The Role of GC-Biased Gene Conversion in Shaping the Fastest Evolving Regions of the Human Genome

Dennis Kostka,*[,1] Melissa J. Hubisz,[2] Adam Siepel,[2] and Katherine S. Pollard[1,3]

[1]Gladstone Institute of Cardiovascular Disease, University of California, San Francisco
[2]Department of Biological Statistics and Computational Biology, Cornell University
[3]Division of Biostatistics & Institute for Human Genetics, University of California, San Francisco

***Corresponding author:** E-mail: dennis.kostka@gladstone.ucsf.edu.

Associate editor: Koichiro Tamura

Abstract

GC-biased gene conversion (gBGC) is a recombination-associated evolutionary process that accelerates the fixation of guanine or cytosine alleles, regardless of their effects on fitness. gBGC can increase the overall rate of substitutions, a hallmark of positive selection. Many fast-evolving genes and noncoding sequences in the human genome have GC-biased substitution patterns, suggesting that gBGC—in contrast to adaptive processes—may have driven the human changes in these sequences. To investigate this hypothesis, we developed a substitution model for DNA sequence evolution that quantifies the nonlinear interacting effects of selection and gBGC on substitution rates and patterns. Based on this model, we used a series of lineage-specific likelihood ratio tests to evaluate sequence alignments for evidence of changes in mode of selection, action of gBGC, or both. With a false positive rate of less than 5% for individual tests, we found that the majority (76%) of previously identified human accelerated regions are best explained without gBGC, whereas a substantial minority (19%) are best explained by the action of gBGC alone. Further, more than half (55%) have substitution rates that significantly exceed local estimates of the neutral rate, suggesting that these regions may have been shaped by positive selection rather than by relaxation of constraint. By distinguishing the effects of gBGC, relaxation of constraint, and positive selection we provide an integrated analysis of the evolutionary forces that shaped the fastest evolving regions of the human genome, which facilitates the design of targeted functional studies of adaptation in humans.

Key words: genome evolution, conserved noncoding elements, lineage-specific adaption, human accelerated regions, GC-biased gene conversion.

Research article

Introduction

Many recent studies have compared the human genome with those of other mammals, with the goal of identifying signatures of adaptive evolution and thereby gaining insight into the genetic basis of human-specific biology (Clark et al. 2003; Pollard, Salama, Lambert, et al. 2006; Prabhakar et al. 2006; Bird et al. 2007; Kim and Pritchard 2007; Kosiol et al. 2008). In protein-coding sequences, a high rate of amino acid-changing substitutions (compared with the rate of synonymous substitutions) is considered to be a hallmark of positive selection. Noncoding regions do not have such a natural partition of substitutions into functional and neutral classes. Instead, researchers have focused on the overall rate of substitutions on a lineage of interest in relation to the expected rate, given the level of conservation between multiple species at the same locus. This method is particularly effective for identifying highly conserved noncoding elements (CNEs) that have experienced a burst of substitutions on a particular lineage. It has been applied to fruit flies (Holloway et al. 2008), specific mammalian lineages (Kim and Pritchard 2007), and humans (Pollard, Salama, Lambert, et al. 2006; Prabhakar et al. 2006; Bird et al. 2007), where the fast-evolving sequences are called human accelerated regions (HARs) or human accelerated conserved noncoding sequences. Experimental investigations have established that some HARs function as RNA genes (Pollard, Salama, Lambert, et al. 2006) and tissue-specific enhancers (Prabhakar et al. 2008). These and other studies currently underway aim to elucidate the impact of human-specific substitutions on the function of HARs.

Focusing on lineage-specific evolution in CNEs has several advantages over analyzing the entire noncoding genome. First, the power to detect substitution rate acceleration is much higher against a background of conservation than when sequences have been evolving close to the neutral rate. Second, power and computational costs are improved by analyzing a small portion (typically 5–10%) of the genome. Finally, the unusually slow rate of evolution (in other species) suggests that CNEs play a functional role across the phylogeny under study. Thus, tests for lineage-specific acceleration in CNEs enable researchers to focus on those parts of the noncoding genome where statistically significant changes in substitution rates can be detected and where these substitutions are most likely to have a functional impact.

This approach, however, detects lineage-specific acceleration but not positive selection per se. This is because substitution rates, while significantly faster than expected given the extreme conservation in other species, might not exceed neutral rates and could therefore reflect the relaxation

of purifying selection and not adaptive evolution. Hence, additional post hoc statistical tests (e.g., comparing rates of substitutions to local neutral rates or tests based on polymorphism data) and functional studies are needed before an adaptive interpretation of an HAR or another accelerated element is appropriate.

Acceleration of substitution rates can also result from processes that do not involve changes in the mode of selection, such as GC-biased gene conversion (gBGC) (Galtier and Duret 2007). gBGC is a nonadaptive recombination-driven process. It is believed to result from a biochemical bias towards guanine or cytosine (GC) alleles in the mismatch repair of heteroduplex DNA during meiotic recombination. An effect of gBGC is to increase the rate of weak (A or T) to strong (G or C) substitutions and to decrease the rate of strong-to-weak substitutions, leading to higher GC-content in affected regions (Duret and Arndt 2008; Duret and Galtier 2009a; Romiguier et al. 2010). Once this process reaches equilibrium, gBGC typically decreases evolutionary rates. However, the initiation of a high rate of gBGC (e.g., because of the origin of a new recombination hot spot) can increase the overall rate of substitutions (Galtier and Duret 2007; Duret and Arndt 2008; Duret and Galtier 2009a) and thereby mimic positive selection. Investigations of the relationship between gBGC and accelerated substitution rates have largely focused on proteins, where gBGC can create spurious signals of positive selection nearby recombination hot spots (Berglund et al. 2009; Galtier et al. 2009). gBGC is estimated to have affected as many as ~20% of genes exhibiting elevated nonsynonymous substitution rates on short branches of the primate phylogeny (Ratnakumar et al. 2010), as well as many of the fastest evolving exons in the human genome (Berglund et al. 2009). GC-biased substitution patterns in some HARs suggest that CNEs may also be the targets of gBGC (Pollard, Salama, King, et al. 2006; Duret and Galtier 2009b; Prabhakar et al. 2009). These findings underscore the need to take gBGC into account in tests for lineage-specific acceleration.

Previous approaches to studying the impact of gBGC on tests for positive selection have generally used one method to identify genomic regions with high substitution rates, followed by a separate method to determine if the observed substitution patterns in these regions are due to gBGC (Dreszer et al. 2007; Berglund et al. 2009; Galtier et al. 2009; Ratnakumar et al. 2010). Duret and Arndt (2008) used a more integrated model to contrast gBGC with neutral evolution but did not investigate the interplay of gBGC and selection. To prioritize HARs for follow-up experiments that assess the functional consequences of human-specific substitutions and to determine the role of gBGC in shaping fast-evolving regions of the human genome, it is desirable to explicitly disentangle the effects of selection and gBGC on CNEs.

Motivated by this challenge, we developed a model for the evolution of nucleotide sequences that simultaneously accounts for both gBGC and selection. This approach enables us to capture the effects of gBGC on substitution rates and GC-content under a range of scenarios, from strong negative selection to strong positive selection. We do not attempt to model the complex process of gBGC in detail but instead focus on capturing its main effects on nucleotide substitution rates and patterns in the framework of statistical phylogenetics. We make use of this integrated model to develop a classification framework based on a series of likelihood ratio tests (LRTs). This enables us to annotate CNEs based on evidence of lineage-specific gBGC, relaxation of constraint, positive selection, and combinations of these forces. We demonstrate the performance of the method on simulated data and then apply it to annotate 202 HARs (Pollard, Salama, King, et al. 2006) with respect to their evolutionary histories. The resulting analyses suggest that substitutions in the majority of HARs cannot be explained by gBGC alone.

