

Probabilistic whole-genome alignments reveal high indel rates in the human and mouse genomes

Gerton Lunter

MRC Functional Genetics Unit, Department of Physiology, Anatomy and Genetics, University of Oxford, South Parks Road, Oxford OX1 3QX, United Kingdom

ABSTRACT

Motivation: The two mutation processes that have the largest impact on genome evolution at small scales are substitutions, and sequence insertions and deletions (indels). While the former have been studied extensively, indels have received less attention, and in particular, the problem of inferring indel rates between pairs of divergent sequence remains unsolved. Here, I describe a novel and accurate method for estimating neutral indel rates between divergent pairs of genomes.

Results: Simulations suggest that new method for estimating indel rates is accurate to within 2%, at divergences corresponding to that of human and mouse. Applying the method to these species, I show that indel rates are up to twice higher than is apparent from alignments, and depend strongly on the local G + C content. These results indicate that at these evolutionary distances, the contribution of indels to sequence divergence is much larger than hitherto appreciated. In particular, the ratio of substitution to indel rates between human and mouse appears to be around $\gamma=8$, rather than the currently accepted value of about $\gamma=14$.

Contact: Gerton.lunter@dpag.ox.ac.uk

1 INTRODUCTION

The study of sequence evolution involves the comparison of homologous characters, and the estimation of their rate of change. A complicating factor in the case of nucleotides and other discrete character is the possibility of back- or repeat-mutation, which causes parsimony methods to underestimate the rate of change. In contrast, probabilistic methods consider all possible sequences of events and weigh them appropriately, and are thus able to give unbiased estimates if sufficient data are available.

In mammals, the rate of indel events is roughly an order of magnitude lower than the rate of substitutions, thus reducing the possibility of ‘collisions’. Moreover, exact back mutations are highly unlikely. These observations would suggest that the parsimony estimate for the rate of indel events (i.e. the density of gaps in an alignment) might be quite accurate. However, this does not appear to be the case. The reason is that alignments are often incorrect, owing to the simultaneous action of substitutions and indels which obscure the homology, and to gap interactions, as explained subsequently.

The alignment problem is often phrased as an optimization problem: find the alignment that maximizes (or minimizes) some score function. It can be shown that this is equivalent, in a probabilistic setting, to finding the maximum likelihood

(ML) placement of gaps (Thorne *et al.*, 1991), a problem that is efficiently solved by the Viterbi algorithm (Durbin *et al.*, 1998). Although this framework allows for the correct treatment of substitutions by using appropriate score matrices, it is essentially a (weighted) parsimony method with regards to gaps. Consequently, as we show elsewhere (Lunter *et al.*, 2007), all alignment methods suffer from various types of bias. These biases consist of certain systematic changes in the placement of gaps in inferred alignments, relative to the gaps’ true positions, and arise through the alignment procedure systematically favouring likely (or higher-scoring) configurations over less likely ones, even though less likely configurations do occur. Two biases in particular, termed ‘gap attraction’ and ‘gap annihilation’ (Fig. 1), each reduce the number of inferred gaps, causing the alignment gap density $\hat{\delta}_{ALIGN}$ to underestimate the true gap density.

This problem can be partially overcome by using probabilistic models. Besides identifying the ML alignment, probabilistic models can also be used to consider all possible alignments, weighted by their posterior probability (Zuker, 1991). This combined information can be used to construct a ‘posterior decoding’ alignment (Durbin *et al.*, 1998; Krogh, 1997), which is more accurate and less affected by alignment biases (Lunter *et al.*, 2007). However, a significant amount of bias remains even in posterior decoding alignments.

When no selection acts on a sequence, indels are expected to be distributed across the genome according to a Poisson process. Previously, we used the observed excess of long ungapped sequences over predictions by this ‘neutral indel model’ (NIM) to distinguish between neutrally evolved and constrained sequence, based on indels alone (Lunter *et al.*, 2006). Gap attraction (Fig. 1A) causes a deviation from the NIM at the other end of the scale, namely, a shortage of gaps close to one another. At intermediate separations, the predictions by the NIM are unaffected by either effect, allowing for an estimate of the gap density $\hat{\delta}_{NIM}$ that is biased neither by gap attraction, nor by the potential admixture of constrained sequence, which otherwise would tend to decrease the neutral indel rate estimates.

Nevertheless, a second bias still affects this estimate. Gap annihilation describes an effect, where pairs of neighbouring and identically-sized gaps (of opposing signature) tend to cancel each other, resulting in contiguous regions of non-homologous pairings within longer ungapped alignment segments (Fig. 1B). While probabilistic aligners are not insensitive to this effect, they can at least detect its occurrence, because alternative potential alignments leave a fingerprint in

