J Mol Evol (2006) 62:168–175
DOI: 10.1007/s00239-005-0207-z

JOURNAL OF MOLECULAR EVOLUTION

A Novel Method Distinguishes Between Mutation Rates and Fixation Biases in Patterns of Single-Nucleotide Substitution

Mikhail Lipatov,[1] Peter F. Arndt,[2] Terence Hwa,[3] Dmitri A. Petrov[1]

[1] Department of Biological Sciences, Stanford University, 371 Serra Mall, Stanford, CA 94305, USA
[2] Max Planck Institute for Molecular Genetics, Ihnestr. 73, 14195, Berlin, Germany
[3] Department of Physics and Center for Theoretical Biological Physics, University of California, San Diego, 9500 Gilman Drive, La Jolla, CA 93092-0374, USA

Received: 6 July 2004 / Accepted: 20 June 2005 [*Reviewing Editor:* Dr. Nicolas Galtier]

Abstract. Analysis of the genome-wide patterns of single-nucleotide substitution reveals that the human GC content structure is out of equilibrium. The substitutions are decreasing the overall GC content (GC), at the same time making its range narrower. Investigation of single-nucleotide polymorphisms (SNPs) revealed that presently the decrease in GC content is due to a uniform mutational preference for A:T pairs, while its projected range is due to a variability in the fixation preference for G:C pairs. However, it is important to determine whether lessons learned about evolutionary processes operating at the present time (that is reflected in the SNP data) can be extended back into the evolutionary past. We describe here a new approach to this problem that utilizes the juxtaposition of forward and reverse substitution rates to determine the relative importance of variability in mutation rates and fixation probabilities in shaping long-term substitutional patterns. We use this approach to demonstrate that the forces shaping GC content structure over the recent past (since the appearance of the SNPs) extend all the way back to the mammalian radiation ~90 million years ago. In addition, we find a small but significant effect that has not been detected in the SNP data—relatively high rates of C:G → A:T germline mutation in low-GC regions of the genome.

Key words: Mutation — Selection — Biased gene conversion — Genome evolution — Single-nucleotide substitutions — GC content — Transitions — Transversions

Introduction

The human genome constitutes a sequence of about 3 billion nucleotide pairs. Each pair can be of only two kinds—A:T and C:G. Accordingly, the most directly measurable feature of this sequence is the local density of C:G nucleotide pairs, i.e., genomic GC content (GC). This density ranges from as low as 35% to as high as 60% among 20-kb genomic sequence fragments, forming a significantly wider distribution than would be expected from a random assignment of A:T and C:G pairs to the sequence (Bernardi 1993; Lander et al. 2001). GC content is highly correlated with other important genomic features (Galtier et al. 2001). In particular, regions with high exon density (e.g., 4% above the genomic average of 1.8%) tend to have unusually high GC content (20% above the genomic average of 38%).

Regions with high density of G:C pairs probably appeared in the amniotic lineage at least as far back as 350 million years ago (MYA) by way of single-nucleotide substitutions that favored G:C pairs in those regions much more strongly than in others (Arndt et al. 2003; Galtier et al. 2001). At a later point in the evolutionary past, however, the patterns of substitutions in these genomic regions stopped being as strongly

Correspondence to: Mikhail Lipatov; email: lipatov@stanford.edu biased in favor of G:C pairs. As a result, their GC content, by then unusually high (40% to 60%), started deteriorating. A recent study based on substitutions in human transposable elements suggests that this deterioration process started ~90 MYA, close to the time of mammalian radiation (Arndt et al. 2003). Its rate is very low, as its progress is limited by how often single-nucleotide substitutions occur in the genome—about 1 per site per billion years (Graur and Li 2000). If the current substitution patterns remain constant, the genome will not reach its equilibrium GC content until approximately 400 million years into the future (Arndt et al. 2003). Should this happen, the GC content will vary much less than it does now—it will be restricted to a range between 35% and 40%.

We hypothesize that the switch in the distribution of single-nucleotide substitution rates 90 MYA signifies a significant, long-term impact on the evolution of the mammalian genome—an impact most readily observed as the lowering and homogenization of genomic GC content. This raises the question of whether the switch happened as a result of a change in selective pressures, a change in the neutral intracellular processes that influence the germline mutation rates, or a change in neutral processes that affect fixation probabilities at nucleotide sites such as biased gene conversion (BGC).

As a first step, we wish to know which of these forces is currently responsible for the trend toward GC content homogenization. To this end, several studies analyzed nearly 6000 single-nucleotide polymorphism sites (SNPs) (Lercher and Hurst 2002; Lercher et al. 2002; Webster and Smith 2004; Webster et al. 2003). These studies revealed that human SNPs are subject to a relatively weak fixation bias in favor of G:C pairs, a bias whose strength is higher in areas of higher GC content. In quantitative terms, the generalized fixation bias coefficient S (same as $N_e s$ in our treatment below) rises from about 0.1 up to about 0.3 as GC content increases from 30% up to 55%.

If G:C alleles are never disfavored during the segregation of G:C/A:T polymorphisms, why is the GC content dropping across the genome? According to the above studies of human SNPs, this is due to the fact that all over the sequence, mutations from G:C to A:T pairs happen two times more often than mutations in the opposite direction. This mutational bias is stronger than the opposing fixation bias, leading to the overall decline in genomic GC content.

The above estimates of fixation-related and mutational biases are based on a relatively limited amount of data—polymorphisms that appeared in the human population recently enough to be segregating at the present time. Here we check whether the same combination of fixation and mutation biases is likely to be responsible for the past 90 million years of genomic evolution.

