Characterization of intron loss events in mammals

Jasmin Coulombe-Huntington and Jacek Majewski[1]

Department of Human Genetics, McGill University, Montreal, Quebec H3A 1A4, Canada

The exon/intron structure of eukaryotic genes differs extensively across species, but the mechanisms and relative rates of intron loss and gain are still poorly understood. Here, we used whole-genome sequence alignments of human, mouse, rat, and dog to perform a genome-wide analysis of intron loss and gain events in >17,000 mammalian genes. We found no evidence for intron gain and 122 cases of intron loss, most of which occurred within the rodent lineage. The majority (68%) of the deleted introns were extremely small (<150 bp), significantly smaller than average. The intron losses occurred almost exclusively within highly expressed, housekeeping genes, supporting the hypothesis that intron loss is mediated via germline recombination of genomic DNA with intronless cDNA. This study constitutes the largest scale analysis for intron dynamics in vertebrates to date and allows us to confirm and extend several hypotheses previously based on much smaller samples. Our results in mammals show that intron gain has not been a factor in the evolution of gene structure during the past 95 Myr and has likely been restricted to more ancient history.

[Supplemental material is available online at www.genome.org.]

Reconstructing the evolutionary history of spliceosomal introns remains one of the most fervently debated topics in eukaryotic evolution (Roy and Gilbert 2006). The long-standing debate over the introns-early versus introns-late hypotheses (Stoltzfus 1994; de Souza et al. 1998) contrasts the ideas of introns either originating in the early RNA world or evolving from an expansion of group II self-splicing introns in an early eukaryotic ancestor (Cavalier-Smith 1991). Understanding the natural history of introns is essential to understanding their function: Are introns simply selfish DNA elements that have been maintained in large genomes akin to retrotransposons, or do they serve a function, such as promoting recombination (Comeron and Kreitman 2000) and alternative splicing (AS) (Kim et al. 2004), resulting in increased proteome diversity and complexity?

Evolutionary investigations of the dynamics of intron gains and losses are generally hampered by the limited availability of high-quality data on the sequence and structure of gene orthologs from diverse species. To date, we have been unable to utilize the entire gene complements of most organisms in question, and the data sets commonly used range from hundreds (Rogozin et al. 2003) to at most a thousand genes (Roy et al. 2003) or several thousand introns (Nielsen et al. 2004).

Here, we make use of the complete, high-quality genomic sequences of four mammalian species—human, mouse, rat, and dog—to investigate intron gain and loss dynamics in mammals. We utilize a gene mapping technique to map annotated reference human genes onto genome-wide, multispecies sequence alignments, allowing us to investigate the predicted intron-exon boundaries of 152,146 introns within 17,242 autosomal genes. A recent study that considered a much smaller number of mammalian genes (Roy et al. 2003) uncovered six differences in intron positions between human and rodents, and suggested that there is no evidence for intron gain, and a very slow rate of intron loss in mammals. Here, we detect >100 cases of intron loss and still no evidence for any intron gain during mammalian evolution. Our large sample size allows us to determine the relative rates of intron losses in mammalian lineages and characterize the types of introns and genes that appear susceptible to loss, providing us with new insight regarding the mechanisms of intron deletion.

[1]Corresponding author.
E-mail jacek.majewski@mcgill.ca; fax (514) 398-1790.
Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.5703406. Freely available online through the *Genome Research* Open Access option.

Results

We used the mapping of annotated human exon–intron boundaries onto the mouse, rat, and dog genomes to detect changes in gene architecture that occurred during the evolution of the four mammalian species. This approach makes use of the highest quality gene annotation (17,242 human genes), but it allows us only to detect either intron loss events that occurred in rodent and dog or intron gain events that occurred in the human lineage. Thus, we also employed the reverse approach: mapping known mouse genes onto mouse/human whole-genome alignments. The latter strategy results in a slightly smaller data set (16,068 mouse genes) but allows us to detect intron losses in the human and intron gains in the mouse genome. The results of the combined analyses are listed in Tables 1 and 2. The name and symbol correspond to the human RefSeq gene where the loss/gain event occurred, except for the events in human, where the symbol refers to the mouse gene. The length for dog, mouse, and rat events is the length of the corresponding intron in human. For human events, the length corresponds to the mouse intron. We classified the results into isolated events (Table 1), i.e., those where a single intron gain/loss event (or multiple nonconsecutive events) occurred in a gene, and concerted events (Table 2), where the change involved multiple successive introns from the same gene. We propose that the single and multiple events may be mediated by slightly different mechanisms (see Discussion), and the two classes were henceforth analyzed separately.

We were able to uncover a total of 120 isolated changes: four occurring in human, 29 in mouse, 46 in rat, 34 in the rodent lineage prior to the mouse/rat divergence, and seven in dog. Remarkably, all of the changes were consistent with a loss, rather than a gain of an intron; i.e., for each case of a deletion of an intron relative to the reference gene structure (either mouse or human), the annotated intron was present in an earlier diverged organism. The loss of each intron was verified by using dog as the