Materials and Methods

Modeling DNA Sequence Evolution under the Joint Action of Selection and gBGC

To investigate the interplay between selection and gBGC, we developed a molecular evolutionary model for lineage-specific changes in the rate and pattern of substitutions. Our approach builds upon the body of literature describing the use of continuous-time Markov chains in phylogenetic models of DNA sequence evolution, as reviewed by Liò and Goldman (1998). In those models, extant and extinct species are related via a binary tree (the species tree), and the likelihood of substitutions along the edges of the tree is governed by a 4×4 rate matrix $\boldsymbol{Q}$, which describes the instantaneous rate at which each nucleotide is substituted by others. In contrast, gBGC is typically modeled as an evolutionary force affecting the way certain alleles fix within a population (Gutz and Leslie 1976; Nagylaki 1983a). To combine the two approaches, we use the weak-mutation model (Golding and Felsenstein 1990) and multiply a neutral rate matrix $\boldsymbol{\mu}$ (describing how mutations arise within a population) with eventual fixation probabilities $\boldsymbol{f}$ to obtain $\boldsymbol{Q}$.

Following Nagylaki (1983a, 1983b), fixation probabilities under the joint action of selection and gBGC can be expressed in terms of two parameters, a selection coefficient $S \in (-\infty, \infty)$ and a gene conversion disparity $B \in [0, \infty)$, both scaled by population size:

$$f_{ij}(S, B) = \frac{1 - \exp\left[-\frac{1}{2N}(S + BI_{ij})\right]}{1 - \exp\left[-(S + BI_{ij})\right]}. \qquad (1)$$

The 4×4 matrix $\boldsymbol{I}$ (defined below) determines which fixation probabilities are affected by gene conversion, and N is the number of breeding individuals. By identifying N with the effective population size N_e, we can expect the fixation probabilities to hold under more general assumptions than the ones used to derive equation (1) (Nagylaki 1983a). We note that equation (1) implies an exponential decrease in the probability of fixation for strong-to-weak mutations under gBGC but only a linear increase for weak-to-strong mutations. Hence, evidence of $B > 0$ will ultimately be based at least as much on the absence of strong-to-weak

substitutions in an alignment as on the presence of weak-to-strong substitutions (see below).

Because gBGC favors strong (*C* or *G*) over weak (*A* or *T*) alleles, equally disfavors weak compared with strong alleles, and does not distinguish between *A* and *T* or between *C* and *G*, we have:

$$I_{ij} = \begin{cases} 1 & \text{for } i \text{ weak and } j \text{ strong} \\ -1 & \text{for } i \text{ strong and } j \text{ weak} \\ 0 & \text{otherwise,} \end{cases}$$

such that $S < 0$ represents purifying selection, $S > 0$ represents positive selection, and $B > 0$ represents a conversion bias towards strong versus weak alleles (gBGC). We note that the parameters *S* and *B* could in general be time dependent (see supplementary text 1, Supplementary Material online). In our approach, they are treated as constants, reflecting average effects over time. We estimate these parameters by combining data across multiple sites (i.e., alignment columns), and the same fixation model applies to all new alleles.

Next, we integrate the fixation probabilities with a continuous-time Markov model for sequence evolution by taking the instantaneous transition rates $\mathbf{Q}$ to be the (element-wise) product of mutation rates $\boldsymbol{\mu}$ and $\boldsymbol{f}(S, B)$ from equation (1) (Golding and Felsenstein 1990; McVean and Vieira 2001; Nielsen and Yang 2003; Kryazhimskiy and Plotkin 2008):

$$\begin{aligned} Q_{ij} &= 2N\mu_{ij}f_{ij}(S, B) \\ &\approx \mu_{ij}\,(S + BI_{ij})/(1 - \exp[-(S + BI_{ij})]). \end{aligned} \quad (2)$$

Here, $i \neq j$, and the diagonal entries of $\mathbf{Q}$ are determined by the constraint that rows of the rate matrix sum to zero. The factor $2N$ in equation (2) scales the mutation rate μ to the population level and assures that $\mathbf{Q} = \boldsymbol{\mu}$ in the absence of gBGC and selection (see below). Equation (2) is similar to the model of Nielsen and Yang (2003), which is concerned with the effects of positive selection in protein-coding DNA sequence. It is also close to the approach of Duret and Arndt (2008), except there, the authors restrict themselves to $S = 0$ (no selection); Harrison and Charlesworth (2011) recently used a similar composite modeling approach to study the effect of biased gene conversion on patterns of codon usage in yeast.

For neutral evolution ($B \to 0$ and $S \to 0$), we note that $2Nf_{ij} \to 1$ for all $i \neq j$ and $\mathbf{Q} \to \boldsymbol{\mu}$. For selection alone ($S \neq 0$ and $B = 0$), equation (2) reduces to $\mathbf{Q} = \boldsymbol{\mu}\,\rho(S)$, with $\rho(S) = S/(1 - \exp[-S])$. In the case of purifying selection, $S < 0$ and $0 \leqslant \rho < 1$, resulting in decreased substitution rates. For positive selection, $S > 0$ and $\rho > 1$, which implies an increase in substitution rates, as described in Pollard et al. (2010). Because the gBGC parameter *B* not only affects the rate of substitutions but also their pattern (see eq. 2), the complete model can be used to disentangle effects of selection and gBGC.

Inferring Substitution Rate Acceleration in the Presence of gBGC

We utilize the model in equation (2) to infer from alignments of DNA sequences of multiple species whether a change in the mode of selection, gBGC, or both has acted along a certain lineage in the species tree. Our goal was to assign alignments to one of four classes: gBGC (C_b, biased class), change in the mode of selection (C_a, accelerated class), both (C_{ab}), or neither (C_0, null class).

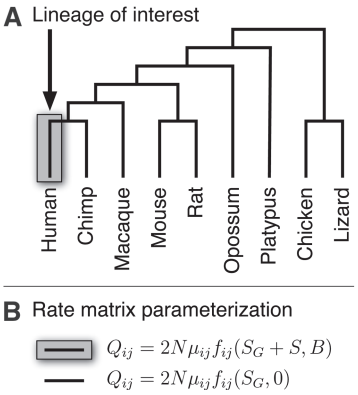

FIG. 1. Lineage-specific evolutionary histories. Panel A: For a branch (or a collection of branches) of interest, we assume a different substitution model compared with the rest of the tree. In addition to global rescaling (via the parameter S_G), a separate selection coefficient (*S*) and gene conversion disparity (*B*) lead to a seminested collection of models for sequence evolution on this lineage (see Materials and Methods). Panel B: Parameterization of the rate matrix $\mathbf{Q}$ for the two different types of branches (eq. 2).

