Strong heterogeneity in mutation rate causes misleading hallmarks of natural selection on indel mutations in the human genome

E. M. Kvikstad, L. Duret

▶ **To cite this version:**

E. M. Kvikstad, L. Duret. Strong heterogeneity in mutation rate causes misleading hallmarks of natural selection on indel mutations in the human genome. Molecular Biology and Evolution, 2014, 31, pp.23-36. 10.1093/molbev/mst185 . hal-02046828

HAL Id: hal-02046828

https://univ-lyon1.hal.science/hal-02046828

Submitted on 7 May 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Strong Heterogeneity in Mutation Rate Causes Misleading Hallmarks of Natural Selection on Indel Mutations in the Human Genome

Erika M. Kvikstad*,‡,[1] and Laurent Duret[1]

[1]Laboratoire de Biométrie et Biologie Evolutive, UMR 5558, CNRS, Université Lyon 1, Villeurbanne, France

‡Present address: Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

***Corresponding author:** E-mail: erika.kvikstad@well.ox.ac.uk.

Associate editor: Matthew Hahn

Abstract

Elucidating the mechanisms of mutation accumulation and fixation is critical to understand the nature of genetic variation and its contribution to genome evolution. Of particular interest is the effect of insertions and deletions (indels) on the evolution of genome landscapes. Recent population-scaled sequencing efforts provide unprecedented data for analyzing the relative impact of selection versus nonadaptive forces operating on indels. Here, we combined McDonald–Kreitman tests with the analysis of derived allele frequency spectra to investigate the dynamics of allele fixation of short (1–50 bp) indels in the human genome. Our analyses revealed apparently higher fixation probabilities for insertions than deletions. However, this fixation bias is not consistent with either selection or biased gene conversion and varies with local mutation rate, being particularly pronounced at indel hotspots. Furthermore, we identified an unprecedented number of loci with evidence for multiple indel events in the primate phylogeny. Even in nonrepetitive sequence contexts (a priori not prone to indel mutations), such loci are 60-fold more frequent than expected according to a model of uniform indel mutation rate. This provides evidence of as yet unidentified cryptic indel hotspots. We propose that indel homoplasy, at known and cryptic hotspots, produces systematic errors in determination of ancestral alleles via parsimony and advise caution interpreting classic selection tests given the strong heterogeneity in indel rates across the genome. These results will have great impact on studies seeking to infer evolutionary forces operating on indels observed in closely related species, because such mutations are traditionally presumed homoplasy-free.

Key words: **indels, natural selection, homoplasy, sequence evolution.**

Introduction

Elucidating the processes that drive the evolution of genome landscapes is critical to understand the functional constraints acting on genome architecture. One long-standing issue is to understand the causes of the huge variation in genome sizes among eukaryotic taxa. It is well known that most of the variability in genome size is due to differences in the amount of noncoding DNA, that is, to variation in gene density. Gene density varies not only between taxa but also within genomes. Notably, in mammals, there is a 16-fold variation in gene density between GC-rich and GC-poor genomic regions (Mouchiroud et al. 1991; Lander et al. 2001). This variation is not only due to differences in the length of intergenic regions but also due to differences in intron lengths (Duret et al. 1995; Lander et al. 2001).

Such variation in genome compactness, affecting both introns and intergenic regions, necessarily results from the differential accumulation of deletions or insertions (indels). Various processes, operating at different scales, can cause indel mutations: mobilization of transposable elements, strand slippage during replication, errors during the repair of DNA double-stranded breaks, unequal crossovers, and so forth. The majority of indels correspond to small insertions and deletions, affecting only a few base pairs. More than 75% of these small indels correspond to tandem duplications or deletions (Zhu et al. 2000; Kondrashov and Rogozin 2004; Taylor et al. 2004; Messer and Arndt 2007; Tanay and Siggia 2008; Montgomery et al. 2013) and most probably result from strand slippage (Streisinger et al. 1966; Levinson and Gutman 1987; reviewed in Garcia-Diaz and Kunkel [2006]). The rate of small indel mutation is therefore strongly dependent on the presence of short tandem repeats and can become extremely high as the number of repeated units increases (e.g., in microsatellites; Ellegren 2000; Webster et al. 2002; Kelkar et al. 2008; Leclercq et al. 2010; Montgomery et al. 2013; reviewed in Ellegren [2004]). Large indels (including transposable element insertions) are much less frequent than small indels, but—due to their size—represent a major contribution to the evolution of genome size (reviewed in Gregory [2005]). What remains unclear is to what extent the accumulation of indels is driven by selective or nonadaptive forces, that is, to what extent the variation in genome compactness has any functional significance (Carvalho and Clark 1999; Comeron and Kreitman 2000; Petrov et al. 2000; Petrov 2001, 2002; Ometto et al. 2005; Presgraves 2006; Pettersson et al. 2009).

Indel mutations might be subject to natural selection either because they directly disrupt the nucleotide sequence

© The Author 2013. Published by Oxford University Press on behalf of the Society for Molecular Biology and Evolution.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Open Access

Article

Fast Track

of a functional element, or because the length (and not necessarily the sequence per se) of the region affected by the indel is functionally important. For example, there is evidence that the length of introns is under selective pressure: A minimum intron size is necessary for efficient splicing (Parsch 2003), but long introns appear to be counterselected in highly expressed genes (possibly to minimize the cost of the transcription (Castillo-Davis et al. 2002). It is also in principle possible that indel mutations are affected by the process of biased gene conversion (BGC), which might increase the probability of fixation of insertion over deletions (or vice versa) in regions of high recombination rate (Lamb 1985; Duret and Galtier 2009; Leushkin and Bazykin 2013). Finally, it is possible that genome-wide variation in gene density predominantly reflects variation in the underlying indel mutation patterns.

Many studies, based on interspecies genome comparisons, have shown that the indel density (events per nucleotide) varies at different scales across the genome (Gu and Li 1995; Britten 2002; Waterston et al. 2002; Britten et al. 2003; Hardison et al. 2003; Arndt and Hwa 2004; Kondrashov and Rogozin 2004; Makova et al. 2004; Taylor et al. 2004; Lunter et al. 2006; Wetterbom et al. 2006; Chen et al. 2007; Clark et al. 2007; de la Chaux et al. 2007; Kvikstad et al. 2007; Lunter 2007; Messer and Arndt 2007; Tanay and Siggia 2008; Chen, Chuang, et al. 2009; Chen, Wu, et al. 2009; Leclercq et al. 2010). Interestingly, Nam and Ellegren (2012) recently reported that both in birds and in human, the ratio of deletion to insertion rate (rDI) is positively correlated with recombination, and hence that recombination leads to genome contraction (but see Discussion). However, it was not investigated whether this variation results from differences in the underlying indel mutation patterns or from fixation biases (i.e., differences in the probability of fixation of insertion and deletion mutations, due to natural selection or BGC). Several approaches, based on the analysis of polymorphism and/or divergence, can be used to distinguish between these hypotheses. First, if a fixation bias tends to favor deletions over insertions, then the distribution of derived allele frequency (DAF) should be shifted toward higher frequencies for deletions as compared with insertions (and vice versa if the fixation bias favors insertions over deletions; Bustamante et al. 2001). Thus, analysis of the derived insertion and, separately, deletion allele frequency spectra can be used to detect and quantify allele frequency differences between the two types of mutations. Indeed, in a recent analysis of approximately 6,000 indels located in genic regions (transcription units and their flanking regions; Bhangale et al. 2005), Sjodin et al. (2010) observed that on average deletions segregate at lower frequencies than insertions and concluded that deletions experience a stronger level of purifying selection in genic regions. A second approach, derived from the McDonald–Kreitman test (McDonald and Kreitman 1991), consists of contrasting patterns of indel polymorphism to divergence: Under the null hypothesis that insertions and deletions share equal fixation probabilities, the rDI is expected to reflect mutation biases alone and thus to be the same for polymorphic and fixed indels. Conversely, if selection (or BGC) tends to favor deletions over insertions, rDI should be larger for fixed than for polymorphic indels (and vice versa if selection favors insertions over deletions). If a fixation bias is detected, one possibility to distinguish selection from BGC consists in analyzing the relationship with recombination. Indeed, given that BGC is directly associated to meiotic recombination, a key prediction of this model is that the fixation bias in favor of deletions (or insertions) should be strongly correlated with crossover rate.