Table 1. Independent intron deletions

RefSeq ID	Position	Size	Species	Symbol	Full name
NM_012207[a]	7	275	Dog	*HNRPH3*	Heterogeneous nuclear ribonucleoprotein H3
NM_025234	6	238	Dog	*WDR61*	WD repeat domain 61
NM_018445	4	96	Dog	*SELS*	Selenoprotein S
NM_032259	8	89	Dog	*WDR24*	WD repeat domain 24
NM_004104	42	76	Dog	*FASN*	Fatty acid synthase
NM_025241	8	86	Dog	*UBXD1*	UBX domain containing 1
NM_002096	8	75	Dog	*GTF2F1*	General transcription factor IIF, polypeptide 1
NM_182752	1	150	Mouse	*FAM79A*	Hypothetical protein LOC127262
NM_003132[a]	5	84	Mouse	*SRM*	Spermidine synthase
NM_006600	5	94	Mouse	*NUDC*	Nuclear distribution gene C homolog
NM_007122[a]	9	245	Mouse	*USF1*	Upstream stimulatory factor 1 isoform 1
NM_004550	6	261	Mouse	*NDUFS2*	NADH dehydrogenase (ubiquinone) Fe-S protein 2
NM_153188	16	130	Mouse	*TNPO1*	Transportin 1
NM_001090	10	96	Mouse	*ABCF1*	ATP-binding cassette, sub-family F, member 1
NM_007355	7	204	Mouse	*HSP90AB1*	Heat shock 90-kDa protein 1, β
NM_138419	6	5966	Mouse	*FAM54A*	DUF729 domain containing 1
NM_007189	4	82	Mouse	*ABCF2*	ATP-binding cassette, subfamily F, member 2
NM_006421	11	112	Mouse	*ARFGEF1*	Brefeldin A–inhibited guanine
NM_001273	39	175	Mouse	*CHD4*	Chromodomain helicase DNA binding protein 4
NM_006191	8	81	Mouse	*PA2G4*	Proliferation-associated 2G4, 38 kDa
NM_004184	8	703	Mouse	*WARS*	Tryptophanyl-tRNA synthetase isoform a
NM_001376	67	95	Mouse	*DYNC1H1*	Dynein, cytoplasmic, heavy polypeptide 1
NM_014030	13	71	Mouse	*GIT1*	G protein–coupled receptor kinase interactor 1
NM_002230	9	196	Mouse	*JUP*	Junction plakoglobin
NM_002805	5	81	Mouse	*PSMC5*	Proteasome 26S ATPase subunit 5
NM_020695	14	85	Mouse	*REXO1*	Transcription elongation factor B polypeptide 3
NM_001961	7	359	Mouse	*EEF2*	Eukaryotic translation elongation factor 2
NM_020230	10	82	Mouse	*PPAN*	Peter Pan homolog
NM_001379	37	268	Mouse	*DNMT1*	DNA (cytosine-5-)-methyltransferase 1
NM_005498	4	113	Mouse	*AP1M2*	Adaptor-related protein complex 1, μ 2 subunit
NM_032377	2	200	Mouse	*ELOF1*	Elongation factor 1 homolog (ELF1, S.)
NM_000516	9	104	Mouse	*GNAS*	Guanine nucleotide binding protein, α
NM_001670[a]	5	968	Mouse	*ARVCF*	Armadillo repeat protein
NM_020755	6	133	Mouse	*SERINC1*	Tumor differentially expressed 2
NM_003086	16	179	Mouse	*SNAPC4*	Small nuclear RNA activating complex,
NM_002046	4	129	Mouse	*GAPDH*	Glyceraldehyde-3-phosphate dehydrogenase
NM_005216[a]	9	123	Rat	*DDOST*	Dolichyl-diphosphooligosaccharide-protein
NM_014409[b]	4	7100	Rat	*TAF5L*	PCAF associated factor 65 β isoform a
NM_003400	14	85	Rat	*XPO1*	Exportin 1
NM_016516	19	114	Rat	*VPS54*	Vacuolar protein sorting 54 isoform 1
NM_001747	9	579	Rat	*CAPG*	Gelsolin-like capping protein
NM_005911[a]	8	635	Rat	*MAT2A*	Methionine adenosyltransferase II, α
NM_014670	7	139	Rat	*BZW1*	Basic leucine zipper and W2 domains 1
NM_004953[a]	18	125	Rat	*EIF4G1*	Eukaryotic translation initiation factor 4
NM_006859	5	82	Rat	*LIAS*	Lipoic acid synthetase isoform 1 precursor
NM_018115	18	108	Rat	*SDAD1*	SDA1 domain containing 1
NM_017676[a]	5	99	Rat	*FLJ20125*	Hypothetical protein LOC54826
NM_002198	4	109	Rat	*IRF1*	Interferon regulatory factor 1
NM_004381[a]	8	813	Rat	*CREBL1*	cAMP responsive element binding protein-like 1
NM_007355	5	136	Rat	*HSP90AB1*	Heat shock 90-kDa protein 1, β
NM_015153	6	81	Rat	*PHF3*	PHD finger protein 3
NM_000971	4	111	Rat	*RPL7*	Ribosomal protein L7
NM_018449	20	716	Rat	*UBAP2*	Ubiquitin associated protein 2
NM_001001973[a]	5	111	Rat	*ATP5C1*	ATP synthase, H+ transporting, mitochondrial F1
NM_003591	13	104	Rat	*CUL2*	Cullin 2
NM_018237	22	95	Rat	*CCAR1*	Cell-cycle and apoptosis regulatory protein 1
NM_003375	5	88	Rat	*VDAC2*	Voltage-dependent anion channel 2
NM_001011663	2	87	Rat	*PCGF6*	Polycomb group ring finger 6 isoform a
NM_005146	11	92	Rat	*SART1*	Squamous cell carcinoma antigen recognized by T
NM_006842	20	102	Rat	*SF3B2*	Splicing factor 3B subunit 2
NM_002898	8	179	Rat	*RBMS2*	RNA binding motif, single-stranded interacting
NM_013449	26	115	Rat	*BAZ2A*	Bromodomain adjacent to zinc finger domain, 2A
NM_007062	4	100	Rat	*PWP1*	Periodic tryptophan protein 1
NM_002271	19	89	Rat	*RANBP5*	RAN binding protein 5
NM_002271	23	84	Rat	*RANBP5*	RAN binding protein 5
NM_007111	6	604	Rat	*TFDP1*	Transcription factor Dp-1
NM_002892	20	217	Rat	*ARID4A*	Retinoblastoma-binding protein 1 isoform I
NM_207661	13	1030	Rat	*FLJ11806*	Nuclear protein UKp68 isoform 3
NM_020990	4	129	Rat	*CKMT1B*	Creatine kinase, mitochondrial 1B precursor
NM_005926	5	101	Rat	*MFAP1*	Microfibrillar-associated protein 1
NM_005881	9	80	Rat	*BCKDK*	Branched chain ketoacid dehydrogenase kinase

(continued)

Table 1. *Continued*

RefSeq ID	Position	Size	Species	Symbol	Full name
NM_000546[a]	6	568	Rat	*TP53*	Tumor protein p53
NM_001961	13	80	Rat	*EEF2*	Eukaryotic translation elongation factor 2
NM_001961	11	1156	Rat	*EEF2*	Eukaryotic translation elongation factor 2
NM_182513	3	118	Rat	*SPBC24*	Spindle pole body component 24 homolog
NM_001436	3	86	Rat	*FBL*	Fibrillarin
NM_032034	4	80	Rat	*SLC4A11*	Solute carrier family 4 member 11
NM_181801[b]	4	108	Rat	*UBE2C*	Ubiquitin-conjugating enzyme E2C isoform 4
NM_007098	26	4829	Rat	*CLTCL1*	Clathrin, heavy polypeptide-like 1 isoform b
NM_014303	5	91	Rat	*PES1*	Pescadillo homolog 1, containing BRCT domain
NM_001379[a]	36	797	Rat	*DNMT1*	DNA (cytosine-5-)-methyltransferase 1
NM_001469	4	264	Rat	*XRCC6*	ATP-dependent DNA helicase II, 70-kDa subunit
NM_024319	1	84	Rodent	*C1orf35*	Hypothetical protein LOC79169
NM_016252	7	77	Rodent	*BIRC6*	Baculoviral IAP repeat-containing 6
NM_014763	1	191	Rodent	*MRPL19*	Mitochondrial ribosomal protein L19
NM_145212	5	383	Rodent	*MRPL30*	Mitochondrial ribosomal protein L30
NM_006773	6	85	Rodent	*DDX18*	DEAD (Asp-Glu-Ala-Asp) box polypeptide 18
NM_012290	15	82	Rodent	*TLK1*	Tousled-like kinase 1
NM_001090	16	142	Rodent	*ABCF1*	ATP-binding cassette, subfamily F, member 1
NM_022551	3	81	Rodent	*RPS18*	Ribosomal protein S18
NM_001634	6	291	Rodent	*AMD1*	S-adenosylmethionine decarboxylase 1 isoform 1
NM_001010	5	105	Rodent	*RPS6*	Ribosomal protein S6
NM_004357	8	104	Rodent	*CD151*	CD151 antigen
NM_015104	5	129	Rodent	*KIAA0404*	Hypothetical protein LOC23130
NM_020680	3	80	Rodent	*SCYL1*	SCY1-like 1
NM_000920	14	492	Rodent	*PC*	Pyruvate carboxylase precursor
NM_001166	2	94	Rodent	*BIRC2*	Baculoviral IAP repeat-containing protein 2
NM_002046[a]	3	90	Rodent	*GAPDH*	Glyceraldehyde-3-phosphate dehydrogenase
NM_002046[a]	6	92	Rodent	*GAPDH*	Glyceraldehyde-3-phosphate dehydrogenase
NM_053275	3	104	Rodent	*RPLP0*	Ribosomal protein P0
NM_001312[a]	2	89	Rodent	*CRIP2*	Cysteine-rich protein 2
NM_001003	3	140	Rodent	*RPLP1*	Ribosomal protein P1 isoform 1
NM_002952[a]	4	79	Rodent	*RPS2*	Ribosomal protein S2
NM_024860[a]	2	72	Rodent	*SETD6*	Hypothetical protein LOC79918
NM_024805[a]	2	619	Rodent	*C18orf22*	Hypothetical protein LOC79863
NM_002819[a]	4	83	Rodent	*PTBP1*	Polypyrimidine tract-binding protein 1 isoform
NM_002695	3	803	Rodent	*POLR2E*	DNA directed RNA polymerase II polypeptide E
NM_003938	13	70	Rodent	*AP3D1*	Adaptor-related protein complex 3, δ1
NM_020170	8	428	Rodent	*NCLN*	Nicalin
NM_003685	12	104	Rodent	*KHSRP*	KH-type splicing regulatory protein
NM_032285	1	150	Rodent	*MGC3207*	Hypothetical protein LOC84245 isoform 2
NM_003333	4	84	Rodent	*UBA52*	Ubiquitin and ribosomal protein L40 precursor
NM_015965	4	236	Rodent	*NDUFA13*	Cell death-regulatory protein GRIM19
NM_000979	5	132	Rodent	*RPL18*	Ribosomal protein L18
NM_005560	68	83	Rodent	*LAMA5*	Laminin α5
NM_033405	10	97	Rodent	*PRIC285*	PPAR-α interacting complex protein 285
NM_008084	5	85	Human	*LOC14433*	Similar to glyceraldehyde-3-phosphate
NM_145370	4	86	Human	*Gps1*	G protein pathway suppressor 1
NM_031170	7	263	Human	*Krt2–8*	Keratin complex 2, basic, gene 8
NM_027350	3	112	Human	*Nars*	Asparaginyl-tRNA synthetase

