Supplementary Tables

Table S1. Alignment near exon-intron boundaries at which the results of *Spaln* differ from Projector annotations

Projector ID	Alignment

Hs.1.ENST
00000271555.2
	`AGCCGGGCTGgtgagtagag ccctttccag-GAGGCTGTGC`
	`AGCCGGGCTG.......... ..........GGAGGCTGTGC`

Hs.13.ENST
00000316546.1
	`GAAGGCGCAGgtggggcgcg aattttgcag-AAATGGAGCA`
	`GAAGGCGCAG.......... ..........GAAATGGAGCA`

Hs.17.ENST
00000225740.1
	`TGTGTCAAAGgtaatcctct ctgtccccag-GCGCCATGAG`
	`TGTGTCAAAG.......... ..........GGCGCCATGAG`

Hs.7.ENST
00000289604.1
	`AGCATTTTAAgtatggggaa tctcctccag-AATCCAGAAA`
	`AGCATTTTAA.......... ..........GAATCCAGAAA`

Hs.7.ENST
00000289604.1
	`GTTCTTAAAGgtagtgtaca tcccatacag-GCTCAAGAAT`
	`GTTCTTAAAG.......... ..........GGCTCAAGAAT`

Hs.8.ENST
00000323173.1
	`TGGACTGCACgtaagtagaa cttttctcag-GTTGGTGACA`
	`TGGACTGCAC.......... ..........GGTTGGTGACA`

Hs.20.ENST
00000201955.1
	`AAACTTTTAGgtgagatgtg ttttgtttag-GTCCTGGGAG`
	`AAACTTTTAG.......... ..........GGTCCTGGGAG`

Hs.20.ENST
00000201955.1
	`TCTTAAGCAGgtctgttgat ttgtttttag-AATTTTGATT`
	`TCTTAAGCAG.......... ..........GAATTTTGATT`

Hs.20.ENST
00000201955.1
	`AGAGATTTTGgtgagtgaaa ctcttttcag-AAGTTTTCTT`
	`AGAGATTTTG.......... ..........GAAGTTTTCTT`

Hs.1.ENST
00000295314.1
	`CACAGGGGAGgtgagaaggg tgttgcccag-GCTCTGTGAG`
	`CACAGGGGAG.......... ..........GGCTCTGTGAG`

Hs.22.ENST
00000215792.1
	`CCCCTACTCTgtcatctcag agtcactcag-GAACACCCTC`
	`CCCCTACTCT.......... ..........GGAACACCCTC`

Hs.2.ENST
00000233767.1
	`TGCGCGCAGAgtgagtggcc cttttcgcag-AAAGTTTCAT`
	`TGCGCGCAGA.......... ..........GAAAGTTTCAT`

Hs.2.ENST
00000233767.1
	`AAGGTGGCAGgtcagtaaat ttccctatag-GATGTCTCAG`
	`AAGGTGGCAG.......... ..........GGATGTCTCAG`

The upper and lower lines in each alignment indicate the genomic sequence and the cDNA sequence, respectively. Introns inferred by *Spaln* are shown in lower-case letters, whereas exons and cDNA sequences are shown in upper-case letters. The consensus dinucleotides at the ends of introns are underlined. In each case, the Projector annotation assigns the “g” at the 3’ end of intron to a part of the following exon.

Table S2. Default parameter values of *Spaln* and *SpalnX*

Parameter	Command Line Option	Value	
		Spaln	*SpalnX*
Nucleotide Match	-ym	2	2
Nucleotide Mismatch	-yn	-6	-2
Gap open penalty	-v	8	6
Gap extension penalty	-u	3	2
Factor given to an exon-boundary signal relative to sequence match score	-yy	4	4
Intron penalty	-yi	28	33
Minimum intron length	-yL	30	30

Table S3. Commonality among genomic loci detected by different methods

Methods	Common	Union	%
B-G	122981	124255	98.90
B-M	118345	124282	95.17
G-M	117621	124213	94.58
B-S	123402	124308	99.23
G-S	122689	124254	98.66
M-S	118167	124282	95.02
B-G-M	117341	124305	94.36
B-G-S	122401	124316	98.43
B-M-S	117811	124322	94.74
G-M-S	117198	124305	94.24
B-G-M-S	116941	124325	94.04

The second column indicates the number of queries that were commonly mapped by the methods shown in the first column, and the third column indicates the cardinality of their union. The fourth column indicates the percentage of commonly mapped queries of the 124355 total cDNAs examined. The methods used for mapping, *Blat*, *Gmap*, *Megablast*, and *Spaln*, are abbreviated by their initials.