For each alignment, we assume a neutral model, M_N, corresponding to $\mathbf{Q} = \mu$ on each branch of the species tree. For example, when analyzing the HARs, we estimate M_N from multiple sequence alignments of untranscribed flanking sequence (see supplementary text 2, Supplementary Material online). Another possibility would be to utilize 4-fold degenerate sites or ancestral repeats if a sufficient number of such sites are available near the locus of interest. Building on M_N, we model various lineage-specific evolutionary histories by defining a seminested collection of models, taking into account the background rate of substitutions at the locus (via a species tree-wide selection coefficient S_G). Each model is subject to different constraints on two parameters (the lineage-specific selection coefficient *S* and the lineage-specific conversion disparity *B*), and the entire analysis is performed with respect to a single lineage (branch) of interest in the species tree. The model parameters are used to calculate $\mathbf{Q}$ (S_G tree wide and *S* and *B* on the lineage of interest) according to equation (2) and taking μ from M_N (fig. 1). Each model corresponds to an annotation class (supplementary fig. S1, Supplementary Material online):

- Null Class (C_0). No lineage-specific processes. This class allows for a global selection coefficient S_G, acting on each branch of the species tree of M_N. This rescaling accounts for tree-wide purifying selection in CNEs. "Constraints": $S = S_G$ and $B = 0$, "Model": M_0.
- gBGC Class (C_b). Lineage-specific GC-biased substitution rate increase compared with other branches. In addition to tree-wide rescaling (via S_G), gBGC acts on the lineage of interest. Constraints: $S = S_G$ and $B \geqslant 0$, Model: M_b.
- Acceleration Class (C_a). Lineage-specific increase in substitution rate, without GC bias. In addition to tree-wide

rescaling (via S_G), unbiased acceleration acts on the lineage of interest. This model covers the scenarios of acceleration due to relaxation of constraint ($S \leqslant 0$) or positive selection ($S > 0$), and we later disentangle these two cases (see below). Constraints: $S \geqslant S_G$ and $B = 0$, Model: M_a.

– Acceleration and gBGC Class (C_{ab}). GC-biased substitution rate increase with additional unbiased substitutions. In addition to tree-wide rescaling (via S_G), unbiased acceleration and gBGC act together on the lineage of interest. Constraints: $S \geqslant S_G$ and $B \geqslant 0$, Model: M_{ab}.

We note that model M_0 is nested within models M_a and M_b and that both M_a and M_b are nested within model M_{ab}. All four models reduce to M_N when $S_G = 0$, $S = 0$, and $B = 0$. $S_G < 0$ with $S = S_G$ corresponds to purifying selection in all aligned species, as expected for CNEs. On the other hand, $S > S_G$ and/or $B > 0$ correspond to lineage-specific changes in substitution rate and/or patterns that cannot be accounted for by globally rescaling the species tree.

A Likelihood Ratio-Based Alignment Classification Procedure

To annotate an alignment D to a certain class, we perform a series of LRTs between the models introduced above. We use the `phyloFit` routine in RPHAST (Hubisz et al. 2011) to obtain likelihoods L_0, L_a, L_b, L_{ab} for models M_0, M_a, M_b, M_{ab}, respectively, by maximizing over the parameters S_G, S, and B within the constraints of each model:

$$\begin{aligned} L_0 &= \max_{S_G} \log P(D|M_0(S_G, S, B)) \\ &\quad \text{subject to} \quad S = S_G, B = 0 \\ L_b &= \max_{S_G,B} \log P(D|M_b(S_G, S, B)) \\ &\quad \text{subject to} \quad S = S_G, B \geqslant 0 \\ L_a &= \max_{S_G,S} \log P(D|M_a(S_G, S, B)) \\ &\quad \text{subject to} \quad S \geqslant S_G, B = 0 \\ L_{ab} &= \max_{S_G,S,B} \log P(D|M_{ab}(S_G, S, B)) \\ &\quad \text{subject to} \quad S \geqslant S_G, B \geqslant 0. \end{aligned} \tag{3}$$

These likelihoods depend on the initial neutral model M_N, but this dependence is not indicated in the notation for simplicity. To perform an LRT between model M_i and model M_j, we compare L_i with L_j, and we reject M_i in favor of M_j if $L_i > L_j + d_{i|j}$, where $d_{i|j} > 0$ is a constant defining the LRT's critical region.

We classify each alignment by a rigorous procedure that uniquely maps a series of LRTs to an annotation class, in a manner that is conservative with respect to annotating selection (supplementary text 3, Supplementary Material online). Briefly, we first compare each of the three lineage-specific models with the null model. In order to further disentangle gBGC from selection, we compare the lineage-specific models with each other. We then use the following three rules to classify an alignment: First, for an alignment to get annotated to a certain class, the associated model has to reject M_0. Second, if the LRTs imply clear preference for one model compared with all others, the corresponding class is chosen. Otherwise (if ties arise), we split them by preferring C_b over C_a and both C_a and C_b over C_{ab}. This approach uniquely maps a series of LRTs to annotation classes (supplementary fig. S3, Supplementary Material online), and supplementary text 3, Supplementary Material online, describes the mapping in more detail.

Preferring C_b over C_a makes our approach conservative with respect to annotating selection because all selection-annotated alignments show a significant preference towards selection as compared with gBGC alone. On the other hand, C_b can contain alignments with equal evidence for M_a and M_b. To identify alignments with strong evidence of gBGC, we split C_b into two subclasses: C_{b+} contains the elements of C_b where the LRT rejects M_a in favor of M_b and C_{b-} contains the other cases.

Critical Regions of the LRTs

In order to apply the LRTs, a critical region needs to be specified for each test. Our classification procedure considers at most seven LRTs, corresponding to all possible comparisons between M_0, M_b, M_a, and M_{ab} where the alternative is not nested within the null hypothesis. Instead of relying on asymptotic results for nested model LRTs, we determine the parameters $d_{i|j}$ via simulation in order to directly account for the sequence properties (e.g., GC-content, gap patterns) and the relatively short lengths of alignments we analyze (see below). The critical regions are defined via

$$\begin{aligned} d_{a|0}: &\sup_{S_G} P(L_a - L_0 > d_{a|0}|M_0) &&= \alpha \\ d_{b|0}: &\sup_{S_G} P(L_b - L_0 > d_{b|0}|M_0) &&= \alpha \\ d_{ab|0}: &\sup_{S_G} P(L_{ab} - L_0 > d_{ab|0}|M_0) &&= \alpha \\ d_{a|b}: &\sup_{S_G,B} P(L_a - L_b > d_{a|b}|M_b) &&= \alpha \\ d_{b|a}: &\sup_{S_G,S} P(L_b - L_a > d_{b|a}|M_a) &&= \alpha \\ d_{ab|b}: &\sup_{S_G,B} P(L_{ab} - L_b > d_{ab|b}|M_b) &&= \alpha \\ d_{ab|a}: &\sup_{S_G,S} P(L_{ab} - L_a > d_{ab|a}|M_a) &&= \alpha, \end{aligned} \tag{4}$$

where we take $\alpha = 0.05$. We approximate the suprema with maxima over a finite grid of parameter values, with the constraints given in equation (3) in effect (see below).

Distinguishing Positive Selection from Relaxation of Purifying Selection

For the acceleration class C_a, we know that the rate of substitutions on the lineage of interest exceeds the rate in M_0 (i.e., the most likely rate in the absence of lineage-specific effects). However, $S > S_G$ does not distinguish the action of positive selection ($S > 0$) from relaxation of constraint ($S < 0$). Note that $S = 0$ corresponds to the branch of interest in M_a having the same length as in the neutral model M_N. To annotate an alignment with respect to the specific type of change in mode of selection, we divide the class C_a

into two subclasses C_{a-} and C_{a+}. These classes correspond to two models M_{a-} and M_{a+}, nested within M_a, that differ from M_a via the parameter constraints $S_G < S \leqslant 0$ for M_{a-} (relaxation of constraint) and $0 < S$ for M_{a+} (positive selection). Note that M_{a-} is only defined for alignments evolving more slowly than the local neutral rate ($S_G < 0$), which is the case in CNEs. We perform an LRT between M_{a+} and M_{a-} and annotate an alignment to C_{a+} if we reject M_{a-} in favor of M_{a+}. Again, we take $\alpha = 0.05$ and determine the critical region via simulation. C_{a+} then contains C_a alignments with faster than neutral substitution rates, with a false positive rate of α. These are candidates for having experienced positive selection on the lineage of interest.