[a]Deletion disrupts a predicted AS event based on EST evidence (ExonWalk).
[b]Deletion disrupts a known AS event (RefSeq).

outgroup for changes occurring in human, mouse, or rat and by using chicken as the outgroup for changes occurring in dog.

Figure 1 shows an example of an intron deletion event occurring in mouse displayed in the vertebrate MultiZ alignment track of the UCSC Genome Browser. This case illustrates common misalignments close to the splice sites, which is the reason we allowed for a 25-bp margin of error in the distance between exon edges in the target species during the search for intron loss and gain (see Methods). To confirm each gain/loss event, we extracted the original genomic sequences from the assemblies and used ClustalW to realign the reference species intron and 100 bp of flanking upstream and downstream sequences with the homologous target species region. The data for all the alignments are available in Supplemental data online. Our analysis shows that at least 117 of the detected intron losses are exact. The remaining three cases are also likely to be exact losses but fall into regions of relatively poor quality genomic sequence and require single base insertion/deletion events in the alignments.

Rates of intron loss/gain

We find a very low rate of intron loss throughout the mammalian evolution and no evidence for intron gain. Based on the total number of donor/acceptor splice site pairs identified in the alignments (146,964, 141,942, 146,727, and 124,474 for mouse, rat, dog, and human, respectively), we determined the rates for intron loss per million years per intron as follows: 5.32×10^{-6} for the mouse–rat common ancestor, 6.58×10^{-6} for mouse, 1.08×10^{-5} for rat, 5.30×10^{-7} for dog, and 4.28×10^{-7} for human. These estimates assume that human and dog lineages

Table 2. Multiple consecutive intron deletions

RefSeq	Pos.	Size	Loss	Symbol	Full name
NM_012311	5	3038	Rodent	*KIN*	KIN, antigenic determinant of recA
NM_012311	6	859	Rodent	*KIN*	KIN, antigenic determinant of recA
NM_012311	7	5485	Rodent	*KIN*	KIN, antigenic determinant of recA
NM_012311	8	3112	Rodent	*KIN*	KIN, antigenic determinant of recA
NM_012311	9	2261	Rodent	*KIN*	KIN, antigenic determinant of recA
NM_012311	10	1166	Rodent	*KIN*	KIN, antigenic determinant of recA
NM_012311	11	2466	Rodent	*KIN*	KIN, antigenic determinant of recA
NM_012311	12	3747	Rodent	*KIN*	KIN, antigenic determinant of recA
NM_005926	3	2154	Mouse	*MFAP1*	Microfibrillar-associated protein 1
NM_005926	2	256	Mouse	*MFAP1*	Microfibrillar-associated protein 1

diverged 95 Mya, human and rodent 75 Mya (Waterston et al. 2002), and mouse and rat 30 Mya (Nei et al. 2001; Springer et al. 2003). In order to assess whether the rates are proportional to generation time, we multiplied these rates by the age of sexual maturity of each organism (⅙, ⅓, 3, and 12 yr for mouse, rat, dog, and human) and normalized the resulting figures, so that the rate for human is equal to one. (Note that we are making a somewhat simplistic assumption that the ages behaved proportionally during the evolution of each lineage). We obtain ratios of 0.21, 0.70, 0.31, and 1 for mouse, rat, dog, and human, respectively. It appears that generation time is not the only factor affecting the rate of intron loss. Other possible factors include the relative activity of reverse transcriptase within each lineage. We note that the rat genome has a higher density of LINE elements, which encode their own reverse transcriptase, than the mouse genome (data not shown), possibly resulting in higher amounts of reverse-transcribed cDNA available for recombination.

Projected sizes of deleted introns

One of the most striking characteristics distinguishing the deleted introns is their extremely small size. The mean size of a human intron is 6259 bp, while the deleted cases were on average 355 bases long (in human). Figure 2 illustrates the difference in projected size distribution of deleted introns and that of all introns. The difference in the distributions is highly statistically significant (Student's *t*-test assuming unequal variances, $T = -57.3$, df = 208, two-tailed $P < 10^{-10}$). Most of the deleted introns (81 out of 120) are <150 bases. We further investigated five cases of unusual intron deletions that exceeded 1000 bp in length (5968, 7100, 1030, 1158, and 4380 nucleotides in genes *FAM54A*, *TAF5L*, *FLJ11806*, *EEF2*, and *CLTCL1*, respectively). Four of those cases occurred in the rat lineage and one in mouse. We identified the corresponding intron in the closest relative

