Influence of Electron–Holes on DNA Sequence-Specific Mutation Rates

Martha Y. Suárez-Villagrán[1,2], Ricardo B.R. Azevedo[3], and John H. Miller Jr[1,2,*]

[1]Department of Physics, University of Houston, Houston

[2]Texas Center for Superconductivity, University of Houston, Houston

[3]Department of Biology and Biochemistry, University of Houston, Houston

*Corresponding author: E-mail: jhmiller@uh.edu.

Accepted: March 20, 2018

Abstract

Biases in mutation rate can influence molecular evolution, yielding rates of evolution that vary widely in different parts of the genome and even among neighboring nucleotides. Here, we explore one possible mechanism of influence on sequence-specific mutation rates, the electron–hole, which can localize and potentially trigger a replication mismatch. A hole is a mobile site of positive charge created during one-electron oxidation by, for example, radiation, contact with a mutagenic agent, or oxidative stress. Its quantum wavelike properties cause it to localize at various sites with probabilities that vary widely, by orders of magnitude, and depend strongly on the local sequence. We find significant correlations between hole probabilities and mutation rates within base triplets, observed in published mutation accumulation experiments on four species of bacteria. We have also computed hole probability spectra for hypervariable segment I of the human mtDNA control region, which contains several mutational hotspots, and for heptanucleotides in noncoding regions of the human genome, whose polymorphism levels have recently been reported. We observe significant correlations between hole probabilities, and context-specific mutation and substitution rates. The correlation with hole probability cannot be explained entirely by CpG methylation in the heptanucleotide data. Peaks in hole probability tend to coincide with mutational hotspots, even in mtDNA where CpG methylation is rare. Our results suggest that hole-enhanced mutational mechanisms, such as oxidation-stabilized tautomerization and base deamination, contribute to molecular evolution.

Key words: mutation rate bias, context-dependent mutation, electron–hole, mitochondrial DNA, hypervariable segment I, human genome, oxidation, tautomer, cancer

Introduction

Rates of evolution and levels of genetic variation vary widely across the genomes of most organisms. For example, certain sites in the control region of mitochondrial DNA (mtDNA) have diverged rapidly between human and chimpanzee, and are highly variable within the human population (Hasegawa et al. 1993; Tamura and Nei 1993; Wakeley 1993; Excoffier and Yang 1999; Meyer et al. 1999; Bandelt et al. 2006; Howell et al. 2007; Rosset et al. 2008; Soares et al. 2009). Many of these sites appear to be mutational "hotspots" (Stoneking 2000).

In mutation accumulation (MA) experiments, multiple replicate populations kept at low effective population size are allowed to accumulate all but the most deleterious mutations over multiple generations (Bateman 1959; Mukai 1964). Thus, MA experiments allow a relatively unfiltered look at the evolutionary consequences of mutation. Recent MA studies have confirmed that mutation rates vary considerably between sites in many organisms (Haag-Liautard et al. 2008; Lee et al. 2012; Zhu et al. 2014; Sung et al. 2015; Behringer and Hall 2016). The mechanisms of hypermutability, however, remain unclear (Galtier et al. 2009). Some variation in mutation rate appears to be sequence-specific (Haag-Liautard et al. 2008; Lee et al. 2012; Zhu et al. 2014; Sung et al. 2015; Behringer and Hall 2016), indicating that it may in part be driven by physicochemical mechanisms.

In this work we address the possibility that DNA's sequence-specific electronic properties (Xu et al. 2007; Shih et al. 2012)—specifically of electron–holes (Carrillo-Nuñez and Schulz 2008; Shih et al. 2008; Bacolla et al. 2013; Suárez Villagrán and Miller 2015)—differentially affect each site's mutability. A hole is a mobile site of positive charge, or

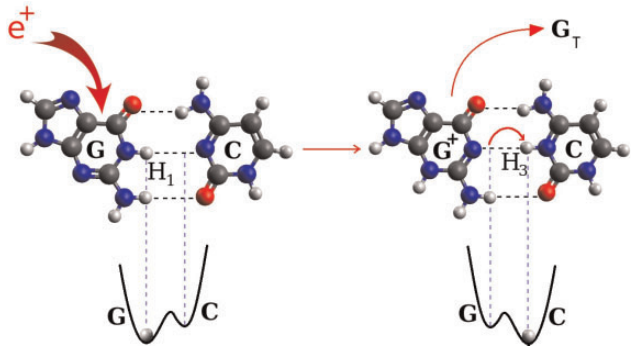

Fig. 1.—Tautomeric hydrogen bond shift in a G:C base pair, before (left) and after (right) G oxidation by a hole causes proton (H_1) migration toward the C (top right). Original (left) and altered (right) double-well potentials are shown at the bottom. Chemical structures were drawn using the Avogadro molecule editor and visualizer (Hanwell et al. 2012).

oxidized state (Cerón-Carrasco et al. 2010), left behind when an electron is removed, for example, by ionizing radiation or contact with an oxidizing agent. A hole exhibits quantum mechanical properties, and its wavefunction spreads out to enable long distance transport in an artificial DNA molecule with repeated base pairs (Meggers et al. 1998; Lewis et al. 2000; Giese et al. 2001; Endres et al. 2004; Giese 2004). The varying base ionization potentials in a natural sequence (Sugiyama and Saito 1996), however, cause the hole wavefunction to localize with higher probability in deeper potential wells, similar to electron localization in a disordered potential (Anderson 1958). Guanine has the lowest ionization potential (Sugiyama and Saito 1996), and thus the highest tendency to trap holes.

When acting on a specific base, a hole can enhance the probability of a base pair replication mismatch through a variety of possible mechanisms (Watson and Crick 1953; Löwdin 1963; Cerón-Carrasco et al. 2010; Bebenek et al. 2011). These include oxidation-induced stabilization of the normally rare, mismatch promoting tautomer (fig. 1) of a given base (Cerón-Carrasco et al. 2010), oxidative base deamination (Kreutzer and Essigmann 1998; Dizdaroglu 2015), formation of wobble pairs, and other mechanisms (see [Modrich 1987; Schroeder et al. 2018] for reviews). Behringer and Hall (2016) reported that fission yeast shows an elevated mutation rate of C:G base pairs. This rate is especially high when C:G is the middle base pair in CCG and TCG, or the respective reverse complementary triplets CGG and CGA. Yeast DNA is not believed to be subject to methylation. (By contrast, our analysis, below, of data by Aggarwala and Voight 2016 shows influence both of methylation of CpG pairs and of holes in human DNA.) Recent studies also suggest that localized holes correlate with mutations involved in cancer and other diseases (Bacolla et al. 2013), and with human variant frequency spikes in the mitochondrial gene *ND1* (Suárez Villagrán and Miller 2015). Thus, computational DNA hole spectroscopy (Suárez Villagrán and Miller 2015) shows promise in the prediction of intrinsic sequence- and site-dependent mutability.