Critical to these population genetic tests is the accurate identification of ancestral and derived states. Polarization, that is, the distinction of an insertion from a deletion, typically relies on the detection of gaps in multiple species sequence alignments. In humans, the inference of indels has been greatly improved by the recent whole genome sequencing of closely related primate species (Chen et al. 2007; Kvikstad et al. 2007; Messer and Arndt 2007; Tanay and Siggia 2008; Chen, Wu, et al. 2009; Leclercq et al. 2010; Ananda et al. 2011). On average, the indel mutation rate is relatively low: The density of polymorphic indels in the human genome is eight times lower than that of single nucleotide polymorphisms (SNPs) (Montgomery et al. 2013), in agreement with estimates of mutation rates inferred from the analysis of disease-causing mutations (Lynch 2010). Hence, given the short evolutionary distances between humans and the other sequenced primate genomes, the probability of homoplasy (i.e., multiple indel mutations at the same locus) is a priori expected to be low for the vast majority of genomic positions. However, there is evidence (based on the analysis of polymorphism and divergence) that indel mutation rate varies very widely across sites, not only at microsatellite loci (Ellegren 2000; Webster et al. 2002; Kelkar et al. 2008; Leclercq et al. 2010; Montgomery et al. 2013; reviewed in Ellegren [2004]) but also in the rest of the genome (Kvikstad et al. 2007; Montgomery et al. 2013). There is therefore a possibility of polarization errors due to homoplasy at indel mutation hotspots. Yet, this risk of error is largely ignored and attention to accurate inference of the evolutionary history of indels has lagged considerably (with notable exceptions, e.g., Chindelevitch et al. 2006; Diallo et al. 2007; Belinky et al. 2010; Hickey and Blanchette 2011).

Here, we analyze the approximately 1.6 million indels recently determined from genome-wide population-scale sequencing projects (The 1000 Genomes Pilot 1 Project; 1000 Genomes Project Consortium 2010; Montgomery et al. 2013) to ascertain whether insertions and deletions are differentially affected by selection pressure or BGC. We demonstrate that the signal of nonneutral evolution is in fact essentially artifactual and can be explained by the highly heterogeneous rates of indel mutation throughout the genome. Indeed, we find that the earlier mentioned standard tests of molecular evolution are very sensitive to indel homoplasy at sites of elevated mutation rates, which leads to errors in the identification of indel event (insertion vs. deletion) and polarization (derived vs. ancestral allele). Given the very strong heterogeneity in indel mutation rate across the genome, classic selection tests, when applied to indels, should be interpreted with caution.

Results

Polymorphic Indels

To evaluate the impact of indels on genome architecture, we analyzed data provided by the 1000 Genomes Pilot 1 Project (The 1000 Genomes Project Consortium 2010; Montgomery et al. 2013). Indels were identified in 179 individuals across three panels of Yoruban (YRI), Northern and Western European (CEU), and a combined Japanese/Chinese (JPTCHB) ancestry, providing an unprecedented genome-wide (autosomal only) catalogue of 1,582,703 indels at the population level (Montgomery et al. 2013).

Polarization

Indels in the 1000 Genomes Pilot 1 project were identified as sequence length differences with respect to the human reference sequence (National Center for Biotechnology Information [NCBI] build 36). To distinguish insertions from deletions, we applied the principle of parsimony to determine ancestral and derived alleles using the primate phylogeny ((((human, chimpanzee), gorilla), orangutan), macaque) of the 44-way multiple species alignments available from University of California Santa Cruz (UCSC) Genome Bioinformatics (Karolchik et al. 2008). We only considered genomic sites for which, in addition to the two human alleles (long and short), at least two of the four nonhuman primate sequences were present in the multiple alignment. To limit the risk of polarization error due to homoplasy, we selected indels for which all sequences were in agreement with a single, human-specific insertion or deletion. This strict procedure resulted in successful polarization of 54% of indels, giving a genome-wide estimate of 274,962 insertions and 578,298 deletions, respectively (Montgomery et al. 2013; see Materials and Methods for details). Indels range in size from 1 to 50 bp and follow the standard power-law frequency distribution of size with the majority (49%) of 1 bp in length. Despite our efforts to improve the reliability of our inference of the ancestral state by using multiple outgroups of closely related species (Belinky et al. 2010), we refer to these variants throughout the text as estimated insertions and deletions, due to the continued possibility of polarization error using the parsimony approach.

Comparison of Estimated DAF for Insertions and Deletions

To detect a potential fixation bias that could contribute to the preferential accumulation of insertions rather than deletions (or vice versa), we sought to evaluate the DAF spectra of insertions and deletions, separately. However, polarization error will confound analysis of indel allele frequencies in two ways: First, errors will lead to misidentification of an insertion as a deletion (and vice versa). Second, polarization error will lead to error in the DAF due to incorrect identification of the ancestral state. Importantly, we refer to the estimated DAF (eDAF, i.e., including polarization errors) throughout the text to distinguish between our estimated frequencies and the true DAF.

Consistent with previous results (Sjodin et al. 2010), analysis of the eDAF spectra for polymorphic insertions and deletions inferred from 1000 Genomes data reveals that genome-wide insertions segregate at significantly higher frequencies than deletions (4.8% difference in mean segregation frequencies in YRI, $P < 10^{-16}$, Mann–Whitney test of differences in frequency distribution; fig.1*A*; see supplementary fig. S1 [Supplementary Material online] for eDAF spectra of all populations). Previously, authors interpreted the higher eDAF of human-specific insertions to result from stronger negative selection pressure on deletions (Sjodin et al. 2010). If the main reason for this apparent selective pressure against deletions is that they tend to disrupt functional elements, then this signature of selective pressure should be absent in nonfunctional sequences. We therefore analyzed indel eDAF spectra in ancestral repeats (ARs), which are generally considered as good neutral markers, essentially devoid of functional elements (Waterston et al. 2002; Hardison et al. 2003; Lunter et al. 2006). Interestingly, we observed that the difference in eDAF between insertions and deletions is the same for indels located in ARs as compared with non-AR noncoding sequences (fig. 1*B*). Hence, the apparent difference in fixation dynamics between insertion and deletion cannot be attributed to the disruption of functional elements. A remaining hypothesis is that the deleterious effect of deletions is due to their impact on the length of sequences. We observed that the difference in eDAF between insertions and deletions is the same for indels located in introns and intergenic regions (supplementary figs. S2 and S3, Supplementary Material online). Thus, this apparent selective pressure cannot be attributed specifically to constraints acting on intron lengths. In other words, according to the selectionist hypothesis, this pattern would reflect a selective pressure against the shortening of the whole genome, which seems a priori unlikely.

An alternative possibility is that the fixation bias in favor of insertions results from BGC (Lamb 1985; Duret and Galtier 2009; Leushkin and Bazykin 2013). For example, if the repair of gaps in heteroduplex DNA tends to favor long alleles over short alleles then one would expect to see an increase in the probability of transmission of long alleles, specifically in regions of high recombination. To test this prediction, we computed the difference in mean eDAF between insertions and deletions for indels located in regions of different recombination, using crossover rates as a proxy measure for total recombination rate. As a positive control, we computed on the same population data set, the difference in mean eDAF between SNPs resulting from AT to GC (WS) changes versus GC to AT (SW) changes. In agreement with previous studies (Spencer 2006; Katzman et al. 2011), we found that the difference in eDAF for WS and SW changes increases steadily with local crossover rates, from approximately 1% in regions of low recombination to approximately 5% in regions of high recombination (fig. 2*A*). The same pattern is observed in all populations and is consistent with BGC in favor of GC alleles (Duret and Galtier 2009). In contrast, the difference between insertion and deletion eDAF shows very little variation with crossover rate. In YRI, there is a weak increase of this difference (from ~4% in regions of low recombination to ~5% in

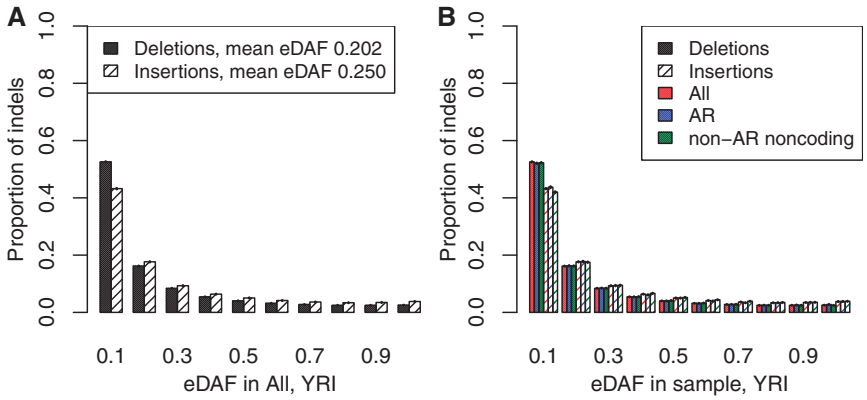

Fig. 1. eDAF spectra for polymorphic deletions (solid bars) and insertions (hash bars) by genome annotation, segregating in the YRI population. (*A*) All indels; (*B*) all indels compared with indels in ARs, and non-AR noncoding sequence. Error bars represent 1 SEM, using a binomial distribution to model the eDAF.