We argue that one way to obtain information about the forces that acted on ancient, no longer segregating polymorphisms is to compare rates of forward (X → Y) and reverse (Y → X) substitutions across the genome. Imagine that there is a cross-genome variation in the fixation bias operating on X/Y polymorphisms. Because both X → Y and Y → X go through the same X/Y polymorphic stage, such variation in fixation biases should affect X → Y and Y → X rates, in opposite ways. In contrast, X → Y and Y → X mutation rates can vary independently across the genome, and thus affect the two substitution rates in an uncorrelated fashion.

We examine the rates of GC-altering substitution over the past 90 million years, and their correlations with genomic GC content (Arndt et al. 2005). We then show that forward (GC-enriching) and reverse (AT-enriching) rates vary together almost exactly the way they should under the influence of a relatively weak ($S \sim 0.4$), GC-favoring fixation bias, whose strength is much like that of the bias that operates on the human SNPs. Specifically, as the GC-enriching substitution rates increase with increasing GC, the AT-enriching rates decrease precisely by the amount we would expect. Thus, the fixation bias variation inferred from the nearly 6000 present-time human SNPs is sufficient to explain the concerted, cross-GC content variation of the forward and reverse substitution rates that resulted from a much larger (by a factor of about 100) number of SNPs which used to segregate in the human lineage.

Weak fixation bias, however, cannot explain one portion of the substitution data—a sharp, peculiar doubling in the rate of C:G → A:T transversions in the regions where GC content is below 40%. This sharp increase is highly correlated with a similar increase in the rate of C:G → G:C transversions in these genomic regions. We argue that the most parsimonious explanation for this phenomenon is the presence of significantly elevated C:G → A:T and C:G → G:C germline mutation rates in regions of low GC content. Note that C:G → A:T transversions happen an order of magnitude less often than GC-altering transitions and that C:G → G:C substitutions do not alter sequence GC content at all. Thus, the mutational heterogeneity we detect has only a minor effect on the evolution of global GC content. Nevertheless, this heterogeneity is significant on the scale of the entire genome and provides an interesting and important insight into its molecular biology and evolution.

Methods

Part 1 of this section describes the heart of our method—that is, the derivation of a fixation bias hypothesized to exclusively account for combined patterns of forward and reverse rates of substitution.

Part 2 explores deviations from this hypothesis, where at least part of the substitutional variation is caused by mutation rates. Part 3 demonstrates how, under any of the above scenarios, we can calculate the ratio of forward to reverse mutation rates given the coefficients of fixation bias. Finally, part 4 describes how we applied our method to the variation of substitution rates across genomic GC content.

1. Calculating the Strength (N_es, Same as S) of a Fixation Bias Responsible for the Differences in Forward and Reverse Substitution Rates Between Two Genomic Regions, Assuming That There Are No Differences in Mutation Rates Between The Two Regions

Statistic F is the ratio of a "forward" substitution rate ($U_{X \to Y}$) in one genomic region ("region 2") to that in another genomic region ("region 1"). Statistic R is the same ratio for "reverse" substitutions ($U_{Y \to X}$):

$$F = \frac{U_{X\to Y}(region\ 2)}{U_{X\to Y}(region\ 1)} \quad \text{and} \quad R = \frac{U_{Y\to X}(region\ 2)}{U_{Y\to X}(region\ 1)}$$

Let us define N_e as the effective number of individuals in a diploid population and s as the fixation bias advantage of an individual with a YY genotype over an individual with an XX genotype. The fixation bias advantage of an XX individual over a YY individual, then, is equal to $-s$, provided that $|s| \ll 1$. We assume codominance, so that the XY heterozygotes have an advantage $s/2$ over the XX homozygotes. Using a well-known dependence of fixation probability on N_es (Gillespie 1998, Eq. 3.23) and a number of simplifying assumptions, we can show that

$$U_{X\to Y} = \mu_{X\to Y} \times \frac{2N_es}{1 - e^{-2N_es}} = \mu_{X\to Y} \times \frac{2S}{1 - e^{-2S}} \tag{1}$$

where $S \equiv N_es$ is a coefficient that reflects the fixation advantage of Y alleles over X alleles in the population, and $\mu_{X \to Y}$ is the rate of mutation from X to Y per nucleotide site. While S, N_e, and s are assumed to be uniform within a given genomic region, between different regions (of which there are only two in this treatment) they can differ freely. We can thus write down Eq. (1) separately for each of the genomic regions we focus on and then form the ratio F:

$$\begin{aligned} F = \frac{U_{X\to Y}(region\ 2)}{U_{X\to Y}(region\ 1)} &= \frac{\mu_{X\to Y}(region\ 2)}{\mu_{X\to Y}(region\ 1)} \\ &\times \frac{S(region\ 2)}{1 - e^{-2S(region\ 2)}} \times \frac{1 - e^{-2S(region\ 1)}}{S(region\ 1)} \end{aligned} \tag{2}$$

where the substitution rate, the rate of mutation, and the fixation advantage of Y alleles over X alleles are all variable between region 1 and region 2.