Table S4. Error rates at the gene level in alignment between cDNA sequences and corresponding genomic segments

Noise (%)	Method										
	Blat	*Exalin*	*Exonerate*	*Gmap*	*GmapX*	*Sim4*	*Spaln*	*SpalnA*	*SpalnS*	*SpalnX*	*SpalnXL*
0	6.13	2.15	2.66	2.15	0.72	2.04	0.41	0.41	2.86	0.51	0.51
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	12.24	2.71	3.39	0.84	0.70	3.53	0.60	0.60	2.92	0.70	0.70
	0.81	1.13	1.56	0.46	0.38	0.81	0.31	0.31	0.06	0.31	0.31
2	20.25	3.39	3.63	1.91	1.77	5.11	0.59	0.70	2.90	0.72	1.04
	1.91	1.61	0.67	2.18	2.18	0.74	0.10	0.10	0.06	0.13	0.85
4	36.16	6.10	6.94	4.86	4.45	14.54	1.70	2.13	3.22	1.86	2.22
	3.13	2.22	0.69	4.03	4.00	2.46	0.78	0.89	0.28	0.63	0.85
6	52.03	15.49	12.20	7.47	6.92	21.01	2.66	3.34	3.41	2.68	6.10
	4.57	6.03	4.26	2.76	2.71	2.15	1.45	1.39	0.27	1.09	3.61
8	66.17	21.30	18.36	14.95	14.22	34.00	4.25	5.13	4.17	4.21	7.65
	2.97	7.00	6.19	5.40	5.40	4.62	1.97	2.13	1.59	1.85	2.30
10	78.66	27.25	23.59	20.01	18.92	49.37	5.73	7.09	3.94	5.15	8.89
	4.47	9.07	5.16	6.00	6.08	4.36	2.20	2.74	0.35	1.58	3.72

Presented are the percentages of incorrectly identified gene structures of the total of 978 human and mouse genes in Projector database under various levels of artificially introduced noise. Each cell indicates the average (upper line) and the standard deviation (lower line) of six experiments.

Table S5. Error rates at the exon level in alignment between cDNA sequences and corresponding genomic segments

Noise (%)	Methods										
	Blat	*Exalin*	*Exonerate*	*Gmap*	*GmapX*	*Sim4*	*Spaln*	*SpalnA*	*SpalnS*	*SpalnX*	*SpalnXL*
0	2.74	0.42	0.44	0.43	0.19	0.45	0.14	0.13	0.62	0.14	0.14
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	4.21	0.50	0.69	0.22	0.18	0.75	0.17	0.16	0.63	0.17	0.17
	0.26	0.17	0.60	0.07	0.05	0.14	0.04	0.05	0.01	0.04	0.04
2	5.82	0.57	0.62	0.33	0.30	1.06	0.16	0.17	0.63	0.17	0.21
	0.34	0.18	0.16	0.23	0.24	0.15	0.02	0.02	0.02	0.01	0.09
4	10.45	0.89	1.16	0.80	0.69	2.83	0.31	0.39	0.68	0.33	0.38
	0.78	0.26	0.18	0.56	0.54	0.42	0.11	0.15	0.05	0.08	0.12
6	15.75	2.06	2.22	1.17	1.03	4.33	0.50	0.62	0.72	0.46	0.83
	1.14	0.69	0.76	0.37	0.34	0.39	0.22	0.24	0.02	0.13	0.42
8	22.63	2.85	3.92	2.27	2.06	7.35	0.69	0.85	0.81	0.64	1.02
	1.71	0.96	1.51	0.62	0.62	0.90	0.24	0.27	0.19	0.21	0.25
10	31.70	3.58	5.55	3.22	2.92	12.13	0.92	1.22	0.80	0.78	1.20
	1.89	1.23	1.70	0.78	0.76	0.93	0.29	0.48	0.05	0.23	0.44

Presented are the percentages of erroneously identified exons in alignment between 978 pairs of cDNA and genomic segments in Projector database under various levels of artificially introduced noise. Each error rate was obtained by the formula: $100 - 200C/(T + P)$, where C, T and P denote the total numbers of correctly identified exons, true exons, and predicted exons, respectively. Each cell indicates the average (upper line) and the standard deviation (lower line) of six experiments.