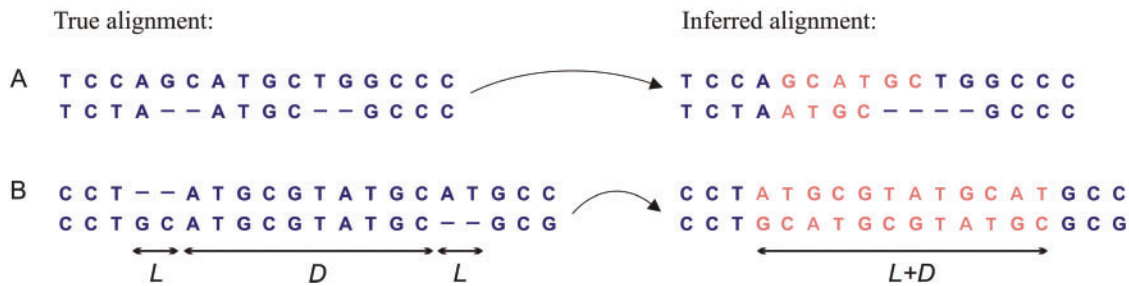


Fig. 1. Examples of gap attraction and gap annihilation in alignments. Two gaps in close proximity in the true alignment (A, left) tend to coalesce in the inferred alignment (A, right), when this represents a more likely evolutionary history despite an apparent increase in divergence. Similarly, two gaps of equal size L but opposite signature, located at a distance D from each other (B, left), will tend to coalesce and cancel in the inferred alignment, resulting in $L + D$ non-homologous aligned nucleotides.

the form of a reduction of the posterior probability assigned to the preferred alignment. This posterior turns out to be a good estimate of the average accuracy of the alignment. In this way, the average posterior quantifies the impact of gap annihilation on alignments, a fact that can be exploited to account for its effect on the gap density estimate.

Here, I describe a novel estimator for the indel rate between pairs of divergent sequences, which uses both the NIM and probabilistic sequence alignments to explicitly account for the biases introduced by both gap attraction and gap annihilation. Simulations show the method to be able to accurately measure indel rates at human–mouse divergence. Applying it to human–mouse alignments, I find higher indel rates than previously reported. These rates depend strongly on G + C content, with high rates both for A + T-rich, and especially for G + C-rich sequence. Indel rates also vary over chromosomes, with the small autosomes experiencing higher rates, and chromosome X experiencing smaller rates. These results imply that alignments are often wrong, particularly for high G + C sequence where functional material is concentrated.

2 METHODS

2.1 Definitions

The symbol δ denotes the *gap density*, the number of gaps in an alignment as a proportion of the number of aligned nucleotides. The *true* gap density refers to the gap density in the true alignment. The term *indel rate* refers to the probability, per unit of time and per nucleotide, of the occurrence of an indel event. This article deals with the human and mouse species only, whose divergence is considered to be one time unit, so that the indel rate is equal to the expected number of indel events per nucleotide. The symbol σ denotes the *substitution rate* (the expected number of substitutions per site and per unit of time). The ratio of substitution rate to indel rate is denoted by γ .

2.2 Marginalized posterior decoding

Whole-genome alignments were computed, using BlastZ human–mouse alignments as a guide to large-scale homology. BlastZ-derived and new guides were covered by tiles measuring 500 bp on each side. Within each tile, a local alignment was computed, built around a five-state HMM, surrounded by 2×2 flanking ‘gap’ states, effectively allowing gaps of arbitrary size at the beginning and end of an alignment. The five core states of the HMM form a generalization of the standard three-state (match, insert, delete) alignment HMM, where the insert and delete

states were doubled to model the geometric mixture distribution of indel lengths. The standard Forward and Backward algorithms were used to compute posteriors for each state at every position. The posteriors referring to insertion states (two flanking states and two core states) were added together and accumulated along the non-emitting coordinate, so that the result refers to the probability of the emitted nucleotide not being aligned, rather than being part of a particular gap [‘Marginalized posterior decoding’, (Lunter *et al.*, 2007)]. A similar procedure was applied to deletion states. Dynamic programming was used to compute the maximum product-of-posteriors across all possible alignments, and a standard traceback algorithm recovered this alignment. New guides were generated if an alignment extended to within 100 bp of a tile’s edge. To ensure that spurious guides would not lead to erroneous contributions, permutation tests were performed to establish significant sequence identity as measured by the Forward algorithm. The resulting alignments were stitched together, and local inconsistencies were resolved by preferring the columns with the largest posterior. The software was written in C++, and the HMM algorithms were generated by a code generator (HMMoC, manuscript in preparation) written in Java.

2.3 Simulating human–mouse sequence

A simulated data set of human–mouse-like sequence was generated generating an ‘ancestral’ sequence from mouse-aligning human sequence, ensuring that various sequence characteristics will resemble the true sequence. This ancestral sequence was then evolved twice, for 0.5 units of time. Indels were generated using G + C-dependent rates, drawing indel sizes from the empirical distribution obtained from G + C-matched BlastZ alignments, and substitutions were added following the Kimura 2-parameter model with transition/transversion ratio of 2 and total substitution rate of 0.436. Inserted sequence was drawn from sequence drawn from gapped BlastZ sequence from the appropriate G + C category. Ancestral sequences were 1000 bp long, and non-homologous 100 nucleotide flanking sequences were added to each end of both descendants to mimic the situation in whole-genome alignments. The simulation software was written in Python.

2.4 Accounting for gap attraction

For small separations, histogram counts of inter-gap lengths are negatively biased because of gap attraction. For larger separations, these counts are positively skewed by an increasingly large proportion of long ungapped segments due to purifying selection. The intermediate region largely follows the neutral indel model. We obtained the slope in this regime by weighted linear regression to the log counts, using as weights the sampling error for each count as determined from

a Bernoulli model. The intermediate region was found by exhaustive search to maximize the coefficient of determination, R^2 , constrained by a minimum region size of 20. The slope parameter s is converted to the estimated gap density via the formula $\hat{\delta}_{NIM} = 1 - e^s$ to account for the transformation to a logarithmic scale.

