of an evolution tree.

gBGC Leads to Shifts in the SFS

The site frequency spectrum is left-shifted when ancestral alleles are favored by GC-biased gene conversion and right-shifted when derived alleles are favored by GC-biased gene conversion. This pattern occurs for all five global populations (Figure 2). Note that SW SNPs (blue) are enriched for low-frequency (DAF = 0.1) alleles and WS SNPs (red) are enriched for high-frequency (DAF > 0.5) alleles. When these allele frequency shifts are quantified by population genetics statistics, we find that SNPs favored by gBGC tend to result in an excess of intermediate-frequency derived alleles. Tajima's D, which is positive when there is an excess of intermediate-frequency alleles, is significantly higher for WS SNPs than SW and unbiased WW or SS SNPs (Table 1, p value $< 10^{-15}$ for all populations using Mann-Whitney U tests). Similarly, Fay and Wu's H, which is negative when there is a lack of rare alleles, is significantly lower for WS than SW and unbiased WW or SS SNPs (p value $< 10^{-15}$ for all populations using Mann-Whitney U tests). The mean frequency of derived alleles also differs for different types of SNPs (p value $< 10^{-15}$ for all populations using Mann-Whitney U tests), with SW SNPs having the lowest derived allele frequencies and WS SNPs having the highest derived allele frequencies. We also note that Hadza and CEU populations have significantly higher values of Tajima's D and lower values of Fay and Wu's H than Pygmy, YRI, and Sandawe populations (Table 1 and Figure 3, p value $< 10^{-15}$ for all comparisons using Mann-Whitney U tests). This pattern is consistent with population growth for each of the latter three populations.

Effects of gBGC Are Stronger in Regions of High Recombination

Allele frequency shifts for WS SNPs are magnified in high-recombination regions of the genome, a pattern that is

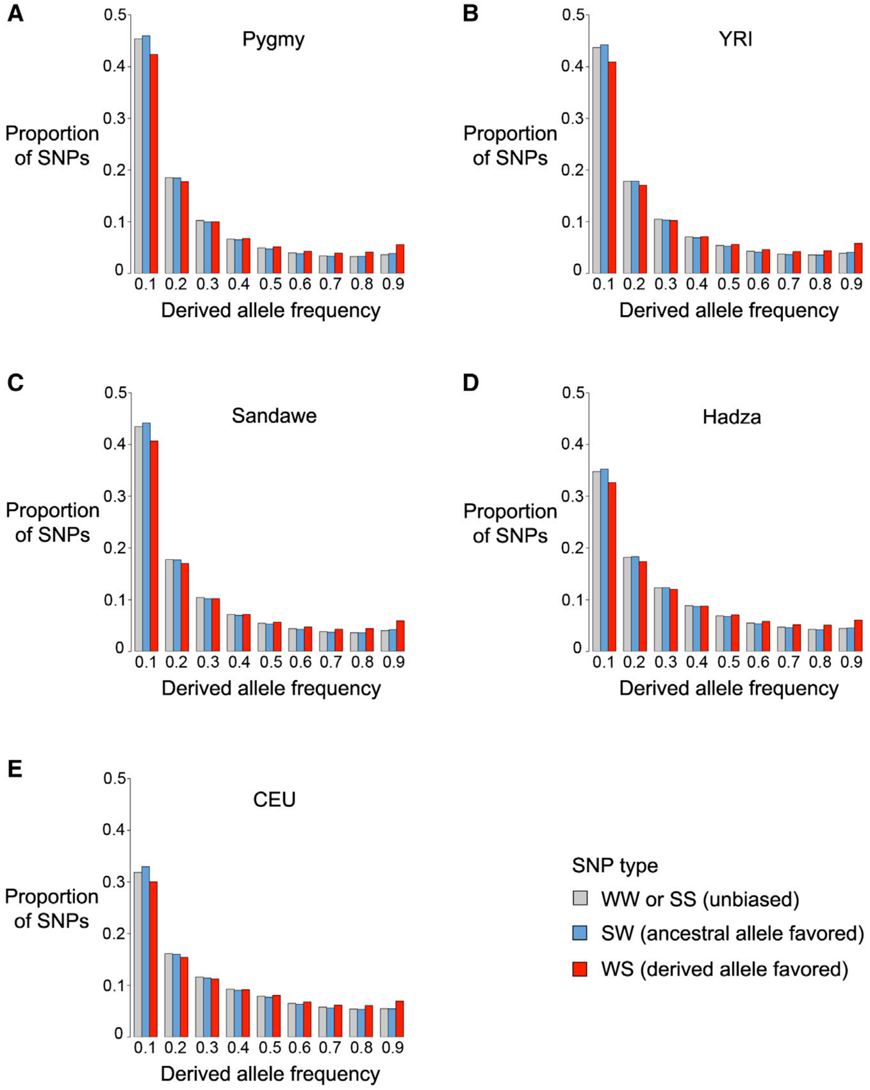


Figure 2. GC-Biased Gene Conversion Modifies the Site Frequency Spectra of Diverse Human Populations
Pygmy (A), YRI (B), and Sandawe (C) populations have an excess of low-frequency derived alleles relative to Hadza (D) and CEU (E) populations. Probability distributions sum to one for each type of SNP: WW or SS (gray), WS (blue), and WS (red). Site frequency spectra of SW SNPs are left-shifted and site frequency spectra of WS SNPs are right-shifted.

consistent with GC-biased gene conversion (Figure 3). The effects of recombination and gBGC on the SFS were quantified using multiple summary statistics: Tajima's D, Fay and Wu's H, mean DAF, and W→S DAF skew. We find that values of Tajima's D are higher for high-recombination regions of the genome for all three types of SNPs (p value $< 10^{-15}$ for Mann-Whitney U tests comparing bottom and top quintile data for all types of SNPs). This effect arises because Tajima's D is defined as the ratio of two random variables that are functions of the number of segregating sites in a sample.[41] Values of Fay and Wu's H are strikingly different for WS SNPs, and we find that values of H are lower in high-recombination regions of the genome (p value $< 10^{-15}$ for Mann-Whitney U tests comparing bottom and top quintile data for WS SNPs). Fay and Wu's H indicate that derived alleles are more likely to be found at intermediate and high frequencies for WS SNPs in high-recombination regions of the genome. On a related note, WS SNPs in high-recombination regions of the genome have a higher mean DAF than WS SNPs in low-recombination regions (p value $< 10^{-15}$ for Mann-Whitney U tests comparing bottom and top quintile data for WS SNPs). Finally, relative amounts of allele frequency skew for different types of SNPs were quantified using a W→S DAF skew statistic. This statistic compares the derived allele frequencies of WS and SW SNPs and is higher than 0.5 when G or C alleles are found at higher frequencies than A or T alleles.[16] We found similar patterns for each population: WS SNPs have a SFS that is skewed more toward high-frequency derived alleles than the SFS of SW SNPs, and this effect is greater in regions of high recombination (Figure 3). Although the deCODE recombination map uses data from Icelandic individuals, recombination rates had similar effects on European and African samples. This pattern probably arises because we used recombination rates averaged over 100 kb intervals and large-scale recombination rates are known to be similar for different populations.[42]

Recombination rates also influence genetic distances and rates of fixation. In the top recombination quintile, PBS statistics for WS SNPs are 4.6% greater than WW or SS SNPs and 9.7% greater than SW SNPs. By contrast, in the lowest recombination quintile, PBS statistics for WS SNPs are 0.6% greater than WW or SS SNPs and 2.4% greater than SW SNPs. Similarly, the strength of gBGC (as inferred by relative rates of fixation) is greater in high-recombination regions of the genome (Figure 4B).