Table S6. Error rates at assigning internal exons in alignment between cDNA sequences and corresponding genomic segments

Noise (%)	Methods										
	Blat	*Exalin*	*Exonerate*	*Gmap*	*GmapX*	*Sim4*	*Spaln*	*SpalnA*	*SpalnS*	*SpalnX*	*SpalnXL*
0	4.36	0.23	0.38	0.38	0.13	0.28	0.08	0.08	0.59	0.07	0.07
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	5.95	0.29	0.73	0.15	0.12	0.66	0.12	0.12	0.61	0.10	0.10
	0.45	0.14	0.82	0.07	0.05	0.16	0.06	0.07	0.02	0.05	0.05
2	7.41	0.31	0.64	0.18	0.14	0.96	0.11	0.13	0.61	0.10	0.10
	0.39	0.03	0.24	0.10	0.08	0.20	0.03	0.03	0.02	0.01	0.01
4	12.19	0.55	1.24	0.49	0.39	2.64	0.26	0.36	0.66	0.27	0.27
	0.84	0.09	0.21	0.13	0.08	0.40	0.11	0.14	0.04	0.08	0.08
6	16.83	0.86	2.40	0.84	0.70	4.18	0.42	0.57	0.68	0.40	0.40
	0.96	0.33	0.73	0.33	0.25	0.43	0.23	0.26	0.03	0.16	0.17
8	23.70	1.21	4.75	1.53	1.33	6.91	0.56	0.76	0.70	0.51	0.51
	1.86	0.28	1.90	0.36	0.26	0.84	0.17	0.22	0.07	0.10	0.10
10	33.72	1.69	6.93	2.62	2.30	11.32	0.86	1.23	0.78	0.77	0.77
	1.41	0.55	2.30	0.43	0.32	0.85	0.28	0.51	0.07	0.28	0.29

Presented are the percentages of erroneously identified internal exons in alignment between 978 pairs of cDNA and genomic segments in Projector database under various levels of artificially introduced noise. Each error rate was obtained by the formula: $100 \cdot C / T$, where C and T denote the total numbers of correctly identified internal exons and true internal exons, respectively. Each cell indicates the average (upper line) and the standard deviation (lower line) of six experiments.

Table S7. Error rates at assigning terminal exons in alignment between cDNA sequences and corresponding genomic segments

Noise (%)	Methods										
	Blat	*Exalin*	*Exonerate*	*Gmap*	*GmapX*	*Sim4*	*Spaln*	*SpalnA*	*SpalnS*	*SpalnX*	*SpalnXL*
0	4.97	0.94	0.63	0.47	0.26	0.84	0.26	0.21	0.63	0.31	0.31
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1	5.78	1.00	0.78	0.36	0.31	1.08	0.28	0.23	0.63	0.33	0.33
	0.17	0.10	0.34	0.07	0.07	0.04	0.03	0.03	0.00	0.03	0.03
2	6.75	1.32	0.74	0.84	0.80	1.46	0.29	0.24	0.65	0.37	0.53
	0.70	0.87	0.18	1.15	1.15	0.23	0.05	0.04	0.03	0.06	0.41
4	10.00	1.95	1.14	2.02	1.89	3.84	0.42	0.43	0.69	0.49	0.72
	1.57	1.04	0.24	2.50	2.50	0.75	0.15	0.18	0.10	0.18	0.40
6	15.86	6.11	2.64	2.70	2.60	5.60	0.67	0.64	0.80	0.61	2.39
	2.40	2.98	1.25	1.43	1.43	1.11	0.18	0.20	0.13	0.14	1.78
8	23.24	8.60	4.57	6.01	5.81	10.16	1.00	1.03	1.14	1.04	2.84
	3.58	4.31	1.76	3.09	3.10	2.13	0.85	0.84	0.96	0.96	1.39
10	30.36	10.34	6.44	7.34	7.17	16.84	0.89	0.89	0.80	0.77	2.72
	5.59	4.55	1.60	4.10	4.10	5.67	0.23	0.23	0.15	0.13	1.63