Now we turn our attention to R—the statistic that summarizes the difference in Y → X substitution rates between region 1 and region 2 (whereas F summarizes the difference in X → Y substitution rates). If the fixation advantage of Y over X is N_es, then the fixation advantage of X over Y is $-N_es$. Using the notation and model assumptions that went into Eq. (1), we obtain an expression for the reverse substitution rate:

$$U_{Y\to X} = \mu_{Y\to X} \times \frac{-2N_es}{1 - e^{2N_es}} = \mu_{Y\to X} \times \frac{-2S}{1 - e^{2S}} \tag{3}$$

Now we can use Eq. (3) to write down $1/R$ much like we used Eq. (1) to write down F:

$$\begin{aligned} 1/R = \frac{U_{Y\to X}(region\ 1)}{U_{Y\to X}(region\ 2)} &= \frac{\mu_{Y\to X}(region\ 1)}{\mu_{Y\to X}(region\ 2)} \\ &\times \frac{S(region\ 1)}{1 - e^{2S(region\ 1)}} \times \frac{1 - e^{2S(region\ 2)}}{S(region\ 2)} \end{aligned} \tag{4}$$

Note that F (Eq. [2]) and R (Eq. [4]) are linked by the values of S in the two genomic regions. However, each also involves a mutation rate ratio ($\mu_{X \to Y}$[region 2]/$\mu_{X \to Y}$[region 1] and $\mu_{Y \to X}$[region 2]/$\mu_{Y \to X}$[region 1], respectively). We assume that there is no a priori reason that the first mutation rate ratio should be related to the second. Therefore, if we allow either of the mutation rates to vary arbitrarily, we cannot infer the values of S from F and R. On the other hand, we can infer S under the constraint of mutation rate uniformity, or the null hypothesis

$$\begin{aligned} H_0 : \mu_{Y\to X}(\text{region } 1) &= \mu_{X\to Y}(\text{region } 2) \text{ and} \\ \mu_{X\to Y}(\text{region } 1) &= \mu_{X\to Y}(\text{region } 2). \end{aligned}$$

H_0 requires that both the forward and the reverse mutation rates are the same in the two regions. Note that it says nothing about the relationship between these rates. We return to that relationship in parts 2 and 3 of the Methods section.

Under H_0, the mutation rate ratios are both equal to 1, and Eqs. (2) and (4) reduce to

$$F = \frac{S_2}{1 - e^{-2S_2}} \times \frac{1 - e^{-2S_1}}{S_1} \tag{5}$$

and

$$1/R = \frac{S_1}{1 - e^{2S_1}} \times \frac{1 - e^{2S_2}}{S_2} \tag{6}$$

where $S_1 \equiv S$(region 1) and $S_2 \equiv S$(region 2). At this point we would like to solve (5) and (6) as a system of equations to obtain S_1 and S_2 in terms of F and R.

First, note that (5) and (6) do not have any solutions when F and $1/R$ are on the opposite sides of unity—i.e., when ($F > 1$ and $1/R < 1$) or ($F < 1$ and $1/R > 1$). In this case, the forward and reverse evolutionary rates change in the same direction, an observation that cannot be explained by a change in the fixation bias alone. Furthermore, we only need to develop a method for finding solutions where both F and $1/R$ are greater than unity, since there is a simple one-to-one correspondence between such solutions and the solutions to the cases where both F and $1/R$ are less than unity. Specifically, this correspondence maps $F \to 1/R$, $R \to 1/F$, $S_1 \to -S_2$, and $S_2 \to -S_1$. What about the boundary cases? No solutions exist either when $F = 1$ and $1/R \neq 1$ or when $F \neq 1$ and $1/R = 1$. In these cases, a change in the forward evolutionary rate is not mirrored by a change in the reverse rate, an observation that, once again, cannot be explained by fixation bias alone. When $F = 1$ and $1/R = 1$, the only solution is $S_1 = S_2 = 0$, corresponding to zero fixation bias in both of the genomic regions.

Thus, we only need to find solutions for $F > 1$ and $1/R > 1$. Note the usefulness of the ratio F/R. After some algebraic manipulation, this ratio simplifies to

$$\frac{F}{R} = e^{2(S_2 - S_1)} \tag{7}$$

Taking the natural logarithm of both sides yields the difference in fixation biases between the genomic regions:

$$\Delta S = S_2 - S_1 = \frac{1}{2} ln\left(\frac{F}{R}\right) \tag{8}$$

Using this equation, we can easily find S_2 if given S_1. S_1, in turn, can be found from F and R using a simple bisection algorithm, provided that we are careful with the two point discontinuities that correspond to either S_1 or S_2 equaling zero. Below, we describe the bisection algorithm we used for this calculation.

Consider the function

$$f(x) = F - \frac{x + \Delta S}{x} \times \frac{1 - e^{-2x}}{1 - \frac{R}{F}e^{-2x}} \tag{9}$$

where F, R, and ΔS are defined above. $f(x)$ has singularities at $x = 0$ and $x = -\Delta S$. Away from these two points, the function is continuous, monotonically increasing, and satisfies $f(x) < 0$ for $x \to -\infty$ and $f(x) > 0$ for $x \to +\infty$. Furthermore, if $f(0) \neq 0$ and

$f(-\Delta S) \neq 0$, this function has exactly one zero, corresponding to the first selection coefficient. In other words, as long as $f(0) \neq 0$ and $f(-\Delta S) \neq 0, f(x) = 0$ if and only if $x = S_1$. We thus apply the standard bisection method to this function in order to find S_1 (Press 1992).