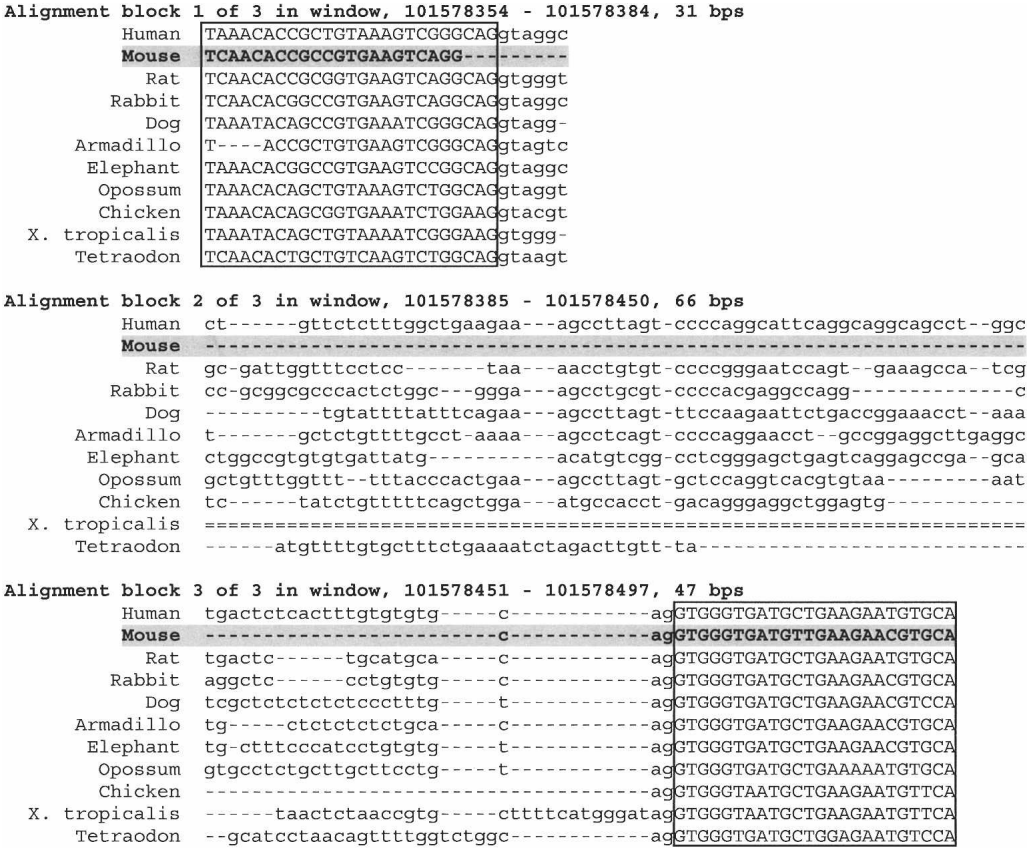

Figure 1. An example of intron loss in the mouse ortholog of the human DYNC1H1 gene, visualized in the UCSC Genome Browser display of multispecies alignments. Uppercase, boxed sequences correspond to exons. Note that the alignment is inexact at the splice sites, resulting in an artifactual 3-bp intron length in mouse, which necessitates an approximate search strategy (described in Methods), and realignment of sequences using an appropriate parameter choice in order to confirm all candidate intron deletions. Realigned sequences of the introns and neighboring exons for all 120 cases of intron loss are provided in Supplemental data.

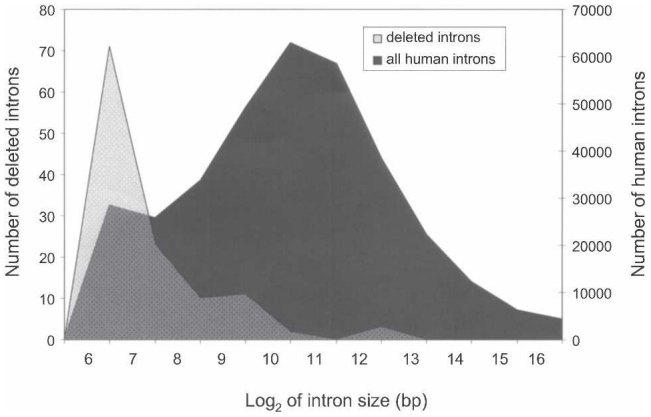

Figure 2. Log_2(size) distribution of all introns (black) versus deleted introns (gray). The deleted introns are unusually short and much shorter than the human genome average.

(mouse, in case the loss occurred in rat, and vice versa) and observed that the introns in the closest relative were actually considerably shorter than in human (984, 4902, 224, 134, and 79 bases, respectively) and hence were likely to be short at the time of their deletion. This suggests that the size of the intron must be an important factor affecting the underlying molecular mechanism of deletion.

Intron phases

Introns can be classified as phase 0 (inserted between two codons), phase 1 (after the first base of a codon), or phase 2 (after the second base). We examined the phase distribution of the 116 deleted introns present in human from Table 1 and compared it to the phase distribution of all introns from the RefSeq data set. The proportions for the deleted introns were 0.52, 0.26, and 0.22 for phases 0, 1, and 2, respectively, while the ratios for the genome average were 0.46, 0.32, and 0.22. The distribution of phases of deleted introns did not differ significantly from the expected ($\chi^2 = 2.24$, $df = 2$, $P = 0.33$). Intron deletions in mammals appear to occur randomly with respect to their phase. This finding contrasts with some earlier observations of phase 0–biased intron loss (Lynch 2002; Roy and Gilbert 2005b).

Positions of deleted introns within genes

We defined the relative position of each intron within a gene as the position, n, of the intron (measured from the 5′ end of the CDS; i.e., for the first CDS intron, $n = 1$), divided by the total number of CDS introns, N. In order to obtain a symmetrical distribution centered about ½, we subtracted ½ from the numerator. Hence, the adjusted relative position, $(n - ½)/N$ has a range between $1/(2N)$ and $1 - 1/(2N)$, and an expectation of ½. We used a χ^2 test to compare the proportion of deletions in the 5′ half of the gene versus the 3′ half. We found that the positions of deleted introns were significantly skewed toward the 3′ half of each gene ($\chi^2 = 7.76$, $df = 1$, $P = 0.0053$). Seventy-three of the intron losses occurred closer to the 3′ end of genes, compared with 43 that were closer to the 5′ end.

Splice site characteristics

We examined the distributions of bases around both splice sites and compared them with the distributions for all introns. We found that the consensus at the 5′ splice site was not significantly different from the control. However, at the 3′ splice site, the two positions following the acceptor AG dinucleotide had a significantly greater frequency of the bases G ($\chi^2 = 3.82$, $df = 1$, $P = 0.05$) and T ($\chi^2 = 4.93$, $df = 1$, $P = 0.03$), respectively. Since the GT at the 3′ splice site is very often preceded by an AG, this stronger consensus sequence may have served to promote a recombination event occurring between the two splice sites, leading to the deletion of the intron.

Expression patterns and ontology of genes undergoing intron loss

We used the EASE (Hosack et al. 2003) interface to classify our genes into GO categories (biological process) and characterize the types of genes that undergo intron deletion events. EASE calculates overrepresentation statistics for each GO category using an EASE score, which approximates a *P*-value by using the upper bound of the distribution of Jackknife Fisher exact probabilities. In Table 3, we list the most overrepresented biological processes (EASE score < 0.05). We note that most of the genes with intron deletions are involved in biosynthesis, metabolism, translation, transcription, and RNA processing. All of the overrepresented categories correspond to ubiquitous housekeeping functions, suggesting that intron deletion events occur predominantly in genes that are both highly expressed and expressed in the germline. In order to further confirm this hypothesis, we utilized microarray expression data available from SymAtlas (Su et al. 2002) to determine the expression intensities and breadths of the candidate genes. Since germline gene expression levels are not known, we used averaged gcRMA (Robust Multichip Average with GC correction) expression over all tissues as a proxy of germline expression (Majewski 2003) and compared the averages of the intron-deleted sample to all genes. The average gcRMA expression level was of 952 overall, and significantly higher, 9560, for the genes with intron deletion (Student's $t = -4.3$, $df = 108$, $P = 3.58 \times 10^{-5)}$. In order to study the breadth of expression, we used MASS 5.0 present/absent calls from >300 tissues and cell