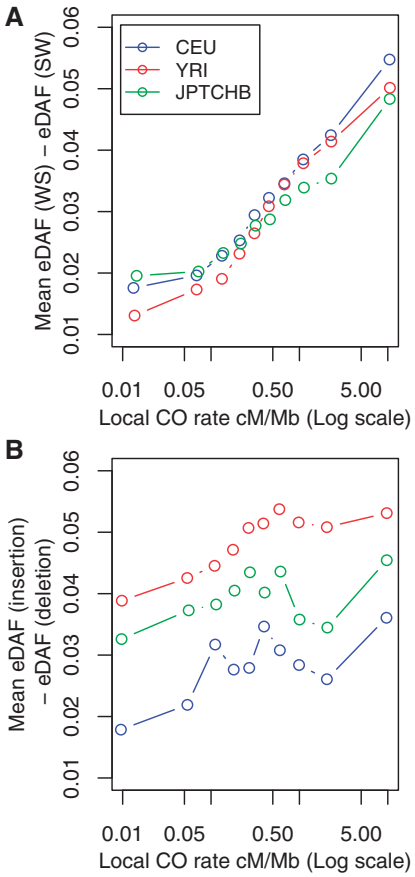

Fig. 2. Difference in mean eDAF as a function of local crossover rates between (*A*) AT to GC (weak to strong: WS) and GC to AT (strong to weak: SW) SNPs and (*B*) insertions and deletions. Sequence variants are shown segregating in the YRI (red), CEU (blue), and JPTCHB (green) populations.

regions of high recombination; fig. 2*B*). However, this trend is barely detectable in the two other populations. Notably, the magnitude of the difference in eDAF in low recombining regions (where the DAF is not expected to be affected by BGC) is greater in all populations than the slight increase with crossover rate (fig. 2*B*). Hence, BGC cannot be the main explanation for the observed difference in eDAF between insertions and deletions.

If both natural selection and BGC fail to explain the observations, it is possible that the difference in eDAF between insertions and deletions is due to an artifact in the identification of mutation events. This hypothesis is further explored later.

eDAF Spectra Are Sensitive to Indel Polarization Errors

The estimation of DAF spectra is highly sensitive to the accurate identification of derived and ancestral states. To illustrate this, let f denote the frequency of an indel polymorphism for which the true derived allele corresponds to a deletion segregating in the population. In the case of indel homoplasy due to the same deletion occurring in the outgroup, this will lead to polarization error via the parsimony approach: At this locus, an indel will erroneously be considered as corresponding to an insertion event segregating at frequency $(1 - f)$. Given that the majority of derived alleles are rare (i.e., f is generally much smaller than 50%), on average, the misidentified alleles will segregate at frequencies $(1 - f)$ much greater than 50%: thus, polarization errors lead to a systematic shift in the inferred eDAF spectra toward higher frequency alleles. Moreover, if the two types of mutation occur at different rates, then one expects more cases of homoplasy for the mutation type that has the highest rate (Eyre-Walker 1998). It has been shown that this artifact has a substantial impact on the eDAF spectra of human SNPs, notably at CpG sites, which leads to spurious signatures of natural selection (Hernandez et al. 2007). In the case of indels, the genome-wide estimate of rDI (2.22) indicates that deletions outnumber insertions. Hence, we expect more cases of deletions misidentified as insertions (i.e., rDI is in fact underestimated) and a subsequent overestimation of the insertion eDAF.

Loci subject to intrinsically high indel mutation rates are expected to experience elevated frequency of indel

homoplasy and hence polarization errors. Because of the potential impact on the inference of derived and ancestral states, we sought to identify indel hotspots, that is, sequence contexts with high indel mutation rate, to reanalyze separately the eDAF of insertions and deletions taking into consideration their local sequence complexity. For instance, the indel mutation rate is on average about eight times lower than the base replacement mutation rate (Lynch 2010; Montgomery et al. 2013). However, the indel density is highly heterogeneous across the genome, varying up to several orders of magnitude higher than the SNP density (Montgomery et al. 2013). Thus, here we define indel hotspots as regions for which the indel density exceeds the SNP density. These regions were identified first by considering repetitive sequences using recent annotations of exact tandem repeats and setting tract length thresholds such that the indel density exceeds the SNP density (see Materials and Methods). Additionally, we considered near-repetitive tracts consisting of repeat units nontandemly arrayed and for which their predicted indel mutability, based on a quantitative model of strand slippage, exceeds the SNP rate (see Materials and Methods for details; Montgomery et al. 2013). Note that, true to their name, indel hotspots account for approximately 4% of the genome, yet are substantially enriched in indels (681,526 events corresponding to 43%; Montgomery et al. 2013), for which only 23% (156,109 events) could be successfully polarized. Nonrepetitive (NR) sequences were defined by excluding such regions (901,178 events, 77% polarized).

We observe strong heterogeneity in the eDAF of insertions and deletions according to their local sequence context. In particular, the differences in insertion versus deletion eDAF are most notable in indel hotspot contexts that account for up to an 8.6% difference in mean segregation frequencies ($P < 10^{-16}$), compared with 2.1% for indels in NR context (fig. 3; see supplementary fig. S4 [Supplementary Material online] for all populations). We observed exactly the same trend in ARs and non-AR regions (supplementary figs. S5 and S6, Supplementary Material online). Thus, the signal of fixation bias appears to be much stronger at sites where the inferred indel mutation rate is higher. Again, this observation is difficult to resolve with selection: If indels were subject to selection because of their impact on the length of sequences, then this selective pressure should not vary with mutation rate; and if indels were counterselected because they disrupt functional elements, then one would have expected a stronger signature of selection in non-AR NR regions, which are enriched in functional elements compared with ARs and to low-complexity or microsatellite repeats.

The observation that the eDAF of indels varies with indel mutation rate is also inconsistent with the BGC model: The process of BGC is independent of the mutation rate, and hence is a priori not expected to be affected by the sequence context. Moreover, whatever the sequence context (indel hotspot or NR), we found no clear covariation of indel eDAF with recombination rate (supplementary fig. S7, Supplementary Material online).

Rather, the apparent fixation bias favoring insertions is consistent with the hypothesis of polarization error. We

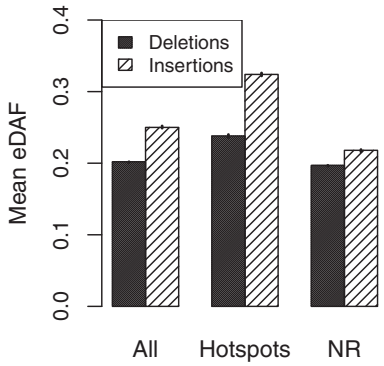

FIG. 3. Heterogeneity in mean eDAF for polymorphic indels located in various DNA contexts. Shown are frequencies of polymorphic deletions (solid bars) and insertions (hashed bars), separately. Contexts: all indels (All), indel hotspots (hotspot), and nonrepetitive (NR; see Materials and Methods for details). Indels are segregating in the YRI population. Error bars represent 95% confidence intervals of the mean.

formulated the expected impact of polarization errors on the eDAF spectra for various rDI (fig. 4A, see Materials and Methods). This model demonstrates that even small levels of polarization error can drive large differences in eDAF between insertions and deletions, and the magnitude of this difference depends on the true relative ratio of insertion versus deletion. For example, given an rDI of 2.75 in the NR portion of the genome, we estimated from our theoretical model that a polarization error rate as low as 2–3% would be sufficient to account for the observed approximately 2% disparity in insertion and deletion eDAF (fig. 4A). In contexts known to have high indel rates, such as indel hotspot loci, we expect even larger extent of polarization error and more pronounced disparity in eDAF; this is consistent with what we observe (fig. 3). Thus, eDAF is sensitive to factors influencing the estimate of the relative proportion of insertion and deletion events, that is, the presence of indel homoplasy.

McDonald–Kreitman Test

Our analysis of indels among NR and hotspot contexts revealed that the eDAF spectra are sensitive to polarization errors, and thus the higher segregation frequencies of insertion versus deletion in NR genome (on average ~2%) may not be an accurate signature of selection. To further investigate the possibility of selection altering fixation of indels, we analyzed patterns of indel divergence in conjunction with polymorphism using a modified McDonald–Kreitman approach (McDonald and Kreitman 1991). In the absence of a bias operating to preferentially increase the allele frequencies of one mutation type over the other, insertions and deletions are expected to share equal fixation probabilities. Thus, the ratio of estimated deletions to estimated insertions (rDI_e; based on polarization using parsimony) at polymorphic loci is expected to equal rDI_e at fixed loci.

To compare rDI_e for polymorphism with divergence, we obtained indels occurring in the human lineage since divergence from chimpanzee using primate multiple-species alignments and a protocol designed to resemble as closely as

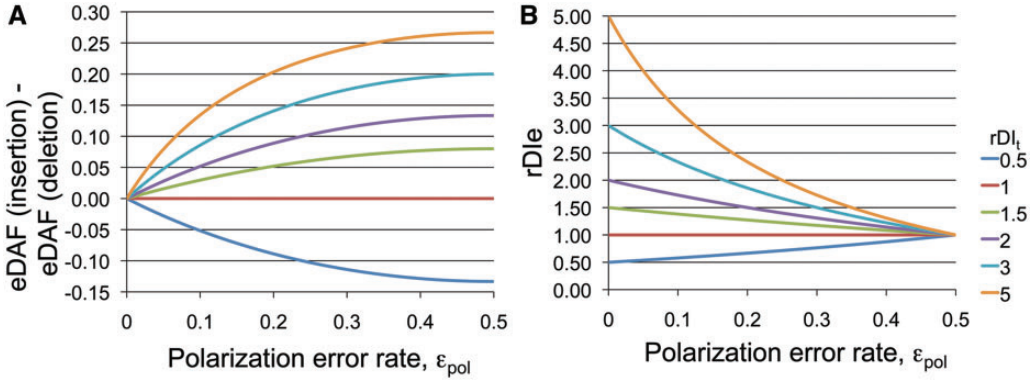

Fig. 4. Theoretical estimation of the difference in insertion and deletion mean eDAF (*A*) and estimated ratio of deletion to insertion (rDI_e; *B*) as a function of polarization errors, for various true ratios of deletion to insertion events (rDI_t).