The bisection method needs to be modified slightly in view of the two singularities. If $f(0) = 0$ or $f(-\Delta S) = 0$, then the zero of the function coincides with the corresponding singularity. However, right next to it, the function is still well-behaved. A practical solution to the problem is to modify the bisection method by making sure $f(x)$ avoids both singularities, and always stays a small step away from each one. That way, if $f(0) = 0$ (i.e., $S_1 = 0$) or $f(-\Delta S) = 0$ (i.e., $S_2 = 0$), the method will converge on the singularity that coincides with the root of the function, and if neither S_2 nor S_1 is zero, the algorithm will simply converge on the root.

This algorithm is implemented using Perl, and can be accessed to solve the system of Eqs. (5) and (6) to find S_1 and S_2 for any given F and R at http://cgi.stanford.edu/~lipatov/forward-reverse/forward-reverse-test.txt.

In conclusion of this section, we briefly discuss the assumptions used in the derivation of Eq. (1). The same assumptions were taken to be true by Sawyer et al., who develop a method for the estimation of selection coefficients from the distribution of polymorphism frequency (Sawyer and Hartl 1992). Their method is subsequently used by several studies to estimate the strength of fixation biases operating on human SNPs (Lercher and Hurst 2002; Webster and Smith 2004). Thus, estimates of N_es presented in these studies are comparable to ours.

2. Calculating the Minimal Contribution of Mutational Variation (Differences in μ_f and μ_r, Same as $\mu_{X \to Y}$ and $\mu_{Y \to X}$) to Substitution Rate Variation (Differences in U_f and U_r, Same as $U_{X \to Y}$ and $U_{Y \to X}$) Under Constraints on the Strength of Fixation Bias (S, Same as N_es)

Consider a case where the magnitudes of S_1 and S_2 in the previous section are higher than some maximum value S_{max}, set by additional evidence. When this is the case, we must conclude that H_0 is not true and that at least some of the difference in evolutionary rates must be due to a difference in the rates of mutation, in addition to a difference in fixation bias. As we lower the threshold of maximum possible fixation bias, and S_{max} approaches zero, we get closer to a situation where all the evolutionary rate variation across genomic regions must be due to a variation in the rates of mutation. Below, we quantify this intuitive observation.

First, we manipulate Eqs. (2) and (4) to express mutation rate ratios in terms of other parameters:

$$\frac{\mu_{f2}}{\mu_{f1}} = \frac{1}{F} \times \frac{S_2}{1 - e^{-2S_2}} \times \frac{1 - e^{-2S_1}}{S_1} \tag{10}$$

$$\frac{\mu_{r1}}{\mu_{r2}} = R \times \frac{S_1}{1 - e^{2S_1}} \times \frac{1 - e^{2S_2}}{S_2} \tag{11}$$

where $S_1 \equiv S(\text{region } 1)$ and $S_2 \equiv S(\text{region } 2)$, $\mu_{f1} \equiv \mu_{X \to Y}(\text{region } 1)$, $\mu_{f2} \equiv \mu_{X \to Y}(\text{region } 2)$, $\mu_{r1} \equiv \mu_{Y \to X}(\text{region } 1)$, $\mu_{r2} \equiv \mu_{Y \to X}(\text{region } 2)$.

Given some values of F and R, we would like to find mutation rate ratios (Eqs. [10] and [11]) that are as close to unity as possible, under the assumption that the magnitudes of both fixation biases are no larger than S_{max}. In other words, if we define a function that normalizes fractions to be greater than 1,

$$normRatio(x) = \max\left\{x, \frac{1}{x}\right\} \tag{12}$$

we want to compute the following:

$$M(S_{\max}) = \min_{|S_1|<S_{\max}, |S_2|<S_{\max}} \left\{ \max \left\{ \begin{array}{l} normRatio\left(\frac{1}{F} \times \frac{S_2}{1-e^{-2S_2}} \times \frac{1-e^{-2S_1}}{S_1}\right), \\ normRatio\left(R \times \frac{S_1}{1-e^{2S_1}} \times \frac{1-e^{2S_2}}{S_2}\right) \end{array} \right\} \right\} \tag{13}$$

That is, the minimal amount by which at least one of the mutation rate ratios (as expressed in Eqs. [10] and [11]) must deviate from unity if selection strength cannot be higher than S_{max}. M(S_{max}) has the following analytical solution for $F < 1/R$:

$$M(S_{\max}) = \begin{cases} \frac{1}{\mathrm{R}e^{2S_{\max}}} & \text{if } \mathrm{S}_{\max} < \frac{1}{4}\ln\left(\frac{F}{R}\right) \\ \left[F \times \left(1 - \frac{1}{2S_{\max}} \ln\left[\frac{F}{R}\right]\right) \times \frac{1-e^{-2S_{\max}}}{1-\frac{F}{R}e^{-2S_{\max}}}\right]^{-1} & \textit{otherwise} \end{cases} \tag{14}$$

However, a simple ratio of mutation rates does not, by itself, tell us anything about the extent to which mutation rate differences explain the differences in evolutionary rates. We might wish to compare the influence of mutational rate variation to that of a variation in fixation bias. In order to do this, we define the following function:

$$Mu(U_1/U_2, \mu_1/\mu_2) = \frac{normRatio\left(\frac{\mu_1}{\mu_2}\right) - 1}{normRatio\left(\frac{\mu_1}{\mu_2}\right) + normRatio\left(\frac{U_1\mu_2}{U_2\mu_1}\right) - 2} \tag{15}$$

Mu(U_1/U_2, μ_1/μ_2) measures the fraction of evolutionary rate nonuniformity caused by the nonuniformity in the rates of mutation. Here, normRatio(μ_1/μ_2) − 1 is the contribution of mutation rate differences to the deviation of U_1/U_2 from unity, and normRatio(U_1*μ_2/U_2*μ_1) − 1 is the contribution of fixation bias differences. Note that this function is equal to 0 for $\mu_1/\mu_2 = 1$ (0% of evolutionary rate differences explained by mutational variation) and equal to 1 for (U_1*μ_2)/(U_2*μ_1) = 1 (100% of evolutionary rate differences explained by mutational variation).

