SUPPLEMENTAL MATERIAL

Supplemental Text 1: The Neutral Indel Model

The Neutral Indel Model predicts that for neutrally evolving genomic sequence the lengths of 'inter-gap segments' (IGS) between adjacent indel events follow a geometric distribution, as a result of a uniform rate of indel events in the absence of selection. This prediction holds regardless of the length of the indels, or whether they are insertions or deletions. The model depends on two assumptions only: first, that indel events are independent of one another, and second, that they occur uniformly across the genome (Lunter et al. 2006). The second of these two assumptions is true only in neutrally evolving sequence, and then only in part, because neutral indel rates vary according to both local G+C content and germ-line history. To achieve a good fit, we stratify the genome into 20 equally-occupied bins according to G+C content in 250 bp windows, separately for each chromosome (see **Materials and Methods**). Simulations show that after this stratification, the resulting distribution fits a geometric remarkably well even in the presence of substantial residual indel rate variation ((Lunter et al. 2006), and results reported here).

Sequence that is subject to purifying selection will contain a deficit of indels, resulting in an excess of longer IGS when compared to the predictions of the Neutral Indel Model. By quantifying this excess, we can estimate the total number of bases which are purged of indel mutations. The majority of IGS are short, and when a genome is dominated by neutrally evolving sequence, as is the case for mammals and other organisms with large (>500 Mb) genomes, these IGS are dominated by neutrally evolving sequence and closely follow the neutral model (e.g. in **Figure 1**, between 25 and 100 bp). This neutral regime is identified by maximising the variance explained by the model (R^2) within a bracket of IGS lengths. The excess of long IGS, attributed to sequence under constraint, is established by first fitting the model to the data in the neutral regime and extrapolating to longer IGS; the justification of this process lies in the accurate fit of AR data across a larger range of IGS lengths. The method estimates $g_{sel} + \Delta$ from the

amount of sequence present in ungapped alignment blocks beyond that expected from a geometric distribution, where Δ represents a contribution of neutral sequence flanking conserved blocks, the amount of which depends on the unknown degree of clustering of functional elements. By varying this assumed degree of clustering between the extremes of complete and no clustering, upper and lower bounds of g_{sel} are obtained, following the methods of Ommeto *et al.* (Ometto et al. 2005); see (Lunter et al. 2006) for details.

Supplemental Text 2: Excess Constrained Sequence in Cattle Transposed Repetitive Elements (TEs)

For the majority of mammalian pairwise alignments, the distributions of indel mutations very closely follow the predictions of the Neutral Indel Model, with only 0.2 – 5.3Mb of mammalian TE sequence estimated to be refractory to indels. The exceptions to this are for estimates based upon alignments in which one of the contributing genomes is that from cattle. Estimates of functional sequence within whole genome alignments involving cattle are similar to what may be expected given the relative divergence from the second species in the pair (**Figure 2**). However, estimates for cattle TEs appear to be inflated. This conclusion holds regardless of whether the cattle genome or the second aligning genome is used to supply TE annotation.

If the cattle genome contains a large number of TEs that appear, when aligned to two other species, to have been purified of indels then these two species should themselves share such functional TE sequence. However, this is not what we observe: we estimate 4.5-6.5Mb of constrained sequence for dog-cattle ARs, 7.4- 10.7Mb for human-cattle ARs, yet only 1.3 – 1.7Mb for ARs between human and dog. This suggests that some cattle-specific anomaly associated with TE-annotated sequence occurs within the cattle genome assembly.

We demonstrated previously that a substantial number of indel errors are present within some mammalian genome sequence assemblies (Meader et al. 2010). Such indel errors often cluster in regions of low read coverage and/or are associated with regions of high G+C content, and inflate counts of short IGSs. As a consequence, these may cause an overestimation of the neutral indel rate, and thus an overestimation of the proportion of the alignment that is identified as being constrained. For those alignments which include the cattle genome sequence, within high G+C bins, there is a greater discrepancy between the whole genome and TE indel rates, than for alignments which do not include the cattle genome sequence (**Supplemental Figure 2**). While this observation may be of biological origin, another explanation is that it is due to an increased density of indel errors within annotated TEs in the current cattle genome assembly. An erroneous (and in particular, higher) estimate of the indel rate in cattle would result in an inflated estimated of the amount of indel-purified sequence; hence, an increased density of indel errors within annotated TEs is compatible with the anomalous TE conservation estimates we obtain. However, a definitive resolution of this issue will require additional genome sequence data, perhaps using alternative sequencing and assembly approaches.