Table 3. Overrepresented GO biological processes

GO biological process	List hits	Population hits	EASE score[a]
Protein biosynthesis	19	650	6.2E − 7
Biosynthesis	26	1199	6.4E − 7
Macromolecule biosynthesis	22	1002	5.7E − 6
Metabolism	75	7637	2.7E − 5
Translation	9	236	2.7E − 4
Pol II promoter transcription	12	477	5.6E − 4
Nucleic acid metabolism	38	3429	3.0E − 3
RNA processing	10	430	3.5E − 3
RNA metabolism	10	460	5.4E − 3
Protein metabolism	31	2696	5.7E − 3
Nucleocytoplasmic transport	5	108	7.3E − 3
Spermine biosynthesis	2	2	1.4E − 2
Translational elongation	3	27	1.6E − 2
Spermine metabolism	2	3	2.1E − 2
Spermidine metabolism	2	3	2.1E − 2
Spermidine biosynthesis	2	3	2.1E − 2
Intracellular transport	10	613	3.0E − 2
Transcription	26	2426	3.0E − 2
Polyamine biosynthesis	2	7	4.9E − 2

[a]A *P*-value approximation, uncorrected for multiple testing, based on the number of hits within a category for our list of 99 genes (which could be identified from our data set by their Locus Link ID), compared with total hits within a population of 13,802 genes (null expectation).

lines and determined the fraction of tissues where expression was positively detected (present). Again, we compared the expression breadth for genes with intron deletions (0.54) to that of all genes (0.26) and found a highly significant difference ($t = -6.9$, $df = 92$, $P = 7.15 \times 10^{-10}$). Thus intron deletions occur preferentially in genes with housekeeping functions, which have experimentally been determined to be both highly and broadly expressed.

Genes experiencing frequent intron loss: *GAPDH*

We performed a detailed analysis of the *GAPDH* gene, where we found evidence of multiple, independent intron losses occurring in mouse, human, and rat. *GAPDH* is a known, very highly expressed housekeeping gene, which supports the hypothesis that expression in the germline is essential for intron loss. We extracted genomic DNA and mRNA *GAPDH* sequences for 15 vertebrate species and used multiple sequence alignment to reconstruct the intron/exon structure of the gene in each species (Fig. 3). A Dollo parsimony approach (assuming a single appearance of the derived character—intron) suggests that there were no gain events throughout vertebrates but numerous losses, including several independent losses of the same intron (intron 9 of the ancestral gene). The result also suggests that the phenomenon of intron loss in vertebrates (at least within this gene) may be accelerated in the mammalian branch.

Does intron loss disrupt AS?

We identified two cases of intron losses disrupting known (RefSeq-confirmed) AS events that alter the predicted amino acid sequence of the gene (for details, see footnote to Table 1). We also detected 20 losses that disrupt predicted (EST-based) events alternatively processed in human. If intron losses occurred randomly, without any regard to preserving AS, the expected number, based on all the RefSeq introns used in this analysis is a disruption of four known events and 17 predicted events. It is unexpected that the number of observed losses that disrupt predicted AS events is actually slightly greater than the null expectation. However, since our ability to predict AS events is highly dependent on the availability of mRNAs and ESTs, and the set of genes undergoing intron losses is extremely highly and broadly expressed, there is likely to be a bias in the annotation of the deleted sample. That is, because of their high expression levels, genes experiencing intron loss have deeper EST coverage and are better annotated with respect to AS than the genome average.

In view of the annotation bias, it is difficult to conclude whether the disrupted predicted AS events are truly functional or constitute an artifact of deep EST coverage and the presence of inadvertent splicing errors. It is also possible that, since AS may be only weakly conserved across species (Pan et al. 2005), a predicted disruption of AS in humans may have no effect on AS in the species where the deletion occurred.

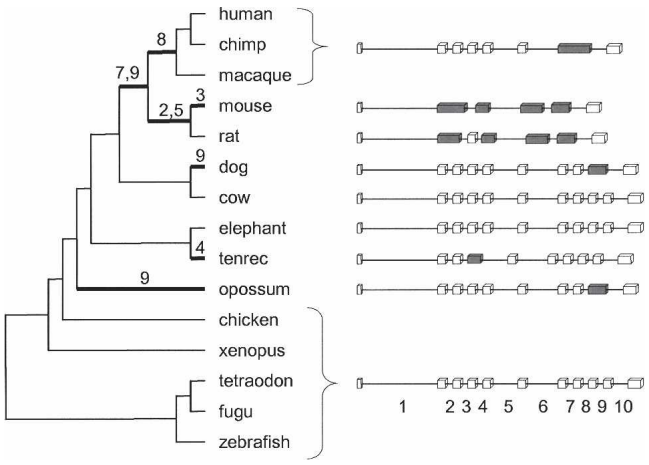

Figure 3. The evolution of the intron–exon structure of the *GAPDH* gene throughout the vertebrate phylogeny. The numbers on the branches indicate the inferred deletion events. The introns are numbered according to their position within the coding sequence of the ancestral gene.

Discussion

We identify >100 cases of intron loss in the four examined mammalian species. Our approach, based on mapping of known human genes to whole-genome sequence alignments of multiple species, allows us to utilize the annotation information from well-studied model species, such as human and mouse, and predict gene structure in other, relatively poorly annotated species. Using our method, we recover all six intron deletion events detected in a smaller scale study (Roy et al. 2003) and extend previous conclusions regarding the patterns of intron loss in mammals. There are several remarkable characteristics of our data set: (1) losses appear to occur almost exclusively for small introns; (2) essentially all of our examples of loss are consistent with an exact deletion event; (3) the loss events are biased toward the 3′ ends of genes, but can be found at all positions; (4) genes that are associated with intron loss events are generally highly expressed and have housekeeping functions; (5) the rate of intron loss is related to (but not fully explained by) the generation time of the organisms and follows a pattern similar to spontaneous mutation; and (6) all of the differences in gene structure are consistent with intron loss events—no detectable intron insertions have occurred in human or mouse since the divergence of their lineages. Below, we discuss some implications of these findings.

Mechanism of intron loss

It has been suggested that intron loss may be mediated either by genomic deletion events or recombination of the genomic locus with a reverse-transcribed, processed mRNA molecule of the gene (Logsdon Jr. et al. 1998). Our analysis suggests that at least 98% (and possibly all) of the observed deletions are exact. In addition, we do not find any evidence for inexact deletions, which would retain a small part of the intron or remove parts of neighboring exons. It has been argued (Roy et al. 2003) that random genomic deletion events would be unlikely to always result in exact intron losses. This is even more evident in our large data set. It would be extremely unlikely that, if intron loss were generally mediated by random deletions, we would not recover any cases of inexact losses. Even in the presence of purifying selection against such potentially deleterious events, it seems plausible that some minor insertion/deletions of the boundary sequence, particularly ones that do not alter the reading frame, would be evolutionarily neutral. Thus the exact character of the detected intron loss events supports the latter model, i.e., recombination with an intronless cDNA of the gene.

The small projected size of the introns provides another insight into the mechanism of loss. It is well documented that genetic recombination events occur less frequently in the pres-

ence of mismatches, insertions, or deletions within the recombining substrates (Majewski and Cohan 1999). We propose that in the cases of intron loss, recombination with cDNA is much more likely if the introns are small, resulting in a high relative effective proportion of sequence identity.