Here, we test the extent to which the localization of electron–holes on DNA sequences explains variation in mutation rates in three systems. First, context-dependent mutation rates in four species of bacteria (Lee et al. 2012; Sung et al. 2012, 2015; Long et al. 2015). Second, site-specific mutation rates in the hypervariable segment I (HVS-I) of the human mtDNA control region (Stoneking 2000; Galtier et al. 2006; Howell et al. 2007; Rosset et al. 2008; van Oven and Kayser 2009). Finally, levels of polymorphism in the middle base within several thousand heptanucleotide permutations in the human genome (Aggarwala and Voight 2016). We conclude that hole probabilities explain part of the variation in mutation rate in all three systems, whereas methylation at CpG pairs plays an additional prominent role in the latter system.

Materials and Methods

Correlation between Holes, Guanine Oxidation, and Ionization Potential

Hole probabilities for randomized base–pair triplets were computed following the computational DNA hole spectroscopy method discussed below. Spearman correlation coefficients were then computed for the average hole probability for the middle base as compared with data reported by Saito et al. (1998), Margolin et al. (2006), including sequence-specific guanine oxidation reactivity and ionization potential.

Bacterial Mutation Rates

Mutation rates in each triplet in *Bacillus subtilis*, *Escherichia coli*, and *Mesoplasma florum* are taken from supplementary tables S5 and S6, Supplementary Material online of Sung et al. (2015). We averaged the mutation rates for triplets in the left and right replichores. The mutation rates in *Pseudomonas fluorescens* are taken from table 2 of Long et al. (2015).

HVS-I Mutation Rates

Mutation rates are taken from Supplemental Table 1, Supplementary Material online ("3,000 samples" column) of Rosset et al. (2008).

Computational DNA Hole Spectroscopy: Model Calculations

We compute hole spectra for specific sequences following (Suárez Villagrán and Miller 2015). DNA is modeled as a two-legged ladder using a tight-binding picture that includes matrix elements representing nearest-neighbor hopping of the hole along each chain ($t_{||}$: parallel hopping) and between chains ($t_{\perp}$: perpendicular hopping; Carrillo-Nuñez and Schulz 2008; Suárez Villagrán and Miller 2015). This tight-binding approach has the advantages that it is computationally tractable and allows us to handle a large number of base pairs

(Carrillo-Nuñez and Schulz 2008). The local hole energy, $\varepsilon_{\ell m}$, is obtained for each site m and chain ℓ using published ionization potentials for the four bases: $\varepsilon_A = 8.24\,\text{eV}$, $\varepsilon_T = 9.14\,\text{eV}$, $\varepsilon_G = 7.75\,\text{eV}$, and $\varepsilon_C = 8.87\,\text{eV}$ (Sugiyama and Saito 1996). This leads to the following tight-binding Hamiltonian (Carrillo-Nuñez and Schulz 2008; Suárez Villagrán and Miller 2015):

$$\widehat{H} = \sum_{m=1}^{N}\left[\sum_{\ell=1}^{2}\left\{\varepsilon_{\ell m}c_{\ell m}^{\dagger}c_{\ell m} + t_{\parallel}\left[c_{\ell,m+1}^{\dagger}c_{\ell,m} + c_{\ell,m-1}^{\dagger}c_{\ell,m}\right]\right\} + t_{\perp}\left\{c_{2,m}^{\dagger}c_{1,m} + c_{1,m}^{\dagger}c_{2,m}\right\}\right], \quad (1)$$

where $c_{\ell m}^{\dagger}$ represents a hole creation operator at site m on chain ℓ.

In matrix form, the Hamiltonian operator, equation (1), becomes a $2N \times 2N$ matrix where the local hole energies lie along the diagonal and the off-diagonal hopping terms lie along either side of the diagonal. We apply periodic boundary conditions by adding hopping matrix elements that couple the first and last sites of each chain. The $2N$ eigenenergies E_n and probability amplitudes $\Psi_n(m,\ell)$ versus site m and strand ℓ for each eigenstate Ψ_n are then computed by diagonalizing the Hamiltonian and normalizing the probability amplitudes within the DNA segment of interest. In this study we use values for $t_{\parallel}$, and $t_{\perp}$ of 1.0, and 0.5 eV, respectively (Carrillo-Nuñez and Schulz 2008; Suárez Villagrán and Miller 2015).

Complete hole probability spectra (relative values) for the two strands are then computed using a pseudothermal distribution of all the states by assuming a Boltzmann distribution:

$$P(m,\ell) = \sum_{i=0}^{N'-1} P_i(m,\ell)\exp\left[-\frac{(E_i - E_0)}{k_B T}\right], \quad (2)$$

where $P_i(m,\ell) = |\Psi_i(m,\ell)|^2$, $N' = 2N$ is the total number of energy eigenstates, E_0 is the lowest eigenenergy, k_B is Boltzmann's constant, T is an effective temperature. This may be higher than the actual temperature due to the non-equilibrium nature of hole creation and transport ($k_B T = 0.05\,\text{eV}$ in the model here). When using actual sequence data, we find that the lowest energy eigenstates are highly localized, each showing a single peak in probability at a given nucleotide site—these lowest energy eigenstates correspond to the largest hole peaks.

In the case of HVS-I and surrounding human mtDNA regions, we use the revised Cambridge reference sequence (rCRS; Andrews et al. 1999). We also compute hole spectra of other common haplotypes—K, J, T2, U5a1a, V, and I (Behar et al. 2007)—for comparison. Hole spectra are computed over the interval 15,989–16,569 and 1–600, which runs continuously along the circular mtDNA molecule. This segment encompasses the entire control region (D-loop) plus ~20 additional base pairs on each side to accommodate periodic boundary conditions.