Strength of gBGC

Using relative rates of fixation, we inferred that the strength of GC-biased gene conversion is akin to weak selection. Per base pair rates of substitution are greater for WS (4.50×10^{-3}) and SW (7.25×10^{-3}) sites compared to WW or SS sites (1.04×10^{-3}). However, mutation rates for WS sites are 3.61 times that of WW or SS sites and mutation rates for SW sites are 7.75 times that of WW or SS sites.[36] After correcting for mutation rate differences, $r_{WS} = 1.198$ and

Table 1. Biased Gene Conversion Modifies Summary Statistics of the Site Frequency Spectrum

Population	Type of SNP	Tajima's D (95% CI)	Fay and Wu's H (95% CI)	Mean DAF (95% CI)
Pygmy	WW or SS	−0.430 (−0.479 to −0.383)	0.264 (0.192 to 0.332)	0.271 (0.265 to 0.277)
	SW	−0.465 (−0.515 to −0.415)	0.248 (0.186 to 0.317)	0.270 (0.264 to 0.276)
	WS	−0.385 (−0.438 to −0.334)	−0.048 (−0.125 to −0.031)	0.297 (0.291 to 0.304)
YRI	WW or SS	−0.357 (−0.411 to −0.300)	0.182 (0.106 to 0.255)	0.282 (0.275 to 0.288)
	SW	−0.389 (−0.444 to −0.333)	0.172 (0.099 to 0.243)	0.281 (0.274 to 0.287)
	WS	−0.319 (−0.374 to −0.268)	−0.123 (−0.206 to −0.041)	0.307 (0.300 to 0.315)
Sandawe	WW or SS	−0.349 (−0.401 to −0.294)	0.159 (0.080 to 0.231)	0.284 (0.278 to 0.291)
	SW	−0.386 (−0.441 to −0.333)	0.152 (0.075 to 0.225)	0.282 (0.276 to 0.289)
	WS	−0.311 (−0.365 to −0.257)	−0.139 (−0.227 to −0.050)	0.309 (0.302 to 0.316)
Hadza	WW or SS	−0.011 (−0.072 to 0.046)	−0.013 (−0.109 to 0.075)	0.319 (0.311 to 0.326)
	SW	−0.036 (−0.094 to 0.022)	−0.003 (−0.087 to 0.081)	0.316 (0.309 to 0.323)
	WS	0.015 (−0.050 to 0.073)	−0.280 (−0.378 to −0.188)	0.340 (0.332 to 0.348)
CEU	WW or SS	0.102 (0.038 to 0.168)	−0.310 (−0.415 to −0.208)	0.348 (0.340 to 0.357)
	SW	0.057 (−0.011 to 0.123)	−0.291 (−0.393 to −0.191)	0.344 (0.336 to 0.352)
	WS	0.120 (0.051 to 0.189)	−0.550 (−0.657 to −0.445)	0.367 (0.358 to 0.376)

SNPs analyzed here are autosomal and non-CpG (a total of 7.54 million fully called SNPs). Values of Fay and Wu's H were normalized as per Zeng et al.[34] 95% confidence intervals of each statistic were found by bootstrapping 10,000 random SNPs a total of 1,000 times.

$r_{SW} = 0.899$. This indicates that evolutionary rates are accelerated at WS sites where derived alleles are favored by gene conversion and decelerated at SW sites where ancestral alleles are favored by gene conversion. Using population genetics theory, we inferred the mathematical relationship between rates of fixation and the population-scaled strength of gBGC (Equations 5 and 6), and the mapping of r to Nb is shown in Figure 4A. Nb was estimated to be 0.0934 for WS sites and 0.0523 for SW sites. These scaled gBGC coefficients are comparable to weak, nearly neutral selection ($|Ns| < 1$). Assuming $N = 10{,}000$, values of b range from 5.23×10^{-6} to 9.34×10^{-6}.

Predicted Disease Burden

β statistics calculated from SFS data (Equation 7) can be used as a proxy of hereditary disease risk. To test the validity of Equation 7, we estimated how inbreeding would increase the probability of observing a homozygote and compared this to known values of increased disease burden from the clinical literature. Using the corrected SFS for all non-CpG SNPs and comparing the relative homozygosity that would arise from first-cousin mating as opposed to random mating, we find that the predicted increase in homozygosity due to first-cousin mating ranges between 2.3% (CEU) and 3.8% (YRI). These values are comparable to clinical estimates of 3.5% excess mortality in the progeny of first cousins.[39] This suggests that it is reasonable to use corrected SFS and Equation 7 to infer how gBGC affects disease burden.

Allele frequency changes due to gBGC have a secondary effect of increasing the risk of hereditary disease. Because most deleterious mutations are recessive[43–46] and derived alleles are more likely to be pathogenic than ancestral alleles[47–49] (but see Di Rienzo and Hudson[50]), we focus on the disease burden of recessive derived alleles. Subsequently, this restriction is relaxed. After correcting for ascertainment bias due to small sample sizes, we weighted SNPs by the probability of observing a homozygote to obtain the mean recessive disease burden for unbiased and biased SNPs. The probability of observing derived homozygotes is similar for biased SW SNPs and unbiased WW or SS SNPs where the ancestral is favored by gene conversion ($\bar{\beta}_{SW}/\bar{\beta}_{WW\text{ or }SS} \approx 1$, p value > 0.3 using a two-tailed Student's t test). These similarities may arise because higher mutation rates for SW SNPs may balance out the effects of gBGC, or because small sample sizes limit our ability to distinguish between allele frequency shifts at the rare end of the SFS. However, SNPs where the derived allele is favored by gene conversion exhibited a strikingly different pattern (Figure 5). In all five populations, WS SNPs have a significantly higher probability of being homozygous than WW or SS SNPs ($\bar{\beta}_{WS}/\bar{\beta}_{WW\text{ or }SS} >> 1$, p value $< 1 \times 10^{-6}$ using a one-tailed Student's t test). This increased homozygosity translates to a 42.2% (Hadza) to 62.8% (Pygmy) increase in the predicted risk of recessive diseases. Because only a subset of all SNPs are WS SNPs, we also quantified the overall effect of gBGC on recessive disease burden by calculating the relative homozygosity of all non-CpG SNPs compared to unbiased SNPs (i.e., $\bar{\beta}_{\text{All non CpG}}/\bar{\beta}_{WW\text{ or }SS}$). Within each population, the overall predicted increase in disease burden due to gBGC ranges from 17.9% to 27.8% (Figure 5A). The probability of observing homozygous derived alleles also varies by population. Comparing different populations, we find that the predicted disease burden of recessive alleles