Supplemental Text 3: Estimates of Constraint in Simulated Genome Sequences

We evolved simulated genome sequences from identical pairs which were initially 200Mb in size, with constant rates of substitution and indel events (see **Materials and Methods**), which were then aligned using BLASTZ. Each simulated genome contained 5% of constrained material, which was refractory to a proportion of indel events. Half of our simulated genome sequence was annotated as 'TEs', but differed in no way from the remaining neutrally evolving sequence. The neutral indel model was then used to estimate the quantity of conserved material between each simulated genome pair. For all simulations except one (see below), the true value of α_{sel} was bounded from below by the lower-bound estimate (**Figure 3**; **Supplemental Table 1**). Also, in most cases, except for extremely small divergences ($d_s < 0.1$)

or very large cryptic indel rate variation (see below), the upper-bound estimate also bounded the true value from below. This was to be expected since the inference model assumes that functional sequence is perfectly intolerant of indel mutations. Indeed, when we simulate sequence with an indel acceptance probability of 0, the upper bound estimate (4.88%) approaches the true amount. From this we conclude that the model's estimates of α_{sel} are robustly conservative.

As expected, decreasing the mean length of the conserved segments that constituted the conserved sequence resulted in a mild reduction in both the lower and upper bounds of α_{sel} estimates (**Supplemental Table 1**). The estimated value of α_{sel} was very modestly affected by variations in the clustering coefficient (the probability that any given conserved segment is directly followed by another), which governs the degree of clustering of conserved segments. In conclusion, therefore, the details of the spatial distribution of conserved sequence across the genome appear not to have a profound influence on the model's ability to estimate $\alpha_{sel.}$

Again as expected, the fraction of indels in constrained sequence that become fixed in the population greatly influences estimation of α_{sel}. When this probability is increased from 0 to 0.2, our lower and upper bound estimates of α_{sel} decrease by 45% (2.23%, from 4.13%) and 35% (3.20%, from 4.88%), respectively. From annotated protein coding sequence, we know that indels are fixed at approximately 10% of the rate observed in neutral sequence (Brandstrom and Ellegren 2007). What the corresponding figure is for non-protein-coding sequence is more difficult to estimate. For our purposes, it is however sufficient to know that our estimates will be conservative whatever the true value of this parameter.

One key assumption underlying our analyses is that neutral sequence accumulates indels uniformly. Despite accounting for G+C content (Lunter et al. 2006), it is inevitable that some cryptic variation in mutation rates will remain. To simulate this effect, we drew indel rates uniformly from an interval taken symmetrically around the mean rate, plus or minus a set percentage. The model is not expected to be

sensitive to cryptic variation below a resolution of roughly the physical distance between neighbouring indels; therefore, to maximize impact, this indel rate was applied within 5kb blocks of the simulated sequence. As expected, introducing cryptic indel rate variation increases the estimated $\alpha_{sel.}$ value. At 40% indel rate variation, the upper bound exceeds the true $\alpha_{sel.}$, while for values over 50% the true value of $\alpha_{sel.}$ is no longer within the estimated bounds. At these high levels of cryptic variation, the model infers a minimum of 1.9-2.4% (for 30% rate variation) and 3.1-4.0% (for 40% rate variation) of TE sequence to be under constraint (**Supplemental Table 1**). These values are, however, inconsistent with earlier findings that among true AR sequence in human and mouse only about 0.1% of sequence appeared to be under constraint (Lowe et al. 2007). Estimates we obtain for other species' pairs are similar (0.0 – 1.4% of ARs, with the exception of alignments involving the cattle genome). This indicates that the true level of unaccounted-for indel rate variation is substantially less than 40%, and therefore estimates of α_{sel} are expected to be conservative.

