SUPPLEMENTARY MATERIAL

Supplementary Table S1. Performance of spliced alignment programs on *gene coding regions* only, for the four vertebrate reference data sets. Columns represent sensitivity and specificity values at the nucleotide, exon and splice junction (intron) levels, the latter when allowing for a margin *V* (0 or 10) of error around the splice junction. Sensitivity was calculated as Sn = TP/(TP+FN) and specificity as Sp = TP/(TP+FP).

Method	Nucleotide		Exon		Splice junction	
	Sn	Sp	Sn	Sp	Sn, V=0(=10)	Sp, V=0(=10)
Mouse: 805 genes, 7,914 exons, 7,109 introns						
sim4	0.817	0.996	0.907	0.992	0.725 (0.755)	0.738 (0.768)
BLAT	0.770	0.986	0.843	0.928	0.070 (0.527)	0.063 (0.477)
tBLAT	0.921	0.984	0.950	0.984	0.186 (0.829)	0.183 (0.813)
GMAP	0.831	0.996	0.799	0.996	0.773 (0.777)	0.976 (0.981)
Exonerate	0.907	0.984	0.879	0.997	0.825 (0.840)	0.950 (0.968)
GeneSeqer	0.687	0.988	0.660	0.923	0.587 (0.595)	0.848 (0.859)
EXALIN	0.949	0.998	0.959	0.997	0.940 (0.952)	0.973 (0.986)
sim4cc	0.982	0.996	0.979	0.998	0.947 (0.957)	0.964 (0.975)
Dog: 43 genes, 379 exons, 336 introns						
sim4	0.888	0.996	0.942	0.995	0.810 (0.833)	0.807 (0.831)
BLAT	0.840	0.987	0.879	0.958	0.065 (0.616)	0.061 (0.578)
tBLAT	0.943	0.981	0.955	0.972	0.173 (0.842)	0.168 (0.818)
GMAP	0.917	0.997	0.881	0.997	0.857 (0.863)	0.983 (0.990)
Exonerate	0.978	0.983	0.953	0.997	0.890 (0.840)	0.934 (0.968)
GeneSeqer	0.727	0.995	0.697	0.978	0.616 (0.619)	0.908 (0.912)
EXALIN	0.990	0.996	0.995	0.992	0.979 (0.985)	0.971 (0.976)
sim4cc	0.995	0.995	0.987	0.997	0.958 (0.961)	0.970 (0.973)
Chicken: 154 genes, 1,597 exons, 1,442 introns						
sim4	0.449	0.991	0.584	0.987	0.285 (0.302)	0.449 (0.475)
BLAT	0.407	0.976	0.431	0.861	0.017 (0.128)	0.027 (0.211)
tBLAT	0.797	0.986	0.837	0.973	0.142 (0.655)	0.157 (0.722)
GMAP	0.553	0.988	0.253	0.990	0.209 (0.212)	0.938 (0.947)
Exonerate	0.602	0.939	0.529	0.997	0.424 (0.436)	0.859 (0.883)
GeneSeqer	0.675	0.986	0.433	0.916	0.374 (0.386)	0.817 (0.842)
EXALIN	0.830	0.998	0.830	0.996	0.791 (0.809)	0.960 (0.982)
sim4cc	0.905	0.981	0.904	0.992	0.801 (0.818)	0.880 (0.899)
Zebrafish: 230 genes, 2,528 exons, 2,296 introns						
sim4	0.141	0.985	0.196	0.991	0.029 (0.031)	0.162 (0.172)
BLAT	0.163	0.966	0.084	0.798	0.001 (0.008)	0.007 (0.067)
tBLAT	0.591	0.984	0.629	0.959	0.086 (0.375)	0.130 (0.567)
GMAP	0.390	0.993	0.023	1.000	0.010 (0.010)	1.000 (1.000)
Exonerate	0.320	0.889	0.245	0.990	0.146 (0.149)	0.811 (0.828)
GeneSeqer	0.623	0.988	0.129	0.940	0.117 (0.118)	0.871 (0.877)
EXALIN	0.536	0.997	0.538	0.994	0.480 (0.494)	0.954 (0.984)
sim4cc	0.710	0.970	0.732	0.985	0.548 (0.568)	0.758 (0.786)

Supplementary Table S2. Performance of spliced alignment programs on plant data. GeneSeqer results are reported for the default version ('GeneSeqer') and for the Arabidopsis-optimized version ('GeneSeqer(A)'); all other programs were run with the default settings. Columns represent sensitivity and specificity values at the nucleotide, exon and splice junction (intron) levels, the latter when allowing for a margin *V* (0 or 10) of error around the splice junction. Sensitivity was calculated as Sn = TP/(TP+FN) and specificity as Sp = TP/(TP+FP). The difference in performance between the Arabidopsis-optimized and the default versions of GeneSeqer underscores the importance of program robustness with different species, possibly yet-uncharacterized genomically, as demanded by the ongoing sequencing of a wide variety of species.