We now define a function similar to $M(S_{max})$:

$$N(S_{\max}) = \min_{|S_1|<S_{\max}, |S_2|<S_{\max}} \left\{ \max \left\{ \begin{array}{l} Mu\left(F, \frac{1}{F} \times \frac{S_2}{1-e^{-2S_2}} \times \frac{1-e^{-2S_1}}{S_1}\right), \\ Mu\left(\frac{1}{R}, R \times \frac{S_1}{1-e^{2S_1}} \times \frac{1-e^{2S_2}}{S_2}\right) \end{array} \right\} \right\} \tag{16}$$

$N(S_{max})$ tells us the minimum degree to which mutation rate variation must explain at least one of the cross-class discrepancies in evolutionary rates (either that in the forward or the reverse rate). We can use it much like $M(S_{max})$ to assess the differences in mutation rates implied by the empirical constraints on selection coefficients.

Although we have not been able to find an analytical solution for $N(S_{max})$, we can compute it using a bisection algorithm. The algorithm's implementation in Perl can be accessed at http://cgi.stanford.edu/~lipatov/forward-reverse/forward-reverse-test.txt.

3. Calculating the Ratio of Forward Mutation Rate to Reverse Mutation Rate (μ_f/μ_r, Same as $\mu_{X \to Y}/\mu_{Y \to X}$) Given the Ratio of Substitution Rates (U_f/U_r, Same as $U_{X \to Y}/U_{Y \to X}$), and the Strength of Fixation Bias (S, Same as N_es)

In parts 1 and 2 above, we were mainly concerned with the differences in forward and reverse substitution rates between genomic regions 1 and 2, as well as the underlying differences in fixation bias and rates of mutation. Here we return to the treatment of a single genomic region and show that, given the strength of a fixation bias operating in the region and the ratio of forward and reverse substitution rates, one can infer the corresponding ratio of forward and reverse mutation rates.

Table 1. Summary of the results for GC-altering transitions

Substitution type	Class 1 (35% GC) sequences	Class 2 (50% GC) sequences	Ratio of Class 2 to Class 1 (or Class 1 to Class 2)
A:T → G:C	2.89 ± 0.02 (U_{f1})	3.23 ± 0.02 (U_{f2})	**1.12** ± **0.01** ($U_{f2}/U_{f1} \equiv F_i$)
G:C → A:T	5.47 ± 0.02 (U_{r1})	4.68 ± 0.03 (U_{r2})	**1.17** ± **0.01** ($U_{r1}/U_{r2} \equiv 1/\mathrm{R}_i$)
Fixation bias coefficient			
S_i (min)	**0.24**	**0.37**	—
S_i (max)	**0.64**	**0.77**	—

Note. Columns 2 and 3 contain normalized A:T → G:C (U_f) and G:C → A:T (U_r) substitution rates for class 1 (35% GC) and class 2 (50% GC) sequences. These values were used to compute the associated statistics F_i and R_i (column 4). F_i and R_i, in turn, were used to compute the limits on S_i—strength of the transition-specific fixation bias that favors G:C nucleotide pairs over A:T pairs, given the assumption that both forward and reverse mutation rates are the same for class 1 and class 2 sequences (Bottom two rows). Details of the latter calculation are contained in part 1 under Methods section; those of the former, in part 4.

We divide Eq. (1) by Eq. (3), to obtain

$$\frac{U_{X\to Y}}{U_{Y\to X}} = -\frac{\mu_{X\to Y}}{\mu_{Y\to X}} \times \frac{1-e^{2S}}{1-e^{-2S}} \tag{17}$$

Upon rearranging (17) and generalizing the notation, we get

$$\frac{\mu_r}{\mu_f} = -\frac{U_r}{U_f} \times \frac{1-e^{2S}}{1-e^{-2S}} \tag{18}$$

where $U_f \equiv U_{\mathrm{X}\to\mathrm{Y}}$, $U_r \equiv U_{\mathrm{Y}\to\mathrm{X}}$, $\mu_f \equiv \mu_{\mathrm{X}\to\mathrm{Y}}$, $\mu_f \equiv \mu_{\mathrm{Y}\to\mathrm{X}}$, and S is the fixation bias favoring "forward" (X → Y) evolution.

Thus, given the rates of forward and reverse substitution and the strength of the underlying fixation bias, we can calculate the ratio of corresponding mutation rates. If S is positive and increasing, Eq. (18) quickly tends to

$$\frac{\mu_r}{\mu_f} = \frac{U_r}{U_f} \times e^{2S} \tag{19}$$

In other words, if the rates of forward and reverse substitution are held constant, the ratio of forward-to-reverse mutation rates is scaled exponentially with the magnitude of the fixation bias coefficient. If, in addition to the substitution rates, we have information about the possible range of μ_f/μ_r, Eq. (18) allows us to place constraints on the possible values of S.