Base–Pair Heptanucleotides in the Human Genome

We computed sequence dependent hole probabilities for 20,000 random sequences of length 2,100, for a total of 42 million base pairs. We then computed the average (from >360 spectra per 7-mer on average) middle base hole probability for each heptanucleotide permutation (out of 16,384) extracted from the randomized sequences. This yielded middle-base hole probabilities for both the reference and complementary 7-mer sequences. The data shown in supplementary table 7, Supplementary Material online of Aggarwala and Voight (2016) focus on the reference sequence only, yielding 8,192 permutations times three autosomal base substitution probabilities, for each of three populations, African, Asian, and European. In order to compare hole probabilities to net substitution probabilities, recently compiled into heptanucleotides (Aggarwala and Voight 2016) using data from the 1000 Genomes Project (The Genomes Project C 2012). In order to draw a comparison to hole probabilities, we first took, from supplementary table 7, Supplementary Material online of Aggarwala and Voight (2016), the average autosomal substitution probability, from the reference to alternate sequence, of the three populations and then the total of the three possible substituted bases (usually dominated by transitions). We compared both our reference and average (of reference and complementary) middle hole probabilities in order to compute the Spearman correlation for the 8,192 different 7-mer sequences.

Results

Hole Probabilities, Like Mutation Rates, Are both Nucleotide- and Context-specific

Despite the widespread assumption that intrinsic mutation rates are independent of local sequence, a growing body of evidence suggests otherwise. For example, recent studies found strong nucleotide-specific mutation rate biases in *B. subtilis* (Sung et al. 2015), *E. coli* (Lee et al. 2012), *M. florum* (Sung et al. 2012, 2015), and *P. fluorescens* (Long et al. 2015). In *M. florum*, for example, G nucleotides showed base-substitution mutation rates 17-fold higher than T nucleotides on average (Sung et al. 2015). The mutation rate biases were also strongly context-dependent. In *B. subtilis*, for example, the T in a GTG trinucleotide showed a base-substitution mutation rate 76-fold higher than that of the middle T in a TTT trinucleotide (Sung et al. 2015). Hole localization is one possible mechanism by which these nucleotide-specific and neighbor-dependent mutation rate biases might emerge. To test this idea, we computed average hole probabilities for the middle nucleotide of the 64 possible base triplets. We then

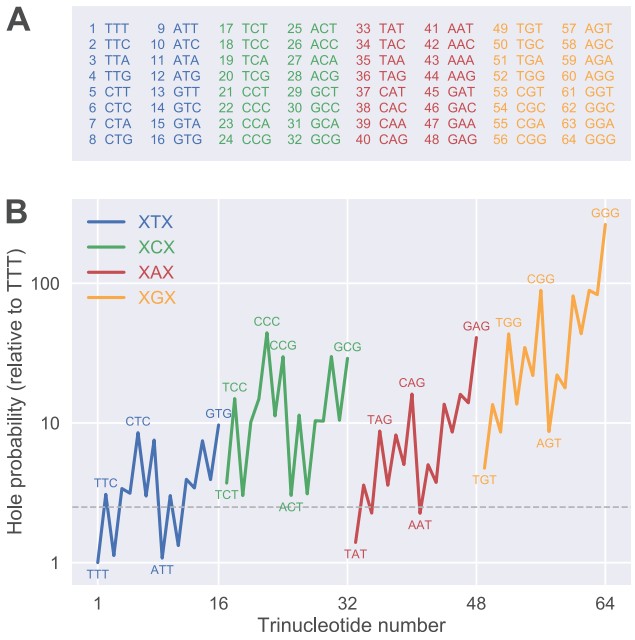

Fig. 2.—Hole probabilities vary over two orders of magnitude. Average hole probability (*B*) of middle nucleotide versus triplet number, as defined in the table (*A*). The highest hole probabilities (*B*) occur in triplets containing one or more G:C base pairs (e.g., 56 = CGG, 62 = GGC, and 64 = GGG). Values are averages of $\sim 2 \times 10^4$ values per triplet and are displayed relative to the average value for TTT on a log scale. Below the dashed line are the seven triplets with the lowest hole probabilities: AAT, ATA, TAA, ATT, TAT, TTA, and TTT.

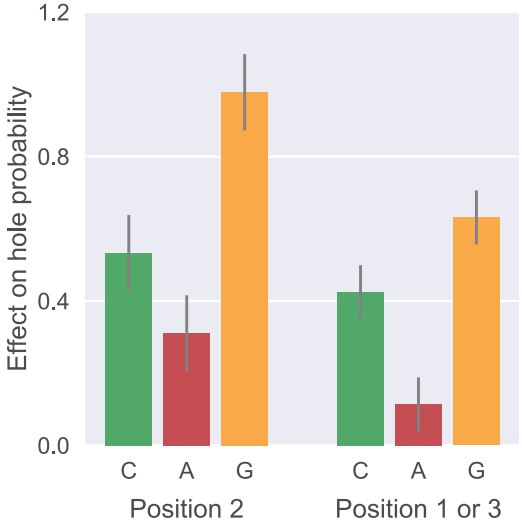

Fig. 3.—A simple additive model explains most of the variation in log hole probability in the middle nucleotide of the 64 triplets shown in figure 2*B*. Values are estimated increases in log hole probability, relative to TTT, for the different nucleotides at each of the three positions. On the basis of general linear model #2 described in the text. Error bars are 95% confidence intervals based on the general linear model. A value of *x* represents a 10^x-fold increase in hole probability. For example, a C at position 2 increases hole probability at position 2 by $10^{0.53} = 3.4$-fold relative to a T.

compared the hole probabilities to empirical estimates of mutation rates in the middle nucleotides of those triplets (Sung et al. 2015).

We began by computing the hole spectra of 1,000 random circular double-stranded DNA sequences of length 1,920, each containing a random sample of 640 trinucleotides drawn with equal probability from the 64 possible trinucleotides. We then evaluated the hole probabilities of the middle nucleotides in the triplets in both strands. The middle nucleotides of different triplets were found to differ by over two orders of magnitude in their average hole probabilities (fig. 2). The middle nucleotide itself has a strong effect, explaining 44% of the variance in log hole probability among triplets shown in figure 2*B* (general linear model #1, with the nucleotide at position 2 as a categorical predictor: $F_{3,60} = 17.53$, $P < 10^{-7}$). The T nucleotide has the lowest hole probability. The nucleotides A, C, and G have hole probabilities 2-, 3.4- and 9.5-fold higher than T, respectively (fig. 3, position 2). Thus, average hole probabilities, like mutation rates, are nucleotide-specific.