Classification of HARs

We use our approach to classify 202 HARs (Pollard, Salama, King, et al. 2006), according to their evolutionary histories. We first estimate a genome-wide strand symmetric model M_{4d} for neutral sequence evolution from 4-fold degenerate sites (supplementary text 2, Supplementary Material online). For each HAR, we obtain multiple sequence alignment data in the form of 28-way Multiz alignments from the University of California, Santa Cruz, CA, Genome Browser (http://genome.ucsc.edu). We retain sequences for human and up to nine vertebrates (chimp, macaque, mouse, rat, dog, opossum, platypus, chicken, and lizard), selected based on genome quality and phylogenetic position. For each HAR, we then estimate a local neutral model M_N by rescaling the species tree of M_{4d} to maximize the likelihood of all untranscribed 28-way alignment blocks 500 kb up- and downstream of the HAR (supplementary text 2, Supplementary Material online).

Having derived a local neutral model M_N for each HAR, we fit the models M_0, M_a, M_b, and M_{ab} to each HAR alignment and compute L_0, L_a, L_b, and L_{ab} (see above). Next, we derive the critical regions for the LRTs underlying our annotation procedure. To do so, we use simulations to approximate the suprema in equation (4). First, we calculate the empirical distribution of "gap patterns" G for each HAR alignment. A gap pattern relates to an alignment column and is a binary annotation of which species have gaps in that column. Applying a gap pattern ensures that the likelihoods of parametrically simulated alignments more accurately reflect those of real multiple sequence alignments. We then use G and M_N to generate data across an evenly spaced grid of parameter values: $S_G = \hat{S}_G, S = \hat{S}_G, \ldots, S_{\max}$, and $B = 0, \ldots, B_{\max}$, where $\hat{S}_G$ is the estimate obtained by fitting M_0 to the HAR alignment (supplementary text 4, Supplementary Material online). For each grid point (S, B), we generate 1,000 alignments as follows. We obtain a model with human-specific acceleration or gBGC by transforming M_N using equation (2) and a parameter combination $(\hat{S}_G, S, B)$. Using this transformed model, we parametrically generate 1,000 ungapped alignments of the same length as the HAR. For each ungapped alignment, we independently sample a gap pattern from G for each alignment column and use these patterns to mask the simulated alignment. We then maximize the likelihood for each of the models M_0, M_a, M_b, and M_{ab} corresponding to the null, acceleration, gBGC, and acceleration plus gBGC classes. Aggregating over alignments, this yields estimates $\hat{d}_{b|0}$, $\hat{d}_{a|0}$, $\hat{d}_{ab|0}$, $\hat{d}_{b|a}(S)$, $\hat{d}_{a|b}(B)$, $\hat{d}_{ab|b}(B)$, and $\hat{d}_{ab|a}(S)$ for each (S, B)-parameter combination. Finally, we determine the classification boundaries by taking maxima (corresponding to the suprema in eq. 4) over the finite grid of estimates, which empirically controls the false positive rate of each test to be not more than 5%. We perform similar calculations to refine class C_a into C_{a+} and C_{a-} and class C_b into C_{b+} and C_{b-}. We note that under this approach, each HAR has its own set of critical regions, reflecting its unique properties in terms of alignment length, substitution rate, and gap patterns.

Results

Jointly Modeling gBGC and Selection

We have developed a molecular evolutionary model and classification procedure to investigate the effects of gBGC and selection in a lineage of interest and applied it to make inferences about the recent evolution of HARs, some of the fastest evolving regions of the human genome (Pollard, Salama, King, et al. 2006). As described in the Materials and Methods, our approach uses the weak-mutation model (Liò and Goldman 1998) and multiplies a neutral rate matrix $\boldsymbol{\mu}$ by fixation probabilities $\boldsymbol{f}$ (eqs. 1 and 2) to obtain a rate matrix $\boldsymbol{Q}$ that accounts for both unbiased acceleration (via a selection coefficient S) and gBGC (via a gene conversion disparity B). This enables us to study the interplay between the two forces. We find that we can accurately recover selection parameters and gBGC disparities from sequence alignments (supplementary text 5, Supplementary Material online). We have made a software implementation of our approach publicly available as part of the R-package RPHAST, an R programming language interface to the open-source comparative, and evolutionary genomics software package PHAST (Hubisz et al. 2011).

We used this model of substitution processes in the presence of selection and/or gBGC to delineate four biologically relevant scenarios for the lineage-specific evolution of DNA sequence (see Materials and Methods). To identify changes in the rate or pattern of substitutions in the human lineage relative to other mammals, we define four classes of DNA sequence alignments: 1) human-specific acceleration (class C_a), 2) human-specific gBGC (class C_b), 3) both (class C_{ab}), or 4) neither (class C_0) (supplementary fig. S1, Supplementary Material online). Since HARs are CNEs in nonhuman species, we further refined the acceleration class into a subclass corresponding to relaxation of purifying selection (C_{a-}) and a subclass with faster than neutral substitution rates, suggestive of positive selection (C_{a+}). We then designed a classification procedure for assigning alignments to classes, based on a series of LRTs. This leads to a natural refinement of the gBGC class C_b into two subclasses C_{b+} and C_{b-}. Alignments are only assigned to the high-confidence gBGC subclass C_{b+} if there is strong evidence of GC-biased substitutions (false positive rate $\leqslant$5%, see Materials and Methods).

Table 1. Evolutionary Classification of HARs.

A							
Class	**Number of HARs**	**Number of Substitutions[a]**	**ΔGC[b]**	**Acceleration[c]**		**Recombination[d]**	
C_0	0	—	—	—	—	—	—
C_{b-}	10	2.30	1.46	19.46	3.02	1.23	0.39
C_{b+}	28	3.68	2.47	41.31	4.83	1.60	0.02
C_{a-}	42	2.48	0.05	11.94	1.00	0.97	0.88
C_{a+}	112	3.32	0.08	29.76	4.63	1.08	0.86
C_{ab}	10	4.50	2.49	46.29	5.97	1.53	0.15
B[e]							
Class	**Number of HARs**	**Number of Substitutions**	**ΔGC**	**Acceleration**		**Recombination**	
C_0	34	0.59	−0.08	1.00	0.18	1.10	0.61
C_{b-}	35	1.31	0.90	14.87	1.87	0.92	0.92
C_{b+}	15	3.00	2.21	45.32	4.45	1.64	0.05
C_{a-}	74	1.89	−0.17	9.33	1.00	1.17	0.43
C_{a+}	42	3.00	−0.08	29.04	4.11	1.17	0.45
C_{ab}	1	3.00	0.74	20.92	2.37	1.78	0.20

[a] Average number of substitutions per HAR.
[b] Average increase in GC-content per HAR, in percentage points.
[c] Average fold change in substitution rate, with respect to M_0 (left column) and M_N(right column), taking the maximum likelihood estimate for the branch length from the model corresponding to the annotated class.
[d] Average male recombination rate in cM per M_b, with *P*-values for a test of higher than expected recombination rate (second column).
[e] Same as panel A, with potential CpG dinucleotides masked in the analysis.