4. Estimates of F_v, $1/R_v$, F_i, and $1/R_i$

Parts 1 and 2 have dealt with a general analysis of F and R. F is a statistic that measures how the rate of "forward" (X → Y) substitution differs between two distinct genomic regions, while R does the same for the rate of "reverse" (Y → X) substitution. We now apply this general method to substitution rate variation across genomic GC content. To do so, we define the two distinct genomic regions as areas with 35% and 50% GC content, and forward and reverse substitutions as those that increase and those that decrease GC content of the sequence. There are two types of GC-altering substitutions: transitions (C:G ⇔ T:A) and transversions (C:G⇔A:T). Accordingly, we analyze the two types of substitutions separately and define two separate sets of statistics:

$$F_i = \frac{U_{T:A\to C:G}(GC=50\%)}{U_{T:A\to C:G}(GC=35\%)}, \quad 1/R_i = \frac{U_{C:G\to T:A}(GC=35\%)}{U_{C:G\to T:A}(GC=50\%)}$$

and

$$F_v = \frac{U_{A:T\to C:G}(GC=50\%)}{U_{A:T\to C:G}(GC=35\%)}, \quad 1/R_v = \frac{U_{C:G\to A:T}(GC=35\%)}{U_{C:G\to A:T}(GC=50\%)}$$

Our estimates of the four substitution rates at each of the two GC content values come from the work of Arndt and colleagues (2005), who plot each of these substitution rates versus genomic GC content and fit the dependence to polynomial regression curves (Arndt et al. 2005, Fig. 5).

We checked that the residual distributions of the C:G → T:A, T:A → C:G, C:G → A:T, and A:T → C:G polynomial regression curves are approximately normal at 35% and 50% GC content. We then generated 5000 values for each substitution rate at both GC content values (a total of $5000 \times 4 \times 2 = 40{,}000$ numbers). Each 5000-value set was drawn from a normal distribution with a mean equal to the value of the corresponding regression curve at the corresponding GC content value and a variance determined from the residuals of that regression at that GC.

We paired up the values from the 35% and 50% number sets for each substitution type, and divided one of the values by the other. This yielded four 5000-value sets, corresponding to each of the above statistics. The quantiles of each of these sets were, in turn, extremely close to those of normal distributions. Consequently, the assumption of normality is inherent in our estimates of 95% confidence intervals on the four statistics (rightmost column in Tables 1 and 2).

We then applied part 1 of Methods to the analysis of F_v, R_v, F_i, and R_i, and found that S_v—the transversion-specific fixation bias coefficient that could explain substitution rate variation (bottom rows in Table 2: transversion-specific fixation biases)—is an order of magnitude stronger than the corresponding transition-specific coefficient $\mathrm{S_i}$ (bottom rows in Table 1: transition-specific fixation biases). By means of part 3 above, we used these fixation bias coefficients to derive the transition and transversion-specific ratios of forward-to-reverse mutation rates (Table 3).

Results and Discussion

Fixation Bias Variation Due to Selection or Biased Gene Conversion Could Cause Most of the C:G → A:T, C:G → T:A, A:T → C:G, and T:A → C:G Substitution Rate Dependencies on GC Content

Arndt and colleagues (2005) estimated GC content (GC) and rates of single-nucleotide substitution for every 1-Mbp region of the human genome. They found that the AT-enriching substitution rates (C:G → A:T and C:G → T:A) decrease, while the GC-enriching substitution rates (A:T → C:G and T:A → C:G) increase with increasing GC. Such combined behavior is expected in the case of a

Table 2. Summary of the results for GC-altering transversions

Substitution type	Class 1 (35% GC) sequences	Class 2 (50% GC) sequences	Ratio of Class 2 to Class 1 (or Class 1 to Class 2)
A:T → C:G	0.79 ± 0.02 (U_{f1})	0.86 ± 0.01 (U_{f2})	**1.09 ± 0.03** ($U_{f2}/U_{f1} \equiv F_v$)
C:G → A:T	1.70 ± 0.01 (U_{r1})	0.98 ± 0.01 (U_{r2})	**1.74 ± 0.02** ($U_{r1}/U_{r2} \equiv 1/\mathrm{R_v}$)
Fixation bias coefficient			
S_v (min)	**2.70**	**3.02**	—
S_v (max)	**4.64**	**4.96**	—

Note. Here U_f represents the rate of A:T → C:G substitutions, and U_r that of C:G → A:T substitutions. Also, F_v, R_v, and S_v are the transversion-specific versions of F_i, R_i, and S_i from Table 1.

Table 3. Mutation rate ratios implied by the substitution rates and fixation bias coefficients in Tables 1 and 2

Substitution type	U_{f1}	U_{r1}	S(min)	S(max)	μ_r/μ_f (min)	μ_r/μ_f (max)
A:T⇔G:C	2.89 ± 0.02	5.47 ± 0.02	0.24	0.64	**3.03**	**6.87**
A:T⇔C:G	0.79 ± 0.02	1.70 ± 0.01	2.70	4.64	**469**	**23,900**

Note. Columns 2 and 3 contain substitution rates; columns 4 and 5, limits on fixation bias coefficients for Class 1 (35% GC content) sequences. These fixation bias coefficients were computed by considering both Class 1 and Class 2 (50% GC content) substitution rates, under the hypothesis that mutation rates do not vary across GC content. Using the information in the columns 2–5, limits on the forward/reverse mutation rate ratios were computed, for both transitions (row 1 in the last two columns; see also Table 1) and transversions (row 2 in the last two columns; see also Table 2). In both cases, we computed the limits on mutation rate ratios by means of Eq. (18) (part 3 in Methods).

changing fixation bias—an increasing preference for the C:G allele during the segregation of C:G/A:T and C:G/T:A polymorphisms. Several forces that could be responsible for such a bias have been proposed, including natural selection (Bernardi 1993, 2000) and biased gene conversion (Galtier et al. 2001; Meunier and Duret 2004).