Table 1. Modified McDonald–Kreitman Tests of the Estimated Deletion to Insertion Ratio (rDI_e) for Indel Polymorphism versus Divergence.

	All		Hotspot[b]		NR[c]	
	N[a]	rDI_e	*N*	rDI_e	*N*	rDI_e
Polymorphism	**276,495**	**2.22**	**50,297**	**0.97**	**226,198**	**2.75**
Divergence	**336,884**	**1.83 (1.3×10^3)***	**74,497**	**0.38 (6.0×10^3)***	**262,387**	**3.0 (1.9×10^2)***

Note.—For each test of polymorphism versus divergence, the rDI_e (chi-square statistic) is provided.

[a]*N*, numbers of indels per category.

[b]Hotspot, indel hotspot loci exhibiting greater than or equal to SNP diversity (see Materials and Methods for details).

[c]NR, nonrepetitive indels defined by excluding hotspots.

*Significant χ^2 statistic ($P < 10^{-16}$).

possible the criteria used for the polarization of polymorphic indels. First, we focused on the approximately half of the human autosomal genome aligning among at least four primate sequences, that is, human and chimpanzee plus a minimum of two outgroups (a subset of the genomic regions for which polymorphism data were polarized). After applying rigorous quality control criteria to reduce potential alignment artifact (see Materials and Methods for detail), we obtained 520,313 high quality indels. We were able to unambiguously polarize 336,884 indels (65%) for which all orthologous sites support one single event in the human lineage—analogous to the polarized polymorphic indels. Human-specific indels range in size from 1 to 50 bp, with the majority 1 bp (49%) as expected.

To account for regional variation in indel density, we compared rDI_e for 276,495 polarized polymorphic indels restricted with the same alignment blocks in which human-specific indel divergence was identified. Note that in terms of eDAF spectra, this subset is representative of the entire set of polymorphic indels (compare supplementary figs. S8 and S9 [Supplementary Material online] with figs. 1 and 3, respectively). Among these indels, we observe significantly larger rDI_e for polymorphism versus divergence (rDI_e polymorphism 2.22, divergence 1.82, $P < 10^{-16}$, chi-square test, one degree of freedom; table 1), which would traditionally be interpreted as evidence of a fixation bias favoring insertions over deletions. However, when we separated the indels into indel hotspot and NR contexts, separately, indel hotspot contexts display an even larger difference in rDI_e among polymorphic and fixed events. By contrast, the rDI_e in NR contexts are consistent with a significant fixation bias favoring deletions over insertions (rDI_e 2.75 for polymorphism and 3.0 for divergence; $P < 10^{-16}$, chi-square test; table 1). Thus, the indels in the hotspot and NR contexts reveal apparently opposite fixation biases, implying that the signature of selection (or BGC) depends on the underlying mutation rate.

Cryptic Indel Mutation Hotspots

In the NR portion of the genome, the eDAF and McDonald–Kreitman analyses provided conflicting results: eDAF spectra support a fixation bias favoring insertions (fig. 3), whereas the McDonald–Kreitman test implies a bias toward fixation of deletions (table 1). Selection is not a satisfactory explanation, nor do these signals appear to be driven by BGC: There is no discernable difference in rDI for polymorphism versus divergence with respect to local fluctuations in crossover (supplementary fig. S10, Supplementary Material online). Given the low indel rate in the NR context of 7.1×10^{-5} indel per NR site per million years (inferred from divergence data; 343,880 indels per 808,782,597 sites, and assuming human–chimpanzee divergence time of 6 My; Locke et al. 2011) and the stringent polarization criteria required to infer the ancestral genome (multiple primates consistent with single indel events), we a priori expected low rates of polarization errors in this best-curated subset of indels. Yet, even in the NR context we noted a substantial proportion of indels (24% divergence; 31% polymorphic) that cannot be unambiguously

polarized, that is, that occur at loci with evidence for at least two indel events in the primate phylogeny. This high frequency of loci subject to multiple indels raises the possibility that additional, cryptic hypermutable sites remain undetected among the polarized indels, contributing to hidden indel homoplasy and polarization errors even within the portion of the genome a priori not identified as hotspot.

Thus, we sought to quantify the presence of cryptic hotspots and evaluate their impact on the reliability of parsimony and the inferences of the McDonald–Kreitman approach. For simplicity, we focused on indels occurring in alignment blocks of human, chimpanzee, orangutan, and rhesus macaque (HCOR; supplementary fig. S11, Supplementary Material online) as these blocks contain the majority (88%) of the indels analyzed in the above comparison of indel polymorphism and divergence. We first defined human insertions and deletions (67,265 and 200,698, respectively) as events occurring along the human branch since divergence from chimpanzee using the orangutan allele to infer the ancestral state. Next, we compared with the rhesus allele at the orthologous position to further distinguish between "simple" events supporting one single indel in the human lineage, that is, unambiguous polarization (orangutan and rhesus agree on ancestral state; 230,086 indels); and complex events at sites with evidence for at least two indels occurring in the primate phylogeny (orangutan and rhesus alleles disagree; 37,877 indels). Using this definition, we observe that a large proportion (14%) of human indels in the NR portion of HCOR alignments correspond to complex indels.

The indel rate estimated by parsimony along the human branch restricted to HCOR blocks is 6.3×10^{-5} indel/bp/My. Given the phylogeny and divergence times for the primate species analyzed here (supplementary fig. S11, Supplementary Material online) and assuming that indel rates are constant across the phylogeny, it is possible to estimate the expected frequency of complex events (see Materials and Methods and supplementary text [Supplementary Material online] for details). Under the assumption that insertion and deletion rates are uniform across the NR regions, the expected frequency of complex events would be only 0.23%, that is, approximately 60 times lower than what we observed. This huge excess clearly indicates that the indel rate is not uniform across NR sites: Sites that are subject to an indel change in the human branch have a strong probability to experience additional indel changes along at least one branch of the phylogeny. Hence, the NR region contains hotspots of indels which do not correspond to simple repeat contexts. Given that we focused our analyses on noncoding regions, which are essentially neutrally evolving, these indel hotspots are unlikely to result from selection, and hence most probably correspond to indel mutation hotspots.

To gain insights on the density and intensity of these cryptic indel mutation hotspots, we considered a simple model of sequence evolution involving two classes of sites, basal and hidden hotspot, evolving at distinct indel mutation rates. This model involves four parameters: two fixed, the true deletion to insertion ratio (rDI_t) and the basal indel mutation rate; and two free, hotspot density (i.e., the proportion of sites located in hotspots) and the hotspot intensity (i.e., the ratio of the indel mutation rate in hotspots relative to the basal rate). We explored the space of parameters that could explain the observed frequency of complex events. According to this model, the observed frequency of complex events could be explained by a small fraction (0.1–2%) of sites located in strong mutation hotspots (approximately 200–1,000 × basal rates). Indel mutation rates are known to be extremely dependent on the presence of local repeat structures and can vary up to 1,000-fold according to the number of copies of tandemly repeated motifs (Ellegren 2000; Webster et al. 2002; Kelkar et al. 2008; Leclercq et al. 2010; Montgomery et al. 2013; reviewed in Ellegren [2004]). Our results suggest that even in NR regions there exist very strong levels of heterogeneity in indel mutation rates.

Errors of the Parsimony Approach due to Strong Heterogeneity of Mutation Rates

We next investigated the impact of this rate heterogeneity on the inference by parsimony of the rate of indel events in human (both for polymorphism and for divergence). Using the range of parameters identified previously (e.g., hotspots at 200- to 1,000-fold basal intensities, 0.1–2% densities), we computed the probabilities of all possible evolutionary scenarios of insertion and deletion across the phylogeny, and then computed the proportion of these scenarios for which the parsimony inference is incorrect (using the same criteria as described previously). There can be two kinds of errors: false discoveries (i.e., indels erroneously identified in the human branch) and false negatives (i.e., indels that occurred in the human branch, but that were not identified as such by parsimony). As shown in table 2, for the range of parameters that we tested, parsimony is expected to lead to a very low false discovery rate (FDR). However, the false negative rate (FNR) is expected to be very high (~15–17% on average), due to the presence of many sites affected by multiple indels, and for which events cannot be unambiguously inferred by parsimony. Interestingly, the FNR is substantially stronger for deletions than for insertions, which leads to underestimation of the ratio of deletion to insertion (table 2). Thus, the values of rDI_e reported in table 1 most probably represent an underestimate of the true rDI.