Method	Nucleotide		Exon		Splice junction	
	Sn	Sp	Sn	Sp	Sn, V=0(=10)	Sp, V=0(=10)
Soybean: 1,878 genes, 12,111 exons, 10,233 introns; Average % identity: 71.4						
sim4	0.229	0.990	0.451	0.994	0.145 (0.158)	0.269 (0.294)
BLAT	0.189	0.993	0.239	0.985	0.070 (0.003)	0.063 (0.011)
tBLAT	0.717	0.991	0.780	0.998	0.138 (0.535)	0.160 (0.619)
GMAP	0.250	0.994	0.120	0.999	0.063 (0.064)	0.951 (0.964)
Exonerate	0.609	0.924	0.481	0.999	0.274 (0.283)	0.771 (0.795)
GeneSeqer	0.437	0.995	0.481	0.985	0.462 (0.478)	0.907 (0.940)
GeneSeqer(A)	0.755	0.997	0.849	0.996	0.832 (0.845)	0.961 (0.976)
EXALIN	0.700	0.999	0.788	1.000	0.737 (0.751)	0.976 (0.994)
sim4cc	0.848	0.985	0.882	0.997	0.755 (0.773)	0.863 (0.885)
Medicago: 1,055 genes, 5,539 exons, 4,484 introns; Average % identity: 70.9						
sim4	0.205	0.990	0.391	0.994	0.120 (0.130)	0.995 (0.319)
BLAT	0.146	0.993	0.182	0.987	0.002 (0.616)	0.039 (0.173)
tBLAT	0.674	0.991	0.740	0.997	0.124 (0.486)	0.155 (0.609)
GMAP	0.218	0.994	0.116	1.000	0.058 (0.059)	0.960 (0.974)
Exonerate	0.602	0.911	0.496	1.000	0.270 (0.283)	0.760 (0.797)
GeneSeqer	0.302	0.996	0.319	0.981	0.306 (0.317)	0.912 (0.942)
GeneSeqer(A)	0.663	0.999	0.739	0.997	0.728 (0.740)	0.908 (0.983)
EXALIN	0.654	0.999	0.740	1.000	0.668 (0.683)	0.970 (0.992)
sim4cc	0.815	0.981	0.850	0.995	0.705 (0.725)	0.847 (0.872)
Rice: 2,873 genes, 21,592 exons, 18,719 introns; Average % identity: 67.4						
sim4	0.092	0.986	0.207	0.995	0.035 (0.038)	0.174 (0.188)
BLAT	0.071	0.994	0.082	0.988	0.001 (0.012)	0.009 (0.138)
tBLAT	0.607	0.991	0.664	0.997	0.100 (0.655)	0.140 (0.722)
GMAP	0.096	0.996	0.034	1.000	0.014 (0.014)	0.952 (0.960)
Exonerate	0.428	0.913	0.303	0.999	0.134 (0.139)	0.735 (0.765)
GeneSeqer	0.189	0.991	0.200	0.970	0.185 (0.192)	0.864 (0.899)
GeneSeqer(A)	0.589	0.996	0.663	0.994	0.645 (0.657)	0.953 (0.970)
EXALIN	0.470	0.998	0.540	1.000	0.459 (0.470)	0.968 (0.992)
sim4cc	0.755	0.971	0.796	0.996	0.590 (0.611)	0.758 (0.786)

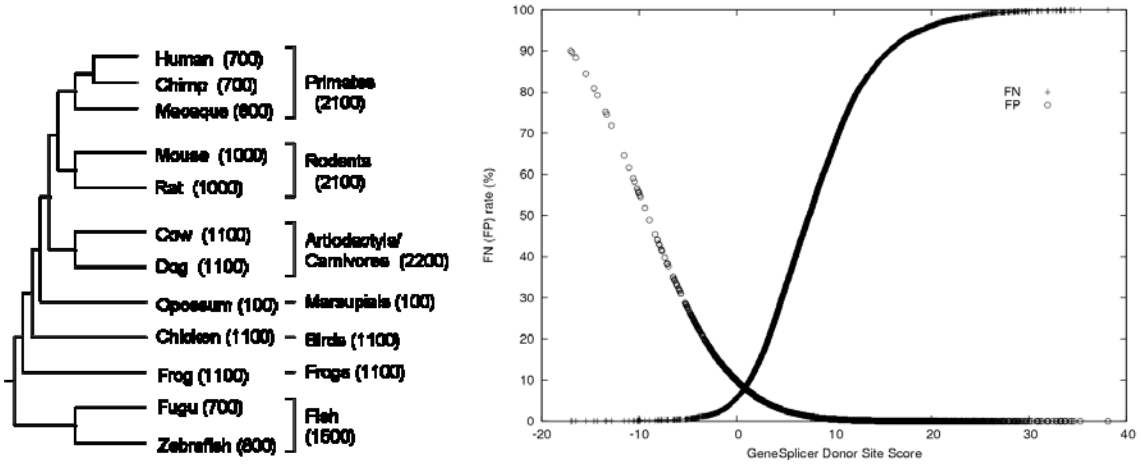
Construction of reference data sets. 6,453 *Glycine max* (soybean), 4,110 *Medicago truncatula* (medicago), 52,281 *Oryza sativa* (rice) mRNA sequences were downloaded from the Genomes section in GenBank. The *Arabidopsis thaliana* TAIR8 annotation (38,963 transcripts) was downloaded from The Arabidopsis Information Resource