Average hole probabilities, like mutation rates, are also context-dependent. For example, the hole probability of the T in GTG is 10-fold higher than that of the middle T in TTT (fig. 2*B*). Taking into account the number of nucleotides of each type in positions 1 and 3 increases the proportion of the variance in log hole probability explained to 92% (general linear model #2, with the nucleotide at position 2 as a categorical predictor, and the total number of nucleotides of each type in positions 1 and 3 as three continuous predictors: $F_{6,57} = 120.3$, $P < 10^{-15}$; comparison to model #1, $\Delta AIC = -121.0$, i.e., model #2 provides a vastly better description of the data in fig. 2*B*). The "contextual" effect of each neighboring nucleotide on the hole probability of the middle nucleotide is similar, but not identical, to that of the middle nucleotide (fig. 3): the total number of nucleotides of each type in a triplet only explains 88% of the variance in log hole probability of the middle nucleotide (general linear model #3, with the total number of nucleotides of each type in positions 1–3 as three continuous predictors: $F_{3,60} = 157.4$, $P < 10^{-15}$; comparison to model #2, $\Delta AIC = +21.6$, i.e., model #2 is better). The contributions of positions 1 and 3 are symmetrical (a general linear model allowing asymmetry shows $\Delta AIC = +6.0$ when compared with the symmetrical model #2, i.e., model #2 is better).

Hole Probabilities of Guanines Are Correlated with Their Oxidation Reactivity

Sequence dependent hole probabilities of guanines are negatively correlated with their ionization potential (Saito et al. 1998) and positively correlated with their

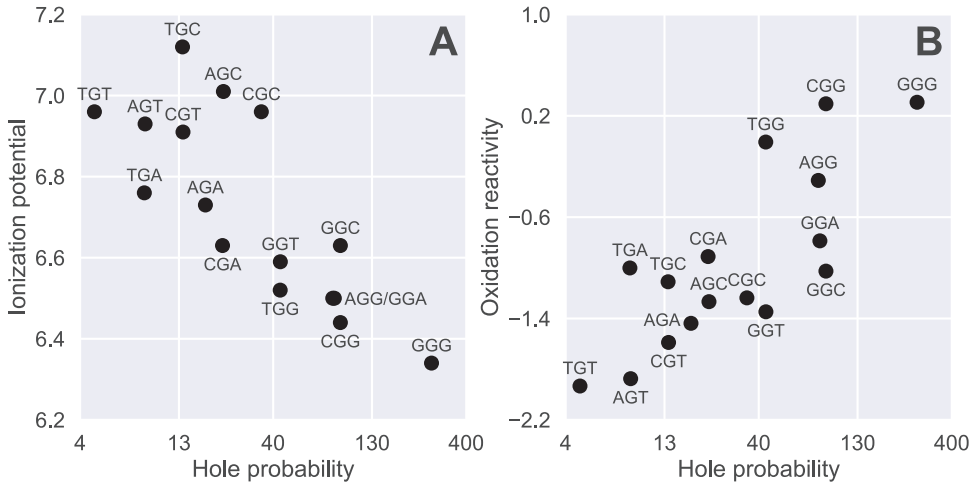

Fig. 4.—Hole probabilities of guanines are correlated with their oxidation reactivity. Relationships between hole probability of guanines in different sequence contexts, and their ionization potential in eV (*A*) (Saito et al. 1998) and reactivity toward riboflavin-mediated photooxidation (*B*) (Margolin et al. 2006). Hole probabilities are the same as in figure 2 (XGX triplets) and are displayed on a log-scale. Oxidation reactivity (*B*) is ln-transformed.

oxidation reactivity (Margolin et al. 2006; fig. 4). Spearman's rank correlation: $\rho = -0.77$ and 0.71, respectively (both $P \leq 0.002$).

Hole Probabilities Explain Some of the Nucleotide-specificity and Context-dependence of Bacterial Mutation Rates

MA is the standard method for studying mutations experimentally (Bateman 1959; Mukai 1964). In a typical MA experiment, several inbred or clonal lines are maintained in isolation at as low an effective population size as possible. This reduces the efficiency of natural selection and allows most mutations to accumulate approximately neutrally. Mutations can be detected by comparing the genomes of MA lines with that of their ancestor.

The mutation rates of the middle nucleotides in the 64 triplets have been estimated in MA experiments in mismatch repair-deficient (MMR^-) strains of *B. subtilis* (Sung et al. 2015), *E. coli* (Lee et al. 2012), *M. florum* (Sung et al. 2012, 2015), and *P. fluorescens* (Long et al. 2015). Hole probabilities are positively correlated with the triplet mutation rates of all species (fig. 5). The correlations are strong for *M. florum* (Spearman's rank correlation coefficient, $\rho = 0.612$, $P < 10^{-5}$) and *B. subtilis* ($\rho = 0.425$, $P = 0.0005$), but weak for *E. coli* ($\rho = 0.190$, $P = 0.13$) and *P. fluorescens* ($\rho = 0.283$, $P = 0.02$; note that these *P*-values are not corrected for multiple tests). An average correlation of $\bar{\rho} = 0.377$ between hole probabilities and the mutation rates of the four species is unlikely to occur by chance alone (two-tailed permutation test: $P < 10^{-5}$, based on 2×10^6 permutations). Notably, the seven triplets predicted to have the lowest hole probabilities (fig. 2) show low mutation rates in all species (fig. 5, rectangles).

Hole Probabilities Explain Some of the Variation among Sites in Mutation Rate in Human mtDNA

The noncoding HVS-I (16,024–16,383) of the human mtDNA control region, appears to contain several mutational hotspots (Excoffier and Yang 1999; Meyer et al. 1999; Stoneking 2000; Bandelt et al. 2006; Rosset et al. 2008; Soares et al. 2009). We now compute hole spectra for the L- and H-strands of the rCRS (Anderson et al. 1981; Andrews et al. 1999) and compare them to the site-specific mutation rates estimated in the study of Rosset et al. (2008) based on 16,609 HVS-I sequences, 37.5% of which had the rCRS haplotype (Behar et al. 2007). Figure 6 shows that some of the variation in mutation rate in HVS-I is explained by variation in hole probability (Spearman's rank correlation coefficients: L-strand, $\rho = 0.132$, $P = 0.01$; H-strand, $\rho = 0.183$, $P = 0.0005$).

Sites with high hole probabilities tend to be highly mutable. The 80 sites with the highest hole probabilities have higher mutation rates (μ) on average than expected by chance (two-tailed permutation tests based on 2×10^5 permutations: L-strand, $\bar{\mu} = 0.253$, $P = 0.0003$; H-strand, $\bar{\mu} = 0.235$, $P = 0.002$). Low hole probabilities, however, are not associated with low mutability in the same way ($P > 0.05$). This asymmetry is not surprising because high mutation rates are expected to be more accurately estimated than low mutation rates. Consistent with this argument, the strength of the correlation between mutation rate and hole probability is highest in regions of relatively high mutation rate (supplementary fig. S1, Supplementary Material online).

