Supplementary file
Laurie et al. 2011

Table 1

Algorithm	*orphan* exons[1]	*over-aligned* exons[2]
Prank+F	2,315	3,024
MAFFT	886	5,318
ClustalW	705	13,789

Table 1. Number of orphan and over-aligned exons identified in alignments by different multiple-alignment algorithms. Alignments corresponded to orthologous coding sequences from five mammalian species (see main manuscript file). [1]Orphan exons are those for which there is no ortholog in the other species; identification by the alignment program means the exon is completely separated out (aligned with gaps). [2]Over-aligned exons are those for which the percentage identity was less than 0.5 across the corresponding part of the alignment, and there were no gaps.

Table 2

	human	macaque	primate ancestral	mouse	rat	rodent ancestral
Prank+F						
Del (deletions)	2494	3884	7883	9847	12381	47784
Ins (insertions)	2285	2973	9582	9911	9236	19971
Del/ Ins ratio	**1.09**	**1.31**	**0.82**	**0.99**	**1.34**	**2.39**
MAFFT						
Del (deletions)	2563	3885	7574	8672	10975	29509
Ins (insertions)	673	1046	3217	3477	3560	9046
Del/ Ins ratio	**3.81**	**3.71**	**2.35**	**2.49**	**3.08**	**3.26**
ClustalW						
Del (deletions)	2906	3789	8476	10435	12852	36601
Ins (insertions)	515	727	3614	2248	2228	7774
Del/ Ins ratio	**5.64**	**5.21**	**2.35**	**4.64**	**5.77**	**4.71**

Table 2. Estimation of insertions and deletions in ancestral repeats using different multiple alignment programs. The sequences correspond to syntenic ancestral repeats from five mammalian species (see main manuscript file). MAFFT and ClustalW estimate a much lower number of insertions than Prank+F.

Table 3

	human	macaque	primate ancestral	mouse	rat	rodent ancestral
Prank+F						
Del (deletions)	91	129	434	307	365	1597
Ins (insertions)	86	105	273	295	248	622
Del/ Ins ratio	**1.06**	**1.23**	**1.59**	**1.04**	**1.47**	**2.57**
MAFFT						
Del (deletions)	91	128	444	316	351	1546
Ins (insertions)	51	73	217	236	185	537
Del/ Ins ratio	**1.78**	**1.75**	**2.05**	**1.34**	**1.90**	**2.88**
ClustalW						
Del (deletions)	99	121	394	294	320	1366
Ins (insertions)	35	61	223	193	162	442
Del/ Ins ratio	**2.83**	**1.98**	**1.77**	**1.52**	**1.98**	**3.09**

Table 3. Estimation of insertions and deletions in coding sequences using different multiple alignment programs. The sequences correspond to 3,126 orthologous coding sequences from five mammalian species (dataset for dN and dS calculation, see main manuscript file). MAFFT and ClustalW estimate a much lower number of insertions than Prank+F.

Table 4

		human	macaque	primate ancestral	mouse	rat	rodent ancestral
Ancestral repeats (ARs)[1]	**Deletions**						
	N	2,494	3,882	7,876	9,839	12,364	47,577
	Length (bp)						
	Mean	2.2	2.5	2.5	3.1	3.1	3.9
	Median	1	1	1	2	2	2
	SD	2.1	2.8	3.1	3.6	3.9	5.0
	Insertions						
	N	2,279	2,938	9,548	9,610	8,912	19,854
	Length (bp)						
	Mean	2.7	3.4	3.3	4.0	3.6	3.2
	Median	1	2	2	2	2	2
	SD	3.9	6.8	4.3	11.3	12.6	4.9
coding Sequences (CDs)[2]	**Deletions**						
	N	214	296	933	686	832	3487
	Length (AAs)						
	Mean	1.6	1.6	1.8	1.9	1.9	1.9
	Median	1	1	1	1	1	1
	SD	1.2	1.6	1.5	1.5	1.6	1.7
	Insertions						
	N	166	216	631	620	558	1297
	Length (AAs)						
	Mean	1.9	2.0	2.0	1.7	1.8	1.9
	Median	1	1	1	1	1	1
	SD	1.7	1.9	1.6	1.6	1.5	1.7