(http://www.arabidopsis.org). The Arabidopsis genome was used as reference. To determine homologous gene pairs for evaluation, for each of the three sets (soybean, medicago and rice) a two-way 'fasta' (Pearson 1990) search was conducted against Arabidopsis transcripts, selecting only 'mutually best' matches. These pairs were filtered to retain only those for which the 'fasta' alignment covered at least 60% of either sequence and both reciprocal alignments had 60% or higher sequence identity. A reference gene annotation for each set was constructed as described in **Methods**.

1. Pearson, W. R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA, *Methods in Enzymology* **183**, 63 - 98.

Supplementary Section S3. Training data for splice signal models.

We collected 10,000 positive (validated splice junctions) and 50,000 negative examples from 12 vertebrate species (human, chimp, macaque, mouse, rat, cow, dog, opossum, chicken, frog, fugu and zebrafish). Training data consisted of 162 bp sequences for GeneSplicer and 52 bp sequences for Glimmer, separately for the donor and the acceptor sites, and centered about the dinucleotide splice signal (GT/AG). Positive examples were extracted from spliced alignments of RefSeq (Pruitt *et al.* 2007) mRNA alignments on their corresponding genomes produced with ESTmapper (Florea *et al.* 2005), excluding experimental sequences (XM accession numbers). Additionally, we requested a gap-free alignment within 5 bp of the splice junctions. Negative examples were randomly extracted from intronic sequences at GT/AG (CT/AC) sites. The samples were collected across the twelve species such that the representation across major clades is roughly uniform (**Supplementary Figure S3, A**) and depending on the number of representative species, and subject to the availability of RefSeq sequences (*e.g.*, opossum). For each of the two splice signal models, the false negative (FN) versus false positive (FP) rates were plotted and used to determine the active score range (**Supplementary Figure S3, B**), which was then normalized to the [0,1] interval.

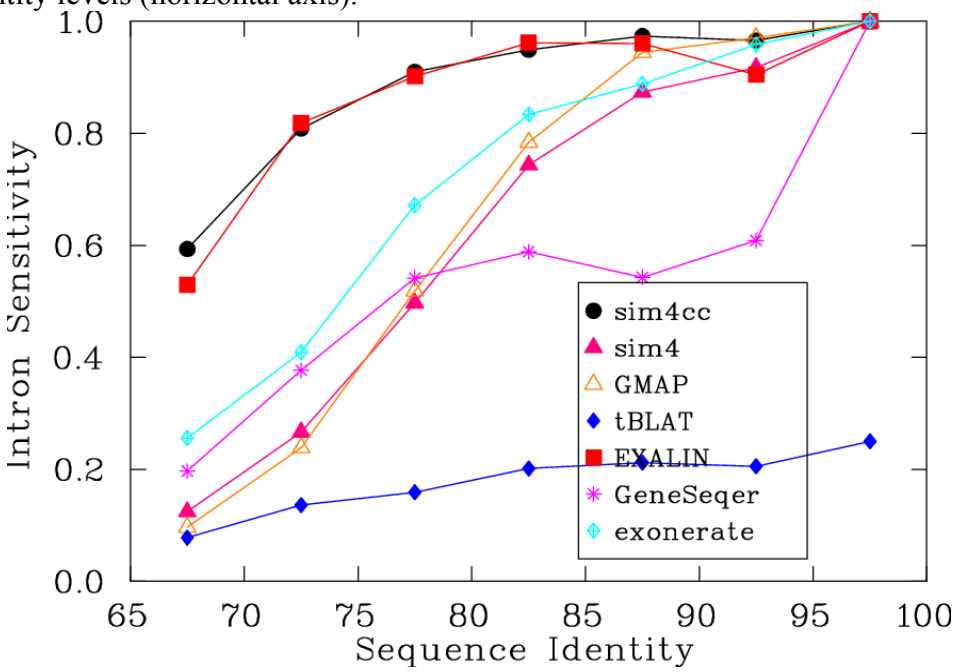


(A) (B)