Presented are the percentages of erroneously identified terminal exons including single exons in alignment between 978 pairs of cDNA and genomic segments in Projector database under various levels of artificially introduced noise. Each error rate was obtained by the formula: $100 \cdot C / T$, where C and T denote the total numbers of correctly identified terminal exons and true terminal exons, respectively. Each cell indicates the average (upper line) and the standard deviation (lower line) of six experiments.

Table S8. Computation time (s) spent in alignments

Noise (%)	Method									
	Blat	*Exalin*	*Exonerate*	*Gmap*	*GmapX*	*Sim4*	*Spaln*	*SpalnA*	*SpalnS*	*SpalnX*	*SpalnXL*
0	166.6 14.4	4131.8 3.3	311.1 0.7	40.5 0.5	40.5 0.2	33.0 0.1	64.5 0.1	75.9 0.1	114.1 0.1	65.9 0.2	66.2 0.2
1	167.7 18.1	4135.0 3.3	308.4 0.8	41.7 0.2	42.1 0.1	33.1 0.2	65.8 0.9	77.7 1.5	114.6 0.2	66.3 0.3	66.5 0.3
2	170.6 16.8	4135.8 1.6	306.5 1.0	42.8 0.4	43.7 0.3	33.1 0.1	68.7 0.5	81.9 0.9	115.6 0.5	67.5 0.4	67.8 0.6
4	166.9 16.9	4135.5 1.5	304.3 1.8	46.4 0.4	48.3 0.4	33.4 0.1	78.2 2.6	96.4 4.0	120.1 1.0	72.2 1.2	72.7 1.6
6	164.1 15.6	4136.7 4.1	301.6 1.8	50.3 0.1	53.2 0.2	33.5 0.2	92.8 3.4	119.1 5.3	127.0 1.7	79.3 1.8	79.9 2.5
8	162.7 13.0	4135.9 4.2	298.4 3.1	53.3 1.6	56.4 1.5	33.8 0.3	117.8 2.4	158.2 4.0	137.6 3.0	90.6 3.3	91.0 3.8
10	159.0 15.3	4135.9 7.0	295.4 2.8	55.1 1.6	58.4 1.6	34.2 0.3	142.6 4.0	197.2 6.6	148.9 4.3	102.5 4.5	102.3 4.3

The CPU time (s) spent in alignment between 978 pairs of cDNA and genomic segments in Projector database under various levels of artificially introduced noise. Each cell indicates the average (upper line) and the standard deviation (lower line) of six experiments.