Table 4. Descriptive statistics of deletion and insertion sizes in ancestral repeats and coding sequences. [1]Number of ARs: 19,631; total length of aligned AR sequence: 4,746,950nt; events size 1-30 bp. [2]Number of CDs: 5,991; total length of aligned CDs : 11,705,952nt; events size 1-10 amino acids. N: number of events; bp: base pairs; AAs: amino acids; SD: standard deviation.

Table 5

	human	macaque	primate ancestral	mouse	rat	rodent ancestral
Tree 1 (n=5,991)						
Del (deletions)	214	296	933	686	832	3,487
Ins (insertions)	166	216	631	620	558	1,297
Del/ Ins ratio	**1.29**	**1.37**	**1.48**	**1.11**	**1.49**	**2.69**
Tree 2 (n=5,991)						
Del (deletions)	232	300	918	704	882	3,425
Ins (insertions)	166	225	597	603	549	1,232
Del/ Ins ratio	**1.40**	**1.33**	**1.54**	**1.17**	**1.61**	**2.78**
Tree 3 (n=4,840)						
Del (deletions)	117	168	489	346	439	1,816
Ins (insertions)	83	126	312	335	322	699
Del/ Ins ratio	**1.41**	**1.33**	**1.57**	**1.03**	**1.36**	**2.60**

Table 5. Estimation of insertions and deletions using different phylogenetic trees as input to Prank+F. Tree 1 corresponds to that given in the text, calculated from a concatenate of a random subset of 150 ortholog sets. Tree 2 corresponds to the distances given by Miller *et al*, 2007 (Genome Res. 17: 1797-1808). Tree 3 corresponds to a six species tree, as for original dataset, but including *Monodelphis domestica* as a further outgroup. Due to some missing orthologs in *Monodelphis*, the data presented is for 4,840 ortholog sets.

Table 6

			human	macaque	primate ancestral	mouse	rat	rodent ancestral
Coding sequence dataset (5,991 ortholog sets)	Number of events	Del (deletions)	214	296	933	686	832	3,487
		Ins (insertions)	166	216	631	620	558	1,297
		Del/ Ins ratio	**1.29**	**1.37***	**1.48*****	**1.11**	**1.49*****	**2.69*****
CodeML coding sequence subset (3,126 ortholog sets)	Number of events	Del (deletions)	91	129	434	307	365	1,597
		Ins (insertions)	86	105	273	295	248	622
		Del/Ins ratio	**1.06**	**1.23**	**1.59*****	**1.04**	**1.47***	**2.57*****

Table 6. Estimated number of insertion and deletion events in orthologous sequences in different mammalian branches. Events considered were of length 1-10 amino acids. Spearman correlation between Del/Ins ratio of the two datasets was rho=0.94, $p<0.05$. Asterisks indicate significant departures from the null hypothesis that insertions and deletions occur with equal frequency (Chi-squared test, *$p<0.01$, ***$p<10^{-4}$). There is no significant difference in the relative frequency of indels in any species between the two datasets (Chi-squared, $p>0.05$).