The hole spectra of other common haplotypes—K, J, T2, U5a1a, V, and I (Behar et al. 2007)—were highly significantly correlated with that of rCRS (mean Pearson's correlation coefficients, L-strand: $\bar{r} = 0.997$; H-strand: $\bar{r} = 0.996$), indicating

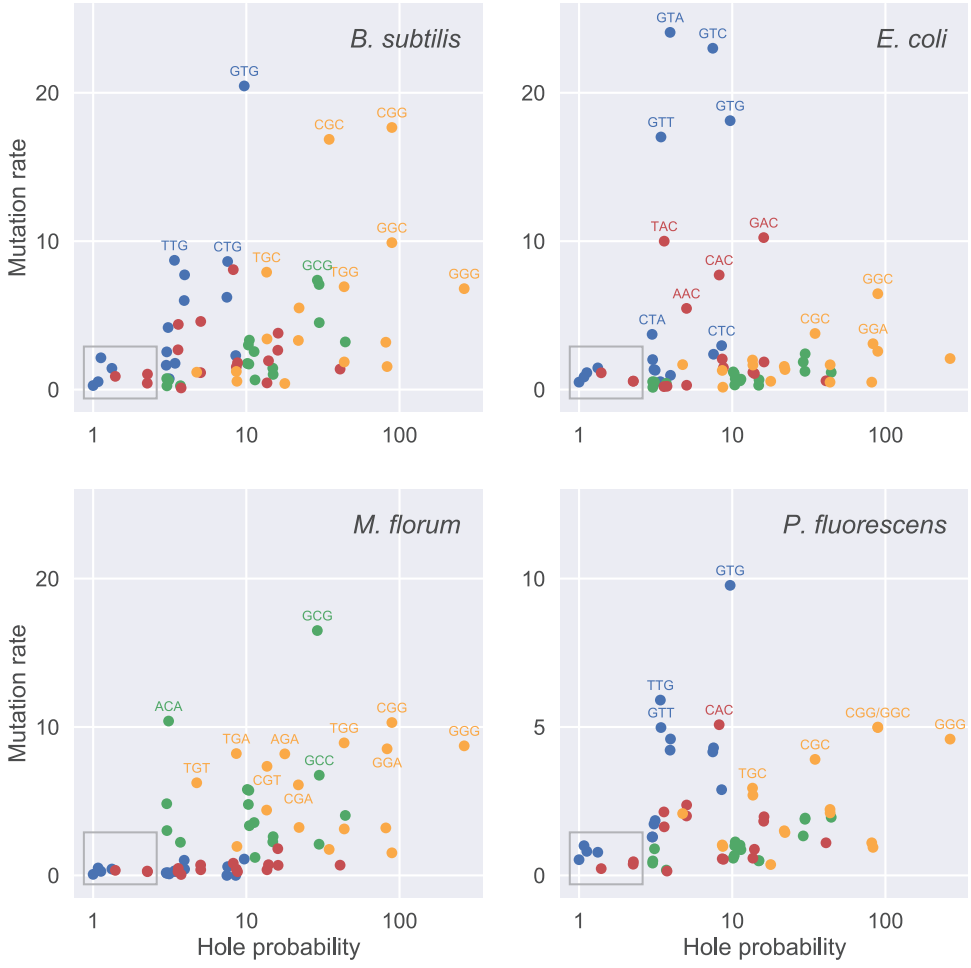

Fig. 5.—Hole probabilities explain some of the nucleotide- and context-specificity of bacterial mutation rates. Scatter plots of mutation rate ($\times 10^{-8}$ per site per generation) versus hole probability in the middle nucleotide of each of the 64 triplets, for four species of bacteria (see supplementary table S1, Supplementary Material online). Hole probabilities are the same as in figure 2 and are displayed relative to TTT on a log scale. The rectangles enclose the seven triplets with the lowest hole probabilities: AAT, ATA, TAA, ATT, TAT, TTA, and TTT (below the dashed line in fig. 2). Note that the *y*-axis of *Pseudomonas fluorescens* is half as long as that of the other three species.

that our results are unlikely to be affected by evolution of the hole spectra.

Hole Probabilities Explain Some of the Variation among Sites in Substitution Rates in Human Nuclear DNA

Levels of single-nucleotide polymorphism vary widely across the human genome (Hodgkinson and Eyre-Walker 2011). A recent study found that the heptanucleotide sequence context of a site accounts for over 80% of the variation in substitution rate in the human genome (Aggarwala and Voight 2016). Figure 7 summarizes the relationship between hole probability and substitution rate for sites with 8,192 different heptanucleotide sequence contexts. Broadly, there is a strong positive correlation between the two variables (Spearman's rank correlation coefficient: $\rho = 0.428$, $P < 10^{-15}$). This correlation is largely driven by the fact that C nucleotides in a CpG dinucleotide have ~12-fold higher substitution probabilities and ~5-fold higher hole probabilities when compared with nucleotides in nonCpG dinucleotides (upper vs. lower cloud in fig. 7). The likely explanation for this difference in substitution rate is that CpG sites are methylated at a much higher rate than nonCpG sites, and 5-methylcytosine undergoes spontaneous deamination to T (Aggarwala and Voight 2016).

To evaluate the additional contribution of hole probability to substitution rate, we analyzed the CpG and nonCpG sites separately using linear regression. Log substitution probability of nonCpG sites increased with log hole probability (slope and 95% confidence interval: $b = 0.079 \pm 0.007$; $P < 10^{-15}$; $r^2 = 7.5\%$). In contrast, log substitution probability of CpG sites *decreased* with log hole probability ($b = -0.017 \pm 0.017$; $P = 0.048$; $r^2 = 0.28\%$). Thus, log hole probability explained some of the variation in log substitution probability for nonCpG sites ($r^2 = 7.5\%$), but not for CpG sites (fig. 7).