Table S9. Performance of three programs in genome mapping and alignment

Observa-tions	Noise (%)	Methods				
		Blat	*Gmap*	*Spaln*	*SpalnX*	*SpalnXL*
%Mapped	0	0.00∓0.00	0.00∓0.00	0.00∓0.00	0.10∓0.00	0.00∓0.00
	1	0.00∓0.00	0.00∓0.00	0.00∓0.00	0.03∓0.05	0.00∓0.00
	2	0.00∓0.00	0.00∓0.00	0.02∓0.04	0.03∓0.08	0.02∓0.04
	4	0.00∓0.00	0.00∓0.00	0.05∓0.09	0.05∓0.09	0.05∓0.09
	6	0.00∓0.00	0.03∓0.05	0.24∓0.14	0.22∓0.12	0.22∓0.12
	8	0.00∓0.00	0.03∓0.05	0.65∓0.11	0.63∓0.15	0.55∓0.21
	10	0.20∓0.17	0.20∓0.11	2.57∓1.03	2.45∓1.08	1.19∓0.50
%Gene	0	2.76∓0.00	2.35∓0.00	0.41∓0.00	0.51∓0.00	0.51∓0.00
	1	9.15∓0.73	1.17∓0.44	0.63∓0.30	0.68∓0.28	0.80∓0.31
	2	18.13∓2.06	2.35∓2.18	0.65∓0.16	0.68∓0.13	1.19∓0.98
	4	34.68∓3.25	5.33∓3.88	1.69∓0.79	1.79∓0.63	2.40∓0.78
	6	51.43∓4.75	8.27∓2.85	3.02∓1.52	2.98∓1.15	7.21∓3.48
	8	65.64∓3.07	16.41∓5.12	5.61∓2.45	5.33∓1.94	9.08∓2.54
	10	77.78∓5.35	23.21∓5.43	9.15∓2.31	8.37∓1.94	10.53∓4.50
%Exon	0	0.47∓0.00	0.45∓0.00	0.07∓0.00	0.06∓0.00	0.07∓0.00
	1	1.99∓0.15	0.24∓0.07	0.10∓0.04	0.10∓0.04	0.11∓0.04
	2	3.98∓0.36	0.38∓0.24	0.11∓0.02	0.09∓0.02	0.15∓0.11
	4	8.84∓0.76	0.83∓0.54	0.24∓0.11	0.25∓0.08	0.33∓0.11
	6	14.71∓1.10	1.27∓0.40	0.44∓0.22	0.42∓0.14	0.88∓0.42
	8	21.92∓1.66	2.47∓0.59	0.78∓0.33	0.73∓0.20	1.08∓0.26
	10	30.94∓2.22	3.76∓0.69	1.34∓0.31	1.28∓0.37	1.22∓0.53
%Internal Exon	0	0.30∓0.00	0.38∓0.00	0.04∓0.00	0.03∓0.00	0.03∓0.00
	1	1.97∓0.22	0.17∓0.07	0.08∓0.05	0.06∓0.05	0.08∓0.05
	2	4.07∓0.39	0.21∓0.12	0.07∓0.02	0.06∓0.01	0.08∓0.04
	4	9.22∓0.80	0.50∓0.12	0.21∓0.10	0.22∓0.07	0.25∓0.08
	6	14.86∓0.90	0.92∓0.34	0.37∓0.24	0.40∓0.18	0.46∓0.19
	8	22.24∓1.76	1.72∓0.38	0.64∓0.23	0.65∓0.13	0.61∓0.07
	10	32.79∓1.17	3.10∓0.38	1.60∓0.43	1.67∓0.65	0.87∓0.31
%Terminal Exon	0	1.10∓0.00	0.63∓0.00	0.16∓0.00	0.16∓0.00	0.21∓0.00
	1	2.00∓0.14	0.54∓0.06	0.19∓0.04	0.21∓0.06	0.24∓0.04
	2	3.59∓0.63	1.04∓1.13	0.22∓0.07	0.20∓0.06	0.39∓0.40
	4	7.33∓1.78	2.26∓2.41	0.30∓0.14	0.31∓0.17	0.58∓0.44
	6	14.14∓2.33	3.11∓1.57	0.62∓0.25	0.54∓0.15	2.41∓1.75
	8	22.08∓3.85	6.74∓2.98	1.27∓1.10	1.20∓0.97	2.90∓1.38
	10	29.03∓6.98	9.37∓3.60	1.38∓0.32	1.32∓0.31	2.67∓1.80
CPU (s)	0	503.4∓0.8	24.9∓0.3	35.5∓0.5	39.8∓0.4	39.6∓0.1
	1	487.2∓19.1	24.6∓1.7	36.9∓1.2	42.0∓1.2	41.3∓1.0
	2	474.7∓12.5	29.9∓5.2	43.1∓4.2	45.9∓2.5	44.8∓2.9
	4	455.9∓16.7	39.0∓3.9	59.5∓8.3	56.4∓6.5	55.5∓6.3
	6	432.7∓18.3	47.0∓4.9	82.1∓16.0	66.6∓9.6	66.7∓10.8
	8	439.3∓27.8	54.0∓7.2	128.8∓18.1	91.8∓14.1	92.0∓12.3
	10	412.1∓24.4	59.2∓5.5	191.9∓57.0	125.1∓35.3	121.1∓33.9

Each cell indicates the fraction of missed genes (%Mapped), fraction of genes with incorrectly identified structures (%Gene), fraction of incorrectly identified exons (%Exon), fraction of incorrectly identified internal exons (%Internal Exon), fraction of incorrectly identified terminal exons including single exons (%Terminal Exon), and the CPU time used upon mapping and alignment of the total of 978 human and mouse CDS queries against the respective genomic sequences. The queries contain artificially introduced noise of the level indicated in the second column. The average and the standard deviation of six experiments are presented.