This classification procedure is designed to aid investigators in prioritizing HARs for experimental and bioinformatic follow-up studies. For example, HARs in the positive selection class C_{a+} are good candidates for functional studies of human-specific adaptation. Those in the relaxation of constraint class C_{a-} may be of interest as examples of loss of function (nonadaptive or adaptive) but would suggest different experiments than putative cases of functional acquisition. HARs in the gBGC class C_{b+} are interesting because they are instances of CNEs that have experienced excess and potentially deleterious substitutions in the human lineage. The acceleration and gBGC class C_{ab} contains HARs with many human-specific substitutions (not all of which are weak-to-strong), so that both evolutionary forces are needed to explain the differences between the human and chimpanzee versions of their sequences. These HARs may shed light on the interplay between selection and gBGC in the human genome.

Selection and gBGC Interact

To investigate the interaction between selection and gBGC, we used our model to explore the expected number of substitutions and the change in GC-content along a lineage of interest under a range of values for the selection coefficient *S* and the gene conversion disparity *B*. We utilized alignment data from 4-fold degenerate sites in the ENCODE regions (ENCODE Project Consortium 2007) to obtain estimates for the neutral mutation rates $\boldsymbol{\mu}$ in equation (2), assuming a general reversible parametrization (supplementary text 2, Supplementary Material online, and Materials and Methods). Then, we calculated the expected change in GC-content and the expected number of substitutions along a single branch (supplementary text 6, Supplementary Material online) for a grid of realistic (S, B) values. The expected number of substitutions under neutral evolution ($S = B = 0$) was set to 0.005 (i.e., one expected substitution per 200 bp, approximately what is observed on the human lineage since the chimp–human ancestor). We investigated $0 \leqslant B \leqslant 15$. This parameter range yielded expected increases in GC-content up to about 5%, a value that is observed in some of the HARs (table 1 and supplementary table S2, Supplementary Material online). The scaled gBGC disparity for the average recombination hot spot has been reported to be 8.7 (Ratnakumar et al. 2010), which is within the range we investigated. We explored selection coefficients of $-4 \leqslant S \leqslant 10$, corresponding to substitution rates as high as ten times the neutral rate, as observed in substitution hot spots (Duret and Arndt 2008). Negative values of *S* enabled us to explore the interplay between gBGC and purifying selection.

We found that the quantitative influence of gBGC on GC-content and substitution rate depends on the presence and level of selection. As expected, higher values of the gBGC disparity *B* lead to increased GC-content for all values of *S* (fig. 2). But this effect is not constant; the interplay between gBGC and selection results in a greater impact of gBGC on GC-content with increasing *S*. Similarly, the effect of gBGC on the rate of substitutions also depends on the value of *S*. While in general the number of expected substitutions is higher for larger values of *B*, the effect of gBGC on the substitution rate is most pronounced for small *S* and decreases as *S* increases. This trend includes cases where the selection coefficient $S = 0$ (no selection) or even $S < 0$ (negative selection), where we found that gBGC can still lead to substantially increased substitution rates at plausible values for the disparity parameter *B*. These findings suggest that it is critical to model gBGC and selection together to avoid spurious

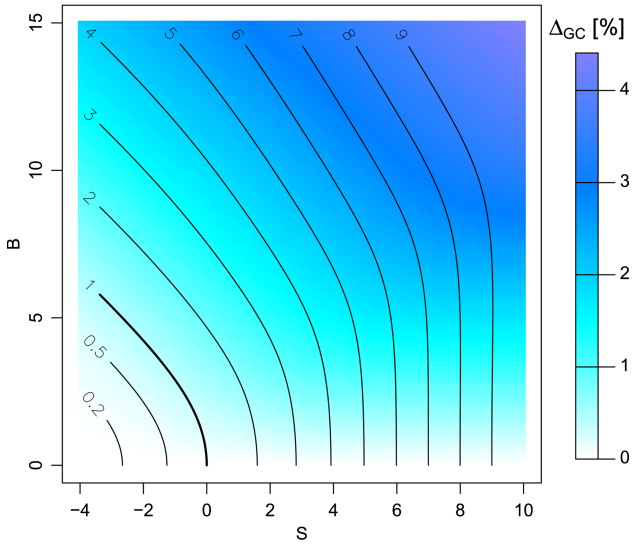

FIG. 2. Effects of gBGC and selection on GC-content and substitution rates. We investigated the substitution process along on a short branch (0.005 expected substitutions under the neutral model) for a range of values for the selection coefficient S and the gBGC disparity B. Each parameter combination has a unique effect on change in GC-content (Δ_{GC}, color scale) and the expected number of substitutions (contour lines, labels are the fold change) compared with a neutral model ($S = 0$ and $B = 0$, bold line). For any fixed selection coefficient S, increasing the gBGC disparity B increases the GC-content and the expected number of substitutions. These effects are nonlinear and depend on the level of selection.

conclusions about positive selection due to gBGC-induced elevations in substitution rates.

Detecting Substitution Rate Acceleration in the Presence of gBGC

We investigated the accuracy of our classification method for inferring the presence of substitution rate acceleration, gBGC, or a combination of both from alignment data. We aimed to determine if the two processes can be disentangled. If so, we also wanted to know whether our method has sufficient power to be applied to short genomic loci, such as HARs. To address these questions, we simulated alignments with different known combinations of selection coefficients and gBGC disparities on the human branch. The underlying neutral model was the same as in the previous section (supplementary text 2, Supplementary Material online), and we focused on the species tree for human, chimp, and macaque. For this simulation study, we chose $S_G = 0$ because we wanted to study the effects of both purifying and positive selection. We applied our classification method to assign each alignment to one of the four classes based on the pattern of substitutions on the human branch.

We were able to annotate 1,000-bp alignments to the correct class in most of the parameter space (fig. 3, Panel A). As expected, power is reduced for shorter alignments (more white area in Panels B and C of fig. 3), so that most nearly null 100-bp alignments are classified as coming from the null class C_0, rather than the correct acceleration or gBGC class. Nevertheless, we could still detect pronounced instances of gBGC, rate acceleration, and combinations of the two. Importantly, alignments generated with gBGC only ($S = S_G$) are almost never falsely annotated as belonging to the selection-only class C_a (low number of false positives), which implies that our assignment to annotation classes is conservative with respect to annotating selection. Also, most alignments generated with weak selection ($S < 2$) and strong gBGC are conservatively classified as C_0 or C_b. Both these features of the classification method are consequences of our choice of mapping the results of individual LRTs to annotation classes (see Materials and Methods). By adjusting this mapping (or the critical regions of some or all the LRTs), the same method could be tuned to detect weaker selection coefficients at the cost of more false positives. Overall, we conclude that our model captures the hallmarks of gBGC, and that our classification method is capable of delineating pronounced effects of selection and gBGC in short genomic elements.

gBGC Accounts for the Substitutions in Only a Minority of HARs

To explore the relative contributions of positive selection, relaxation of constraint, and gBGC to the fastest evolving regions of the human genome, we applied our classification procedure to the 202 HARs (Pollard, Salama, King, et al. 2006). We were curious to see whether the abundance of weak-to-strong substitutions in the HARs (see table 2) is indicative of gBGC being a dominating force. To that end, we carefully established the critical region for each LRT to be specific for each HAR, taking into account differences in local neutral rate, alignment length, gap pattern, and GC-content (see Materials and Methods). We find that 154 of the 202 HARs are assigned to the acceleration class C_a, 38 to the gBGC class C_b (28 to C_{b+} and 10 to C_{b-}), and 10 to the acceleration and gBGC class C_{ab} (table 1).

In other words, even when strictly controlling false positives, the majority of HARs (76%) can best be explained by a model with a change in the selection coefficient but no gBGC. Of the 154 HARs in C_a, 112 are assigned to the positive selection class C_{a+} because they have substitution rates that significantly exceed the local neutral rate. The remaining 42 are assigned to the relaxation of constraint class C_{a-}. We compared our method for class assignment with model selection via the Bayesian information criterion (BIC) and via the Akaike information criterion (AIC). All three selection criteria agree rather well (table 3). BIC tends to favor simpler models compared with AIC, as is expected (Hastie et al. 2009), whereas our LRT-based method results in a classification that is intermediate in complexity.

