–SUPPLEMENT–

Computational discovery of human coding and non-coding transcripts with conserved splice sites

Dominic Rose [a,f,ℓ], Michael Hiller [j], Katharina Schutt [b,c,d], Jörg Hackermüller [a,c], Rolf Backofen [f,g,h], Peter F. Stadler [a,e,c,i,k]

[a] *Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany.*

[b] *LIFE – Leipzig Research Center for Civilization Diseases, University of Leipzig, Germany.*

[c] *Fraunhofer Institut for Cell Therapy and Immunology, AG RNomics, Leipzig, Germany.*

[d] *Department of Molecular Immunology, University of Leipzig, Germany.*

[e] *Interdisciplinary Center of Bioinformatics, University of Leipzig, Germany.*

[f] *Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany.*

[g] *Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Germany.*

[h] *Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Germany.*

[i] *Institute for Theoretical Chemistry, University of Vienna, Austria.*

[j] *Department of Developmental Biology, Stanford University, USA.*

[k] *Sante Fe Institute, Santa Fe, USA*

[ℓ] *Corresponding author:* `dominic@bioinf.uni-leipzig.de`

Supplement

Supplemental Figures and Tables

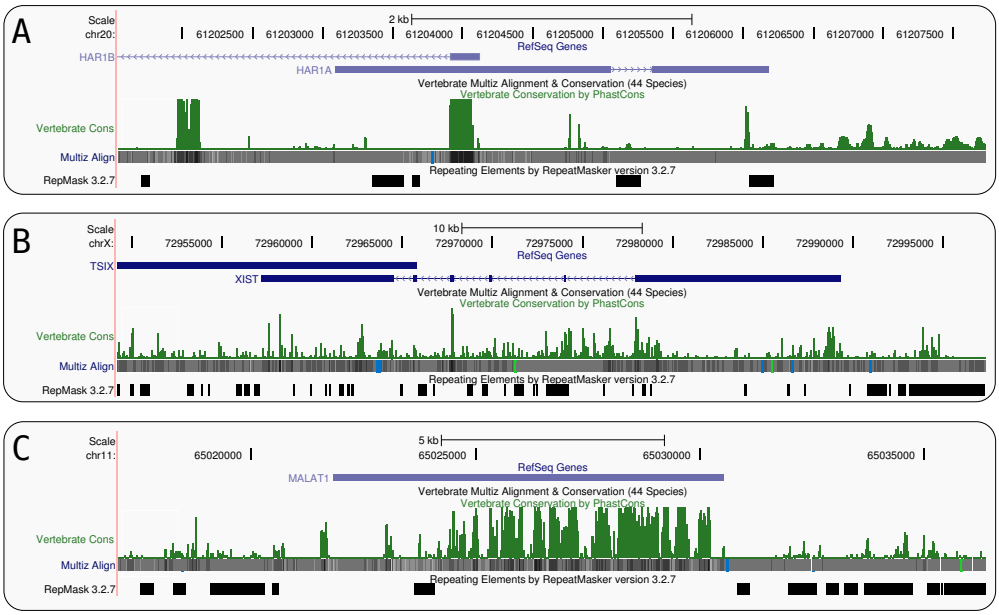

Fig. S1. **Sequence conservation of lncRNAs.** The figure shows three representative human lncRNAs with different sequence conservation among vertebrates as indicated by the `PhastCons` profile. (A) *HAR1A*, chr20:61,203,089-61,206,182. (B) *XIST*, chrX:72,957,220-72,989,313 . (C) *MALAT-1*, chr11:65,021,809-65,030,513. Long ncRNAs are generally poorly conserved which mostly excludes sequence conservation as key feature for their computational prediction.

	coding	5'UTR	3'UTR	non-coding	ESTs
Donor	226,686	14,167	533	516	621,355
Acceptor	226,272	11,274	407	527	619,807

Table S1
Existing hg18 splice site annotation. The table gives of the data basis at the time of starting the work. Numbers refer to the union of the `RefSeq`, `UCSC Genes`, and mRNA `UCSC Genome Browser` tracks. We distinguish between coding, UTR, non-coding and EST-confirmed sites. Minimal required intron length: 20 nt.