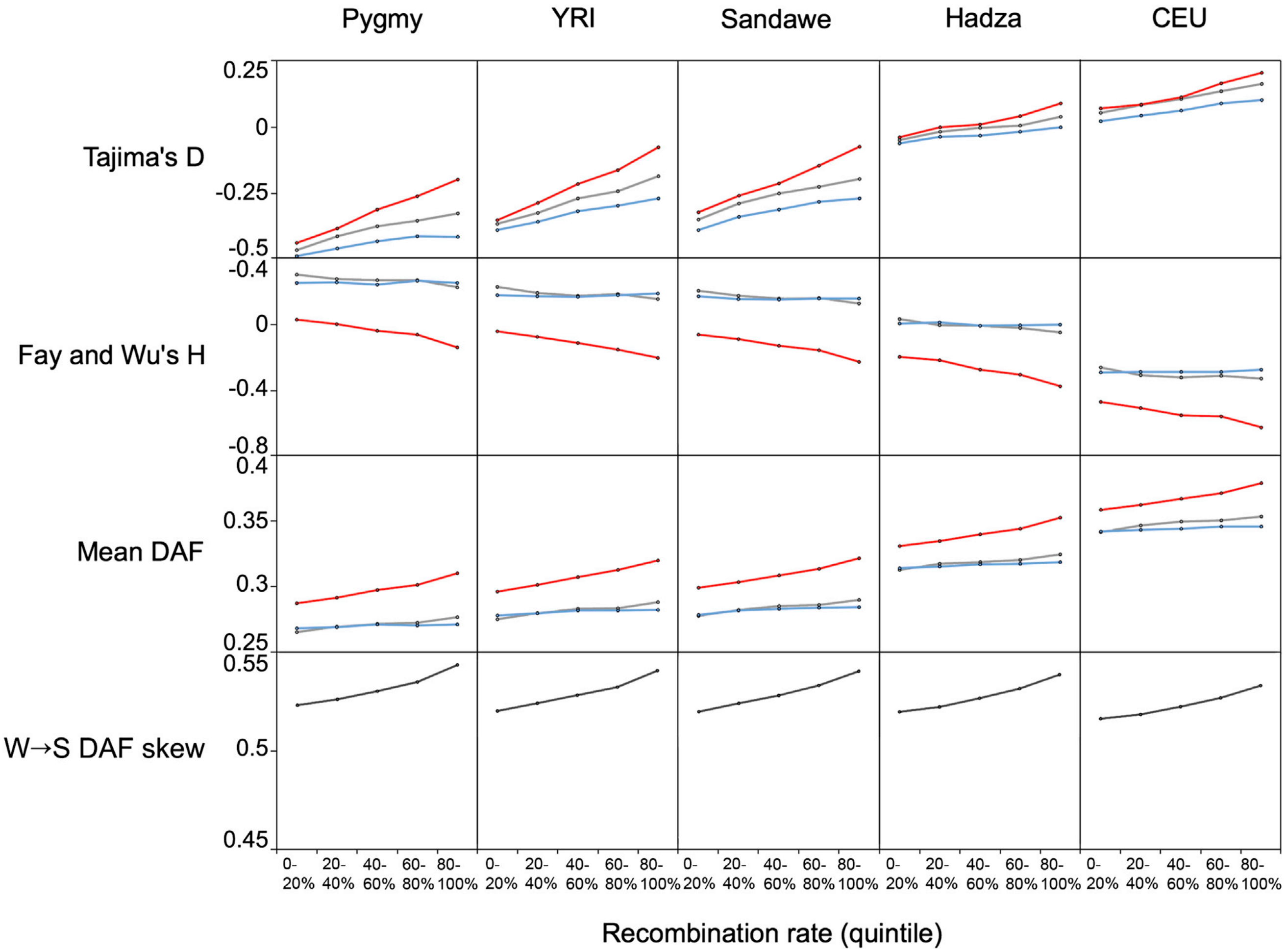


Figure 3. The Effects of Biased Gene Conversion Are Stronger in High-Recombination Regions of the Genome
SNPs were divided into quintile (20%) bins based on recombination rates from the 2010 deCODE data set. Four population genetic statistics were calculated for each population and recombination bin: Tajima's D, Fay and Wu's H, mean derived allele frequency (DAF), and a measure of W→S DAF skew. For the first three of these statistics, values were calculated separately for WW or SS (gray), SW (blue), and WS (red) SNPs.

is lowest for YRI genomes and highest for Hadza and CEU genomes (Figure 5B). This pattern is consistent with the ancestors of modern-day Hadza and Europeans having a lower effective population size due to population bottlenecks, reducing the efficacy of natural selection to eliminate deleterious mutations in each of these populations.

We relax the assumption that disease alleles are recessive and derived and find that predicted disease burden is still increased by gBGC, albeit to a lesser extent. The effects of gBGC on disease burden are strongest for recessive disease alleles and weakest for dominant alleles (Table 2). This occurs because the SFS is weighted toward rare alleles and small increases in the frequency of derived alleles lead to relatively large homozygosity increases. We also find that the effects of gBGC are stronger if 100% of disease alleles are derived as opposed to 90% derived and 10% ancestral (Table 2). It is likely that the effects of gBGC in different populations are modulated by demographic phenomena like population bottlenecks, admixture, and the explosive growth of modern human populations.[51]

Because selection against deleterious alleles can also modify the SFS and affect the risk of hereditary disease, we calculated the effects of gBGC for genetic systems that include mutation, selection, and gene conversion. Here, recessive disease alleles are removed by natural selection when homozygous and favored by gBGC when heterozygous, leading to a form of balancing selection. Using gBGC coefficients from Figure 4A and mutation rates from the 2012 deCODE data set,[35] we calculated equilibrium allele frequencies for a theoretical model of mutation-selection-conversion balance. Figure 5 shows equilibrium allele frequencies for biased WS SNPs (red), biased SW SNPs (blue). and unbiased WW or SS SNPs (gray). Due to elevated mutation rates, equilibrium allele frequencies for WS and SW SNPs are higher than unbiased WW or SS SNPs (Figure 5C). When selection is weak, equilibrium allele frequencies are driven by a balance between gBGC and mutation, and when selection is strong, equilibrium allele frequencies are driven by a balance between selection and mutation. Because otherwise deleterious alleles can be