3 RESULTS

3.1 Accounting for gap attraction: the neutral indel model

Under the twin assumptions of uniform indel rates and the absence of selection, indels are expected to be uniformly distributed across the genome (Lunter *et al.*, 2006). This observation implies that after the sequence has evolved for some time, the distance between successive indels along the genome shows a geometric distribution: the probability δ of any nucleotide to be affected by an indel, conditional on its left (or right) neighbour having survived, is constant, and the probability of observing no indels across L contiguous sites is just $(1 - \delta)^L$, predicting a geometric distribution of the distance between neighbouring indel events. Note that this argument does not imply that the fates of neighbouring nucleotides are independent (which, indeed, is not true when indels affect multiple nucleotides at once), since δ is a *conditional* probability.

To see how the prediction that inter-indel distances are geometrically distributed carries over to alignment gaps, the effects of gap attraction and gap annihilation must be considered. Gap attraction refers to the tendency of neighbouring gap pairs to coalesce, reducing the gap count by one, at the cost of an increased apparent divergence (and reduced accuracy) near gaps (Fig. 1A). Compared to the predicted geometric distribution, this causes a reduction of the number of closely-spaced gap pairs.

Gap annihilation refers to the tendency of two neighbouring gaps of equal size but opposite signature (e.g. an insertion and a deletion in the same lineage, or two insertions in different lineages) to coalesce, leaving no gap (Fig. 1B). This systematic removal of a proportion of gap pairs results in a reduction of the effective δ parameter, which governs the slope of the geometric distribution. In addition, gap attraction causes another small decline of the number of closely-spaced gap pairs. This additional reduction occurs because, conditional on observing a gap (which thus has escaped gap annihilation), the probability of finding an identically-sized gap of opposite signature in its vicinity is reduced. As this effect is conditional on the gap size and signature, it is much weaker than the corresponding effect due to gap attraction. However, since gap annihilation removes two rather than one gap, it has a longer range.

A histogram of inter-gap distances in human–mouse BlastZ alignments (Blanchette *et al.*, 2004; Schwartz *et al.*, 2003) is shown in Figure 2. In logarithmic coordinates, a geometric distribution traces a straight line, and a very good fit is apparent for intermediate inter-gap distances (~ 20 –50 nt). The predicted shortage for smaller distances is also apparent, while the excess of long inter-gap distances is due to an admixture of constrained sequence (Lunter *et al.*, 2006).

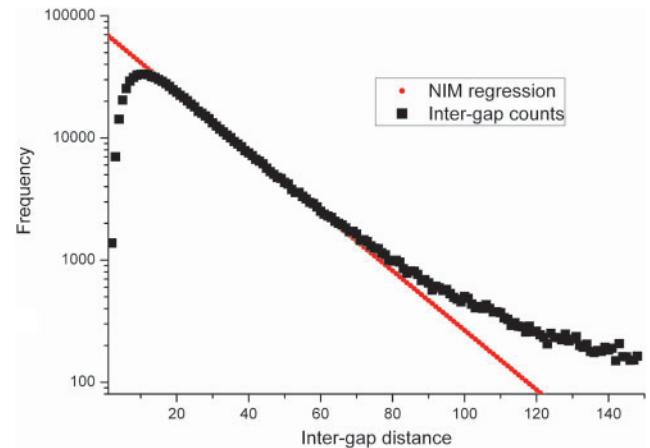


Fig. 2. The distribution of inter-gap distances reveals the effects of gap attraction. Shown is the histogram of inter-gap distances for BlastZ-aligned human–mouse data. Under neutral evolution and uniform indel rates, the perfect alignment would give rise to a straight line (neutral indel model [NIM] regression line, red) in the log–linear coordinates used here. The departure at short distances (<20 bp) is caused by gap attraction, an alignment artefact. The departure at longer inter-gap distances is caused by purifying selection on functional sequence.

By identifying the largest range of inter-gap distances for which a geometric accurately predicts the histogram counts, and using linear regression to estimate the slope parameter, an estimate $\hat{\delta}_{NIM}$ is obtained that is unaffected by both gap attraction and admixture of constrained sequence, and is only partially affected by gap annihilation.

3.2 Accounting for gap annihilation: posteriors

Gap annihilation occurs when the proposed alternative alignment, having two fewer gaps, represents a more likely evolutionary scenario despite an increased number of inferred substitutions along the resulting nonhomologous aligned segment. Probabilistic aligners can consider all possible alignments simultaneously, and identify any uncertainty by competing alignments through a reduction in the posterior probability assigned to the inferred alignment. A simulation experiment shows that the average accuracy of alignments (defined as the proportion of nucleotides that get aligned correctly) is closely approximated by the average posterior probability (Fig. 3). While the posterior is of limited help to distinguish the correct alignment from among the many possibilities in individual cases, it does quantify the extent to which gap annihilation affects the alignments.

Here, I derive an improved estimator $\hat{\delta}_{PROB}$ that corrects for gap annihilation using the average posterior. To derive this estimator, I need to make three assumptions. First I assume that the asymptotic accuracy of alignments away from gaps, α , is known. Second, the reduction of α below unity is assumed to be due to gap annihilation only, so that higher-order interactions between indels may be ignored. Third, I assume *small gaps*, more precisely, that the average gap size is smaller than the average distance between gaps.