We draw two main conclusions from these simulations. First, both the upper and lower bound estimates of α_{sel} are expected to be conservative estimates of the proportion of sequence under purifying selection. Only when the simulated divergence drops below d_s= 0.1 does the upper bound estimate exceed the true value (**Figure 3**). This conclusion depends strongly on the fixation probability of indels within conserved sequence, relative to substitutions, which was estimated at 10% based on observations in protein-coding sequence (Brandstrom and Ellegren 2007). While ultra-conserved non-coding sequence certainly has a lower indel fixation probability, most conserved non-coding sequence appears to be more accepting of indels than is coding sequence, which is consistent with our conclusion. However estimates of average conservation in non-coding conserved sequence necessarily depend on criteria for conservation, making this conclusion in part circular; in particular, the conclusion that upper bound estimates for α_{sel} in fact represent conservative *lower* bounds would be arguable. Nevertheless, because the lower bound estimate

exceeds the true value only for clearly unreasonable extents of cryptic variation, α_{sel} estimates are robustly conservative.

The second conclusion is that α_{sel} estimates, in particular the lower bound, show only a minimal dependence on the true divergence between species, across a divergence range where alignments are possible and indel densities are sufficiently high (d_s= 0.05-0.65). This is in marked contrast to the large (2.5- to 3-fold) range of g_{sel} estimates we observe between diverse pairs of mammalian species (e.g. mouse-rat 189.0-258.4 Mb, mouse-dog 73.7-83.1 Mb, **Figure 2**), which our simulations show cannot be attributed to a d_S –dependence in the estimation procedure.

Supplemental Text 4: Estimating substitution rates in partitions of the *Drosophila* genomes

We sought to identify indel-purified segments (IPSs) defined as long ungapped sequences whose absence of indels is likely to reflect the purging of deleterious indel alleles in ancestral populations. IPSs thus are predicted to represent functional sequence. We predict sets of IPSs at a given false-discovery rate (FDR), the predicted proportion of neutral segments in the set, weighted by sequence length, by imposing lower thresholds on IGS lengths. This procedure was undertaken independently for each G+C category, and for the X chromosome and autosomal chromosomes separately, while maximising the total amount of identified segments under selection using the method of Lagrange multipliers.

We are able to identify 54.6Mb of putatively functional IPS regions in the *D. melanogaster* genome with a 1% false discovery rate (FDR) and 65.4Mb of IPS with a 10% FDR. These IPS regions, when aligned to *D. simulans*, exhibit a ~3-fold lower nucleotide substitution rate (median values of 0.037 and 0.042 substitutions per base for the 1% and 10% FDR sets, respectively), compared to estimates of neutral substitution rates based on short intronic segments (0.12 substitutions per base) and synonymous protein coding bases (0.13 subs per base) (Haddrill et al. 2005), and a similar rate to alignable coding sequence (0.032 substitutions per base). 15.4Mb (66.9%) of protein-coding sequence, and 62% of annotated

miRNAs intersected with the 1% FDR set of putatively-functional IPS, which are similar proportions to those identified from human-mouse alignments (Lunter et al. 2006).

Supplemental Text 5: Analysis of the Human-Macaque Whole Genome Alignment

Short IGS in whole genome alignments of neutrally evolving sequence between two primate species' pairs may not approximate a geometric distribution due to factors both artefactual and biological. Firstly, genome sequence assemblies are not perfect representations of genome sequences, and regions of poorer quality contain clusters of indel errors; indeed, the Neutral Indel Model can be usefully employed in quantifying the accuracy of genome assemblies (Meader et al. 2010). Such clusters may not be observed for pairs of more distantly-related species because true indel events may greatly outnumber indel errors. For pairs of more closely-related species (such as macaque and human) these clustered errors inflate the numbers of short IGS between 1-100 bp in length. A high number of short IGS mainly reflects errors in draft (e.g. macaque), rather than finished (e.g. human), genome assemblies (Meader et al. 2010). In order to account for this effect arising from assembly error, IGS limits over which the regression is fitted were increased (see **Supplemental Table 2**).