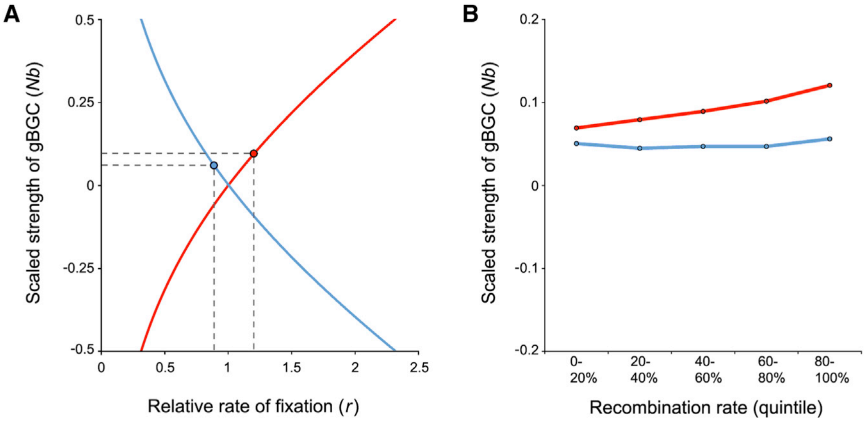


Figure 4. Estimated Strength of Biased Gene Conversion

After correcting for mutation rate differences, relative rates of fixation were used to estimate the population-scaled strength of gBGC.

(A) Curves describe the mathematical relationship between *Nb* and *r* for WS sites (red, Equation 5) and SW sites (blue, Equation 6). Circles denote genome-wide estimates of *Nb* from empirical data (0.0934 for WS sites and 0.0523 for SW sites).

(B) Effects of different recombination rate quintiles on the population-scaled strength of gBGC.

pushed to intermediate frequencies by biased gene conversion, recessive disease alleles are more likely to be homozygous at WS SNPs. This increased hereditary disease burden is magnified when selection is weak.

Discussion

The Population Genetic Effects of gBGC in Humans

Using high-coverage whole-genome sequencing data from multiple populations, we have demonstrated that GC-biased gene conversion modifies evolutionary distances and confirmed that allele frequency shifts are greater in high-recombination regions of the human genome. High-coverage sequence data minimizes the confounding effects of genotyping error, and by studying multiple populations we were able to show that the population genetics effects of gBGC are robust to demographic history. Statistical differences between different types of SNPs in a single population are comparable to statistical differences for the same type of SNP in different populations (Table 1). Because allele frequency distributions and population genetics statistics differ for variants that are favored or unfavored by gBGC, demographic inference is likely to be inaccurate if biased SNPs are analyzed. For example, computational tools like δaδi[52] rely on accurate SFS to infer demographic history. Modified values of F_{ST} can also lead to misestimates of population split times. Inclusion of WS SNPs results in lower values of Fay and Wu's H, and this can be misinterpreted as evidence of a recent population bottleneck (Table 1 and Figure 3). We also find that gBGC behaves like natural selection: the SFS of SW SNPs is left-shifted, a pattern that mimics negative selection, and the SFS of WS SNPs is right-shifted, a pattern that mimics positive selection (Figure 2). Because of this, studies that ignore gBGC may overestimate the effects of selection. Similarly, phylogenetic data from multiple primates indicate that gBGC results in elevated d_N/d_S ratios, a pattern that can be misinterpreted as selection.[4] gBGC decreases the allele frequencies of derived A or T alleles and increases the allele frequencies of derived G or C alleles. Furthermore, because A/T and G/C alleles differ in their chance of being passed to the next generation, the effects of gBGC are similar to meiotic drive. Classical population genetics theory does not focus on the molecular nature of alleles (i.e., whether variants involve adenine, thymine, guanine, or cytosine). Instead, it traditionally describes populations in terms of allele frequencies.[53,54] Our findings underscore the need for theoretical population genetics to include molecular phenomena such as biased gene conversion.

gBGC Is a Weak, but Important, Evolutionary Force

Comparisons between the relative rates of fixation of biased and unbiased sites reveal that the population-scaled strength of gBGC is on the order of $Nb \approx 0.0523$ to 0.0934 (Figure 4A). This indicates that gene conversion is a relatively weak force on a genomic scale—comparable to a nearly neutral allele under weak selection. Assuming an effective population size of 10,000 individuals, gBGC coefficients range between 5.23×10^{-6} and 9.34×10^{-6}. This means that WS or SW heterozygotes have a 50.000364% chance of passing on a G or C allele to their offspring, and gBGC results in non-Mendelian inheritance. As a point of comparison, gBGC coefficients are approximately 600 times greater than the genome-wide mutation rate parameter.[35] Even a modest amount of bias can have a noticeable effect on the genetics of populations because the effects of gBGC are compounded over evolutionary time. For example, we find that F_{ST} and PBS statistics are greater for SNPs favored by gBGC. We also note that *Nb* is greater in high-recombination regions of the genome (Figure 4B) and that the strength of gene conversion is known to be greater in recombination hotspots.[40]

This present study marks the first time that the evolutionary strength of gBGC in humans has been quantified using high-coverage whole-genome sequencing data. Our estimates of *Nb* for the top recombination quintile (0.056 for SW sites and 0.121 for WS sites) are roughly comparable to prior estimates that use data from chromosome 20 (0.325).[9,17] Aside from methodological differences, we note that estimates from prior studies are complicated by the use of SFS data from admixed African American samples. Our genome-wide estimates of *Nb* were also smaller

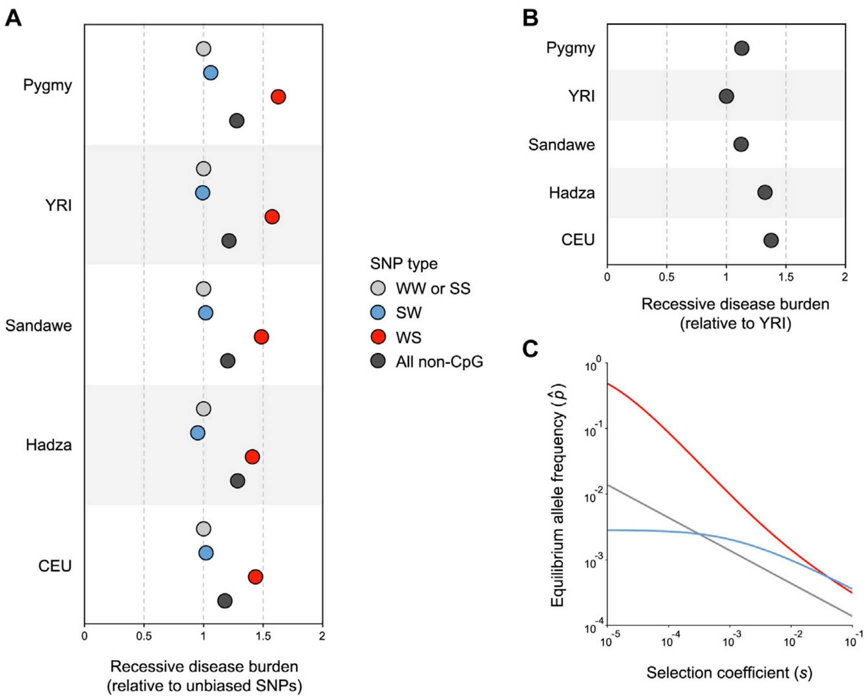


Figure 5. Predicted Disease Burden of Recessive Alleles and Mutation-Selection-Conversion Balance