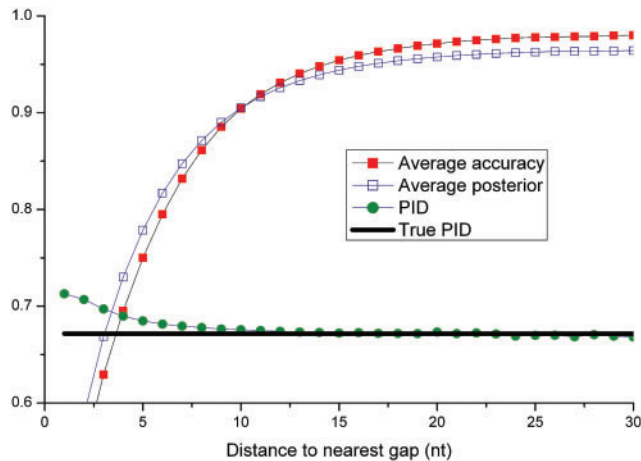


Fig. 3. Alignment accuracy is well predicted by the posterior. Shown are the average accuracy (filled squares) and average posterior (open squares) by the distance from the nearest gap, computed from simulated and realigned sequence at human–mouse divergence. For reference, the true and aligned proportion sequence identity (PID) is also shown. While the accuracy cannot be computed for real sequence, the posterior can, and approximates the accuracy reasonably well. The accuracy increases with increasing distance to gaps, but does not tend to one because of the effects of gap annihilation. At distances above 25 nucleotides from the nearest gap the curve plateaus, and the average posterior at this distance was used as a proxy for the asymptotic accuracy.

Suppose that in an alignment of a pair of sequences containing H homologous nucleotide pairs, and on average δ^{-1} homologous nucleotides between true gaps, a proportion p of gaps are affected by gap annihilation. Two annihilating gaps of (equal) size L at distance D from each other cause $L + D$ nucleotides to become wrongly aligned (Fig. 1B). Gap annihilation tends to affect gaps near to each other, so $D < \delta^{-1}$ in expectation, and from the small gap assumption it follows that $L < \delta^{-1}$. The number of affected gap pairs in the sequence is $\frac{1}{2}pH\delta$. Invoking the second assumption, the proportion of affected nucleotides can be calculated as $1 - \alpha = \frac{1}{2}pH\delta(L + D)/H < p$. The true and observed gap densities (corrected for gap attraction) are related by $\hat{\delta}_{NIM} = (1 - p)\delta$, and it follows that $\delta = \hat{\delta}_{NIM}/(1 - p) > \hat{\delta}_{NIM}/\alpha$. In conclusion, $\hat{\delta}_{PROB} := \hat{\delta}_{NIM}/\alpha$ is an improved and still conservative estimate of the true gap density.

In practice, α may be approximated by the average posterior probability for columns at a reasonably large distance from gaps. From Figure 3 it appears that 25 nucleotides is a reasonable choice (requiring an aligned segment of at least 51 nt long). This choice balances the influence of gap attraction with the reduced number of data points further away from gaps.

The effectiveness of the combined correction for gap attraction and gap annihilation can be shown by simulation. As a first experiment, sequences were generated under a Jukes–Cantor model with varying divergence (up to $\sigma = 0.9$ substitutions per site), and with an indel rate chosen such that the gap density remained at a constant ratio of $\gamma = 40/3 \approx 13$ (Silva and Kondrashov, 2002) to the substitution rate (Fig. 4). The alignment gap density $\hat{\delta}_{ALIGN}$ shows an increasingly large

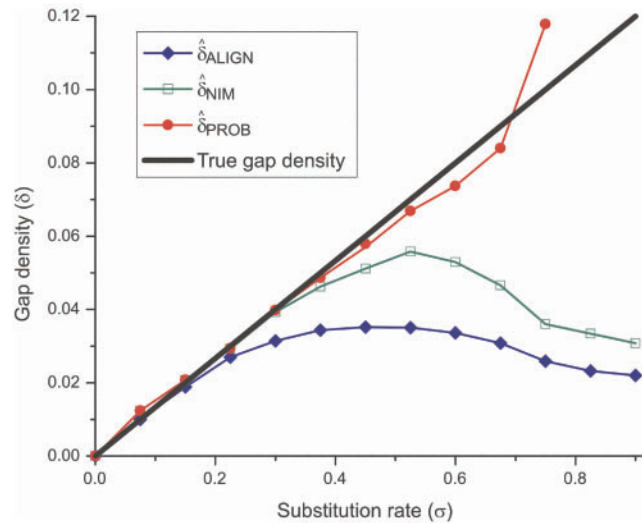


Fig. 4. Estimated gap densities in simulated sequence at various divergences. Shown are: $\hat{\delta}_{ALIGN}$, the density of gaps in the alignment; $\hat{\delta}_{NIM}$, the estimate corrected for gap attraction using the inter-gap histogram slope; and $\hat{\delta}_{PROB}$, the same estimate additionally corrected for gap annihilation. The black diagonal line shows the true gap density as used in the simulations.

deviation starting at $\sigma = 0.2$, and plateaus around $\sigma = 0.5$, after which the gap density *decreases* with increasing divergence. Correcting for gap attraction ($\hat{\delta}_{NIM}$) accounts for virtually all bias up to $\sigma = 0.4$, but shows a large increase of the bias beyond $\sigma = 0.5$. Correcting for both gap attraction and gap annihilation yields $\hat{\delta}_{PROB}$, which is virtually unbiased up to about $\sigma = 0.7$, after which this estimator too breaks down.