Secondly, primate-specific TEs are known to contain homo-nucleotide runs, that appear at specific intervals from one another, and which are prone to indel mutation (Batzer and Deininger 2002). The high frequency of indel mutations in these runs results in high counts of IGS lengths reflecting the physical separation between these runs. To ensure that this anomaly in the human-macaque IGS frequency distribution did not affect our estimate of α_{sel} for these species, we only considered regions of the macaque-human alignment that are not annotated as being TEs. Consequently, for this species pair we are unable to estimate the amount of functional sequence lying within their TEs.

References:

Batzer MA and Deininger PL. 2002. Alu repeats and human genomic diversity. *Nat Rev Genet* **3:** 370-379.

Brandstrom M and Ellegren H. 2007. The genomic landscape of short insertion and deletion polymorphisms in the chicken (Gallus gallus) Genome: a high frequency of deletions in tandem duplicates. *Genetics* **176:** 1691-1701.

Haddrill PR, Charlesworth B, Halligan DL, Andolfatto P. 2005. Patterns of intron sequence evolution in Drosophila are dependent upon length and GC content. *Genome Biol* **6:** R67.

Lowe CB, Bejerano G, Haussler D. 2007. Thousands of human mobile element fragments undergo strong purifying selection near developmental genes. *Proc Natl Acad Sci U S A* **104:** 8005-8010.

Lunter G, Ponting CP, Hein J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput Biol* **2:** e5.

Meader S, Hillier LW, Locke D, Ponting CP, Lunter G. 2010. Genome Assembly Quality: Assessment and Improvement Using the Neutral Indel Model. *Genome Res***:** In press.

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:** 1521-1527.

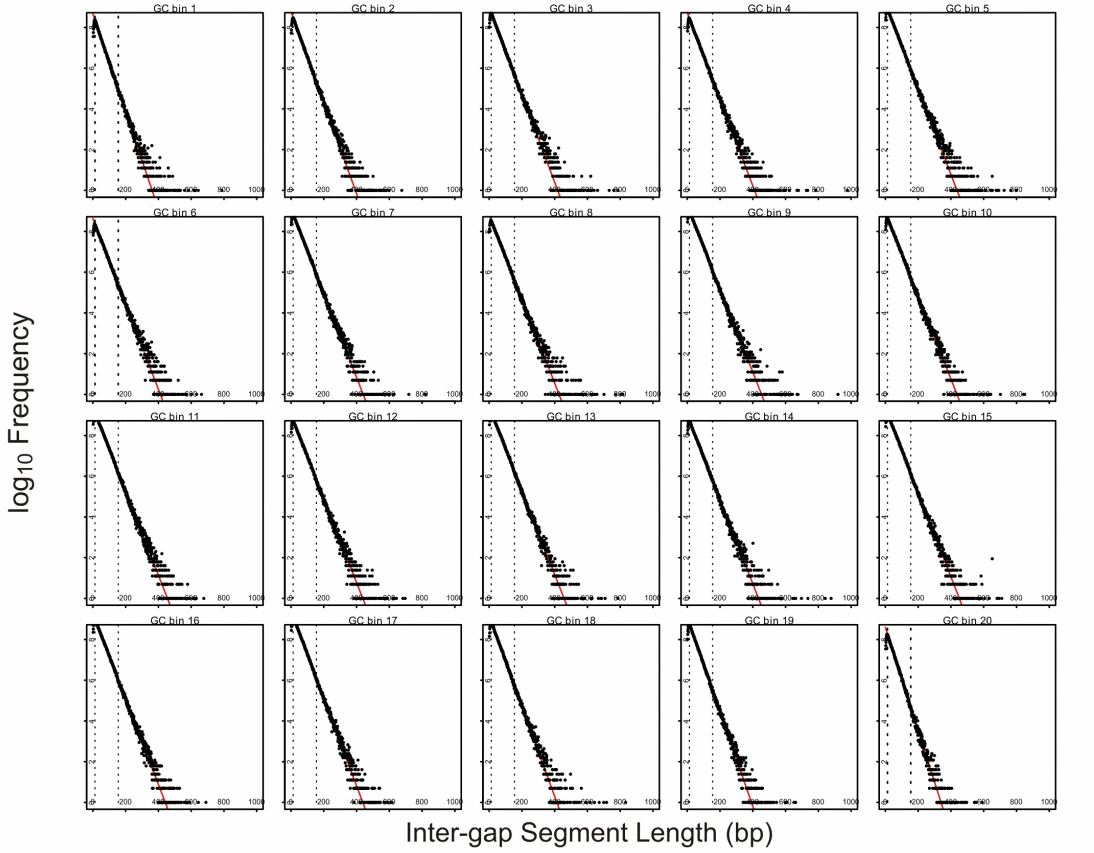
Supplemental Figure 1: Genomic distributions of inter-gap segment lengths in mouse-rat (A) ancestral repeat and (B) whole genome alignments.