(A) After correcting for small sample sizes using *trueFS*, the disease burden of recessive alleles was estimated from the corrected site frequency spectrum and the probability of observing homozygous individuals. Values for each population are normalized relative to unbiased WW or SS SNPs. The overall effect of gBGC (dark gray) was found by weighting the probability of observing WW or SS, SW, or WS SNPs.

(B) Relative disease burden of recessive alleles compared across different populations. Values shown are for all non-CpG SNPs and are normalized relative to YRI.

(C) The joint effects of mutation, selection, and GC-biased gene conversion on equilibrium allele frequencies. Equation 8 was used to generate equilibrium allele frequencies ($\hat{p}$) for deleterious recessive alleles at WS (red), SW (blue), and WW or SS SNPs (gray). Parameter values used: $b_{WS} = 9.34 \times 10^{-6}$, $b_{SW} = -5.23 \times 10^{-6}$ (negative because gBGC favors S alleles), $b_{WWorSS} = 0$, $\mu_{WS} = 6.89 \times 10^{-9}$, $\mu_{SW} = 1.48 \times 10^{-8}$, and $\mu_{WWorSS} = 1.91 \times 10^{-9}$.

than a previous study that focused on GC-rich coding regions[21] and a pair of studies that used phylogenetic methods to estimate the strength of gBGC.[55,56] Minor differences in the estimated strength of gBGC at WS and SW sites in Figure 4 may be due to ancestral state misidentification. Consider the situation where cytosine in a hypermutable CpG dinucleotide mutates to thymine and reaches fixation in chimpanzee and orangutan lineages, causing an ancestral C to be misinferred as T. If this site remains unchanged in the human lineage, it will incorrectly appear to be a WS substitution, and if this site also fixes in the human lineage it will incorrectly appear to be unchanged as opposed to a SW substitution. Taken together, these two scenarios suggest that the ancestral state misidentification may cause the strength of gBGC to be overestimated at WS sites and underestimated at SW sites. We also note that the effects of recombination rate on *Nb* appear to be greater for WS sites than SW sites. This pattern can arise if mutation rates for different types of sites are not independent of recombination rates. We anticipate that estimates of *Nb* in future studies will benefit from higher-resolution recombination maps (including maps of actual gene conversion events[57]) and more accurate mutation rates.

GC-biased gene conversion can be an important mechanism for loss of genetic variation and divergence of isolated populations, and it is known that small biases in gene conversion can dramatically affect fixation probabilities and segregation times.[15] Indeed, human accelerated regions of the genome are enriched for the signatures of biased gene conversion.[18] In contrast to selection, gBGC is sequence dependent and it can act genome-wide, while only a small fraction of base pairs are likely to actually be under selection. One key difference between gBGC and selection is that gene conversion tracts are on the order of a few hundred base pairs,[2] while genetic hitchhiking can influence linked variation up to 1 Mb away if selection is strong ($s > 0.01$).[58]

The Curse of the Converted

What sort of processes can cause disease alleles to be common? Our results indicate that biased gene conversion should be added to the list more familiar causes (population bottlenecks, evolutionary tradeoffs, and recessivity). We find that hereditary disease burden can be reasonably captured by the SFS and Equation 7. Allele frequency shifts due to gBGC result in a *curse of the converted*, whereby WS SNPs are more likely to result in genetic diseases. The increase in predicted disease burden can be substantial (+42.2% to 62.8% for recessive derived alleles). To our knowledge, this marks the first time that relative increases in hereditary disease risks due to gBGC have been quantified for human populations. The increased disease risk due to allelic gene conversion found here parallels the increased disease risk that arises from interlocus gene conversion between paralogs.[59] Similarly, gBGC tracts identified using a phylogenetic hidden Markov model (HMM) appear to be enriched for disease-associated polymorphisms.[60]

The *curse of the converted* is stronger for recessive disease alleles. We find that biased WS SNPs are more likely to be homozygous than biased SW or unbiased WW or SS SNPs, and this leads to an increased recessive disease burden (Figure 5). Increases in allele frequency at WS SNPS will have a disproportionate effect compared to

Table 2. Relative Disease Burden of Biased WS SNPs Compared to Unbiased WW or SS SNPs

Scenario	Pygmy	YRI	Sandawe	Hadza	CEU
Recessive alleles ($\pi = 1$, $h = 0$, $d = 1$, $F = 0$)	1.628	1.576	1.485	1.422	1.436
Alleles with intermediate dominance ($\pi = 1$, $h = 0.5$, $d = 1$, $F = 0$)	1.390	1.357	1.307	1.268	1.315
Dominant alleles ($\pi = 1$, $h = 1$, $d = 1$, $F = 0$)	1.294	1.268	1.229	1.198	1.251
Recessive alleles, 90% derived, 10% ancestral ($\pi = 1$, $h = 0$, $d = 0.9$, $F = 0$)	1.206	1.180	1.166	1.157	1.170

Corrected SFS and Equations A10, A11, A12, A13, A14, and A15 were used to obtain the relative disease burden ($\overline{\beta}_{WS}/\overline{\beta}_{WW\text{ or }SS}$) for each scenario. Parameters: π (penetrance), h (dominance coefficient), d (proportion of disease alleles that are derived), and F (inbreeding coefficient).

decreases in allele frequency at SW SNPs because the probability of observing a recessive homozygote is a function of the square of allele frequency. The overall effect of gBGC is to increase the homozygosity of derived alleles (and consequently increase the disease burden of recessive alleles). This is shown in Figure 5, as the set of all non-CpG SNPs has an increased recessive disease burden compared to unbiased WW or SS SNPs. The effects of gBGC on hereditary disease risk also apply to scenarios where disease alleles are not recessive, albeit to a lesser degree (Table 2). Intermediate dominance ($h = 0.5$) yields increases in risk that are two-thirds that of recessive alleles, and complete dominance yields increases in risk that are half that of recessive alleles. Similarly, the effects of gBGC on hereditary disease risk are reduced if a fraction of disease-causing alleles are ancestral.