Since gap attraction and gap annihilation are both second-order interaction effects, it may be expected that they contribute proportionally across the divergence range. However, from Figure 4 it appears that gap annihilation becomes important at larger divergences than gap attraction does. A possible cause is the longer range of gap annihilation, causing higher-order interaction effects to contribute more strongly than for gap attraction.

3.3 Probabilistic realignment

To measure indel rates using the proposed method, alignments must be annotated with their posterior column probabilities. These were generated for alignments of the human and mouse genomes using a new whole-genome realignment algorithm. The algorithm takes existing whole-genome alignments as guides to large-scale homologies, and uses banded local probabilistic alignments on an initial set of tiles covering the guide alignment, together with an iterative ‘tile-and-stitch’ procedure, to compute a new alignment. A manuscript describing this alignment algorithm is in preparation.

The algorithm combines several known and some novel features. It uses local alignments to ensure that no spurious non-homologous flanking sequence is included. As an additional precaution, a permutation test is used to determine whether the sequences covered by a tile show significant

similarity, providing robustness to erroneous guides. The algorithm computes posterior probabilities for all possible alignment columns covered by a tile. Besides providing an annotation, these posteriors are used to determine the alignment by maximizing the product of posteriors of contributing columns. Elsewhere, we have shown that this algorithm performs better than standard Viterbi decoding (Lunter *et al.*, 2007). The algorithm was parameterized by a Baum–Welch training stage, and all parameters (in particular, those parameterizing the gap density) were made dependent on the local G + C content. Finally, the indel length spectrum, which is characterized by an abundance of small indels but a relatively heavy tail, is modelled using a geometric mixture distribution. In a simulation experiment, we elsewhere show that these features improve the sensitivity of alignments for sequence at roughly human–mouse divergence from 84% to 88%, while not increasing the false-positive fraction (Lunter *et al.*, 2007). To large extent, this improvement is due to the posterior decoding algorithm, which reduced the impact of both gap attraction and gap annihilation.

To generate the alignments and posteriors, I used existing BlastZ human–mouse alignments as a guide (Schwartz *et al.*, 2003). Parameters were estimated on human chromosome 10 data, using one iteration of the Baum–Welch algorithm (Durbin *et al.*, 1998). Using these parameters, the human genome was realigned to mouse on a 72-processor Linux cluster in 48 h.

The realigned data showed slightly lower divergence in neutrally evolving regions than the corresponding BlastZ alignments (realigned PID 67.4%; BlastZ PID 65.9%, corresponding to substitution rates $\sigma = 0.436$ and $\sigma = 0.456$ for a Kimura 2-parameter model with a transition/transversion ratio of 2.0.). This difference is most likely due to improvements in the alignment, specifically, the insertion of more gaps. While this alone would increase the PID, simulations we performed previously (data now shown) give reason to suppose the difference is at least partly due to the improved quality of the alignments.

3.4 Estimates of the human–mouse indel rate

To estimate the human–mouse indel rate, the genome was divided into 20 equally-sized fractions depending on the local (250 bp) G + C content. Distinct alignment parameters were used for each of these fractions, and after calculating the whole-genome alignment, inter-gap distance histograms were collected for each fraction separately. The three gap density estimates $\hat{\delta}_{ALIGN}$, $\hat{\delta}_{NIM}$ and $\hat{\delta}_{PROB}$ are shown in Figure 5A.

Note that these are estimates for the average density of gaps in alignments, rather than the rate of indel events per unit of time. To convert the final gap density estimate into an indel rate, the evolution of sequences under a particular indel rate was simulated (starting from an initial estimate), using the empirical indel size spectrum obtained from BlastZ-aligned sequence in the same G + C-band. After which the gap density was measured on the simulated true alignment, and compared to the estimated value. This procedure was iterated until the appropriate gap density was obtained. The resulting correction is small, with the indel rate being 2–3% higher than the

corresponding gap density in the range of divergences considered. This can be understood by considering that, on the one hand, the number of gaps is less than the number of indels because of overlapping indels merging into single gaps in the final, true alignment. On the other hand, the indel rate is measured relative to all (extant) nucleotides, while gap densities are measured relative to the smaller fraction of aligning nucleotides. These effects oppose each other, and apparently nearly cancel.

The estimated indel rate in autosomes varies from 0.055 for the AT-richest 5% of the genome, via around 0.0505 for intermediate G + C contents, to a peak of 0.067 indels per site for sequence with G + C content above 57%. Indel rates also vary per chromosome, with the smaller chromosomes exhibiting higher rates, especially towards high G + C sequence [data not shown, but see (Lunter *et al.*, 2006)]. For example, chromosome 1 has experienced a rate of 0.055, 0.049 and 0.065 indels per site for low, intermediate and high G + C respectively, while for chromosome 21 the corresponding numbers are 0.055, 0.050 and 0.075. Chromosome X is an outlier, with much reduced rates (0.047, 0.041 and 0.054 for the three G + C bands). The average rate on chromosome Y was 0.054; insufficient data were available to make G + C-specific estimates.