Table 7

		human	macaque	primate ancestral	mouse	rat	rodent ancestral
Deletions	**Yes**						
	Mean	0.25	0.27	0.19	0.21	0.19	0.14
	Median	0.19	0.22	0.15	0.16	0.15	0.11
	SD	0.21	0.23	0.16	0.21	0.14	0.10
	NO						
	Mean	0.16	0.15	0.14	0.13	0.12	0.09
	Median	0.10	0.09	0.09	0.08	0.08	0.06
	SD	0.20	0.18	0.19	0.16	0.15	0.08
Insertions	**Yes**						
	Mean	0.22	0.22	0.17	0.18	0.16	0.13
	Median	0.13	0.14	0.15	0.13	0.13	0.11
	SD	0.27	0.24	0.15	0.18	0.14	0.10
	NO						
	Mean	0.16	0.15	0.14	0.13	0.12	0.10
	Median	0.10	0.09	0.09	0.08	0.08	0.07
	SD	0.20	0.18	0.19	0.16	0.15	0.09
Deletions or Insertions	**Yes**						
	Mean	0.24	0.24	0.18	0.19	0.17	0.13
	Median	0.17	0.18	0.15	0.14	0.13	0.11
	SD	0.24	0.24	0.16	0.20	0.14	0.10
	NO						
	Mean	0.16	0.15	0.14	0.12	0.12	0.08
	Median	0.10	0.09	0.09	0.08	0.08	0.06
	SD	0.19	0.18	0.19	0.16	0.15	0.08

Table 7. Non-synonymous to synonymous (dN/dS) ratio for mammalian proteins that have at least one indel event as indicated. Differences for each category are highly significant ($p<<0.01$) with the exception of the values for insertions in macaque (significant at $p<0.01$), and insertions in Human (not significant, $p=0.18$), Kolmogorov-Smirnov test.

Table 8

			human	macaque	primate ancestral	mouse	rat	rodent ancestral
Ancestral repeat dataset (19,631 sequences)	Number of events	Deletions	2,494	3,882	7,876	9,839	12,364	47,577
		Insertions	2,279	2,938	9,548	9,610	8,912	19,854
		Del/Ins ratio	**1.09***	**1.32****	**0.82****	**1.02**	**1.39****	**2.40****
3'UTRs dataset (746 sequences)	Number of events	Deletions	290	327	1003	972	1228	4440
		Insertions	330	274	1296	988	886	2044
		Del/Ins ratio	**0.88**	**1.19**	**0.77****	**0.98**	**1.39****	**2.17****

Table 8. Number of insertion and deletion events in orthologous non-coding sequences in different mammalian branches. Events considered were of length 1-30 nucleotides. Spearman correlation between Del/Ins ratio of the two datasets was rho =0.94, $p<0.05$. *Del/Ins different from 1 at $p<0.05$, chi-squared test, 1df; **Del/Ins different from 1 at $p<10^{-4}$, chi-squared test, 1df.

Table 9

	human	macaque	primate ancestral	mouse	rat	rodent ancestral
Del (deletions)	330	536	1,369	1,055	1,279	4,996
Ins (insertions)	253	357	889	859	801	1,827
Del/ Ins ratio	**1.30**	**1.50**	**1.54**	**1.23**	**1.60**	**2.73**

Table 9. Number of insertions and deletions observed without pre-alignment filtering. Data given are for 10,129 orthologous proteins obtained from Ensembl Version 55. Alignments were generated using Prank+F with branch lengths as described in Miller *et al*, 2007 (Genome Res. 17: 1797-1808). Following alignment, events observed adjacent to exon boundaries, and those occurring in regions with low sequence identity were excluded, as described in the text.

Table 10

Specied	Human	Mouse	Macaque	Rat	Cow
Golden Path	3,093,120,360	2,716,965,481	3,097,179,960	2,718,897,321	3,033,353,239

Table 10. Golden path genome length for species used in this study.

Table 11

	All Proteins		Proteins containing a low-complexity region		Proteins containing an AA tandem repeat ≥4	
	n	Median (mean)	n	Median (mean)	n	Median (mean)
Human	5991	548 (679)	5157	589 (722)	2872	677 (824)
Macaque	5991	532 (658)	5078	577 (703)	2749	662 (806)
Mouse	5991	544 (675)	5103	591 (720)	2479	678 (830)
Rat	5991	538 (664)	5098	586 (709)	2707	673 (817)

Table 11. Protein length in different datasets. All proteins: all proteins in the 1:1 orthologous protein dataset; Proteins containing a low-complexity region: subset of proteins containing at least one low-complexity regions as identified by SEG with default parameters; Proteins containing an AA tandem repeat: subset of proteins containing at least one perfect amino acid tandem repeat of length 4 or longer.