The predicted disease burden also varies for different global populations. Specifically, genomes from bottlenecked Hadza and CEU populations are more likely to be homozygous for derived alleles than genomes from Pygmy, YRI, and Sandawe populations (Figure 5). This result is consistent with a previous study that used high-coverage whole-genome sequences to find that non-African genomes contain more damaging homozygous alleles than African genomes.[49] Similarly, a smaller proportion of genetic variation found in African and African American genomes involves deleterious nonsynonymous mutations.[23,61] By contrast, data from the 1000 Genomes Project suggest that non-African individuals do not have an excess of loss-of-function mutations.[62]

A general pattern is that the effects of gBGC are dampened by selection against disease alleles. Genetic systems where selection against recessive alleles is balanced by gene conversion and mutation have equilibrium allele frequencies that are substantially different than genetic systems with just selection and mutation (Figure 5C). The effects of gBGC on hereditary disease risk are robust to weak selection. These effects are greater for recessive alleles that are nearly neutral, and if selection is sufficiently weak, gBGC can result in the fixation of deleterious G or C alleles.[15] However, if selection is sufficiently strong ($s > 0.01$), the effects of gBGC on disease burden are likely to be minimal.

In conclusion, GC-biased gene conversion shapes patterns of diversity in human genomes, and it contributes to substantially increased risks of hereditary disease. These effects are stronger in high-recombination regions of the genomes and are observed in multiple populations. Genetic data obtained from high-coverage whole-genome sequencing suggest that realistic models of evolution should incorporate the details of molecular genetic phenomena like gene conversion.

Appendix A

Population Branch Statistics

Equations for population branch statistics (PBS) were obtained using pairwise genetic distances between populations (as quantified by F_{ST}) and the topology in Figure 1.

$$PBS_{Pygmy} = \frac{F_{ST(Pygmy,YRI)} + F_{ST(Pygmy,CEU)} - F_{ST(YRI,CEU)}}{2} \quad \text{(Equation A1)}$$

$$PBS_{YRI} = \frac{F_{ST(Pygmy,YRI)} + F_{ST(YRI,CEU)} - F_{ST(Pygmy,CEU)}}{2} \quad \text{(Equation A2)}$$

$$PBS_{((YRI,Pygmy),(CEU,(Hadza,Sandawe)))} = \frac{2F_{ST(Pygmy,CEU)} + F_{ST(YRI,Hadza)} + F_{ST(YRI,Sandawe)}}{4} - \frac{2F_{(Pygmy,YRI)} + F_{ST(CEU,Hadza)} + F_{ST(CEU,Sandawe)}}{4} \quad \text{(Equation A3)}$$

$$PBS_{CEU} = \frac{2F_{ST(YRI,CEU)} + F_{ST(CEU,Hadza)} + F_{ST(CEU,Sandawe)}}{4} - \frac{F_{ST(YRI,Hadza)} + F_{ST(YRI,Sandawe)}}{4} \quad \text{(Equation A4)}$$

$$PBS_{(((YRI,Pygmy),CEU),(Hadza,Sandawe))} = \frac{2F_{ST(CEU,Sandawe)} + F_{ST(Pygmy,Hadza)} + F_{ST(YRI,Hadza)}}{4} - \frac{2F_{(Sandawe,Hadza)} + F_{ST(Pygmy,CEU)} + F_{ST(YRI,CEU)}}{4} \quad \text{(Equation A5)}$$

$$PBS_{Hadza} = \frac{F_{ST(CEU,Hadza)} + F_{ST(Hadza,Sandawe)} - F_{ST(CEU,Sandawe)}}{2} \quad \text{(Equation A6)}$$

$$PBS_{Sandawe} = \frac{F_{ST(CEU,Sandawe)} + F_{ST(Hadza,Sandawe)} - F_{ST(CEU,Hadza)}}{2} \quad \text{(Equation A7)}$$

Disease Burden of Ancestral and Derived SNPs

Disease burden (β) is influenced by penetrance (π), derived allele frequency (p), the probability that disease alleles are derived alleles (d), the inbreeding coefficient (F), and the dominance coefficient (h). When ancestral alleles are pathogenic:

$$\beta_{d=0} = \pi\left[\left((1-p)^2 + p(1-p)F\right) + (2p(1-p)(1-F))h\right]. \tag{Equation A8}$$

When derived alleles are pathogenic:

$$\beta_{d=1} = \pi\left[\left(p^2 + p(1-p)F\right) + (2p(1-p)(1-F))h\right]. \tag{Equation A9}$$

Mean Disease Burden in Special Cases

Simplified equations for the mean disease burden per SNP can be found by considering special cases of Equation 7. Unless otherwise specified, these equations assume penetrance to be complete, derived alleles to be pathogenic, and populations to be outbred. Mean disease burden per SNP when disease alleles are recessive:

$$\overline{\beta}_{h=0,\ d=1,\ F=0} = \sum_{i=1}^{j}\left[p_i^2\right]\Big/ j. \tag{Equation A10}$$

Mean disease burden per SNP when disease alleles have intermediate dominance:

$$\overline{\beta}_{h=0.5,\ d=1,\ F=0} = \sum_{i=1}^{j}\left[p_i\right]\Big/ j. \tag{Equation A11}$$

Mean disease burden per SNP when disease alleles are dominant:

$$\overline{\beta}_{h=1,\ d=1,\ F=0} = \sum_{i=1}^{j}\left[p_i(2-p_i)\right]\Big/ j. \tag{Equation A12}$$

Mean disease burden per SNP when disease alleles are recessive (90% of pathogenic alleles derived and 10% of pathogenic alleles ancestral):

$$\overline{\beta}_{h=0,\ d=0.9,\ F=0} = \sum_{i=1}^{j}\left[0.9p_i^2 + 0.1(1-p_i)^2\right]\Big/ j. \tag{Equation A13}$$

Mean disease burden per SNP when disease alleles are recessive and there is first-cousin mating:

$$\overline{\beta}_{h=0,\ d=1,\ F=0.0625} = \sum_{i=1}^{j}\left[p_i^2 + 0.0625p_i(1-p_i)\right]\Big/ j. \tag{Equation A14}$$

Relative Disease Burden

The relative disease burden of different types of SNPs can be obtained by dividing the mean disease burden of a particular type of SNP by the mean disease burden of unbiased WW or SS SNPs.

$$\text{Relative disease burden of } WS \text{ SNPs} = \overline{\beta}_{WS}\Big/\overline{\beta}_{WW \text{ or } SS} \tag{Equation A15}$$

$$\text{Relative disease burden of } SW \text{ SNPs} = \overline{\beta}_{SW}\Big/\overline{\beta}_{WW \text{ or } SS} \tag{Equation A16}$$

$$\text{Overall effect of gBGC on disease burden} = \overline{\beta}_{\text{All non-CpG}}\Big/\overline{\beta}_{WW \text{ or } SS} \tag{Equation A17}$$