Supplementary Figure

Spaln:

```
   20449 AGCTGCTTTGGTGGCAAGAAAGCAAAAGgtgggtgggtgaggcggagatgtgtctctttc
    2040 AGCTGCTTTGGAGGCAAGAAAGCAAAAG
         AGCTGCTTTGGwGGCAAGAAAGCAAAAG

;; skip 480 nt's

   20989 cctccggccgcccagAGGAGTGGTTCCATCAAGgtgagtccctgtttgttctgctgtcca
    2068                AGGCGCGGTACCATCAAG
                        AGGmGyGGTwCCATCAAG

;; skip 720 nt's

   21769 cgtaccacccttctgtttcagGGAAGAAGGATGCAGAGATGGACCGGAACTTTGACACAC
    2086                      GGAAGACGGATGCAGATTTGGACCGGAACTTTGACACAC
                              GGAAGAmGGATGCAGAkwTGGACCGGAACTTTGACACAC

   21829 TGGACCTGCCTAAACGAACGGAGGCTGCAAAAGgtgaatggataaagaggcagagctggg
    2125 TGGGCCTGTCGAAACGAACGGAGGCTGCAAAAG
         TGGrCCTGyCkAAACGAACGGAGGCTGCAAAAG

---------------------------------------------------------------------
```

Gmap:

```
   20449 AGCTGCTTTGGTGGCAAGAAAGCAAAAGGTGGGTGGGTGAGGCGGAGATG...CACCCTT
         ||||||||||| |||||||||||||||---------| |  ||||  ]]]...]]]|| |
    2040 AGCTGCTTTGGAGGCAAGAAAGCAAAA         GAGGCGCGGTA  1281   CCAT

   21781 CTGTTTCAGGGAAGAAGGATGCAGAGATGGACCGGAACTTTGACACACTGGACCTGCCTA
         | -----|||||||| |||||||||  |||||||||||||||||||||||| |||| | |
    2082 CA     AGGGAAGACGGATGCAGATTTGGACCGGAACTTTGACACACTGGGCCTGTCGA

   21841 AACGAACGGAGGCTGCAAAAGGTG...TAGGCTTCGAGCTGCTGTACCAGCCGGAGGTGG
         |||||||||||||||||||||>>>...>>>||||||||||||||||||||||||||||||
    2137 AACGAACGGAGGCTGCAAAAG   969   GCTTCGAGCTGCTGTACCAGCCGGAGGTGG

---------------------------------------------------------------------
```

Exonerate:

```
  2040 : AGCTGCTTTGGAGGCAAGAAAGCAAAAGAGGCGCGG  >>>> Target Intron 7 >
         ||||||||||| ||||||||||||||||  | | ||++         1299 bp
 20449 : AGCTGCTTTGGTGGCAAGAAAGCAAAAGGTGGGTGGgt......................

  2077 : >>>  TACCATCAAGGGAAGACGGATGCAGATTTGGACCGGAACTTTGACACACTGGGCC
            -+|  ||   | | | || | | | |||  |||||||||||||||||||||||| ||
 20488 : ...tgTTTCAGGGAAGAAGGATGCA-G-AGA--TGGACCGGAACTTTGACACACTGGACC

  2131 : TGTCGAAACGAACGGAGGCTGCAAAAG  >>>> Target Intron 8 >>>>  GCTT
         || | ||||||||||||||||||||||++          969 bp         ++||||
 21835 : TGCCTAAACGAACGGAGGCTGCAAAAGgt.........................agGCTT
```

Figure S1. An example of variation in alignments produced by different methods. The cDNA of the projector entry Mm.16.ENST.0000044480 was subjected to artificial mutation at 10% of the sites and aligned with the corresponding genomic segment by *spaln*, *gmap*, and *exonerate*. The portion of the results surrounding the 18 nt short exon is presented with slight modifications in format.