We analyzed the 647 individual substitutions in the 202 HARs to compare substitution rates and GC-content across the six classes. To quantify differences in substitution rates between classes, we calculated odds ratios (ORs) and performed Fisher's exact tests on the individual alignment columns (supplementary text 7, Supplementary Material online). Substitution rates are similar for C_{a+} and C_{b+} and they exceed those of the C_{a-} and C_{b-}, respectively (table 1). Since more HARs are assigned to the acceleration classes

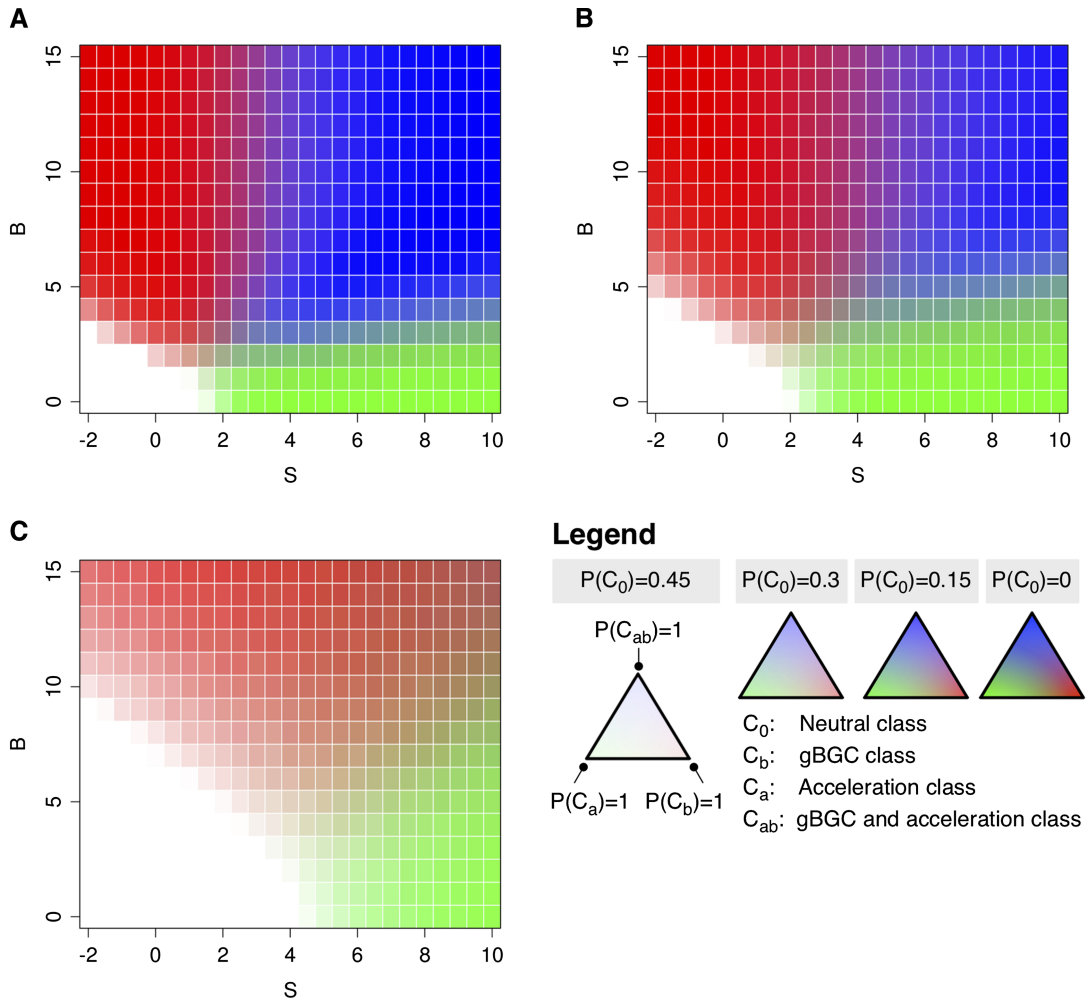

FIG. 3. Inferring acceleration in the presence of gBGC. We used simulations to determine the frequency with which alignments generated from models with a wide range of levels of selection (S) and gBGC (B) are assigned to each class by our methodology. Increasing brightness corresponds to a decreasing fraction of the null class (C_0). For the other three classes, the color representation corresponds to a point on the probability 2-simplex (red = gBGC, green = acceleration, blue = both). Because our classification procedure is conservative with respect to annotating selection, the red area is larger than the green area in each plot. Panel A: 1,000-bp alignments. Power is high and relatively few nonnull alignments are assigned to C_0 (white/light grid points). Panel B: 500-bp alignments. Power is slightly reduced. Panel C: 100-bp alignments. Power is significantly lower (more white/light grid points), but we are still able to correctly annotate most of the extreme instances of substitution rate acceleration in the presence of gBGC.

(and these classes have a high average number of substitutions), 74% (476 of 647) of all human-specific substitutions in HARs belong to C_a and 57% (394 of 647) belong to C_{a+}, which is a significant association (OR $= 1.26, P = 0.004$). As expected, substitutions in class C_b also significantly increase GC-content on average (OR $= 45.66$, $P < 1 \times 10^{-15}$), whereas GC-increasing substitutions are depleted in class C_a (OR $= 0.07$, $P < 1 \times 10^{-15}$).

Pollard, Salama, King, et al. (2006) found a positive correlation between acceleration level and the proportion of inferred human substitutions that were weak-to-strong across the 202 HARs, with the most striking evidence of gBGC in extremely accelerated HARs (HAR1–HAR5). Consistent with those results, we find that the proportion of HARs in class C_{b+} increases with acceleration level and is equal to the proportion of HARs in class C_{a+} for HAR1–HAR5 (supplementary table S3, Supplementary Material online). We assign HAR1 and HAR3 to C_{b+}, reflecting the fact that their weak-to-strong-biased substitution patterns are best explained by the gBGC model. This finding suggests the pursuit of experimental studies that explore possible

Table 2. Substitutions of Each Type across All HARs.

Type	Number of Substitutions (%)
Weak-to-strong	369 (57)
Strong-to-weak	187 (29)
Weak-to-weak	46 (7)
Strong-to-strong	45 (7)

Table 3. Number of HARs Annotated to Each Class via BIC, LRTs, and AIC.

Class	BIC	LRTs	AIC
C_b	37	38	35
C_a	164	154	149
C_{ab}	1	10	18

human-specific losses of function for these two elements. In contrast, HAR4 and HAR5 have unbiased substitution patterns and are assigned to class C_{a+}. Interestingly, our method assigns HAR2—about which there has been a lively debate regarding evolutionary history (Prabhakar et al. 2008; Prabhakar et al. 2009; Duret and Galtier 2009b)—to the mixed selection and gBGC class C_{ab} because it harbors mostly weak-to-strong substitutions but also one strong-to-strong substitution (which is unlikely under M_b, because HAR2 is under strong purifying selection with $S_G \ll 0$). Evidence of positive selection is strongest among the highly, but not extremely, accelerated HARs (66% of HAR6–HAR49 are assigned to class C_{a+}), suggesting prioritization of these HARs for studies of functional adaptation in humans. Supplementary table S2, Supplementary Material online, gives further details about evolutionary rates and patterns in each HAR.