3.5 Simulations

The second set of simulation experiments served to further justify the proposed measure for indel rates, and in particular, to justify the rates measured on human–mouse sequence. The simulated sequences were generated to closely resemble actual human and mouse-aligning sequence, with respect to their sequence content, gap length distribution, substitution rate and indel rate. This was achieved by evolving actual human sequence, so that features such as dependencies between sites and tandem repeat contents are approximately matched to the real sequence. The simulations also accounted for the variation of each of these aspects with G + C content. Segmental duplications (Cheng *et al.*, 2005) and other large-scale events were not simulated, since these mutations are rare and have a negligible effect on the overall indel rate.

The generated sequences were probabilistically realigned, using the same parameters as for the real data (rather than the simulation parameters). Indel rates were computed from the resulting alignment as before (Fig. 5C). The original indel rate is recovered almost exactly across the range of G + C values, with a RMS deviation from the true rate of 1.6%, and no apparent systematic bias.

To determine whether either the alignment parameters or the G + C content itself has a large contribution to the inferred indel rates, I generated a second set of simulated sequence, with the same sequence content but a G + C-independent indel rate of 0.0505 indels per site. These sequences were realigned using the same alignment parameters as before, which presumably would bias towards high indel rates for the extreme G + C contents. The results (Fig. 5D) suggest that the inferred indel rates may indeed be slightly biased by the alignment parameters, but this bias is small compared to the variation in the parameters used by the alignment algorithm. This accords with the observation that alignments show good robustness