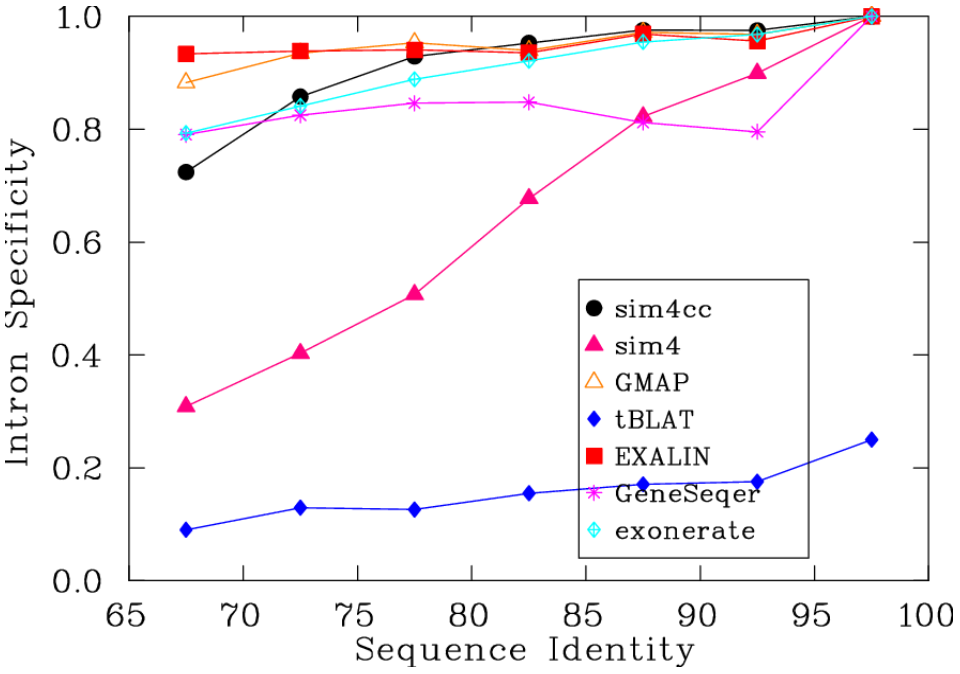
Supplementary Figure S3. (A) Clade-based sampling of positive examples for splice site model training. (The phylogenetic tree is not drawn to scale.) For each species, negative examples were 5-fold the number of positive examples. (B) GeneSplicer FN and FP rates with varying score cutoffs, shown as percentages. For instance, a score cutoff of -1.25 will have a 3% FN rate (i.e., missing 3% of known examples), but will predict 13.05% false positives (FP).

1. Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **35**, D61-65.
2. Florea, L., Di Francesco, V., Miller, J., Turner, R., Yao, A., Harris, M., Walenz, B., Mobarry, C., Merkulov, G.V., Charlab, R. *et al.* (2005) Gene and alternative splicing annotation with AIR. *Genome Res*, **15**, 54-66.

Supplementary Figure S4. Performance of spliced alignment programs (splice junction sensitivity (A) and specificity (B); vertical axis) with varying sequence identity levels (horizontal axis).



(A)



(B)

Supplementary Table S5. Identification of novel poplar 'genes' from *Fagus grandifolia* 454 unigene sequences. Candidate alignments with ≥80% genomic coverage and ≤20% overlap with the existing annotation were initially selected. Potential paralogous matches, identified by matches elsewhere in the genome with ≥50% genomic coverage at greater or similar percent sequence identity ('genome-based'), or with ≥50% overlap with the annotated transcripts ('gene-based'), were removed from the initial set to produce a final set of 'novel' genes.

Method	**Candidates**		**'Paralogs'**			**Novel 'genes'**
	Alignments	**'Genes'**	**Genome-based**	**Gene-based**	**All**	
EM	212	115	5	17	18	97
GMAPX	234	100	15	37	42	58

Supplementary Table S6. Performance of spliced alignment programs on the 21 microexons (≤25 bp) in the vertebrate data set. Numbers in parentheses represent predicted exons that overlap the true exon range, but whose either or both endpoints are incorrect.

	Mouse	**Dog**	**Chicken**	**Zebrafish**	**Total**
Exons	**17**	**2**	**0**	**2**	**21**
sim4	3+(1)	0	n.a.	0	3+(1)
BLAT	(1)	0	n.a.	0	(1)
tBLAT	(4)	0	n.a.	0	(4)
GMAP	3	1	n.a.	0	4
Exonerate	2	0	n.a.	0	2
GeneSeqer	11+(1)	2	n.a.	0	13+(1)
EXALIN	15	1+(1)	n.a.	0	16+(1)
sim4cc	6	0	n.a.	0	6