Frequencies of inter-gap segments (IGS, black dots) are shown on a log10 scale. Alignments were partitioned into 20 bins based upon the G+C content of the mouse genome (see **Materials and Methods**). The red lines represent the predictions of the Neutral Indel Model, a geometric distribution of IGS lengths, which were calibrated using short IGS (boundaries of IGS used for calibration are marked by broken lines). For ancestral repeat alignments the data fit accurately the predictions of the Neutral Indel Model. For whole genome alignments there is an excess of longer IGS (>100bp) compared to the predictions of the Neutral Indel Model.

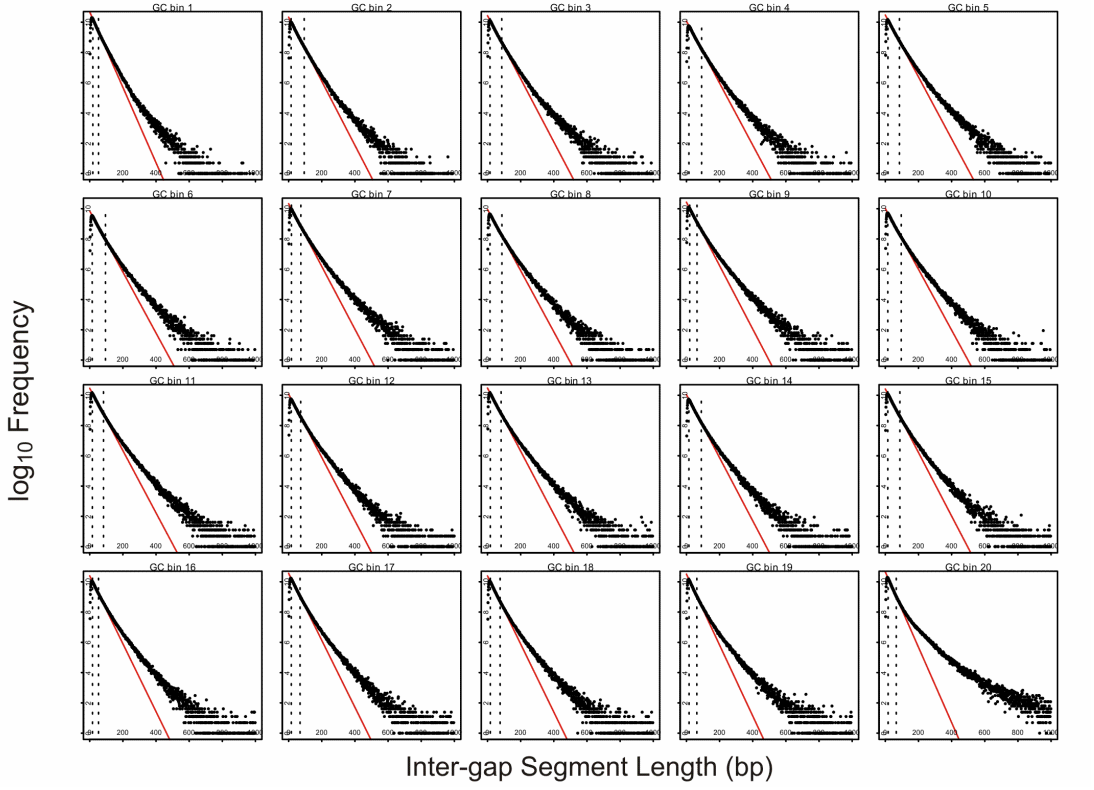
Supplemental Figure 2: Differences in neutral indel rates estimated between whole genome alignments and between ancestral repeat alignments