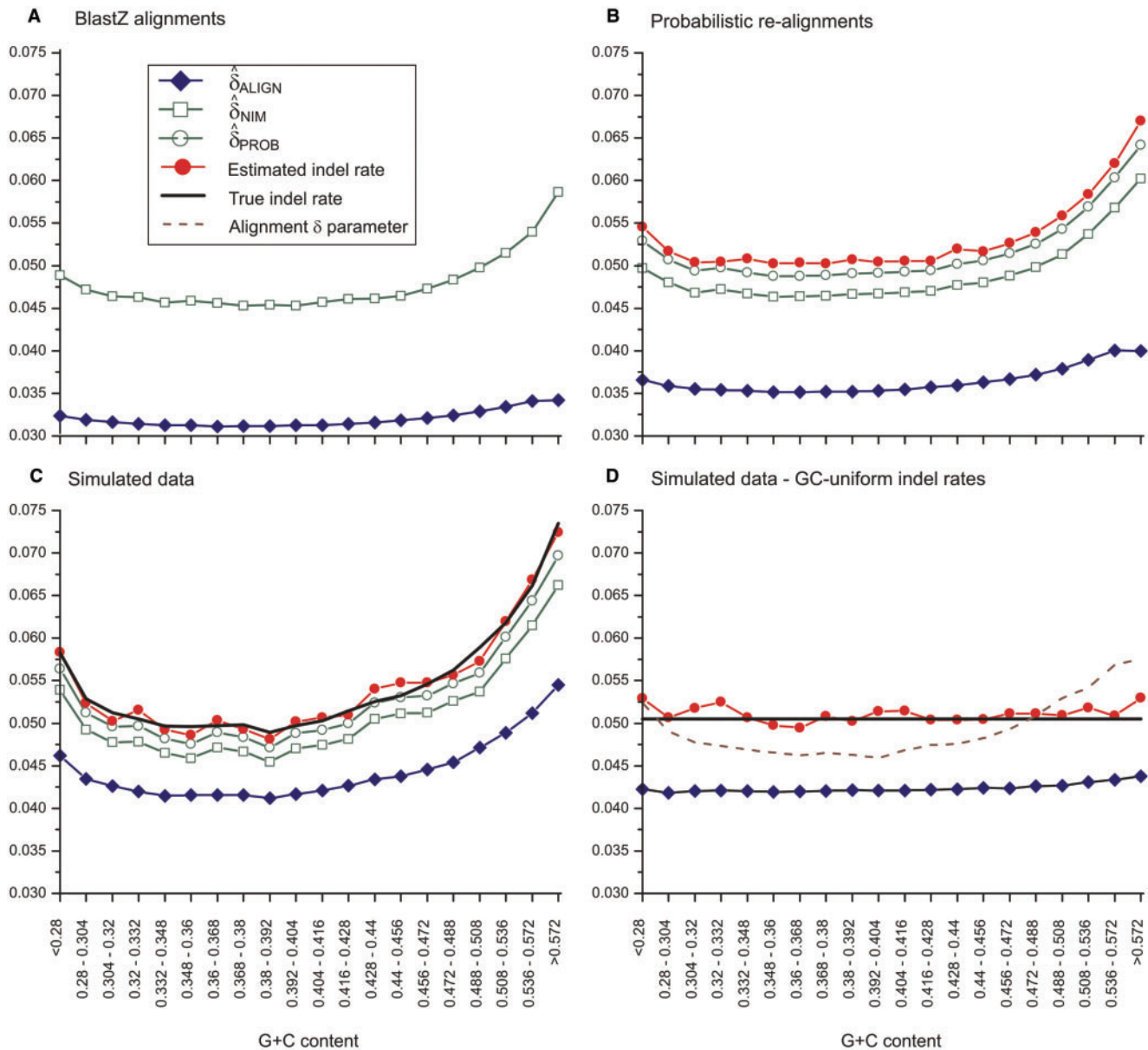


Fig. 5. Estimates of indel rates and gap densities. Shown are: (A) gap density estimates from BlastZ alignments; (B) gap density and indel rate estimates from probabilistic realigned sequence; (C) estimates from simulated data using the indel rate estimates from realigned human–mouse data; (D) estimates from simulated data using uniform indel rate across the G + C spectrum (G + C fractions on the X-axes). The four panels show a subset of the following measures: $\hat{\delta}_{ALIGN}$, gap density on the alignment (blue diamonds); $\hat{\delta}_{NIM}$, gap density estimate corrected for gap attraction (green open squares); $\hat{\delta}_{PROB}$, gap density estimate corrected for both gap attraction and gap annihilation (green open circles); the estimated indel rate (red filled circles). For simulated data, the true indel rate is shown in black. For comparison, panel (D) also shows the gap density parameters used by the alignment algorithm (dashed curve).

against modest deviations of the evolutionary parameters from their true values (Lunter *et al.*, 2007).

The gap densities within alignments are substantially higher in simulated sequence than in alignments of real human and mouse data (Fig. 5A and C). This difference is offset by a lower asymptotic posterior probability within human–mouse alignments (91.2–93.6% for real sequence, 96.4–95.0% for

simulated sequence), suggesting that real sequence is more difficult to align than simulated sequence. It seems possible that this effect is caused by correlations between inserted (or deleted) sequence with neighbouring sequence, itself resulting from duplications and deletions through polymerase slippage. Such possible correlations were not modelled in our simulations, or in the alignment procedure.

4 DISCUSSION

This article introduces a new method for measuring indel rates between pairs of divergent genomes. Simulations suggest that at the level of divergence observed between human and mouse, the method could be accurate to within 2%.

The indel rate measurements on human and mouse form the main findings of this paper: the average indel rate (autosomal average, 0.053 indels per site) is much higher than is apparent from alignments, and indel rates vary strongly with local G + C content and between chromosomes. The inferred average autosomal rate compares reasonably well with an earlier estimate of 0.056 events per site based on an analysis of intron size evolution, using gene orthology assignments instead of alignments (Ogurtsov *et al.*, 2004).

Posterior alignment probabilities were required to infer these rates, and for this reason the human and mouse genomes were realigned. Earlier work suggests that these alignments are an improvement upon existing whole-genome alignments, and show a reduction of alignment biases. While these reduced biases do result in an increased alignment gap density $\hat{\delta}_{ALIGN}$, the high indel rates we report are not only, or even mainly, the result of improved alignments, since the gap density estimates based on BlastZ and probabilistic alignments are nearly identical after accounting for gap attraction (cf. $\hat{\delta}_{NIM}$ in Fig. 5A and B). The new alignments exhibit a modestly reduced substitution rate compared to the rate inferred from existing alignments. Together with the increased indel rates, these results imply that the ratio of substitution rate to indel rate between human and mouse, γ , is about 8 (0.436/0.053), much lower than the value $\gamma = 14$ (0.456/0.032) which would be inferred from existing alignments directly (Britten *et al.*, 2003; Ogurtsov *et al.*, 2004; Silva and Kondrashov, 2002).

The discrepancy between the new indel rate measurements and the density of gaps in existing alignments varies, from 60% (0.0311 observed gaps per aligned nucleotide, versus 0.0505 inferred indels per site) for intermediate G + C values, to almost a factor 2 for the 5% G + C-richest fraction of the genome (0.0343 observed versus 0.0670 inferred). The indel rate variation with local G + C content is striking, with sequence in the highest G + C category experiencing 33% more indels than those of intermediate G + C (0.0670/0.0505). Interestingly, also AT-rich sequence appears to be preferentially targeted by indels (8% increase compared to intermediate G + C).

The strong dependence of indel rates with G + C begs a question: does G + C content drive indel rates, or do they co-vary through a third causal variable? One process that could explain a direct link is polymerase slippage. This process is thought to be induced by local sequence similarity (e.g. tandem repeats), and sequences with extreme high or low G + C content is statistically more likely to contain repetitive structures through the effective reduction of the alphabet (Lunter *et al.*, 2006). However, this argument is less well able to explain the pronounced increase of indel rates for high G + C, which in mammals corresponds to ~57%, close to an even distribution of nucleotides. Therefore, other factors must influence the indel rates at the high end of the G + C content spectrum.

Several mutational processes are strongly associated with high G + C, including copy number variations (CNVs)

(Nguyen *et al.*, personal communication). Could it be that the mechanism causing CNVs may, at smaller scales, also be responsible for small indels? Alternatively, it has been suggested that, as in the case for yeast (Infante *et al.*, 2003), CNVs might be driven by nonhomologous recombination, which is known to associate with high G + C. It has been argued that mammalian recombination is mutagenic, although the issue is still under debate (Hellmann *et al.*, 2003; Spencer *et al.*, 2006). If recombination induces small indels, this would explain the strong association of indel rates with high G + C.