Figure 1

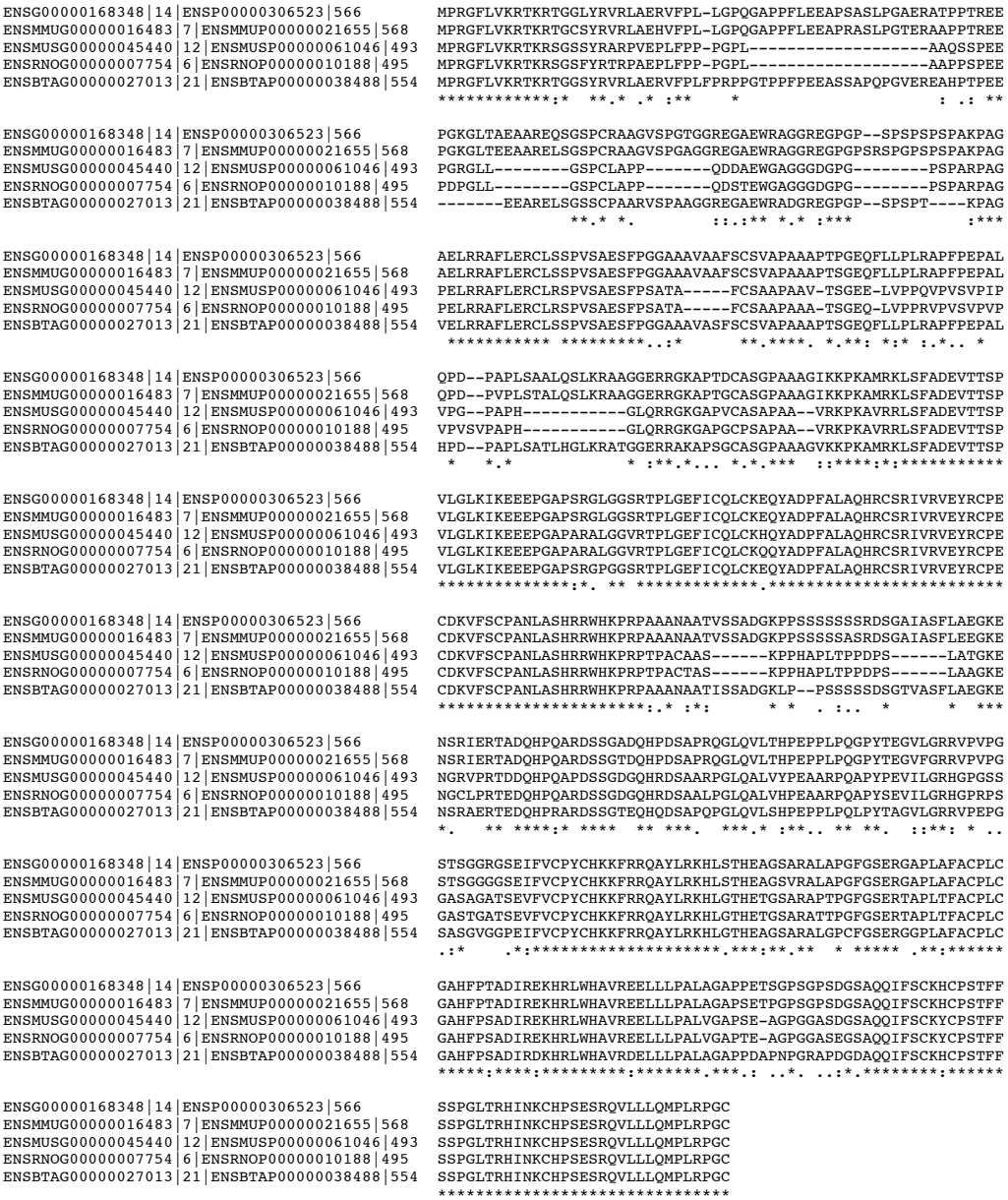

Figure 1. Multiple sequence alignment of insulinoma-associated 2 protein. Sequences are from human (ENSP00000306523), macaque (ENSMMUP00000021655), mouse (ENSMUSP00000061046), rat (ENSRNOP00000010188) and cow (ENSBTAP00000038488). The protein is encoded by a single exon. It is approximately 10% shorter in the rodents than in the other species (494 versus 554 in cow and 567 in primates) due to 8 short deletions (totalling 29 amino acids) and two larger deletions of 11 and 19 amino acids, respectively.