It is important to notice that the error rate of parsimony depends on the length of branches in the phylogenetic tree relating the sequences included in the multiple alignment (supplementary fig. S11 and text for details, Supplementary Material online). Because of this, the error rate differs for the analysis of polymorphism and divergence data: Both the FDR and FNR are significantly larger for divergence than for polymorphism ($P < 2.2e{-}16$, Welch's t test; table 2). These differences in parsimony error rates between polymorphism and divergence indels can lead to significant differences in the rDI_e, producing false positive signatures of fixation bias. And indeed, among all combinations of parameters that fit with the observed proportion of complex events, nearly one half (41%) produce significant differences (assessed by chi-square tests, $P < 0.05$, one degree of freedom) in rDI_e between indel

Table 2. Expected Parsimony Error Rates due to Indel Rate Heterogeneity, Estimated from Models of Indel Sequence Evolution.

	Mean	Range (Lower)	Range (Upper)
Deletions			
FNR[a]			
Divergence	0.170	0.064	0.469
Polymorphism	0.149	0.061	0.327
FDR[b]			
Divergence	3.28e−04	6.30e−05	1.18e−03
Polymorphism	8.36e−05	2.39e−05	2.96e−04
Insertions			
FNR			
Divergence	0.023	0	0.273
Polymorphism	0.020	0	0.219
FDR			
Divergence	3.42e−03	1.07e−03	6.49e−03
Polymorphism	3.37e−04	2.26e−04	6.82e−04
rDI_e[c] ($simple$[d])			
$rDI_t - rDI_e$			
Divergence	1.217	−2.220	5.439
$rDI_e - rDI_e$			
Divergence–polymorphism	0.067	−4.44e−16	0.214

[a]cFNR, False negative rate.
[b]dFDR, False discovery rate.
[c]rDI_e, parsimony estimated deletion to insertion ratio.
[d]Simple, indels at sites with unambiguous polarization using orangutan and rhesus sequences to determine ancestral state.

polymorphism and divergence. In most cases (90%), this bias leads to an apparent fixation bias in favor of deletions. Thus, the McDonald–Kreitman test is sensitive to hidden homoplasy due to cryptic indel hotspots, and is not reliable for indels inferred by parsimony in the presence of intense rate heterogeneity.

Discussion

To determine the relative impact of selection or neutral processes on the evolution of gene density, we sought to ascertain whether the dynamics of allele fixation were different for insertions and deletions. We analyzed a substantial number (853,260) of polymorphic insertions and deletions inferred from the 1000 Genomes pilot project for which the derived and ancestral states could be determined by parsimony via comparisons with multiple primate sequences.

Genome-wide, the eDAF spectra apparently indicate that insertions segregate at higher frequencies than deletions (fig. 1) and McDonald–Kreitman analysis reveals that the rDI_e is larger for indel polymorphism than divergence (table 1). Thus, both results suggest an apparent bias favoring the fixation of insertions over deletions. Although these results are consistent with previous reports of stronger selection acting on deletions versus insertions in transcribed sequences (Sjodin et al. 2010), here, we assayed indels in genes and intergenic regions; hence, the apparent insertion and deletion segregation bias is a more global property. Thus, we investigated the potential for selection to alter indel allele frequencies by testing two key predictions. First, if deletions were more strongly counterselected than insertions because they disrupt functional elements, then one would expect a stronger signature of selection in regions of the genome likely to harbor functional elements. We observe that the mean allele frequencies of insertions are higher than deletions not only genome-wide but also for indels located in ARs and in noncoding regions (fig. 1*B*); thus, a role for selection against indels that disrupt functional elements is unlikely. Second, if deletions were subject to selection pressure because of their impact on the length of sequences, then one might a priori expect differences between introns and intergenic regions. Contrary to this prediction, the apparent signature of selection is of the same magnitude in introns and intergenic regions. Thus, we found no evidence consistent with a differential strength of selection acting on deletions versus insertions to account for the apparent segregation distortion of indel allele frequencies, contrary to a recent publication (Sjodin et al. 2010).

An alternative hypothesis to explain the difference in eDAF between insertions and deletions is that the dynamics of fixation of indel mutations is affected by BGC. Leushkin and Bazykin (2013) recently reported that in *Drosophila*, the rDI ratio (inferred from the analysis of small indels within noncoding regions) is negatively correlated with recombination, specifically for fixed indels. Furthermore, the analysis of polymorphic indels showed that this correlation is not due to a mutagenic effect of recombination, but to an increased rate of fixation of insertions in regions of high recombination. These observations suggest that in *Drosophila*, BGC might favor insertions over deletions (Leushkin and Bazykin 2013). The authors also observed a negative correlation between rDI (measured on fixed small indels) and recombination in human and in yeast (although the correlation is much weaker than in *Drosophila*). They therefore suggested that BGC might be affecting indels in these species as well.

Our analyses also revealed a weak negative correlation between rDI and crossover rates in human (supplementary fig. S10, Supplementary Material online). However, this correlation is the same for fixed and polymorphic indels (supplementary fig. S10, Supplementary Material online). Furthermore, although the eDAF of insertions is higher than that of deletions, this apparent bias in favor of insertion shows no significant covariation with crossover rates (fig. 2*B*), contrary to SNPs, for which we found clear signatures of BGC (fig. 2*A*). Thus, if BGC is affecting indels in humans, as proposed by Leushkin and Bazykin (2013), its impact appears to be much weaker than on SNPs. In any case, the difference in eDAF between insertions and deletions is almost as strong in regions of very low crossover rates as in regions of high crossover rates (fig. 2*B*), and hence this apparent fixation bias cannot be explained by crossover-associated BGC. It should be noted that crossovers represent only a small fraction of all recombination events. However, there is evidence that in mammals, rates of crossover are strongly correlated with those of double-strand break formation (i.e., total recombination rate, including both crossovers and noncrossovers; Smagulova et al. 2011; Brunschwig et al. 2012). Hence, to

reconcile our observations with the hypothesis that indels are affected by BGC, one would have to assume that indel–BGC is a process specific to the mechanism of noncrossover resolution, which seems a priori unlikely.

It should be noted that contradictory results have been published regarding the relationship between rDI and recombination in human. Notably, in contrast to our results and those of Leushkin and Bazykin (2013), Nam and Ellegren (2012) reported a positive correlation between rDI and recombination in human, in two independent data sets of small indels (a set of indels detected within transposable elements and a set of polymorphic indels in unique sequences). The discrepancies between these studies are probably due to differences in the method of indel detection (species represented in the multiple alignments, filtering of low quality alignments, etc.). In any case, this indicates that the weak correlations detected between rDI and recombination in human should be interpreted with caution.

It is worth noting that the signal of fixation bias in favor of insertions (based on the analysis of eDAF spectra and on the McDonald–Kreitman test) is most prominent in indel hotspots—regions of the genome that display elevated mutation rates. Furthermore, in the NR, that is, "cold" portion of the genome, these classic population genetic tests provide contradictory evidence: eDAF spectra support the preferential fixation of insertions (fig. 3), whereas the McDonald–Kreitman test supports the preferential fixation of deletions (table 1). These results indicate that the intensity and even the direction of the apparent signal of fixation bias depend on the underlying mutation rate, which is obviously not consistent, neither with selection nor with BGC.

Instead, we propose that the data are consistent with an alternative hypothesis wherein parsimony errors due to hidden homoplasy drive false signatures of natural selection. Indel hotspot contexts show the strongest signature of a fixation bias, according to both segregation frequencies of polymorphic indels and tests of comparison of indel polymorphism versus divergence. Sites experiencing such elevated mutation rate are prone to recurrent and parallel mutations, both within a given lineage and among closely related species, that is, homoplasy. Such homoplasy will contribute to errors in polarization of indels, which systematically bias both the determination of ancestral versus derived alleles, and hence the identification of insertion versus deletion mutations.

Indeed, here we highlight two important aspects of the impact of indel mutation rate heterogeneity on the parsimony inference of indels. First, we reveal strong indel rate heterogeneity even among nonrepetitive sequences. The frequency of NR sites subject to multiple indel events across the primate phylogeny is 60 times higher than expected according to a model of uniform mutation rate. We estimated that a modest fraction of unidentified cryptic hotspots (up to 2% of NR sites) but at intense mutation rates (200–1,000 × basal rate) could account for the observed fraction of unpolarized complex events among high quality indels. Microsatellites, for example, tandem repeats of simple/low complexity sequence are estimated to account for approximately 3% of the human genome (Lander et al. 2001) and known to display intense rates, frequently up to several orders of magnitude higher than basal (reviewed in Ellegren 2004). Thus, such an estimate of as yet unidentified cryptic hotspots is not unreasonable.