We also find that genes susceptible to intron loss tend to be involved in housekeeping functions and expressed at relatively high levels. Again, high level of expression most likely results in relatively high levels of reverse-transcribed copies of the gene, leading to an increased probability of recombination. A similar effect has been demonstrated for the frequency of processed pseudogenes (Zhang et al. 2003). Furthermore, in order for the recombination events to result in intron losses that are transmitted to the next generation and have a chance to increase in frequency in the population, the loss events must occur in the germline, as opposed to somatic cells. Thus, germline expression of the gene would be an essential condition for intron loss. In accordance with this prediction, we find that our intron-deleted data set is highly enriched in housekeeping (ubiquitously expressed) genes. Thus both the expression levels and the expression patterns of the genes support the recombination-mediated model of intron loss.

Finally, we find that the position of the lost introns is significantly biased toward the 3′ ends of the genes. This is in accordance with recent studies of lower eukaryotes (Sverdlov et al. 2004; Roy and Gilbert 2005b) and again supports intron loss being mediated by recombination, since reverse transcription of the mRNA is believed to occur preferentially from the 3′ end (Weiner et al. 1986). However, this result may also reflect the bias in distribution of intron sizes (first introns are generally longer and more difficult to remove by recombination) and selective pressures against deleting potentially regulatory regions, which may be present close to the 5′ termini of genes (Majewski and Ott 2002).

Multiple intron losses in single genes

We identify several cases of multiple intron loss in single genes, which we classify as two distinct types of events. The first, concerted type, is exemplified by a loss of multiple successive introns: eight successive introns of the *HsKin17* gene in the rodent lineage (before the rat–mouse divergence) and two introns of the *MFAP1* gene. The introns lost in the concerted events are generally considerably longer than those involved in the usual single-loss cases. We propose that this type of rearrangement event occurs as a result of recombination with nearly complete cDNA, whereas in the more typical cases the recombining cDNAs may be incomplete or fragmented. The overall greater available length of substrate involved in the concerted losses may promote recombination despite the presence of long unmatched intronic regions.

The second, punctuated type of multiple losses, involves the more typical short introns and is exemplified by the *GAPDH* and *EEF2* genes. Those genes exhibit evidence of several independent intron losses, occurring either in the same species, or distinct lineages. A similar phenomenon has been previously observed in the *white* gene in insects (Krzywinski and Besansky 2002). For our most extreme case, *GAPDH* (seven introns within the human CDS), we manually examined the genomic species alignments of other mammalian and vertebrates species, which we had not considered in the genome-wide study. We found evidence for three independent losses of the ancestral intron 9 (dog, primate/rodent lineage, possum), and a total of nine losses that occurred in this gene in mammals (Fig. 3). The fish, frog, chicken, elephant, and cow have retained the same ancestral structure. *GAPDH* is a known, extremely highly expressed housekeeping gene, often used as a control in gene expression studies due to its consistent elevated expression in all cell types. Consistent with general observations for highly expressed genes, *GAPDH* has multiple annotated pseudogenes (Kent et al. 2003; Zhang et al. 2003), further supporting the hypothesis that high levels of potentially reverse-transcribed mRNA in germ cells lead to an elevated probability of not only retrotransposition but also intron loss.

Selection favoring intron loss?

The preferential intron loss in highly expressed housekeeping genes is also consistent with selection for transcription efficiency favoring the resulting short transcript (Castillo-Davis et al. 2002). While the selection pressure and the increased likelihood of recombination in highly transcribed genes are not mutually exclusive and, in theory, may both contribute to the association of intron loss and expression levels, it seems unlikely that selection is a major force responsible for the observed intron losses. Most of the deleted introns are extremely short (~100 bp), while much longer introns are present in the corresponding genes and had not been deleted. Selection alone would favor the loss of longer introns. In the example of the *GAPDH* gene, a loss of an 82-bp intron from a 3783-bp transcript would result in only a very modest 2% decrease in the time of transcription. In comparison, loss of the first intron fully contained within the CDS (~1700 bp) could result in a 45% reduction. We propose that the availability of cDNA and the length of unmatched intronic sequences in the recombining strands are the primary limiting factors in the process of intron loss. Once the gene conversion event occurs, selection may be an additional force increasing the probability of fixation of such events. However, because of low effective population sizes, genetic drift, rather than selection (Jeffares et al. 2006; Roy and Gilbert 2006), is much more likely to be the main determinant of fixation rates in mammals.

Rates of intron loss and gain

The rate of intron loss in mammals appears extremely slow. The fastest genome-wide rate, in the rat lineage is approximately one intron loss per 1.53 Myr. This corresponds to one intron loss per 217,644 Myr, per intronic site (based on 141,942 introns identified in the human/rat comparison). We note that the rates are not clocklike and appear to be dependent on the generation time of each lineage: rodent > dog > human, but they are also likely to be affected by other factors, such as the effective population size (in the absence of selection, the rate of fixation of intron losses is expected to be inversely proportional to the population size). At the current rate, it would take $>10^{12}$ yr for the human genome to shed half of its introns. Hence, intron loss/gain does not appear to be a major factor in mammalian evolution.

Since we have not detected any cases of intron gain, we estimate the process to proceed considerably slower than intron loss. Our approach would allow us to detect intron gain events occurring in the mouse/rodent lineage, after its divergence from the human lineage. Since no cases of intron gain (and 63 losses) were found, the genome-wide rate of gain is at least 60-fold slower than the rate of loss. Although some claims to intron gain in mammals have been reported (Veeramachaneni and Makalow-

ski 2005), more recent analyses (Shepelev and Fedorov 2006), including ours, do not confirm such findings. Our observations support the premise that modern introns are evolutionarily inert and, having expanded through early eukaryotic genomes (or being inherited through earlier yet ancestors), have been gradually, albeit very slowly, disappearing within mammalian lineages over at least the past 100 Myr.

Our results in mammals bear similarity to two recent studies of intron dynamics. Cho et al. (2004) investigated closely related nematode species and found frequent and recurrent intron loss and a much slower (fivefold to 10-fold) rate of putative intron gain. Roy and Hartl (2006) studied two relatively closely related Plasmodium species and found a very slow rate of intron loss and possibly no gain during ~100 Myr of evolution. On the other hand, these results contrast with estimates from fungi (Nielsen et al. 2004) and more distantly related eukaryotic clades (Fedorov et al. 2002; Rogozin et al. 2003; Babenko et al. 2004; Roy and Gilbert 2005c; Yoshihama et al. 2006), which suggest prevalence of intron gain over intron loss over periods >200 Myr.

The combined data on intron dynamics implies that introns have not been actively proliferating during the past 100 Myr of evolution in any of the studied species. Most of inferred intron gains must have occurred significantly earlier. However, ours and other studies in closely related species show a pattern of recurrent intron loss, particularly in highly expressed housekeeping genes that are most likely to recombine with their reverse transcribed cDNA. Because of their high degree of conservation at both sequence and functional levels, such genes are the ones most often used in gene structure comparisons (Rogozin et al. 2003; Yoshihama et al. 2006). The process of loss is likely to be much more accelerated in organisms with high reproductive rates and large population sizes where selection for reduced transcript length, rather than genetic drift, will lead to fixation of losses in populations. It is also likely to be rapid in species with short introns, where recombination with intronless substrates is more efficient. In the presence of recurrent parallel loss, studies using parsimony methods and large evolutionary distances (Rogozin et al. 2003; Coghlan and Wolfe 2004; Yoshihama et al. 2006) will certainly underestimate the rate of loss and overestimate the rate of intron gains. Even maximum likelihood approaches (Nguyen et al. 2005; Roy and Gilbert 2005a; Csuros 2006) are not fully immune to this error and will overestimate the gain rate particularly if they fail to account for the rate variation among sites (Yang 1996), which we have shown to occur in the case of intron loss. While we are still very far from resolving the debate over the age and origin of spliceosomal introns, our analysis suggests that many studies may mistake parallel intron losses for gains, and that as a result most introns may be significantly older than we currently believe.