Similarly, the relative disease burden of different populations can be compared:

$$\text{Pygmy disease burden relative to YRI} = \overline{\beta}_{\text{All non-CpG, Pygmy}}\Big/\overline{\beta}_{\text{All non-CpG, YRI}} \tag{Equation A18}$$

$$\text{Sandawe disease burden relative to YRI} = \overline{\beta}_{\text{All non-CpG, Sandawe}}\Big/\overline{\beta}_{\text{All non-CpG, YRI}} \tag{Equation A19}$$

$$\text{Hadza disease burden relative to YRI} = \overline{\beta}_{\text{All non-CpG, Hadza}}\Big/\overline{\beta}_{\text{All non-CpG, YRI}} \tag{Equation A20}$$

$$\text{CEU disease burden relative to YRI} = \overline{\beta}_{\text{All non-CpG, CEU}}\Big/\overline{\beta}_{\text{All non-CpG, YRI}} \tag{Equation A21}$$

Acknowledgments

This work was supported by an NIH grant (DP1 ES022577) to S.A.T. and an NIH NRSA postdoctoral fellowship (F32HG006648) to J.L. We thank Katie Pollard, Amy Williams, Laurent Duret, Wen-Ya Ko, Matthew Hansen, other members of the S.A.T. lab, and two anonymous reviewers for their helpful comments and suggestions.

Received: January 24, 2014
Accepted: September 10, 2014
Published: October 2, 2014

References

1. Duret, L., and Galtier, N. (2009). Biased gene conversion and the evolution of mammalian genomic landscapes. Annu. Rev. Genomics Hum. Genet. *10*, 285–311.
2. Padhukasahasram, B., and Rannala, B. (2013). Meiotic gene-conversion rate and tract length variation in the human genome. Eur. J. Hum. Genet. Published online February 27, 2013. http://dx.doi.org/10.1038/ejhg.2013.30.
3. Chen, J.M., Cooper, D.N., Chuzhanova, N., Férec, C., and Patrinos, G.P. (2007). Gene conversion: mechanisms, evolution and human disease. Nat. Rev. Genet. *8*, 762–775.
4. Ratnakumar, A., Mousset, S., Glémin, S., Berglund, J., Galtier, N., Duret, L., and Webster, M.T. (2010). Detecting positive selection within genomes: the problem of biased gene conversion. Philos. Trans. R. Soc. Lond. B Biol. Sci. *365*, 2571–2580.
5. Ko, W.Y., Kaercher, K.A., Giombini, E., Marcatili, P., Froment, A., Ibrahim, M., Lema, G., Nyambo, T.B., Omar, S.A., Wambebe, C., et al. (2011). Effects of natural selection and gene

conversion on the evolution of human glycophorins coding for MNS blood polymorphisms in malaria-endemic African populations. Am. J. Hum. Genet. *88*, 741–754.

6. Brown, T.C., and Jiricny, J. (1989). Repair of base-base mismatches in simian and human cells. Genome *31*, 578–583.
7. Brown, T.C., and Jiricny, J. (1987). A specific mismatch repair event protects mammalian cells from loss of 5-methylcytosine. Cell *50*, 945–950.
8. Duret, L., and Arndt, P.F. (2008). The impact of recombination on nucleotide substitutions in the human genome. PLoS Genet. *4*, e1000071.
9. Spencer, C.C., Deloukas, P., Hunt, S., Mullikin, J., Myers, S., Silverman, B., Donnelly, P., Bentley, D., and McVean, G. (2006). The influence of recombination on human genetic diversity. PLoS Genet. *2*, e148.
10. Webster, M.T., and Smith, N.G. (2004). Fixation biases affecting human SNPs. Trends Genet. *20*, 122–126.
11. Frisse, L., Hudson, R.R., Bartoszewicz, A., Wall, J.D., Donfack, J., and Di Rienzo, A. (2001). Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. Am. J. Hum. Genet. *69*, 831–843.
12. Hellenthal, G., and Stephens, M. (2006). Insights into recombination from population genetic variation. Curr. Opin. Genet. Dev. *16*, 565–572.
13. Boulton, A., Myers, R.S., and Redfield, R.J. (1997). The hotspot conversion paradox and the evolution of meiotic recombination. Proc. Natl. Acad. Sci. USA *94*, 8058–8063.
14. Coop, G., and Myers, S.R. (2007). Live hot, die young: transmission distortion in recombination hotspots. PLoS Genet. *3*, e35.
15. Nagylaki, T. (1983). Evolution of a finite population under gene conversion. Proc. Natl. Acad. Sci. USA *80*, 6278–6281.
16. Katzman, S., Capra, J.A., Haussler, D., and Pollard, K.S. (2011). Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. Genome Biol. Evol. *3*, 614–626.
17. Necşulea, A., Popa, A., Cooper, D.N., Stenson, P.D., Mouchiroud, D., Gautier, C., and Duret, L. (2011). Meiotic recombination favors the spreading of deleterious mutations in human populations. Hum. Mutat. *32*, 198–206.
18. Dreszer, T.R., Wall, G.D., Haussler, D., and Pollard, K.S. (2007). Biased clustered substitutions in the human genome: the footprints of male-driven biased gene conversion. Genome Res. *17*, 1420–1430.
19. Pollard, K.S., Salama, S.R., Lambert, N., Lambot, M.A., Coppens, S., Pedersen, J.S., Katzman, S., King, B., Onodera, C., Siepel, A., et al. (2006). An RNA gene expressed during cortical development evolved rapidly in humans. Nature *443*, 167–172.
20. Galtier, N., and Duret, L. (2007). Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. Trends Genet. *23*, 273–277.
21. Galtier, N., Duret, L., Glémin, S., and Ranwez, V. (2009). GC-biased gene conversion promotes the fixation of deleterious amino acid changes in primates. Trends Genet. *25*, 1–5.
22. Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., et al. (2010). Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. Science *327*, 78–81.
23. Lachance, J., Vernot, B., Elbers, C.C., Ferwerda, B., Froment, A., Bodo, J.M., Lema, G., Fu, W., Nyambo, T.B., Rebbeck, T.R., et al. (2012). Evolutionary history and adaptation from high-coverage whole-genome sequences of diverse African hunter-gatherers. Cell *150*, 457–469.
24. Lam, H.Y., Clark, M.J., Chen, R., Chen, R., Natsoulis, G., O'Huallachain, M., Dewey, F.E., Habegger, L., Ashley, E.A., Gerstein, M.B., et al. (2012). Performance comparison of whole-genome sequencing platforms. Nat. Biotechnol. *30*, 78–82.
25. Gaffney, D.J., and Keightley, P.D. (2008). Effect of the assignment of ancestral CpG state on the estimation of nucleotide substitution rates in mammals. BMC Evol. Biol. *8*, 265.
26. Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution *38*, 1358–1370.
27. Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.
28. Schlebusch, C.M., Skoglund, P., Sjödin, P., Gattepaille, L.M., Hernandez, D., Jay, F., Li, S., De Jongh, M., Singleton, A., Blum, M.G., et al. (2012). Genomic variation in seven Khoe-San groups reveals adaptation and complex African history. Science *338*, 374–379.
29. Shriver, M.D., Kennedy, G.C., Parra, E.J., Lawson, H.A., Sonpar, V., Huang, J., Akey, J.M., and Jones, K.W. (2004). The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum. Genomics *1*, 274–286.
30. Sarich, V.M., and Wilson, A.C. (1973). Generation time and genomic evolution in primates. Science *179*, 1144–1147.
31. R Core Team (2013). R: A Language and Environment for Statistical Computing (Vienna: R Foundation for Statistical Computing).
32. Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.
33. Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.
34. Zeng, K., Fu, Y.X., Shi, S., and Wu, C.I. (2006). Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics *174*, 1431–1439.
35. Kong, A., Thorleifsson, G., Gudbjartsson, D.F., Masson, G., Sigurdsson, A., Jonasdottir, A., Walters, G.B., Jonasdottir, A., Gylfason, A., Kristinsson, K.T., et al. (2010). Fine-scale recombination rate differences between sexes, populations and individuals. Nature *467*, 1099–1103.
36. Kimura, M. (1962). On the probability of fixation of mutant genes in a population. Genetics *47*, 713–719.
37. Lachance, J., and Tishkoff, S.A. (2013). SNP ascertainment bias in population genetic analyses: why it is important, and how to correct it. Bioessays *35*, 780–786.
38. Nielsen, R., Hubisz, M.J., and Clark, A.G. (2004). Reconstituting the frequency spectrum of ascertained single-nucleotide polymorphism data. Genetics *168*, 2373–2382.
39. Bittles, A.H., and Black, M.L. (2010). Evolution in health and medicine Sackler colloquium: Consanguinity, human evolution, and complex diseases. Proc. Natl. Acad. Sci. USA *107* (*Suppl 1*), 1779–1786.
40. Glémin, S. (2010). Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. Genetics *185*, 939–959.
41. Thornton, K. (2005). Recombination and the properties of Tajima's D in the context of approximate-likelihood calculation. Genetics *171*, 2143–2148.