Second, our models further demonstrate that indel rate heterogeneity due to cryptic hotspots leads to systematic errors in the parsimony inference. The eDAF spectra of insertions and deletions are highly sensitive to polarization error due to false positives. Variation in the mutation rate causing changes in the relative ratio of deletion to insertion leads to parsimony errors in determining the ancestral versus derived alleles, and hence systematically bias the eDAF spectra. A partial solution to reduce false positive errors is to include additional closely related outgroup species when identifying indels, as suggested by a recent analysis of indel homoplasy inferred from sequence data among metazoan taxa (Belinky et al. 2010). Indeed, requiring at least four primate species allowed us to identify approximately a third of the indels (24% divergence; 31% polymorphic) as loci with evidence for at least two indel events occurring at a single genomic site over the course of approximately 33 My spanning the primate phylogeny analyzed here (Locke et al. 2011). Notably, such sites will go undetected and contribute to error using a trio approach, common among indel studies. It is puzzling that we continue to observe a signal of segregation bias for indels located in the NR genome even after increasing the phylogenetic sampling (supplementary fig. S8, Supplementary Material online), given our low estimates of false positives based on the models of indel sequence evolution (table 2). We note that our model of indel sequence evolution is very simplistic, assuming all hotspots have the same intensity, which is unlikely to be the case. Additionally, our theoretical model of eDAF assumes a constant polarization error rate (due to false positives), when it might be the case that the FDR differs for insertions and deletions; moreover, we do not know the true underlying rDI, and the difference in eDAF between insertions and deletions scales dramatically with rDI_t (fig. 4*A*). It is possible that additional factors, such as the sensitivity of indel calling methods, also differ for insertions versus deletions and thus lead to issues with interpretation of the indel DAF spectra.

In addition, our investigation of parsimony inference uncovered a strikingly high FNR in the presence of indel mutation rate heterogeneity, particularly for deletions that occur more frequently than insertions. The large FNR contributes to systematic underestimation of the rates of deletions and insertions, as well as their relative frequencies (e.g., rDI). Notably, parsimony errors differ over short versus long time scales, such that estimates of the indel mutation rate for polymorphism and divergence will be biased to different degrees. Thus, the McDonald–Kreitman test is sensitive to polarization error due to false negatives, and comparisons of indel polymorphism to divergence in the manner presented here (and in Sjodin et al. 2010), or by using nucleotide substitutions as a neutral proxy (e.g., Podlaha et al. 2005; Chen, Chuang, et al. 2009) should be interpreted with caution.

In principle, maximum likelihood methods to infer insertions from deletions would constitute an improvement upon parsimony approaches by providing unbiased estimators of

insertion and deletion rates, and allowing the estimation of error. However, to date such implementations have remained challenging due to the complexity of modeling indel sequence evolution (Diallo et al. 2007; Hickey and Blanchette 2011). Models have been developed to reconstruct the most likely scenario of insertions and deletions, taking into account the differences in indel prevalence according to size (Diallo et al. 2007). However, these models assume that indel rates are uniform across sites. Recent improvements to alignment accuracy via probabilistic indel models of sequence evolution have tackled the nature of the context dependency of indel rate variation (Hickey and Blanchette 2011), yet implementation is currently limited to pairwise sequence alignments and thus cannot be applied to infer ancestral states. Additionally, point mutations may transform a highly indel-prone tandem-repeated context (e.g., microsatellite) into a cold region—and vice versa (Kelkar et al. 2011, and references therein). Thus, indel rates can vary temporally across the phylogeny. The development of a probabilistic indel model of sequence evolution taking into account the heterogeneity of indel mutation rates, both in space and in time, would greatly increase our confidence in the estimates of indel rates from comparative genomic approaches, and hence downstream applications of molecular evolutionary analyses. However, these developments remain technically very challenging. In the meantime, one should remain cautious when interpreting classical selection tests on indel data inferred by parsimony.

Maximum likelihood approaches that allow us to determine the fraction of sites likely error-prone will not, however, identify the particular events themselves. Thus, a first step to improve filtering of homoplasic sites remains characterization of complex events and determination of the underlying sources of cryptic indel mutation rate heterogeneity. Our model indicates that the proportion of cryptic hotspots in the NR genome is nonnegligible, yet it does not provide information on which sites these correspond to. Initial inspection of complex events does not reveal any striking properties: complex indels are distributed uniformly across the genome (data not shown), and they are slightly more AT-rich than simple indels (complex 35% G + C content, simple 41% G + C, $P < 10\,e{-}14$ Welch's two-sample test). Nor are they overrepresented in microsatellite, simple repeat, or low complexity annotations of the human genome provided by other sources (e.g., Repeat Masker; Smit et al. 1996–2004), as would be expected if fine-tuning of indel hotspot criteria were the explanation for their escape from detection. Identification of cryptic hotspots thus remains an important challenge for future studies.

Classification of the sources of cryptic indel mutation hotspots will not only be important for correct identification of ancestral indel states, but also it will enrich our understanding of the mechanisms of indel mutagenesis. Given divergence times in the range of 6–8 My (Chimpanzee Sequencing and Analysis Consortium 2005), and the overall low rate of indel mutation (approximately one tenth that of point mutation, e.g., Montgomery et al. 2013), indel homoplasy is traditionally regarded as weak to nonexistent and not considered in analyses of indel evolution. Our models confirm that when indel mutation rates are uniform, parsimony errors are indeed extremely low and thus, on average, indel homoplasy will not be a concern. However, a growing number of studies have illustrated an unprecedented heterogeneity in indel rates genome-wide (e.g., Kvikstad et al. 2007; Leclercq et al. 2010; Ananda et al. 2011; Montgomery et al. 2013)—and thus elucidating the sources of indel variation will be important for understanding dynamics of indel mutation as well as fixation.

Materials and Methods

Polarization of Polymorphic Indels

Indels in the 1000 Genomes Pilot 1 project were identified across three population groups: 59 YRI individuals, 60 CEU individuals, and 60 JPTCHB individuals. Indels were polarized using the principle of parsimony to determine ancestral status among the primate species (human hg18, chimpanzee panTro2, gorilla gorGor1, orangutan ponAbe2, and rhesus macaque rheMac2) available in the 44-way multiple species alignments at UCSC Genome Bioinformatics (Karolchik et al. 2008). Successfully polarized sites required a minimum of two of the four nonhuman primate sequences to align to the two human alleles, and all outgroup sequences to be in agreement with a single, human-specific insertion or deletion. As exact tandem duplications/deletions account for a large majority of indels (discussed earlier), and gap placement in these regions can be arbitrary, the number of tandem copies of the motif present in the indel sequence in each of the aligning genomes was used to identify ancestral states. This strict procedure resulted in successful polarization of approximately 50% of indels, giving a genome-wide (autosomal only) estimate of 274,962 insertions and 578,298 deletions, respectively (Montgomery et al. 2013). Data sets of indels for analysis of allele frequencies included only those events for which the minor allele was observed in at least one individual in at least one population (to exclude indels that might correspond to errors in the reference genome): This corresponded to 622,379 indels (YRI), 417,190 indels (CEU), and 355,525 indels (JPTCHB).

Identification and Polarization of Indel Divergence

As we required two outgroups to polarize the polymorphic indel data set (discussed earlier), to identify putative indels occurring in the human lineage since divergence with chimpanzee, we focused on alignment blocks of the autosomal genome containing human and chimpanzee plus a minimum of two additional primate sequences. We applied rigorous quality control criteria to reduce potential alignment artifact by excluding overlapping MAF blocks, overlapping and adjacent indel events (Kvikstad et al. 2007). This protocol (implemented in GALAXY; Blankenberg et al. 2007) resulted in 520,313 indels, of which 336,884 (65%) were unambiguously polarized as one single human-specific insertion/deletion.

Theoretical Predictions of the Impact of Polarization Errors on eDAF and rDI_e

We computed the impact of polarization error on the estimated ratio of deletion to insertion (rDI_e) as follows. Let rDI_t

denote the ratio of true deletion (D_t) to true insertion (I_t), and ε_{pol} the polarization error. One can compute the estimated number of deletions (D_e):

$$D_e = D_t(1-\varepsilon_{pol}) + I_t\varepsilon_{pol} \quad (1)$$

and likewise the estimated number of insertion (I_e):

$$I_e = I_t(1-\varepsilon_{pol}) + D_t\varepsilon_{pol} \quad (2)$$

From the above equations, the estimated rDI_e can be expressed as follows:

$$rDI_e = \frac{rDI_t(1-\varepsilon_{pol}) + \varepsilon_{pol}}{1-\varepsilon_{pol}(1-rDI_t)}. \quad (3)$$

Similarly, one can derive the relationship between the estimated mean DAF, $\overline{eDAF}$, for deletions (insertions) and polarization error. For a Wright–Fischer population of n chromosomes, the expected number of mutations present in i copies and segregating at a given DAF, $f = i/n$, can be approximated from the Poisson random distribution by k/f, where k represents a constant (Fu 1995). Thus, for example, the true number of deletions segregating at frequency, $D_t(f)$, can be approximated by

$$D_t(f) = \frac{k}{f}, \quad (4)$$

and the true number of insertions can be formulated in terms of the number of true deletions and the rDI_t:

$$I_t(f) = \frac{D_t(f)}{rDI_t}. \quad (5)$$

It follows that at each frequency, the estimated number of deletions will be the sum of deletions inferred correctly at that frequency and insertions, misidentified due to the presence of polarization error, segregating at frequency $1-f$:

$$D_e(f) = D_t(f)(1-\varepsilon_{pol}) + I_t(1-f)\varepsilon_{pol}. \quad (6)$$

By substituting equations (4) and (5), we can reduce equation (6) to the following:

$$D_e(f) = \frac{k}{f}(1-\varepsilon_{pol}) + \frac{D_t(1-f)}{rDI_t}\varepsilon_{pol}, \quad (7)$$

and further algebraic reduction yields,

$$D_e(f) = k\left(\frac{1-\varepsilon_{pol}}{f} + \frac{\varepsilon_{pol}}{rDI_t(1-f)}\right). \quad (8)$$