The presence of a fixation bias that distinguishes between C:G and A:T pairs is evidenced by the skewed frequency distributions of human single-nucleotide polymorphisms (SNPs) (Lercher and Hurst 2002; Lercher et al. 2002; Webster and Smith 2004; Webster et al. 2003). Specifically, the average frequency of AT → GC (i.e., A:T → G:C plus A:T → C:G) SNPs in the human genome is slightly higher than that of the GC → AT SNPs (20.6% versus 18.4%), which is consistent with a relatively weak ($\mathrm{N_e s}$ ~ 0.2) fixation bias in favor of C:G pairs (Webster and Smith 2004). Here we extend the scope of Webster and Smith's findings and check whether a fixation bias similar in magnitude to the one they detect can account for the opposing behaviors of GC-increasing and AT-increasing substitutions (Arndt et al. 2005).

As mentioned above, the fact that changing GC content drives the GC-increasing and AT-increasing substitution rates in opposite directions implies that fixation bias could be solely responsible for the trend. In part 1 under Methods, we show that, in general, forward (X → Y) and reverse (Y → X) substitution rate variation can yield the strength of the underlying fixation bias, given that X → Y and Y → X mutation rates stay constant. We apply this method to the patterns of substitution rates found by Arndt et al. (2005) and find that the transition rate variation (C:G → T:A and T:A → C:G) can be explained by a fixation bias of the same order of magnitude as that found by Webster and Smith (2004) (transition-specific $N_e s$ ~ 0.4; see Table 1).

A visual comparison of the variation in transition (C:G⇔T:A) rates with GC content with that in the transversion (C:G⇔A:T) rates suggests that the two behave similarly, with the exception of a sharp increase in the rates of C:G → A:T transversions at GC content below 40%. Quantitatively, when we apply part 1 under Methods to the transversions, the resulting estimates of the fixation bias coefficient are as low as the ones we find for transitions, provided that we exclude genomic areas with low GC (transversion-specific $N_e s$ ~ 0.4 for 40% < GC < 60%; results not shown).

In sum, the dependence of C:G → T:A and T:A → C:G transitions on GC content across its full range, and that of C:G → A:T and A:T → C:G transversions for GC above 40%, can both be explained by changes in a fixation bias with $N_e s$ ~ 0.4. This bias necessarily favors C:G, (as opposed to A:T) pairs, since GC content alters the AT-enriching substitution rates more dramatically than the GC-enriching rates. The magnitudes of this fixation bias, its direction, and even its cross-GC content increase all agree with the studies of human SNPs (Lercher and Hurst 2002; Webster and Smith 2004).

We can reflect on the nature of the putative fixation bias by comparing its effect on GC-enriching transitions with that on GC-enriching transversions, since neither substitution type exhibits the anomalous behavior we detected for AT-enriching transversions (see above). According to the observed values of F_i (Table 1) and F_v (Table 2), although the fixation bias is somewhat more pronounced among the transitions than it is among the transversions, the difference is not significant, as the 95% confidence intervals for the two statistics overlap substantially. Consequently, our results do not provide any evidence of a fixation bias that operates more strongly on GC-altering transitions than on the transversions. This observation is consistent with the hypothesis that the fixation bias is caused by selection for high GC content, since such a force is not expected to treat transitions and transversions differently.

Recently, however, gene conversion has been proposed as a likely candidate for the force behind the human SNP fixation bias. At first sight, our findings are not in favor of this hypothesis, since biased gene conversion has no a priori reason to affect transversions and transitions in the same way. However, recent experimental studies show that, at least in one biological instance, the frequency of transversion-type single nucleotide gene conversion is not significantly different from that of transition-type conversion (Birdsell 2002), suggesting that biased gene conversion may also explain our results.

Sharply Elevated C:G → A:T and C:G → G:C Mutation Rates in Areas of Low GC Content

If we include genomic areas with low GC content in our analysis of transversions, the variation in their rates could still be explained in terms of a changing fixation bias. However, the strength of such a bias would have to be an order of magnitude larger (transversion-specific $N_e s \sim 4.0$ for $30\% < \mathrm{GC} < 60\%$; see Table 2). Such a large fixation bias favoring C:G pairs must be counterbalanced by a nearly 500-fold transversion-specific mutational bias in favor of A:T pairs (see Table 3 and part 3 of Methods). The magntidues of the transversion-specific fixation-related and mutational biases are both highly unlikely to be this strong.

We conclude that the fixation bias currently acting on human SNPs could not have caused the observed variation of transversion rates across the entire range of genomic GC content. At least part of this variation must be due to differences in the rates of mutation. Such differences are significant but can stay confined to just one of the two mutation rates. Specifically, if $N_e s$ is constrained to be below 0.4 (see, for instance, our estimates of transition-specific $N_e s$ in Table 1), C:G → A:T mutation rate has to be at least 1.2-fold higher at low GC content, compared to that at high GC content [i.e., $M(S_{max}) = 1.2$ when $S_{max} = 0.4$: see part 2 of Methods]. This mutational effect accounts for at least 36% of the cross-GC content differences in the corresponding rate of substitutions [i.e., $N(S_{max}) = 0.36$ when $S_{max} = 0.4$].