HAR Classification Is Not Driven by CpG Effects

To account for the effect of high substitution rates in cytosine-phosphate-guanine (CpG) dinucleotides, we repeated our analyses after conservatively dropping all substitution columns that might correspond to human- and chimp-ancestral CpG sites (i.e., masking all substitutions except class 1 sites, as defined by Meunier and Duret 2004). Across all HARs, 264 of the estimated 647 human substitutions are masked. Of the masked substitutions, 173 were inferred to be weak-to-strong and 63 were inferred to be strong-to-weak in our original analysis. Because CpG masking decreases the number of strong-to-weak substitutions and the total number of substitutions, it reduces our power to distinguish selection from gBGC and neutral evolution, substantially affecting the classification of HARs. Fourteen HARs have no substitutions left after masking and cannot be accurately classified. As expected, the proportion of the remaining 188 HARs annotated to the neutral class C_0 is higher after CpG masking (34 versus 0 without masking). There is also an increase in the number of HARs in the gBGC class C_b (50 versus 38 without masking). Nonetheless, the majority of HARs (62%; 116 of 188 with substitutions) are still assigned to acceleration classes, 42 to the positive selection class C_{a+}, and 74 to the relaxation of constraint class C_{a-}. Relative rates of substitutions and changes in GC-content between classes are qualitatively similar to those in the unmasked analysis. Thus, our primary findings cannot be explained by CpG effects (table 1).

Recombination Rates Are Elevated near HARs in the gBGC Class

Our classification method does not explicitly demonstrate the action of selection or gBGC. In particular, HARs in classes C_b and C_{ab} may have been shaped by GC-biased fixation pressures other than gBGC, such as selection for higher GC-content. To address this distinction, Duret and Arndt (2008) investigated the association between equilibrium GC-content (GC*) and recombination rates. They reported a strong correlation between GC* and male recombination across the human genome and concluded that gBGC is a likely explanation for this phenomenon. Therefore, we analyzed male and female population-averaged recombination rates (Kong et al. 2002, 2010) in the regions around each HAR and compared these rates between different classes of HARs. For each class, we tested for an association with recombination rate by a bootstrap procedure that accounts for size differences between the classes (supplementary text 8, Supplementary Material online). Using the sex-specific recombination maps from Kong et al. (2002), we find that HARs in class C_{b+} tend to be located in regions with higher recombination rates than other HARs (table 1). Further, this association is more pronounced for male than female recombination (supplementary table S2, Supplementary Material online), consistent with earlier reports (Duret and Arndt 2008). We observe similar patterns with the Kong et al. (2010) recombination map, although the magnitudes of these effects and their sex bias do differ between data sets (supplementary table S1, Supplementary Material online). Thus, we find an association between recombination rate and GC-biased substitution patterns, which is consistent with gBGC playing a role in shaping the HARs assigned to the class C_{b+} by our methods.

Discussion

This study describes a new approach to disentangling the forces that have shaped the fastest evolving regions of the human genome. To address the question of how many human-specific accelerated CNEs can be explained by gBGC versus by selection, we developed a nucleotide substitution model that jointly accounts for selection and gBGC in an integrated framework. Using this model and criteria based on likelihood ratio statistics, we classified 202 HARs (Pollard, Salama, King, et al. 2006) according to evidence of changes in the mode of selection and/or gBGC. We find that substitution patterns in 76% of HARs are best explained by acceleration alone (class C_a). Further refining our annotation with respect to rate acceleration, we find that 55% of HARs have evolved too rapidly for relaxation of purifying selection to be a likely explanation (class C_{a+}) and 21% are consistent with relaxation of constraint (class C_{a-}). Nonetheless, a substantial minority of HARs are classified as having evolved under gBGC alone (class C_b: 19%, class C_{b+}: 14%). Our classification provides candidates for HARs with particular evolutionary histories, but further functional evidence and/or analyses of polymorphism data are needed before drawing any definitive conclusions about the action of selection or gBGC in a particular HAR.

One line of evidence supporting the hypothesis that gBGC shaped the substitution patterns in HARs in the C_{b+} class is our finding that male recombination rates are significantly higher than expected by chance near these HARs. Thus, we conclude that a sizable minority of the fastest evolving regions in the human genome may have been

shaped by gBGC. However, to directly infer a causal role of gBGC, additional complementary data would be needed to rule out other explanations, such as mutagenic effects of recombination itself or effects related to DNA melting temperature (Duret and Arndt 2008).

In our analysis, we focus on a preannotated set of short genomic regions, as opposed to conducting a genome-wide screen on the megabase scale (Dreszer et al. 2007; Duret and Arndt 2008). All our findings pertain to the 202 HARs (Pollard, Salama, King, et al. 2006) and should not be assumed to necessarily generalize to the whole genome or to other sets of CNEs. In particular, our estimates of the proportion of HARs influenced by selection or gBGC may not represent the genome-wide prevalences of these forces. HARs constitute a highly biased set of CNEs, selected on the basis of unusually high substitution rates on the human branch, and one could expect different results in a more balanced sample of CNEs. For instance, by applying our classification method to a random subset of 1,000 of the candidate chimp-, mouse-, and rat-conserved sequences from which the HARs were identified, we find less evidence of selection, as expected (supplementary text 9, Supplementary Material online). We also annotate a smaller percentage of candidate regions to the gBGC class C_b, although the ratio of regions in C_b compared with C_a is somewhat higher than in the HARs, suggesting that the prevalence of gBGC among CNEs could be higher genome wide than the 19% we estimate from HARs. Although the analyses presented here were not designed to estimate the prevalence of gBGC across the human genome, our approach could potentially be used in a future genome-wide study to address this question. Nonetheless, the fact that some HARs can be explained by gBGC alone does provide new evidence that nonadaptive forces could have important effects on average substitution patterns across the genome and may affect a subset of highly conserved sequences. Our study also does not provide information about the genome-wide association between recombination rates and the evolution of GC-content. But our results are consistent with previous work showing correlations between GC-biased substitutions and recombination rates in the human genome (Dreszer et al. 2007; Duret and Arndt 2008) and the genomes of other Metazoans (Capra and Pollard 2011).

Since HARs are short alignments (on the order of 100 bp) and the human branch is relatively short (approximately 0.5 substitutions per 100 bp under neutrality), phylogenomic inference is inherently limited by the small number of informative sites (i.e., sites with substitutions) present in the alignment data. We therefore consciously used a simple model with a constant gBGC disparity, as in Duret and Arndt (2008), which does not directly model mutagenic effects of gBGC and does not account for variation in substitution rates or patterns across alignment positions. Additionally, we make the common assumption of independence between alignment columns and do not account for dinucleotide effects. Also, our model does not allow for increased substitution rates due to relaxation of purifying selection for specific types of alleles or due to mutational biases other than weak-to-strong (Eyre-Walker 1992; Takano-Shimizu 1999; Lawrie et al. 2011). But we expect that our selection-associated models (M_a and M_{ab}) are not overly sensitive to such effects because they do not contain parameters that can directly account for the resulting biased substitution patterns. We did explore the possibility of relaxing the assumption of constant gBGC disparity by means of a mixture model and found no major impact on our classification results (supplementary text 1, Supplementary Material online). Although more complex models would certainly be more realistic, their use would probably lead to overfitting and parameter estimates with inflated variance. One compelling reason to model dinucleotides is CpG hypermutability, and we addressed this possible confounder by masking potential ancestral CpG sites. All our qualitative findings were robust to CpG masking.