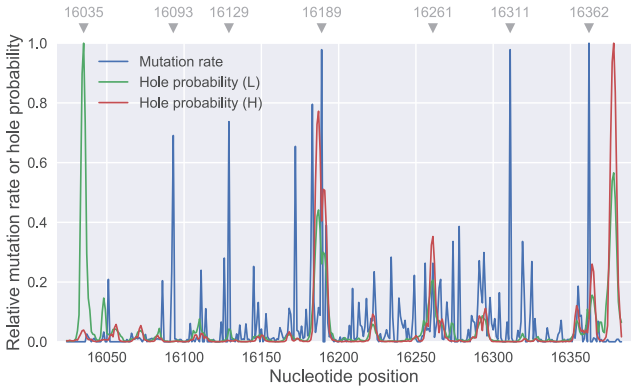

Fig. 6.—Hole probabilities explain some of the variation in mutation rate among sites in the human mitochondrial HVS-I. Mutation rates (per site per million years) were obtained from Rosset et al. (2008). Hole probabilities were estimated separately for both strands (data in supplementary table S2, Supplementary Material online). Both hole probabilities and mutation rates have been rescaled to have a maximum of 1. Certain nucleotide positions are highlighted at the top.

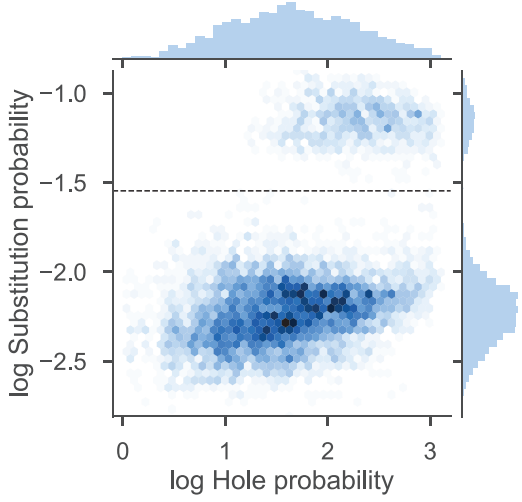

Fig. 7.—Hole probabilities explain some of the variation among sites in substitution rates in humans. Hexagonal bin plot of substitution probability (Aggarwala and Voight 2016) against hole probabilities for the middle bases of noncoding human DNA heptanucleotides. Both variables are $\log_{10}$ transformed. Hole probabilities are reported relative to the lowest value (that of TATAATA). The dashed line indicates the maximum log substitution rate (–1.546) shown by nonCpG sites. Only 4 out of 1024 (0.4%) heptanucleotides, where positions 4 and 5 are a CpG dinucleotide, have log substitution rates <–1.546.

Discussion

There is a growing body of evidence that the rate of substitution at a site depends on its immediate sequence context (Hodgkinson and Eyre-Walker 2011). One likely mechanism for these patterns is context-dependent mutation rate biases (Haag-Liautard et al. 2008; Lee et al. 2012; Sung et al. 2015; Behringer and Hall 2016). Our results support the hypothesis that localized electron–holes can affect site-specific mutability, perhaps by triggering base–pair substitutions.

DNA repair mechanisms are expected to counteract the effects of electron–holes on mutation. Thus, we expect that hole-related mutational mechanisms should be easier to detect when DNA repair mechanisms are impaired. Our data provide tentative support for this prediction. The strongest evidence for a correlation between hole probability and mutation rate was found in mismatch repair-deficient strains of bacteria. Hole probabilities were at least as good a predictor of context-dependent mutation rates in a given species of bacteria (mean Spearman's rank correlation coefficient, $\bar{\rho} = 0.377$, $n = 4$), as the rates of one species were at predicting those of other species ($\bar{\rho} = 0.338$, $n = 6$).

Another complication is that other mutational mechanisms, such as DNA methylation, could be confounded with hole probability. *Escherichia coli* shows two main types of methylation: methylation by the Dam methylase affects the A in GATC sequences and methylation by the Dcm methylase affects the second C in CCAGG and CCTGG sequences. These methylation sites do not appear to show an increased mutation rate in *E. coli*. The A in GAT triplets, which include Dam methylation sites, has a mutation rate of 2.1×10^{-8} per site per generation, and this is lower than the average for all A nucleotides (2.8×10^{-8} per site per generation); similarly, the second C in CCA and CCT triplets, which include Dcm methylation sites, have mutation rates of 0.59×10^{-8} and 0.29×10^{-8} per site per generation, respectively, and these are lower than the average for all C nucleotides (0.86×10^{-8} per site per generation). Interestingly, the A in GAC triplets has the highest mutation rate of all triplets with A in the middle position (10.2×10^{-8} per site per generation), and the A in GACC is methylated when Dam is overexpressed (Clark et al. 2012). Thus, it is possible that Dam methylation contributes to some of the nucleotide-specificity and context-dependence of mutation rates in *E. coli*. This could explain why *E. coli* shows the lowest correlation between mutation rate and hole probability. DNA methylation is unlikely to explain the strongest correlations reported here: both *B. subtilis* and *M. florum* appear to lack both Dam and Dcm methylation (Dreiseikelmann and Wackernagel 1981).

DNA methylation is also rare in human mtDNA (Liu et al. 2016). However, the results for the mtDNA control region were somewhat less conclusive than those for bacteria. Some of the strongest mutational hotspots in HVS-I correlate with peaks in the hole spectrum (e.g., positions 16,189, 16,192 and 16,261; fig. 6). But there are exceptions: some mutational hotspots have low hole probabilities (e.g., positions 16,093, 16,172, and 16,311), and some mutational coldspots have high hole probabilities (e.g., positions 16,033–16,036 and 16,377–16,379 (fig. 6). The mutation rate estimates of Rosset et al. (2008) are indirect and, therefore, may reflect the action of natural selection since they are based on sequences from live individuals. For example,

position 16,034 has a high hole probability but low (germline) mutation rate. However, somatic mutations at that position 16,034 have been found in prostate and ovarian tumors [MITOMAP, also see Chen et al. 2002; Brandon et al. 2006; Yu 2012; Samuels et al. 2013], indicating that mutations at this site may experience purifying selection.

DNA methylation is common in the human nuclear genome and has a major effect on substitution rates (Aggarwala and Voight 2016). Interestingly, we were able to detect an effect of holes independent of DNA methylation: log hole probability explained $r^2 = 7.5\%$ of the variation in log substitution probability for nonCpG sites. Future work is needed to compute hole probabilities for *methylated* DNA to assess whether or not holes may have some influence on this process in humans.