42. Wegmann, D., Kessner, D.E., Veeramah, K.R., Mathias, R.A., Nicolae, D.L., Yanek, L.R., Sun, Y.V., Torgerson, D.G., Rafaels, N., Mosley, T., et al. (2011). Recombination rates in admixed individuals identified by ancestry-based inference. Nat. Genet. *43*, 847–853.
43. Wilkie, A.O. (1994). The molecular basis of genetic dominance. J. Med. Genet. *31*, 89–98.
44. Haldane, J.B.S. (1927). A mathematical theory of natural and artificial selection, part V: selection and mutation. Math. Proc. Camb. Philos. Soc. *23*, 838–844.
45. Mukai, T., Chigusa, S.I., Mettler, L.E., and Crow, J.F. (1972). Mutation rate and dominance of genes affecting viability in *Drosophila melanogaster.* Genetics *72*, 335–355.
46. Halligan, D.L., and Keightley, P.D. (2009). Spontaneous mutation accumulation studies in evolutionary genetics. Annu. Rev. Ecol. Evol. Syst. *40*, 151–172.
47. Lachance, J. (2010). Disease-associated alleles in genome-wide association studies are enriched for derived low frequency alleles relative to HapMap and neutral expectations. BMC Med. Genomics *3*, 57.
48. Kumar, S., Sanderford, M., Gray, V.E., Ye, J., and Liu, L. (2012). Evolutionary diagnosis method for variants in personal exomes. Nat. Methods *9*, 855–856.
49. Torkamani, A., Pham, P., Libiger, O., Bansal, V., Zhang, G., Scott-Van Zeeland, A.A., Tewhey, R., Topol, E.J., and Schork, N.J. (2012). Clinical implications of human population differences in genome-wide rates of functional genotypes. Front. Genet. *3*, 211.
50. Di Rienzo, A., and Hudson, R.R. (2005). An evolutionary framework for common diseases: the ancestral-susceptibility model. TRENDS in Genetics *21*, 596–601.
51. Keinan, A., and Clark, A.G. (2012). Recent explosive human population growth has resulted in an excess of rare genetic variants. Science *336*, 740–743.
52. Gutenkunst, R.N., Hernandez, R.D., Williamson, S.H., and Bustamante, C.D. (2009). Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet. *5*, e1000695.
53. Crow, J.F., and Kimura, M. (1970). Introduction to Population Genetics Theory (New York: Harper & Row).
54. Charlesworth, B., and Charlesworth, D. (2010). Elements of Evolutionary Genetics (Greenwood Village: Roberts & Company).
55. Lartillot, N. (2013). Phylogenetic patterns of GC-biased gene conversion in placental mammals and the evolutionary dynamics of recombination landscapes. Mol. Biol. Evol. *30*, 489–502.
56. De Maio, N., Schlötterer, C., and Kosiol, C. (2013). Linking great apes genome evolution across time scales using polymorphism-aware phylogenetic models. Mol. Biol. Evol. *30*, 2249–2262.
57. Comeron, J.M., Ratnappan, R., and Bailin, S. (2012). The many landscapes of recombination in *Drosophila melanogaster.* PLoS Genet. *8*, e1002905.
58. Pritchard, J.K., and Przeworski, M. (2001). Linkage disequilibrium in humans: models and data. Am. J. Hum. Genet. *69*, 1–14.
59. Casola, C., Zekonyte, U., Phillips, A.D., Cooper, D.N., and Hahn, M.W. (2012). Interlocus gene conversion events introduce deleterious mutations into at least 1% of human genes associated with inherited disease. Genome Res. *22*, 429–435.
60. Capra, J.A., Hubisz, M.J., Kostka, D., Pollard, K.S., and Siepel, A. (2013). A model-based analysis of GC-biased gene conversion in the human and chimpanzee genomes. PLoS Genet. *9*, e1003684.
61. Lohmueller, K.E., Indap, A.R., Schmidt, S., Boyko, A.R., Hernandez, R.D., Hubisz, M.J., Sninsky, J.J., White, T.J., Sunyaev, S.R., Nielsen, R., et al. (2008). Proportionally more deleterious genetic variation in European than in African populations. Nature *451*, 994–997.
62. MacArthur, D.G., Balasubramanian, S., Frankish, A., Huang, N., Morris, J., Walter, K., Jostins, L., Habegger, L., Pickrell, J.K., Montgomery, S.B., et al.; 1000 Genomes Project Consortium (2012). A systematic survey of loss-of-function variants in human protein-coding genes. Science *335*, 823–828.