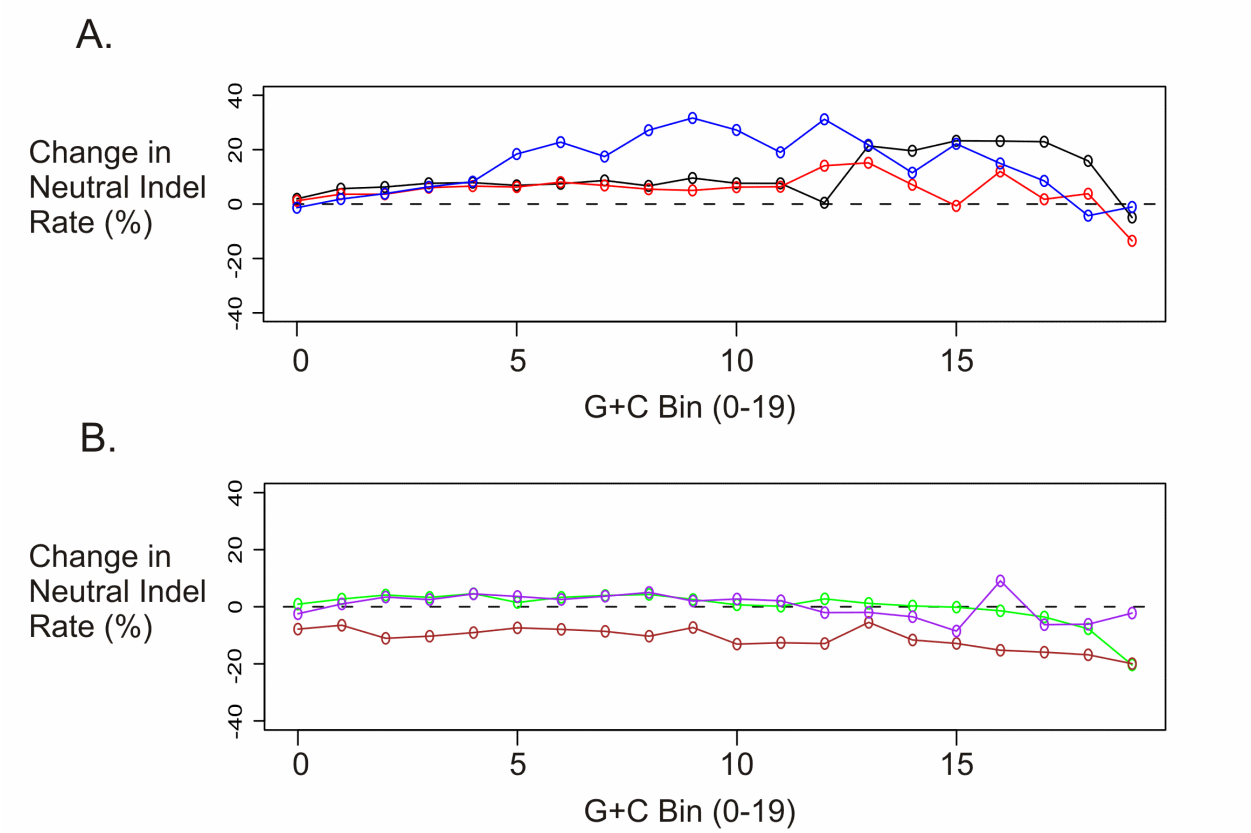
Percentage differences in estimates of indel rates in ancestral repeat alignments relative to whole genome alignments for three alignments of the cattle genome (A) and three alignments which did not involve cattle (B). The baseline of no change is indicated with a broken line. Where the cattle genome was included (mouse-cattle in blue; human-cattle in black; dog-cattle in red) there is an increase in the estimated neutral indel rate, particularly in higher G+C bins. When the cattle genome is not included (human-horse in purple; mouse-dog in brown; human-dog in green) there is only a marginal increase in the estimated neutral indel rate.