The intrinsic mutability, finally, does not scale linearly with hole probability even if hole localization is a major mutation-triggering event. A more comprehensive tool, capable of estimating site-dependent mutability from a known sequence, would use computational DNA hole spectroscopy as a starting point but would also need to incorporate one or more base mismatch mechanisms. These might include oxidation-stabilized tautomerism (Watson and Crick 1953; Löwdin 1963; Cerón-Carrasco et al. 2010; Bebenek et al. 2011), base deamination (Krokan et al. 2002; Bacolla et al. 2014), or other mechanisms some, but not necessarily all, potentially influenced by holes.

Any hole-enhanced mutation mechanism would likely be influenced by: 1) the probability of hole localization and oxidation of a given base; 2) any potential barrier in the oxidized state for forming a base–pair mismatch; and 3) the energetics of DNA polymerases and repair enzymes involved in replication. These would be affected by the specific base pair and its position within a sequence. Although we generally find the highest hole probabilities on segments with several G:C base pairs in a row, A:T base pairs may have more favorable energetics for creating a mismatch once a specific base becomes oxidized (Lewis et al. 2014). Intriguingly, we sometimes find, both in figure 6 (e.g., position 16,362) and in our previous work (Suárez Villagrán and Miller 2015), that a peak in mutation rate (or allele frequency) occurs on an A:T site near the *edge* of a hole peak rather than on a G:C site in the middle. A plausible hypothesis is that the *mutation* probability is enhanced by a smaller barrier to create a mismatch at an A:T pair, whose *hole* probability is enhanced by the adjacent hole peak centered on adjacent G:C sites. Lewis et al. (2014) find that the ranking of redox potentials between G and A, when going from normal to rare tautomer forms, reverses: from $G < A$ to $A_T < G_T$.

In summary, the results presented here support the hypothesis that physical mutation mechanisms, such as those triggered by sequence-specific hole localization, play important roles in molecular evolution. Unraveling the relative importance of holes versus other physicochemical mutation and repair mechanisms, however, remains a challenge for the future. Mutability models that build on computational DNA hole spectroscopy, but incorporate other factors, could ultimately lead to better understanding of both evolution and the emergence of somatic disease states such as cancer.

Supplementary Material

Supplementary data are available at *Genome Biology and Evolution* online.

Acknowledgments

This work was supported by the State of Texas through the Texas Center for Superconductivity at the University of Houston (M.Y.S.V. and J.H.M.), and by the National Institutes of Health (grant number R01 GM101352 to R.B.R.A.).

Literature Cited

Aggarwala V, Voight BF. 2016. An expanded sequence context model broadly explains variability in polymorphism levels across the human genome. Nat Genet. 48(4):349.

Anderson PW. 1958. Absence of diffusion in certain random lattices. Phys Rev. 109(5):1492–1505.

Anderson S, et al. 1981. Sequence and organization of the human mitochondrial genome. Nature 290(5806):457–465.

Andrews RM, et al. 1999. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet. 23(2):147.

Bacolla A, Cooper D, Vasquez K. 2014. Mechanisms of base substitution mutagenesis in cancer genomes. Genes 5(1):108–146.

Bacolla A, et al. 2013. Guanine holes are prominent targets for mutation in cancer and inherited disease. PLoS Genet. 9(9):e1003816.

Bandelt H-J, Kong Q-P, Richards M, Macaulay V. 2006. Estimation of mutation rates and coalescence times: some caveats. In: Bandelt H-J, Macaulay V, Richards M, editors. Human mitochondrial DNA and the evolution of homo sapiens. Berlin, Heidelberg (Germany): Springer. p. 47–90.

Bateman AJ. 1959. The viability of near-normal irradiated chromosomes. Int J Radiat Biol Relat Stud Phys Chem Med. 1(2):170–180.

Bebenek K, Pedersen LC, Kunkel TA. 2011. Replication infidelity via a mismatch with Watson–Crick geometry. Proc Natl Acad Sci USA. 108(5):1862–1867.

Behar DM, et al. 2007. The genographic project public participation mitochondrial DNA database. PLOS Genet. 3(6):e104.

Behringer MG, Hall DW. 2016. Genome wide estimates of mutation rates and spectrum in Schizosaccharomyces pombe indicate CpG sites are highly mutagenic despite the absence of DNA methylation. G3: Genes Genomes Genet. 6:149–160.

Brandon M, Baldi P, Wallace D. 2006. Mitochondrial mutations in cancer. Oncogene 25(34):4647–4662.

Carrillo-Nuñez H, Schulz PA. 2008. Localization of electronic states in finite ladder models: participation ratio and localization length as measures of the wave-function extension. Phys Rev B. 78(23):235404.

Cerón-Carrasco JP, Requena A, Perpète EA, Michaux C, Jacquemin D. 2010. Theoretical study of the tautomerism in the one-electron oxidized guanine–cytosine base pair. J Phys Chem B. 114(42):13439–13445.

Chen JZ, Gokden N, Greene GF, Mukunyadzi P, Kadlubar FF. 2002. Extensive somatic mitochondrial mutations in primary prostate cancer using laser capture microdissection. Cancer Res. 62(22):6470–6474.

Clark TA, et al. 2012. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. Nucl Acids Res. 40:e29.