We wish to note a formal possibility that the mutations which cause a sharp rise in the rate of C:G → A:T substitution at low GC content are strongly adaptive, and thus unlikely to be detected in the polymorphism datasets. Nevertheless, we consider this sort of scenario to be highly improbable, especially since the effect is confined to just one of the four GC-altering substitution types.

There are other possible reasons why this effect was not detected among the SNPs. One is that the heterogeneity in mutation rates was present for much of the past 90 million years but no longer persists in the genome. Another possibility is that the mutational patterns differ between the ~50% of the genome that consists of transposable elements (Lander et al. 2001) and the remaining nonfunctional sequence. The final, and we believe the most likely, possibility is that the detected effect is confined to a single, infrequent substitution type (C:G → A:T). Such events are very rarely observed in the SNP data, where we may not yet have enough statistical power to detect this mutational heterogeneity. In contrast, the divergence data of Anrdt et al. (2005) are based on many more (×100) events and thus have the power to detect even very subtle effects.

The sharp increase in C:G → A:T substitution rates at low (<40%) GC content is similar to an analogous effect among the C:G → G:C rates (Arndt et al. 2005, Fig. 5). In fact, these two substitution rates correlate with each other better than any other two rates (Arndt et al. 2005, Table 1). This correlation is the only one that remains significant after the effects of GC content on all the rates of substitution are removed statistically (Arndt et al. 2005, Table 3). Consider that neither the C:G → G:C nor the C:G → A:T variation with GC content is likely to be caused by fixation biases alone: the first—because there is no known genome-wide mechanism that distinguishes between a C:G and a G:C allele, and the second—due to our findings based on the analysis of forward/reverse substitutions. Consequently, the evidence presented in this article in combination with the results of previous studies suggests that a sizable portion of the genome (at least 15% of the sequence) is simultaneously subject to elevated C:G → G:C and C:G → A:T mutation rates.

Conclusion

Our analysis of the dependencies of single-nucleotide substitution rates on human GC content reveals significantly elevated C:G → A:T germline mutation rates in areas of low (< 40%) GC content. This difference is no less than 1.2-fold, contributing at least 36% to a sharp increase in the rate of C:G → A:T substitution at low GC.

A similar, sharp increase at low GC content is observed in the rate of C:G → G:C substitutions. C:G → A:T and C:G → G:C rates are strongly correlated, and neither is likely to vary due to fixation biases alone. We thus postulate that a common mechanism causes a significant increase in C:G → A:T and C:G → G:C mutation rates in GC-poor regions of the human genome.

Once we control for the unusually high C:G → A:T mutation rates at low GC content, GC-enriching and AT-enriching substitutions change across GC content precisely the way we would expect from the action of the fixation bias detected among human SNPs. In particular, the remaining correlations of all single-nucleotide substitution rates with GC content bear the very specific footprint of a weak ($S \sim 0.4$) fixation bias in favor of G:C pairs—a bias whose strength goes up with increasing GC.

The method we use to detect mutational heterogeneity across genomic GC content is both novel and generally applicable to the analysis of biological evolution. It is based on the observation that in a biallelic system, forward and reverse evolutionary rates are coupled by the action of fixation-related pressures on the same polymorphism. Comparison of the two evolutionary rates represents a rich source of information about these pressures. Our technique constitutes an example of how such information can be extracted to reach conclusions about the roles of past evolutionary forces in shaping the structure of genomic sequence.

Acknowledgments. We would like to thank Jerel Davis, Nadia Singh, Marc Feldman, Nicolas Galtier, and the two anonymous reviewers for comments that went a long way to improve the manuscript. M.L. is funded by a fellowship from the Achievement Rewards for College Scientists Foundation, through the Stanford Graduate Fellowship program. T.H. is grateful for financial support by the NSF through Grants 0211308, 0083704, 0216576, and 0225630. D.P. is grateful for the financial support by the Alfred P. Sloan Foundation, Hellman Faculty Award, Terman Award from the Stanford University, and NSF Grant DEB-0317171.

References

Arndt PF, Petrov DA, Hwa T (2003) Distinct changes of genomic biases in nucleotide substitution at the time of mammalian radiation. Mol Biol Evol 20:1887–1896

Arndt PF, Hwa T, Petrov DA (2005) Substantial regional variation in substitution rates in the human genome: importance of GC content, gene density, and telomere-specific effects. J Mol Evol 60:748–763

Bernardi G (1993) The vertebrate genome: isochores and evolution. Mol Biol Evol 10:186–204

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

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

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

Gillespie JH (1998) Population genetics: a concise guide. Johns Hopkins University Press, Baltimore, MD

Graur D, Li W-H (2000) Fundamentals of molecular evolution. Sinauer, Sunderland, MA

Lander ES, Linton LM, Birren B, et al. (2001) Initial sequencing and analysis of the human genome. Nature 409:860–921

Lercher MJ, Hurst LD (2002) Can mutation or fixation biases explain the allele frequency distribution of human single nucleotide polymorphisms (SNPs)? Gene 300:53–58

Lercher MJ, Smith NG, Eyre-Walker A, Hurst LD (2002) The evolution of isochores: evidence from SNP frequency distributions. Genetics 162:1805–1810

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

Press WH (1992) Numerical recipes in C: the art of scientific computing. Cambridge University Press, Cambridge, NY

Sawyer SA, Hartl DL (1992) Population genetics of polymorphism and divergence. Genetics 132:1161–1176

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

Webster MT, Smith NG, Ellegren H (2003) Compositional evolution of noncoding DNA in the human and chimpanzee genomes. Mol Biol Evol 20:278–286