Methods

DNA sequences and interspecies alignments

We used the RefSeq annotation of the human genomic sequence to extract coding sequences of human genes (Hinrichs et al. 2006). Only the sequences which could be in silico translated into their predicted protein were retained. This strategy resulted in a high confidence, nonredundant data set of 17,242 human autosomal genes, containing 152,146 distinct introns within their coding sequence. We based our analysis on the four available highest-quality mammalian genome assemblies: human (hg17), mouse (mm7), rat (rn3), and dog (canFam2). We mapped the well-annotated human genes onto the genome-wide alignments present within the 17-way MultiZ (Blanchette et al. 2004) alignment tracks in order to determine the intron–exon structures in the target species. We considered only introns that were flanked by coding, or partially coding, exons, since noncoding UTR sequences are poorly conserved (and often not conserved) among species and provide poor anchors for detecting splice sites within alignments. We also performed the reverse analysis by mapping a set of 16,068 mouse RefSeq genes (129,336 CDS introns) onto the mouse vs. human genomic sequence alignments.

We used the following criteria to detect intron loss events in the target sequence (or gain in the reference sequence): (1) for each reference species intron, we identified the positions of both the donor and acceptor splice sites within the MultiZ alignment; (2) within the target species, we flagged an intron as potentially lost if the distance between the donor and acceptor sites was lower than a predetermined cutoff of 25 bp. The latter condition was necessary since alignments are often imperfect at the exon–intron boundary (Fig. 1). In particular, especially in the case of intron loss events, the last 2 bp of an exon, which have an AG consensus, tend to align with the downstream intronic acceptor site (also AG), but more serious misalignments are also common. Nevertheless, allowing a margin of 25 bp did not introduce any false—positive results (as manually verified in the final curated results), since sequences <25 bp cannot be efficiently spliced in mammals (Lim and Burge 2001) and correspond to imperfect alignments, rather than actual introns. Using the above first pass search criteria, we identified 623 cases of potential intron loss/gain.

Since the genome assemblies and the resulting alignment contain numerous sequencing, assembly, and alignment artifacts, all potential intron loss events were further filtered based on the quality of the underlying alignment. In the process of constructing the BLASTZ alignments, gaps in the sequences may be filled in using secondary (non-syntenic) sequences. This significantly increases the proportion of aligned sequences but also results in an increased probability of introducing alignment errors. Thus, only potential intron loss cases that mapped to the highest confidence, top, syntenic, long (encompassing at least two neighboring genes) alignment nets (Kent et al. 2003) were retained for further analysis. Cases occurring in genes that were aligned to multiple or nonsyntenic portions of target genomes, which could potentially constitute alignments to duplicate genes or pseudogenes, were rejected. This strategy resulted in 157 cases of intron loss/gain, of which 35 occurred in both rat and mouse, for a total of 122 events.

For all the candidates, we extracted the sequence of 100 bp flanking the intronic site from the genomic sequence assembly and used ClustalW (Thompson et al. 1994) with high gap opening penalty (80) and low gap extension penalty (zero) to align it to the human intron-containing sequence and visualize the detailed evidence for intron loss. After performing some minor supervised adjustments, mainly correcting the misalignment of the terminal AG of an upstream exon with the downstream acceptor site (see above), this allowed us to confirm the deletion events and demonstrate that essentially all of the events are cases of exact deletion, with no alteration to the coding sequence. All of the 120 isolated intron loss/gain events were successfully validated using the sequence alignments (Supplemental data).

Characterization of genes involved in loss events

In order to functionally classify the genes involved in intron loss events, we used the EASE (Hosack et al. 2003) interface to the

Gene Ontology annotation. We identified the GO categories with the highest support—lowest EASE score—for overrepresentation by the genes within our list, compared with all known genes.

In order to approximate expression levels and expression breadth of the genes, we used microarray expression data from SymAtlas (Su et al. 2002). Although the relevant variable is the expression level in the germline, this information is currently not available. As a proxy for gene expression levels, we used the mean values of gcRMA summaries across all tissues studied. Note that because of developmental history of germ cells, testes- and ovary-specific expression levels may not be the appropriate indicator of germline expression, and a global average expression may provide a better estimate (Majewski 2003), particularly in the case of housekeeping genes. As an estimate of expression breadth, we used the present/absent calls from the MASS 5.0 summaries and, for each gene, calculated the percentage of tissues where expression was detected.

Intron loss and AS

In order to study the relationship of intron loss and AS, we cross-referenced the set of lost intron positions with alternative gene isoforms present in the RefSeq data set. We identified all the introns where the deletion in the target species disrupts a known AS event in human. For example, deletion of an intron would prevent alternative usage of the adjacent exons (cassette events), as well as alternative (cryptic) splice site usage of the adjacent splice sites. We further limited the AS events of interest to only those that altered the predicted amino acid sequence of the gene. In order to obtain a background genome-wide estimate for the probability of any intron loss disrupting AS, we also determined how many introns from our entire input data set border alternatively spliced exons that would be disrupted by a deletion.

While the RefSeq set of genes is manually curated and highly accurate, it contains relatively few alternatively spliced isoforms. Hence we also analyzed predicted AS events from the ExonWalk annotation of the UCSC Database. Briefly, the ExonWalk program merges EST and cDNA evidence together to predict full-length isoforms, including alternative transcripts. To predict transcripts that are biologically functional, rather than the result of technical or biological noise, ExonWalk requires that every intron and exon be: (1) present in cDNA libraries of another organism (i.e., also present in mouse), (2) have three separate cDNA GenBank entries supporting it, or (3) be evolving like a coding exon as determined by the Exoniphy program (Siepel and Haussler 2004). Once the transcripts are predicted an open reading frame finder is used to find the best open reading frame. Transcripts that are targets for nonsense mediated decay are filtered. We further filtered out all predicted transcripts that did not begin with an ATG and did not end with a stop codon.

Acknowledgments

This research was supported by funds from the Canadian Institute of Health Research and the Canada Research Chairs program.

References

Babenko, V.N., Rogozin, I.B., Mekhedov, S.L., and Koonin, E.V. 2004. Prevalence of intron gain over intron loss in the evolution of paralogous gene families. *Nucleic Acids Res.* **32:** 3724–3733.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res.* **14:** 708–715.

Castillo-Davis, C.I., Mekhedov, S.L., Hartl, D.L., Koonin, E.V., and Kondrashov, F.A. 2002. Selection for short introns in highly expressed genes. *Nat. Genet.* **31:** 415–418.

Cavalier-Smith, T. 1991. Intron phylogeny: A new hypothesis. *Trends Genet.* **7:** 145–148.

Cho, S., Jin, S.W., Cohen, A., and Ellis, R.E. 2004. A phylogeny of *Caenorhabditis* reveals frequent loss of introns during nematode evolution. *Genome Res.* **14:** 1207–1220.

Coghlan, A. and Wolfe, K.H. 2004. Origins of recently gained introns in *Caenorhabditis*. *Proc. Natl. Acad. Sci.* **101:** 11362–11367.

Comeron, J.M. and Kreitman, M. 2000. The correlation between intron length and recombination in *Drosophila*. Dynamic equilibrium between mutational and selective forces. *Genetics* **156:** 1175–1190.