Could transposable elements (TEs) account for the observed high indel rates, and in particular their G + C-dependence? TEs have marked G + C preferences, and have been very active in both the mouse and human lineages (Lander *et al.*, 2001). However, of the 34.6 million gaps in the human–mouse BlastZ alignment (hg18/mm8), only about 5.6% (1.94 million) are TE-associated (loosely defined as being over 100 nt long, and appearing opposite TE-annotated sequence for at least 75% of its length). Moreover, the proportion of TE-associated gaps does not peak in the highest (and lowest) GC-category, but rather around 40–46% G + C, contrasting with the U-shaped curve of the total indel rate. TEs thus appear to have a limited contribution to the overall indel rate.

Chromosome size also appears to explain some of the indel rate variation, with smaller chromosomes showing markedly increased indel rates [particularly in the higher G + C bands, (Lunter *et al.*, 2006)], possibly related to higher indel rates in subtelomeres which are proportionally larger in small chromosomes, or a larger per-site recombination rate due to obligate crossovers. The X chromosome stands out with a much reduced indel rate (0.041–0.054 depending on G + C). The X spends only 1/3 of the time in the male germline, and either a lower rate of spontaneous, replication-associated or recombination-associated indels in the female lineage may explain this reduction. The indel rate on the Y (0.054) is similar to that of the autosomes, agreeing with previous observations (Makova *et al.*, 2004).

The fact that indels are associated with high-G + C isochores might have implications for the evolution of functional sequence. In particular, protein-coding genes are strongly enriched within high-G + C isochores, and their codon sequence and promoters thus appear to be under rather stronger mutation pressure from indels than might have been appreciated hitherto. This might, in part, explain the observed large turnover of promoter sequence, although selection is also likely to play a role (Taylor *et al.*, 2006).

These results are important from another perspective as well. If up to 50% of indels are not represented as gaps in alignment, then this implies that alignments are very often wrong. Using the posterior probability as a proxy for alignment accuracy, it appears that far away from gaps (where accuracy is the highest), only between 94% and 95% of sites are correctly aligned, with again the high-G + C regions being hardest hit. On average the situation is worse: alignments from simulated sequence is correct in only 79% of aligned sites, with an average posterior of 81%, and realigned human–mouse sequence shows similarly low average posteriors (82%). Real sequence shows lower gap densities in the alignment than simulated sequences, suggesting that alignment biases are more important for real

than for simulated sequence, possibly because of duplications and slippage driving indels, rather than the gain and loss of random sequence as assumed by all models used for alignment. In conclusion, alignments of reasonably divergent sequences should be treated with caution. Inferences based on alignments should take account of local alignment uncertainty, and the particular biases existing in current 'best' alignments.

ACKNOWLEDGEMENTS

The author thanks Martin Goodson, Leo Goodstadt, Chris Ponting and Caleb Webber for helpful discussions, and an anonymous reviewer for pointing out a subtle issue with gap annihilation. The author was funded by the Medical Research Council, UK.

Conflict of interest: None declared.

REFERENCES

- Blanchette, M. *et al.* (2004) Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.*, **14**, 708–715.
- Britten, R. *et al.* (2003) Majority of divergence between closely related DNA samples is due to indels. *Proc. Natl Acad. Sci. USA*, **100**, 4661–4665.
- Cheng, Z. *et al.* (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature*, **437**, 88–93.
- Durbin, R. *et al.* (1998) *Biological Sequence Analysis*. Cambridge University Press, Cambridge.
- Hellmann, I. *et al.* (2003) A neutral explanation for the correlation of diversity with recombination rates in humans. *Am. J. Hum. Genet.*, **72**, 1527–1535.
- Infante, J. *et al.* (2003) Genome-wide amplifications caused by chromosomal rearrangements play a major role in the adaptive evolution of natural yeast. *Genetics*, **165**, 1745–1759.
- Krogh, A. (1997) Two methods for improving performance of an HMM and their application for gene finding. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **5**, 179–186.
- Lander, E. *et al.* (2001) Initial sequencing and analysis of the human genome. *Nature*, **409**, 860–921.
- Lunter, G. *et al.* (2006) Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.*, **2**, e5.
- Lunter, G. *et al.* (2007) Uncertainty in homology inferences: assessing and improving genomic sequence alignment. (submitted).
- Makova, K.D. *et al.* (2004) Insertions and deletions are male biased too: a whole-genome analysis in rodents. *Genome Res.*, **14**, 567–573.
- Ogurtsov, A.Y. *et al.* (2004) Indel-based evolutionary distance and mouse-human divergence. *Genome Res.*, **14**, 1610–1616.
- Schwartz, S. *et al.* (2003) Human-mouse alignments with BLASTZ. *Genome Res.*, **13**, 103–107.
- Silva, J.C. and Kondrashov, A.S. (2002) Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet.*, **18**, 544–547.
- Spencer, C.C. *et al.* (2006) The influence of recombination on human genetic diversity. *PLoS Genet.*, **2**, e148.
- Taylor, M.S. *et al.* (2006) Heterotachy in mammalian promoter evolution. *PLoS Genet.*, **2**, e30.
- Thorne, J.L. *et al.* (1991) An evolutionary model for maximum likelihood alignment of DNA sequences. *J. Mol. Evol.*, **33**, 114–124.
- Zuker, M. (1991) Suboptimal sequence alignment in molecular biology. Alignment with error analysis. *J. Mol. Biol.*, **221**, 403–420.