Dizdaroglu M. 2015. Oxidatively induced DNA damage and its repair in cancer. Mutat Res/Rev Mutat Res. 763:212–245.
Dreiseikelmann B, Wackernagel W. 1981. Absence in *Bacillus subtilis* and *Staphylococcus aureus* of the sequence-specific deoxyribonucleic acid methylation that is conferred in *Escherichia coli* K-12 by the Dam and Dcm enzymes. J Bacteriol. 147(1):259–261.
Endres RG, Cox DL, Singh RRP. 2004. Colloquium: the quest for high-conductance DNA. Rev Mod Phys. 76(1):195–214.
Excoffier L, Yang Z. 1999. Substitution rate variation among sites in mitochondrial hypervariable region I of humans and chimpanzees. Molec Biol Evol. 16(10):1357–1368.
Galtier N, Enard D, Radondy Y, Bazin E, Belkhir K. 2006. Mutation hot spots in mammalian mitochondrial DNA. Genome Res. 16(2):215–222.
Galtier N, Nabholz B, Glémin S, Hurst GD. 2009. Mitochondrial DNA as a marker of molecular diversity: a reappraisal. Molec Ecol. 18(22):4541–4550.
Giese B. 2004. Hole injection and hole transfer through DNA: the hopping mechanism. In: Schuster GB, editor. Long-range charge transfer in DNA I. Berlin, Heidelberg (Germany): Springer. p. 27–44.
Giese B, Amaudrut J, Köhler AK, Spormann M, Wessely S. 2001. Direct observation of hole transfer through DNA by hopping between adenine bases and by tunnelling. Nature 412(6844):318–320.
Haag-Liautard C, et al. 2008. Direct estimation of the mitochondrial DNA mutation rate in *Drosophila melanogaster*. PLOS Biol. 6(8):e204.
Hanwell M, et al. 2012. Avogadro: an advanced semantic chemical editor, visualization, and analysis platform. J Cheminform. 4(1):17.
Hasegawa M, Di Rienzo A, Kocher TD, Wilson AC. 1993. Toward a more accurate time scale for the human mitochondrial DNA tree. J Molec Evol. 37(4):347–354.
Hodgkinson A, Eyre-Walker A. 2011. Variation in the mutation rate across mammalian genomes. Nat Rev Genet. 12(11):756–766.
Howell N, Elson JL, Howell C, Turnbull DM. 2007. Relative rates of evolution in the coding and control regions of African mtDNAs. Molec Biol Evol. 24(10):2213–2221.
Kreutzer DA, Essigmann JM. 1998. Oxidized, deaminated cytosines are a source of C → T transitions in vivo. Proc Natl Acad Sci U S A. 95(7):3578–3582.
Krokan HE, Drabløs F, Slupphaug G. 2002. Uracil in DNA – occurrence, consequences and repair. Oncogene 21(58):8935–8948.
Lee H, Popodi E, Tang H, Foster PL. 2012. Rate and molecular spectrum of spontaneous mutations in the bacterium *Escherichia coli* as determined by whole-genome sequencing. Proc Natl Acad Sci U S A. 109(41):E2774–E2783.
Lewis FD, et al. 2000. Direct measurement of hole transport dynamics in DNA. Nature 406(6791):51–53.
Lewis K, Copeland K, Hill G. 2014. One-electron redox properties of DNA nucleobases and common tautomers. Int J Quant Chem. 114(24):1678–1684.
Liu B, et al. 2016. CpG methylation patterns of human mitochondrial DNA. Sci Rep. 6:23421.
Long H, et al. 2015. Mutation rate, spectrum, topology, and context-dependency in the DNA mismatch repair-deficient *Pseudomonas fluorescens* ATCC948. Genome Biol Evol. 7(1):262–271.
Löwdin P-O. 1963. Proton tunneling in DNA and its biological implications. Rev Mod Phys. 35(3):724–732.
Margolin Y, Cloutier J-F, Shafirovich V, Geacintov NE, Dedon PC. 2006. Paradoxical hotspots for guanine oxidation by a chemical mediator of inflammation. Nat Chem Biol. 2(7):365.
Meggers E, Michel-Beyerle ME, Giese B. 1998. Sequence dependent long range hole transport in DNA. J Am Chem Soc. 120(49):12950–12955.
Meyer S, Weiss G, von Haeseler A. 1999. Pattern of nucleotide substitution and rate heterogeneity in the hypervariable regions I and II of human mtDNA. Genetics 152(3):1103–1110.
Modrich P. 1987. DNA mismatch correction. Annu Rev Biochem. 56:435–466.
Mukai T. 1964. The genetic structure of natural populations of *Drosophila melanogaster*. I. Spontaneous mutation rate of polygenes controlling viability. Genetics 50:1–19.
Rosset S, et al. 2008. Maximum-likelihood estimation of site-specific mutation rates in human mitochondrial DNA from partial phylogenetic classification. Genetics 180(3):1511–1524.
Saito I, et al. 1998. Mapping of the hot spots for DNA damage by one-electron oxidation: efficacy of GG doublets and GGG triplets as a trap in long-range hole migration. J Am Chem Soc. 120(48):12686–12687.
Samuels DC, et al. 2013. Recurrent tissue-specific mtDNA mutations are common in humans. PLoS Genet. 9(11):e1003929.
Schroeder JW, Yeesin P, Simmons LA, Wang JD. 2018. Sources of spontaneous mutagenesis in bacteria. Crit Rev Biochem Molec Biol. 53(1):29–48.
Shih C-T, Roche S, Römer RA. 2008. Point-mutation effects on charge-transport properties of the tumor-suppressor gene p53. Phys Rev Lett. 100(1):018105.
Shih C-T, Wells SA, Hsu C-L, Cheng Y-Y, Römer RA. 2012. The interplay of mutations and electronic properties in disease-related genes. Sci Rep. 2:272.
Soares P, et al. 2009. Correcting for purifying selection: an improved human mitochondrial molecular clock. Am J Hum Genet. 84(6):740–759.
Stoneking M. 2000. Hypervariable sites in the mtDNA control region are mutational hotspots. Am J Hum Genet. 67(4):1029–1032.
Suárez Villagrán MY, Miller JH. 2015. Computational DNA hole spectroscopy: a new tool to predict mutation hotspots, critical base pairs, and disease 'driver' mutations. Sci Rep. 5:13571.
Sugiyama H, Saito I. 1996. Theoretical studies of GG-specific photocleavage of DNA via electron transfer: significant lowering of ionization potential and 5′-localization of HOMO of stacked GG bases in B-form DNA. J Am Chem Soc. 118(30):7063–7068.
Sung W, et al. 2015. Asymmetric context-dependent mutation patterns revealed through mutation–accumulation experiments. Molec Biol Evol. 32(7):1672–1683.
Sung W, Ackerman MS, Miller SF, Doak TG, Lynch M. 2012. Drift-barrier hypothesis and mutation-rate evolution. Proc Natl Acad Sci U S A. 109(45):18488–18492.
Tamura K, Nei M. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molec Biol Evol. 10(3):512–526.
The Genomes Project C. 2012. An integrated map of genetic variation from 1,092 human genomes. Nature 491:56.
van Oven M, Kayser M. 2009. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. Hum Mutat. 30(2):E386–E394.
Wakeley J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. J Molec Evol. 37(6):613–623.
Watson JD, Crick FHC. 1953. Genetical implications of the structure of deoxyribonucleic acid. Nature 171(4361):964–967.
Xu M, Endres RG, Arakawa Y. 2007. The electronic properties of DNA bases. Small 3(9):1539–1543.
Yu M. 2012. Somatic mitochondrial DNA mutations in human cancers. Adv Clin Chem. 57:100.
Zhu YO, Siegal ML, Hall DW, Petrov DA. 2014. Precise estimates of mutation rate and spectrum in yeast. Proc Natl Acad Sci U S A. 111(22):E2310–E2318.

Associate editor: Bill Martin