Csuros, M. 2006. On the estimation of intron evolution. *PLoS Comput. Biol.* **2:** e84; author reply e83.

de Souza, S.J., Long, M., Klein, R.J., Roy, S., Lin, S., and Gilbert, W. 1998. Toward a resolution of the introns early/late debate: Only phase zero introns are correlated with the structure of ancient proteins. *Proc. Natl. Acad. Sci.* **95:** 5094–5099.

Fedorov, A., Merican, A.F., and Gilbert, W. 2002. Large-scale comparison of intron positions among animal, plant, and fungal genes. *Proc. Natl. Acad. Sci.* **99:** 16128–16133.

Hinrichs, A.S., Karolchik, D., Baertsch, R., Barber, G.P., Bejerano, G., Clawson, H., Diekhans, M., Furey, T.S., Harte, R.A., Hsu, F., et al. 2006. The UCSC Genome Browser Database: Update 2006. *Nucleic Acids Res.* **34:** D590–D598.

Hosack, D.A., Dennis Jr., G., Sherman, B.T., Lane, H.C., and Lempicki, R.A. 2003. Identifying biological themes within lists of genes with EASE. *Genome Biol.* **4:** R70.

Jeffares, D.C., Mourier, T., and Penny, D. 2006. The biology of intron gain and loss. *Trends Genet.* **22:** 16–22.

Kent, W.J., Baertsch, R., Hinrichs, A., Miller, W., and Haussler, D. 2003. Evolution's cauldron: Duplication, deletion, and rearrangement in the mouse and human genomes. *Proc. Natl. Acad. Sci.* **100:** 11484–11489.

Kim, H., Klein, R., Majewski, J., and Ott, J. 2004. Estimating rates of alternative splicing in mammals and invertebrates. *Nat. Genet.* **36:** 915–916; author reply 916–917.

Krzywinski, J. and Besansky, N.J. 2002. Frequent intron loss in the white gene: A cautionary tale for phylogeneticists. *Mol. Biol. Evol.* **19:** 362–366.

Lim, L.P. and Burge, C.B. 2001. A computational analysis of sequence features involved in recognition of short introns. *Proc. Natl. Acad. Sci.* **98:** 11193–11198.

Logsdon Jr., J.M., Stoltzfus, A., and Doolittle, W.F. 1998. Molecular evolution: Recent cases of spliceosomal intron gain? *Curr. Biol.* **8:** R560–R563.

Lynch, M. 2002. Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci.* **99:** 6118–6123.

Majewski, J. 2003. Dependence of mutational asymmetry on gene-expression levels in the human genome. *Am. J. Hum. Genet.* **73:** 688–692.

Majewski, J. and Cohan, F.M. 1999. DNA sequence similarity requirements for interspecific recombination in *Bacillus*. *Genetics* **153:** 1525–1533.

Majewski, J. and Ott, J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* **12:** 1827–1836.

Nei, M., Xu, P., and Glazko, G. 2001. Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms. *Proc. Natl. Acad. Sci.* **98:** 2497–2502.

Nguyen, H.D., Yoshihama, M., and Kenmochi, N. 2005. New maximum likelihood estimators for eukaryotic intron evolution. *PLoS Comput. Biol.* **1:** e79.

Nielsen, C.B., Friedman, B., Birren, B., Burge, C.B., and Galagan, J.E. 2004. Patterns of intron gain and loss in fungi. *PLoS Biol.* **2:** e422.

Pan, Q., Bakowski, M.A., Morris, Q., Zhang, W., Frey, B.J., Hughes, T.R., and Blencowe, B.J. 2005. Alternative splicing of conserved exons is frequently species-specific in human and mouse. *Trends Genet.* **21:** 73–77.

Rogozin, I.B., Wolf, Y.I., Sorokin, A.V., Mirkin, B.G., and Koonin, E.V. 2003. Remarkable interkingdom conservation of intron positions and massive, lineage-specific intron loss and gain in eukaryotic evolution. *Curr. Biol.* **13:** 1512–1517.

Roy, S.W. and Gilbert, W. 2005a. Complex early genes. *Proc. Natl. Acad. Sci.* **102:** 1986–1991.

Roy, S.W. and Gilbert, W. 2005b. The pattern of intron loss. *Proc. Natl. Acad. Sci.* **102:** 713–718.

Roy, S.W. and Gilbert, W. 2005c. Rates of intron loss and gain:

Implications for early eukaryotic evolution. *Proc. Natl. Acad. Sci.* **102:** 5773–5778.
Roy, S.W. and Gilbert, W. 2006. The evolution of spliceosomal introns: Patterns, puzzles and progress. *Nat. Rev. Genet.* **7:** 211–221.
Roy, S.W. and Hartl, D.L. 2006. Very little intron loss/gain in Plasmodium: Intron loss/gain mutation rates and intron number. *Genome Res.* **16:** 750–756.
Roy, S.W., Fedorov, A., and Gilbert, W. 2003. Large-scale comparison of intron positions in mammalian genes shows intron loss but no gain. *Proc. Natl. Acad. Sci.* **100:** 7158–7162.
Shepelev, V. and Fedorov, A. 2006. Advances in the Exon-Intron Database (EID). *Brief. Bioinform.* **7:** 178–185.
Siepel, A. and Haussler, D. 2004. Combining phylogenetic and hidden Markov models in biosequence analysis. *J. Comput. Biol.* **11:** 413–428.
Springer, M.S., Murphy, W.J., Eizirik, E., and O'Brien, S.J. 2003. Placental mammal diversification and the cretaceous-tertiary boundary. *Proc. Natl. Acad. Sci.* **100:** 1056–1061.
Stoltzfus, A. 1994. Origin of introns—early or late. *Nature* **369:** 526–527; author reply 527–528.
Su, A.I., Cooke, M.P., Ching, K.A., Hakak, Y., Walker, J.R., Wiltshire, T., Orth, A.P., Vega, R.G., Sapinoso, L.M., Moqrich, A., et al. 2002. Large-scale analysis of the human and mouse transcriptomes. *Proc. Natl. Acad. Sci.* **99:** 4465–4470.
Sverdlov, A.V., Babenko, V.N., Rogozin, I.B., and Koonin, E.V. 2004. Preferential loss and gain of introns in 3′ portions of genes suggests a reverse-transcription mechanism of intron insertion. *Gene* **338:** 85–91.
Thompson, J.D., Higgins, D.G., and Gibson, T.J. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22:** 4673–4680.
Veeramachaneni, V. and Makalowski, W. 2005. DED: Database of Evolutionary Distances. *Nucleic Acids Res.* **33:** D442–D446.
Waterston, R.H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J.F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., et al. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.
Weiner, A.M., Deininger, P.L., and Efstratiadis, A. 1986. Nonviral retroposons: Genes, pseudogenes, and transposable elements generated by the reverse flow of genetic information. *Annu. Rev. Biochem.* **55:** 631–661.
Yang, Z. 1996. Maximum-likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* **42:** 587–596.
Yoshihama, M., Nakao, A., Nguyen, H.D., and Kenmochi, N. 2006. Analysis of ribosomal protein gene structures: Implications for intron evolution. *PLoS Genet.* **2:** e25.
Zhang, Z., Harrison, P.M., Liu, Y., and Gerstein, M. 2003. Millions of years of evolution preserved: A comprehensive catalog of the processed pseudogenes in the human genome. *Genome Res.* **13:** 2541–2558.

Received July 4, 2006; accepted in revised form September 5, 2006.