Figure 2

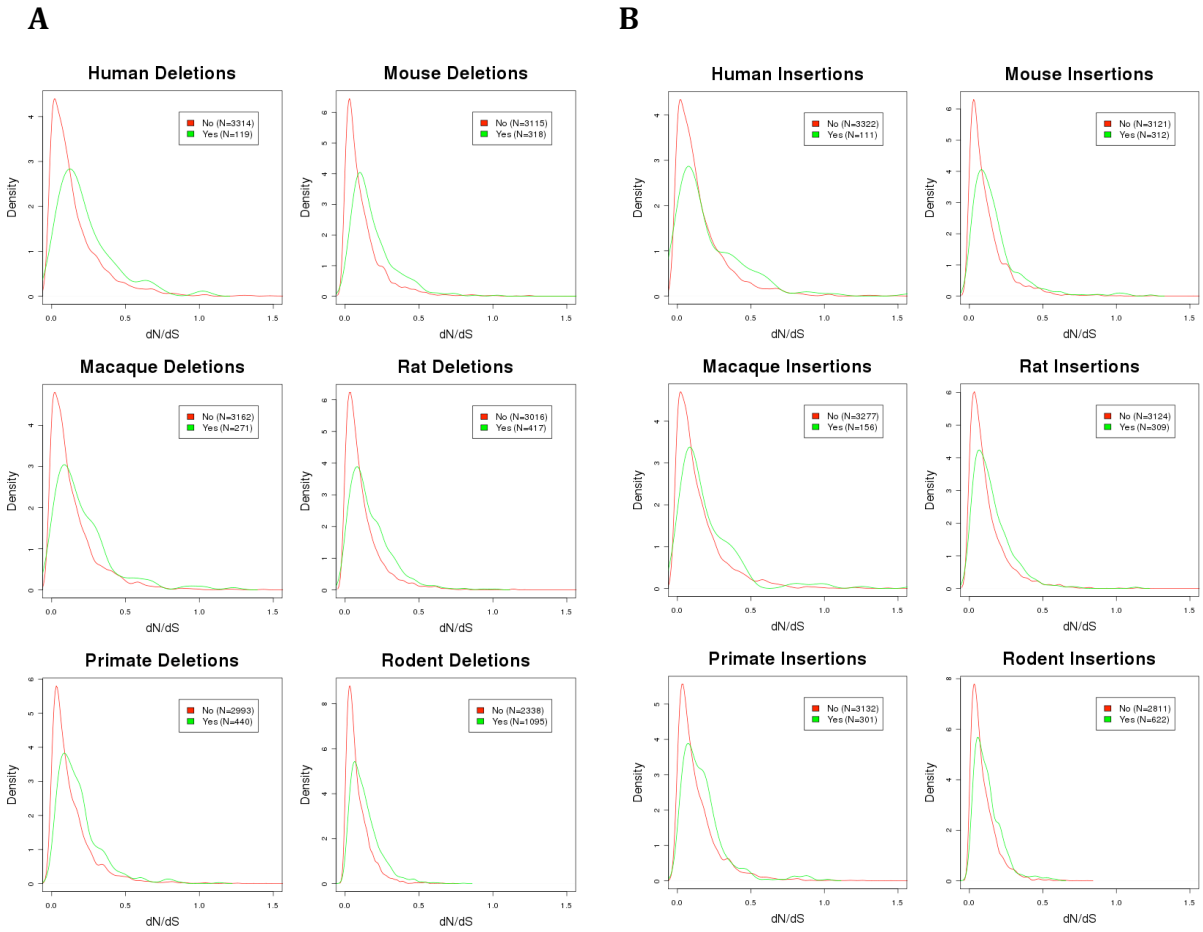

Figure 2. dN/dS distribution for proteins which have (Yes), or do not have, (No) an indel event. The distributions in each graph are all significantly different except that of human insertions (see Supplementary Table 5 for averages and p-values). Proteins that have at least one indel event have higher dN/dS than proteins with no indel events.

Supplementary Methods

Sequences

Ancestral Repeats The *extract pairwise MAF blocks* feature of Galaxy was used to extract regions syntenic to human ancestral repeats, where available, in the macaque, mouse, rat, and cow genomes, through the UCSC genome browser. An in-house Perl script was used to stitch the ends of contiguous regions of pairwise output, providing a final set of 19,631 orthologous ancestral repeat regions, which were then re-aligned with $Prank_{+F}$, using the same distance tree as for the coding sequences.

Coding Sequences Where more than one transcript per gene was available we used the longest transcript, as provided by default by Ensembl. Ortholog sets were removed in cases where any protein sequence contained an "x", indicating that the residue had not been clearly defined, or where the length of the shortest protein sequence in the set was less than 50% of the length of the longest protein in the same set or where the protein length was less than 100 amino acids. Further sets were removed where calculated values of dS or dN were less than 0.01, or dN was greater than 2 for any branch. Keeping only those sets for which there was concordance in identification of 1 to 1 orthologs with Ensembl Version 55, resulted in a further reduction in sets.

Post-alignment filters

Initial appraisal of alignments by eye indicated that there were a relatively large number of cases where parts of the alignment did not appear to represent truly orthologous sequences, in particular with respect to macaque sequences. Detailed examination of such dubious alignments by returning to Ensembl and investigating the exonic structure of the underlying protein sequences indicated that many individual macaque exons described in Ensembl are likely to have annotation errors. This was also observed, though to a lesser frequency, in some rat sequences. Alignment of incorrectly