Likewise, we can estimate the number of insertions as follows:

$$I_e(f) = I_t(f)(1-\varepsilon_{pol}) + D_t(1-f)\varepsilon_{pol}, \quad (9)$$

and simplify algebraically to

$$I_e(f) = k\left(\frac{1-\varepsilon_{pol}}{rDI_t f} + \frac{\varepsilon_{pol}}{1-f}\right). \quad (10)$$

The deletion (insertion) $\overline{eDAF}$ for various rDI_t can then be obtained by taking the sum over all frequencies of the product of the proportion of estimated deletions at that frequency and the DAF, for example,

$$\overline{eDAF}(\text{deletion}) = \sum_{f=1/n}^{(n-1)/n} \frac{D_e(f)f}{D_e}. \quad (11)$$

Indel Hotspot Contexts

Throughout the text, we define indel hotspots as repetitive and near-repetitive tracts for which the indel density exceeds the SNP density, determined as follows. Repetitive tracts were defined as DNA sequence segments that consist of at least two tandem direct repeats of any DNA segment (unit) of length 1–24 bp. The last segment does not need to be complete, so that the length of the repetitive tract can be any number of nucleotides, but at least twice the unit length. Indel hotspots include repetitive regions with a tract length exceeding a threshold arbitrarily chosen so that sites had an indel diversity at least equal to the diversity due to SNPs (having tract lengths of: mononucleotide runs ≥6 bp; dinucleotide runs ≥9 bp; trinucleotide runs ≥11 bp; tetranucleotide runs ≥13 bp; pentanucleotide runs ≥14 bp; hexanucleotide runs ≥16 bp; and runs of 18 bp or more for unit lengths 7–24). Sites of near-repetitive genome sequence, where local indel densities were predicted to exceed the SNP density, were determined by a probabilistic version of Streisinger's classic strand slippage model of local indel mutation rate (Streisinger et al. 1966; Levinson and Gutman 1987) that takes into consideration nontandemly arrayed direct repeats of up to 20 bp in length and interspersed at distances of up to 20 bp (see Montgomery et al. 2013 for details). NR sequences were defined by excluding such regions.

Genomic Annotations

ARs were defined as DNA elements, LTRs, LINEs, and SINEs ancestral to the human–macaque divergence (Kvikstad et al. 2007). Non-AR noncoding regions were defined by excluding Gencode v3b annotations of genes (Harrow et al. 2006), GERP++ annotations of evolutionary constrained regions (Davydov et al. 2010), and ARs. Introns and intergenic regions were similarly defined using Gencode v3b gene annotations.

Analysis of SNPs

Data sets of SNPs and their allele frequency in YRI, CEU, and JPTCHB populations were downloaded from the 1000 Genome Pilot Project (The 1000 Genomes Project Consortium 2010: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/pilot_data/paper_data_sets/a_map_of_human_variation/, last accessed October 24, 2013). To polarize SNPs, we used information about the ancestral allele included in VCF files, which derives from the four-way EPO alignments (for more details, see ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ancestral_alignments/README, last accessed October 24, 2013). We excluded polymorphic sites with more than two alleles (0.01% of all SNPs). We also excluded SNPs for which the indicated ancestral allele was

unknown, unsure (i.e., in lower case in the VCF file) or did not match any of the two alleles. These three categories correspond, respectively, to 4%, 9%, and 0.6% of the initial data set. Finally, we excluded SNPs in a CpG context (i.e., for which one of the two alleles was part of a CpG dinucleotide), to avoid problems of homoplasy at these hypermutable sites. The final data sets include, respectively, 4,992,265 SNPs (CEU), 6,913,465 SNPs (YRI), and 3,969,354 SNPs (JPTCHB).

Recombination Data

Local crossover rates were obtained from HapMap Phase 2 data (The International HapMap Consortium 2007) and calculated in 5-kb windows centered on each polymorphic site (indel or SNP).

Model of Indel Sequence Evolution

We considered a model of indel sequence evolution, according to which a sequence site can have two indel states, insertions and deletions denoted X and Y. The transition rates between these two states are denoted by u (rate of deletion) and v (rate of insertion), respectively. Given the equilibrium frequencies of $v/(u+v)$ and $u/(u+v)$ for X and Y, respectively, and the following instantaneous rate matrix Q:

$$Q = \begin{pmatrix} -u & u \\ v & -v \end{pmatrix}, \tag{12}$$

one can derive the transition probability matrix P

$$P = e^{Qt} = \frac{1}{u+v}\begin{pmatrix} v + ue^{-(u+v)t} & u(1-e^{-(u+v)t}) \\ v(1-e^{-(u+v)t}) & u + ve^{-(u+v)t} \end{pmatrix}, \tag{13}$$

which specifies the probabilities of observing any of the possible transitions among states at time t conditional on the probability of the state X or Y at time $t = 0$. Assuming that u and v are constant over time, the transition probability matrix P can be used to calculate the probabilities of every possible evolutionary scenario along each branch of the phylogenetic tree (see supplementary text [Supplementary Material online] for details).

The phylogeny and divergence times that we used as parameters of the model (supplementary fig. S11, Supplementary Material online) correspond to the subset of four species (human, chimpanzee, orangutan, and rhesus macaque, HCOR) that are present in the alignments used to detect the majority (88%) of indels. The transition rates u and v estimated by parsimony in the human branch for NR regions in HCOR blocks (708,052,484 bp) are 4.7×10^{-5} deletion/bp/My and 1.6×10^{-5} insertions/bp/My, respectively (given an estimated human–chimpanzee divergence time of 6 My; Locke et al. 2011).

Supplementary Material

Supplementary text and figures S1–S11 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Manolo Gouy for help with derivation of the indel two state model; Sylvain Mousset for insightful discussions; and Kateryna Makova and Gerton Lunter, the associate editor and an anonymous reviewer for comments and suggestions on the manuscript. This work was supported by the Agence Nationale de la Recherche grants ABS4NGS: ANR-11-BINF-0001-06 to L.D. and the European Molecular Biology Organization (EMBO) Long-Term Fellowship grant ALTF 354-2010 to E.M.K.

References

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

Ananda G, Chiaromonte F, Makova KD. 2011. A genome-wide view of mutation rate co-variation using multivariate analyses. *Genome Biol.* 12(3):R27.

Arndt PF, Hwa T. 2004. Regional and time-resolved mutation patterns of the human genome. *Bioinformatics* 20(10):1482–1485.

Belinky F, Cohen O, Huchon D. 2010. Large-scale parsimony analysis of metazoan indels in protein-coding genes. *Mol Biol Evol.* 27(2):441–451.

Bhangale TR, Rieder MJ, Livingston RJ, Nickerson DA. 2005. Comprehensive identification and characterization of diallelic insertion-deletion polymorphisms in 330 human candidate genes. *Hum Mol Genet.* 14(1):59–69.

Blankenberg D, Taylor J, Schenck I, et al. (13 co-authors). 2007. A framework for collaborative analysis of ENCODE data: making large-scale analyses biologist-friendly. *Genome Res.* 17(6):960–964.

Britten RJ. 2002. Divergence between samples of chimpanzee and human DNA sequences is 5%, counting indels. *Proc Natl Acad Sci U S A.* 99(21):133633–133635.

Britten RJ, Rowen L, Williams J, Cameron RA. 2003. Majority of divergence between closely related DNA samples is due to indels. *Proc Natl Acad Sci U S A.* 100(8):4661–4665.

Brunschwig H, Levi L, Ben-David E, Williams RW, Yakir B, Shifman S. 2012. Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics* 191(3):757–764.

Bustamante CD, Wakeley J, Sawyer S, Hartl DL. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159(4):1779–1788.

Carvalho AB, Clark AG. 1999. Intron size and natural selection. *Nature* 401(6751):344.

Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA. 2002. Selection for short introns in highly expressed genes. *Nat Genet.* 31(4):415–418.

Chen FC, Chen CJ, Li WH, Chuang TJ. 2007. Human-specific insertions and deletions inferred from mammalian genome sequences. *Genome Res.* 17(1):16–22.

Chen CH, Chuang TJ, Liao BY, Chen FC. 2009. Scanning for the signatures of positive selection for human-specific insertions and deletions. *Genome Biol Evol.* 1:415–419.

Chen J-Q, Wu Y, Yang H, Bergelson J, Kreitman M, Tian D. 2009. Variation in the ratio of nucleotide substitution and indel rates across genomes in mammals and bacteria. *Mol Biol Evol.* 26(7):1523–1531.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Chindelevitch L, Li Z, Blais E, Blanchette M. 2006. On the inference of parsimonious indel evolutionary scenarios. *J Bioinfm Comput Biol.* 4(3):721–744.