A. Mouse-Rat Ancestral Repeats



B. Mouse-Rat Whole Genome




A.
Change in Neutral Indel Rate (%)
40
20
0
-20
-40
0
5
10
15
G+C Bin (0-19)
B.
Change in Neutral Indel Rate (%)
40
20
0
-20
-40
0
5
10
15
G+C Bin (0-19)

Supplemental Table 1: Estimates of constraint in simulated genome sequences containing known quantities of sequence which is refractory to indels.

Indel Acceptence in Conserved Sequence	Mean Length of Conserved Elements (bp)	Clustering Coefficient	Mean Length of Intervening Material in Clustered Conservered Seq. (bp)	Divergence (dS)	Residual Indel Rate Variation (%)	Estimate of Constrained Sequence (αsel)		Constrained Ancestral Repeats	
						Lower (%)	Upper (%)	Lower (%)	Upper (%)
Static Indel Acceptence									
0	60	0.5	15	0.4	0	4.1	4.9	NA	NA
0.05	60	0.5	15	0.4	0	3.8	4.7	NA	NA
0.1	60	0.5	15	0.4	0	3.0	3.6	NA	NA
0.15	60	0.5	15	0.4	0	2.6	3.1	NA	NA
0.2	60	0.5	15	0.4	0	2.2	3.2	NA	NA
Variable Indel Acceptence - Drawn from a gamma distribution for each block									
0.1	60	0.5	15	0.15	0	3.6	5.1	NA	NA
0.1	60	0.5	15	0.4	0	3.2	3.9	NA	NA
0.1	60	0.5	15	0.65	0	2.8	3.4	NA	NA
Varying Clustering									
0.1	60	0	15	0.4	0	2.9	3.6	NA	NA
0.1	60	0	25	0.4	0	3.2	3.8	NA	NA
0.1	60	0	35	0.4	0	3.3	4.0	NA	NA
0.1	60	0.5	15	0.4	0	2.9	3.6	NA	NA
0.1	60	0.5	25	0.4	0	3.0	3.6	NA	NA
0.1	60	0.5	35	0.4	0	3.1	3.7	NA	NA
0.1	60	0.95	15	0.4	0	2.9	3.7	NA	NA
0.1	60	0.95	25	0.4	0	2.8	3.4	NA	NA
0.1	60	0.95	35	0.4	0	3.0	3.6	NA	NA
Evolutionary Distance									
0.1	60	0.5	15	0.05	0	3.1	6.0	NA	NA
0.1	60	0.5	15	0.1	0	3.3	5.0	NA	NA
0.1	60	0.5	15	0.15	0	3.3	4.2	NA	NA
0.1	60	0.5	15	0.275	0	3.2	4.0	NA	NA
0.1	60	0.5	15	0.4	0	3.1	3.7	NA	NA
0.1	60	0.5	15	0.525	0	3.1	3.6	NA	NA
0.1	60	0.5	15	0.65	0	3.0	3.6	NA	NA
Residual Indel Variation									
0.1	60	0.5	15	0.4	0	3.0	3.6	0	0
0.1	60	0.5	15	0.4	10	2.9	3.5	0	0
0.1	60	0.5	15	0.4	20	3.5	4.3	0.3	0.4
0.1	60	0.5	15	0.4	30	4.2	5.2	0.7	0.9
0.1	60	0.5	15	0.4	40	4.8	6.0	1.9	2.4
0.1	60	0.5	15	0.4	50	6.4	8.1	3.1	4
Varying Functional Length									
0.1	30	0.5	15	0.4	0	2.2	2.8	NA	NA
0.1	60	0.5	15	0.4	0	3.1	3.7	NA	NA
0.1	90	0.5	15	0.4	0	3.3	3.9	NA	NA

Supplemental Table 2: Estimates of constrained sequence in alignments of human and macaque genome sequences.

IGS Limits for Neutral Indel Model	Constrained Sequence (Mb)	Median Estimated Indel Rate (Indels per base)
80-150	175.4 – 271.4	0.067
90-160	160.5 – 248.9	0.066
100 -170	153.5 – 237.9	0.065
120 - 190	127.3 – 196.7	0.060