defined exons will result in gaps that will appear to be indel events. In order to limit the incorporation of such spurious indels in downstream analyses, the positions of all exons in all proteins were mapped onto the MSAs and any exons showing less than 50% similarity were excluded from further indel analysis. In addition, indels found immediately adjacent to exon boundaries were discounted, as they are more likely to represent annotation errors than *bona fide* indels.

Analysis of sequence context of indels

Low complexity regions were identified using SEG with default parameters, while the parameters '4,0,0' were used to identify pure amino-acid tracts of length for or more. For observed insertion events, the area that corresponded to the insertion had to be completely within a low-complexity region or amino acid tandem repeat in the same species. For deletions, the orthologous region in the protein of the sister species was used, since this is expected to provide the best approximation of the sequence background upon which the deletion occurred.

Comparison of insertions and deletions in ancestral repeats and coding sequences

To assess the strength of purifying selection in eliminating amino acid insertions and deletions from coding sequences we identified all events of a size that is a multiple of 3nt in ancestral repeats and compared them to the same type of event in coding sequences. The observed proportion of deletion versus insertion events in ancestral repeats was used to estimate the expected number of deletions and insertions in coding sequences. For example in the rat branch deletions represented 54% of all events. Taking into account the total number of events in CDs (1,390) we would expect 746 deletions and 644 insertions. The observed numbers, 832 deletions and 558 insertions, are significantly different to those expected using a chi-square test with 1 d.f. ($p<10^{-3}$).

Estimation of nucleotide substitution rates

Following completion of this step, further ortholog sets were removed if the number of complete columns in the alignment was less than 100, or if the value calculated for dS was less than 0.01, or if either dN or dS was greater than 2.0, as estimates of dN/dS have been shown to be unreliable in such cases (Toll-Riera et al. 2010).