Clark T, Andrew T, Cooper G, Margulies E, Mullikin J, Balding D. 2007. Functional constraint and small insertions and deletions in the ENCODE regions of the human genome. *Genome Biol.* 8:R180.

Comeron JM, Kreitman M. 2000. The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. *Genetics* 156(3):1175–1190.
Davydov EV, Goode DL, Sirota M, Cooper GM, Sidow A, Batzoglou S. 2010. Identifying a high fraction of the human genome to be under selective constraint using GERP++. *PLoS Comput Biol.* 6(12):e1001025.
de la Chaux N, Messer PW, Arndt PF. 2007. DNA indels in coding regions reveal selective constraints on protein evolution in the human lineage. *BMC Evol Biol.* 7:191.
Diallo AB, Makarenkov V, Blanchette M. 2007. Exact and heuristic algorithms for the indel maximum likelihood problem. *J Comput Biol.* 14(4):446–461.
Duret L, Galtier N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet.* 10(1):285–311.
Duret L, Mouchiroud D, Gautier C. 1995. Statistical analysis of vertebrate sequences reveals that long genes are scarce in GC-rich isochores. *J Mol Evol.* 40:308–317.
Ellegren H. 2000. Heterogeneous mutation processes in human microsatellite DNA sequences. *Nat Genet.* 24(4):400–402.
Ellegren H. 2004. Microsatellites: simple sequences with complex evolution. *Nat Rev Genet.* 5(6):435–445.
Eyre-Walker A. 1998. Problems with parsimony in sequences of biased base composition. *J Mol Evol.* 47(6):686–690.
Fu YX. 1995. Statistical properties of segregating sites. *Theor Popul Biol.* 48(2):172–197.
Garcia-Diaz M, Kunkel TA. 2006. Mechanism of a genetic glissando: structural biology of indel mutations. *Trends Biochem Sci.* 31(4):206–214.
Gregory TR. 2005. Synergy between sequence and size in large-scale genomics. *Nat Rev Genet.* 6(9):699–708.
Gu X, Li W-H. 1995. The size distribution of insertions and deletions in human and rodent pseudogenes suggests the logarithmic gap penalty for sequence alignment. *J Mol Evol.* 40:464–473.
Hardison RC, Roskin KM, Yang S, et al. (18 co-authors). 2003. Covariation in frequencies of substitution, deletion, transposition, and recombination during eutherian evolution. *Genome Res.* 13(1):13–26.
Harrow J, Denoeud F, Frankish A, et al. (15 co-authors). 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* 7(1 Suppl):S4.1–9.
Hernandez RD, Williamson SH, Bustamante CD. 2007. Context dependence, ancestral misidentification, and spurious signatures of natural selection. *Mol Biol Evol.* 24(8):1792–1800.
Hickey G, Blanchette M. 2011. A probabilistic model for sequence alignment with context-sensitive indels. *J Comput Biol.* 18(11):1449–1464.
International HapMap Consortium. 2007. A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449:851–861.
Karolchik D, Kuhn R, Baertsch R, et al. (24 co-authors). 2008. The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Res.* 36: D773–D779.
Katzman S, Capra JA, Haussler D, Pollard KS. 2011. Ongoing GC-biased evolution is widespread in the human genome and enriched near recombination hot spots. *Genome Biol Evol.* 3:614–626.
Kelkar YD, Eckert KA, Chiaromonte F, Makova KD. 2011. A matter of life or death: how microsatellites emerge in and vanish from the human genome. *Genome Res.* 21:2038–2048.
Kelkar YD, Tyekucheva S, Chiaromonte F, Makova KD. 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. *Genome Res.* 18:30–38.
Kondrashov A, Rogozin I. 2004. Context of deletions and insertions in human coding sequences. *Hum Mut.* 23:177–185.
Kvikstad E, Tyekucheva S, Chiaromonte F, Makova KD. 2007. A macaque's-eye view of human insertions and deletions: differences in mechanisms. *PLoS Comput Biol.* 3(9):1772–1782.
Lamb BC. 1985. The effects of mispair and nonpair correction in hybrid DNA on base ratios (G + C content) and total amounts of DNA. *Mol Biol Evol.* 2(2):175–188.
Lander ES, Linton LM, Birren B, et al. (235 co-authors). 2001. Initial sequencing and analysis of the human genome. *Nature* 409(6822):860–921.
Leclercq S, Rivals E, Jarne P. 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: a comparative genomic approach. *Genome Biol Evol.* 2:325–335.
Leushkin EV, Bazykin GA. 2013. Short indels are subject to insertion-biased gene conversion. *Evolution* 67:2604–2613.
Levinson G, Gutman GA. 1987. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. *Mol Biol Evol.* 4(3):203–221.
Locke DP, Hillier LW, Warren WC, et al. (121 co-authors). 2011. Comparative and demographic analysis of orang-utan genomes. *Nature* 469(7331):529–533.
Lunter G. 2007. Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes. *Bioinformatics* 23:i289–i296.
Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol.* 2(1):e5.
Lynch M. 2010. Rate, molecular spectrum, and consequences of human mutation. *Proc Natl Acad Sci U S A.* 107(3):961–968.
Makova KD, Yang S, Chiaromonte F. 2004. Indels are male-biased too: a whole-genome analysis in rodents. *Genome Res.* 14:567–573.
McDonald J, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in *Drosophila*. *Nature* 351:652–654.
Messer PW, Arndt PF. 2007. The majority of recent short DNA insertions in the human genome are tandem duplications. *Mol Biol Evol.* 24: 1190–1197.
Montgomery SB, Goode DL, Kvikstad E, et al. (21 co-authors). 2013. The origin, evolution, and functional impact of short insertion-deletion variants identified in 179 human genomes. *Genome Res.* 23(5):749–761.
Mouchiroud D, D'Onofrio G, Aissani B, Macaya G, Gautier C, Bernardi G. 1991. The distribution of genes in the human genome. *Gene* 100: 181–187.
Nam K, Ellegren H. 2012. Recombination drives vertebrate genome contraction. *PLoS Genet.* 8(5):e1002680.
Ometto L, Stephan W, De Lorenzo D. 2005. Insertion/deletion and nucleotide polymorphism data reveal constraints in *Drosophila melanogaster* introns and intergenic regions. *Genetics* 169(3):1521–1527.
Parsch J. 2003. Selective constraints on intron evolution in *Drosophila*. *Genetics* 165(4):1843–1851.
Petrov DA. 2001. Evolution of genome size: new approaches to an old problem. *Trends Genet.* 17(1):23–28.
Petrov DA. 2002. Mutational equilibrium model of genome size evolution. *Theor Popul Biol.* 61(4):531–544.
Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL. 2000. Evidence for DNA loss as a determinant of genome size. *Science* 287(5455):1060–1062.
Pettersson M, Kurland CG, Berg O. 2009. Deletion rate evolution and its effect on genome size and coding density. *Mol Biol Evol.* 26(6):1421–1430.
Podlaha O, Webb DM, Tucker PK, Zhang J. 2005. Positive selection for indel substitutions in the rodent sperm protein catsper 1. *Mol Biol Evol.* 22(9):1845–1852.
Presgraves DC. 2006. Intron length evolution in *Drosophila*. *Mol Biol Evol.* 23(11):2203–2213.
Sjodin P, Bataillon T, Schierup MH. 2010. Insertion and deletion processes in recent human history. *PLoS One* 5(1):e8650.
Smagulova F, Gregoretti IV, Brick K, Khil P, Camerini-Otero RD, Petukhova GV. 2011. Genome-wide analysis reveals novel molecular features of mouse recombination hotspots. *Nature* 472(7343):375–378.
Smit AFA, Hubley R, Green P. 1996–2010. RepeatMasker. open-3.1.2 ed. [cited 2013 Oct 24]. Available from: http://www.repeatmasker.org.
Spencer CC. 2006. Human polymorphism around recombination hotspots. *Biochem Soc Trans.* 34(Pt 4):535–536.
Streisinger G, Okada Y, Emrich J, Newton J, Tsugita A, Terzaghi E, Inouye M. 1966. Frameshift mutations and the genetic code. *Cold Spring Harb Symp Quant Biol.* 31:77–84.

Tanay A, Siggia ED. 2008. Sequence context affects the rate of short insertions and deletions in flies and primates. *Genome Biol.* 9(2):R37.

Taylor MS, Ponting CP, Copley RR. 2004. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res.* 14(4):555–566.

Waterston RH, Lindblad-Toh K, Birney E, et al. (222 co-authors). 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420(6915):520–562.

Webster MT, Smith NG, Ellegren H. 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. *Proc Natl Acad Sci U S A.* 99(13):8748–8753.

Wetterbom A, Sevov M, Cavelier L, Bergstrom TF. 2006. Comparative genomic analysis of human and chimpanzee indicates a key role for indels in primate evolution. *J Mol Evol.* 63(5):682–690.

Zhu Y, Strassmann JE, Queller DC. 2000. Insertions, substitutions, and the origin of microsatellites. *Genet Res.* 76(3):227–236.