This study provides a statistically motivated classification method based on a series of LRTs. This method enabled us to estimate the proportion of HARs affected by changes to the mode of selection and it is straightforward to generalize. Our classification also lends an intuitive quantitative interpretation of the contributions of selection and/or gBGC to the evolution of individual HARs by enabling us to assign each HAR to an evolutionary class (with or without each force) and to estimate the size of the HAR's selection coefficient and/or gBGC disparity (supplementary table S2, Supplementary Material online). These features of the methodology are especially desirable when prioritizing follow-up analyses and experiments. We note that inferences about selection and gBGC inevitably depend on the assumed neutral model, which we estimated (in the form of M_N, see Materials and Methods) from 4-fold degenerate sites and untranscribed sequence flanking each HAR. Although estimating true mutation rates is difficult, we believe this approach yields reasonable estimates, accounts for major known biases, and does not confound our findings. Also, our inferences about selection are based on substitution rates between reference sequences of multiple species and are therefore more indirect than inferences based on polymorphism data, which can be used to detect recent positive selection near HARs (Katzman et al. 2010).

We have presented a general approach for modeling substitution patterns that can be combined with many other models in statistical phylogenetics. The models we have discussed are implemented in RPHAST (Hubisz et al. 2011), an open-source software package for comparative and evolutionary genomics, which is publicly available at http://compgen.bscb.cornell.edu/rphast/. One extension of our current method is to integrate the joint effects of selection and gBGC into a codon model. Such a model could be used to develop tests for positive selection in protein-coding regions that account for gBGC or other substitution biases. Work is in progress to support such models in RPHAST.

Supplementary Material

Supplementary texts, tables S1–S5, and figures S1–S4 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We thank Jeff Wall for helpful discussions. This work was supported by National Institutes of Health (NIGMS) grant GM82901 to D.K., M.J.H., A.S., and K.S.P. A.S. was supported by a David and Lucile Packard Fellowship for Science and Engineering.

References

Berglund J, Pollard KS, Webster MT. 2009. Hotspots of biased nucleotide substitutions in human genes. *PLoS Biol*. 7(1): e1000026.

Bird C, Stranger B, Liu M, Thomas D, Ingle C, Beazley C, Miller W, Hurles M, Dermitzakis E. 2007. Fast-evolving noncoding sequences in the human genome. *Genome Biol*. 8(6):R118.

Capra JA, Pollard KS. 2011. Substitution patterns are GC-biased in divergent sequences across the metazoans. *Genome Biol Evol*. 3: 516–527.

Clark AG, Glanowski S, Nielsen R, et al. (17 co-authors). 2003. Inferring nonneutral evolution from human-chimp-mouse orthologous gene trios. *Science* 302(5652):1960–1963.

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(10):1420–1430.

Duret L, Arndt PF. 2008. The impact of recombination on nucleotide substitutions in the human genome. *PLoS Genet*. 4(5):e1000071.

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

Duret L, Galtier N. 2009b. Comment on "Human-specific gain of function in a developmental enhancer". *Science* 323(5915):714.

ENCODE Project Consortium. 2007. Identification and analysis of functional elements in 1% of the human genome by the encode pilot project. *Nature* 447(7146):799–816.

Eyre-Walker A. 1992. The effect of constraint on the rate of evolution in neutral models with biased mutation. *Genetics* 131(1): 233–234.

Galtier N, Duret L. 2007. Adaptation or biased gene conversion? Extending the null hypothesis of molecular evolution. *Trends Genet*. 23(6):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):1–5.

Golding B, Felsenstein J. 1990. A maximum likelihood approach to the detection of selection from a phylogeny. *J Mol Evol*. 31(6): 511–523.

Gutz H, Leslie JF. 1976. Gene conversion: a hitherto overlooked parameter in population genetics. *Genetics* 83(4):861–866.

Harrison RJ, Charlesworth B. 2011. Biased gene conversion affects patterns of codon usage and amino acid usage in the saccharomyces sensu stricto group of yeasts. *Mol Biol Evol*. 28(1):117–129.

Hastie T, Tibshirani R, Friedman J. 2009. The elements of statistical learning. 2nd ed. Springer Series in Statistics. New York: Springer.

Holloway AK, Begun DJ, Siepel A, Pollard KS. 2008. Accelerated sequence divergence of conserved genomic elements in Drosophila melanogaster. *Genome Res*. 18(10):1592–1601.

Hubisz MJ, Pollard KS, Siepel A. 2011. PHAST and RPHAST: phylogenetic analysis with space/time models. *Brief Bioinform*. 12(1): 41–51.

Katzman S, Kern AD, Pollard KS, Salama SR, Haussler D. 2010. GC-biased evolution near human accelerated regions. *PLoS Genet*. 6(5):e1000960.

Kim SY, Pritchard JK. 2007. Adaptive evolution of conserved noncoding elements in mammals. *PLoS Genet*. 3(9):1572–1586.

Kong A, Gudbjartsson DF, Sainz J, et al. (16 co-authors). 2002. A high-resolution recombination map of the human genome. *Nat Genet*. 31(3):241–247.

Kong A, Thorleifsson G, Gudbjartsson DF, et al. (15 co-authors). 2010. Fine-scale recombination rate differences between sexes, populations and individuals. *Nature* 467(7319):1099–1103.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A. 2008. Patterns of positive selection in six mammalian genomes. *PLoS Genet*. 4(8):e1000144.

Kryazhimskiy S, Plotkin JB. 2008. The population genetics of dN/dS. *PLoS Genet*. 4(12):e1000304.

Lawrie DS, Petrov DA, Messer PW. 2011. Faster than neutral evolution of constrained sequences: the complex interplay of mutational biases and weak selection. *Genome Biol Evol*. 3:383–395.

Liò P, Goldman N. 1998. Models of molecular evolution and phylogeny. *Genome Res*. 8(12):1233–1244.

McVean GA, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics* 157(1):245–257.

Meunier J, Duret L. 2004. Recombination drives the evolution of GC-content in the human genome. *Mol Biol Evol*. 21(6):984–990.

Nagylaki T. 1983a. Evolution of a finite population under gene conversion. *Proc Natl Acad Sci U S A*. 80(20):6278–6281.

Nagylaki T. 1983b. Evolution of a large population under gene conversion. *Proc Natl Acad Sci U S A*. 80(19):5941–5945.

Nielsen R, Yang Z. 2003. Estimating the distribution of selection coefficients from phylogenetic data with applications to mitochondrial and viral DNA. *Mol Biol Evol*. 20(8):1231–1239.

Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. 2010. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 20(1):110–121.

Pollard KS, Salama SR, King B, et al. (13 co-authors). 2006. Forces shaping the fastest evolving regions in the human genome. *PLoS Genet*. 2(10):e168.

Pollard KS, Salama SR, Lambert N, et al. (16 co-authors). 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443(7108):167–172.

Prabhakar S, Noonan JP, Paabo S, Rubin EM. 2006. Accelerated evolution of conserved noncoding sequences in humans. *Science* 314(5800):786.

Prabhakar S, Visel A, Akiyama JA, et al. (13 co-authors). 2008. Human-specific gain of function in a developmental enhancer. *Science* 321(5894):1346–1350.

Prabhakar S, Visel A, Akiyama JA, et al. (13 co-authors). 2009. Response to Comment on "Human-Specific Gain of Function in a Developmental Enhancer." *Science* 323(5915):714.

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. *Philos Trans R Soc Lond B Biol Sci*. 365(1552):2571–2580.

Romiguier J, Ranwez V, Douzery EJP, Galtier N. 2010. Contrasting GC-content dynamics across 33 mammalian genomes: relationship with life-history traits and chromosome sizes. *Genome Res*. 20(8):1001–1009.

Takano-Shimizu T. 1999. Local recombination and mutation effects on molecular evolution in Drosophila. *Genetics* 153(3):1285–1296.