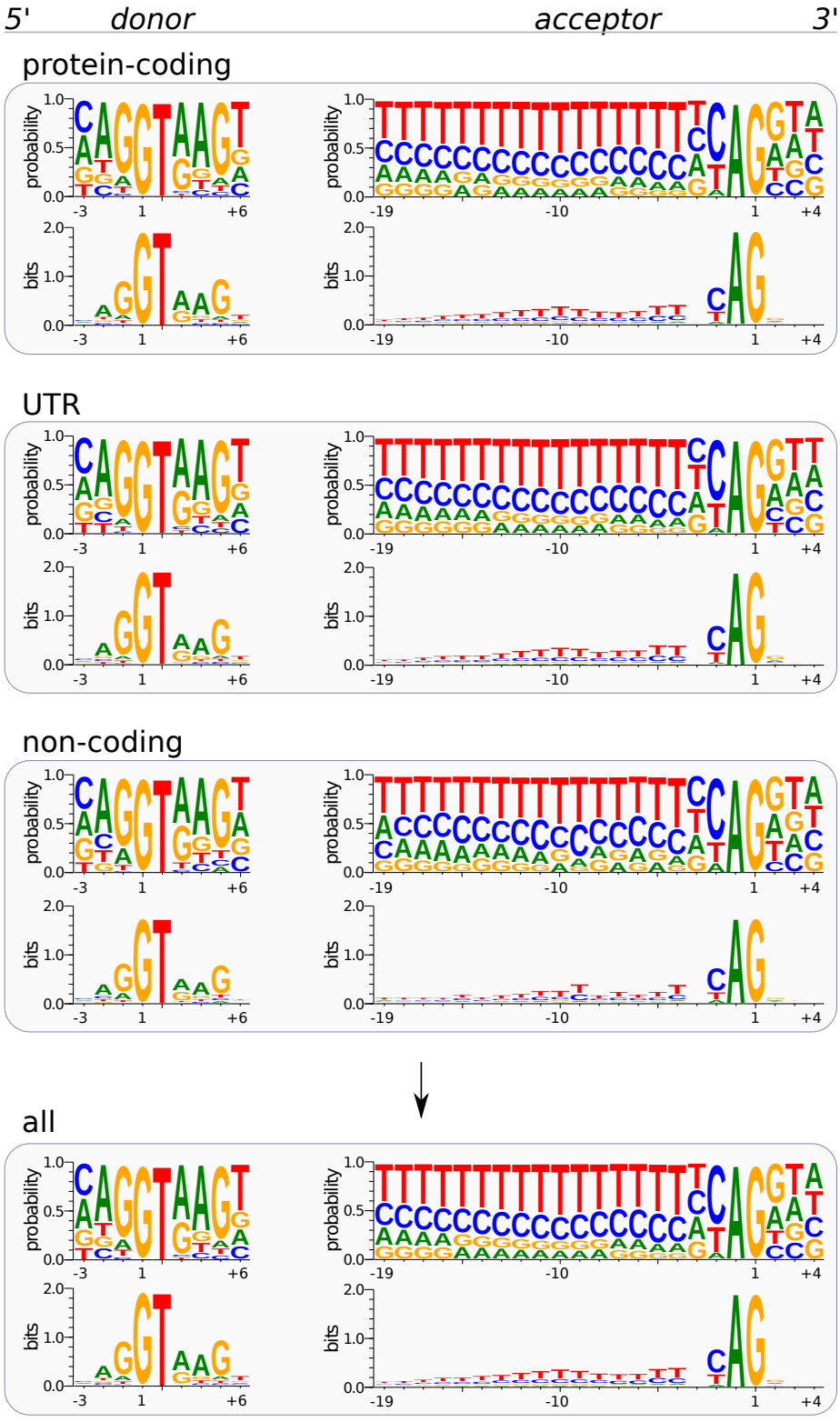

Fig. S2. Sequence logos. The figure illustrates nucleotide frequencies of real splice sites. It shows that splice sites of non-coding genes are highly similar to sites of protein-coding genes or UTRs. Therefore, we trained our SVM models with a mixture of all real splice sites. We expect that our approach yields novel non-coding splice sites although the majority of training samples belongs to protein-coding genes.

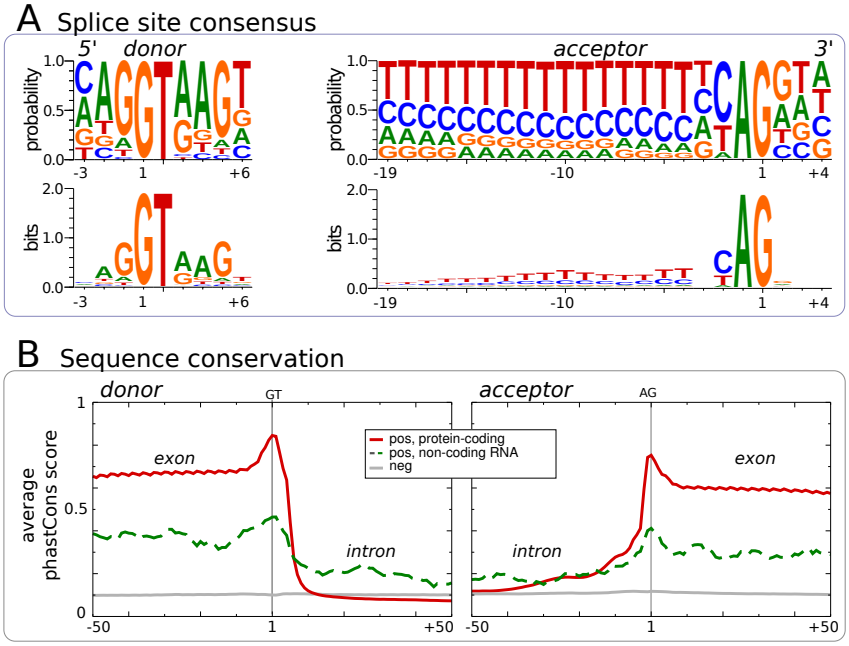

Fig. S3. Sequence conservation of splice sites. (A) The sequence logo shows the preferred nucleotides at each position around real splice sites. (B) The `PhastCons` score profile for the $[-50, 50]$ interval shows that the average sequence conservation decreases downstream of the donor GT and increases upstream of the acceptor AG, both for ncRNAs and coding genes. In contrast to ncRNAs, protein-coding genes exhibit a much higher exonic conservation level. The figure is based on our training set which consists of "positive" and "negative" samples representing real (pos) and pseudo (neg) splice sites.

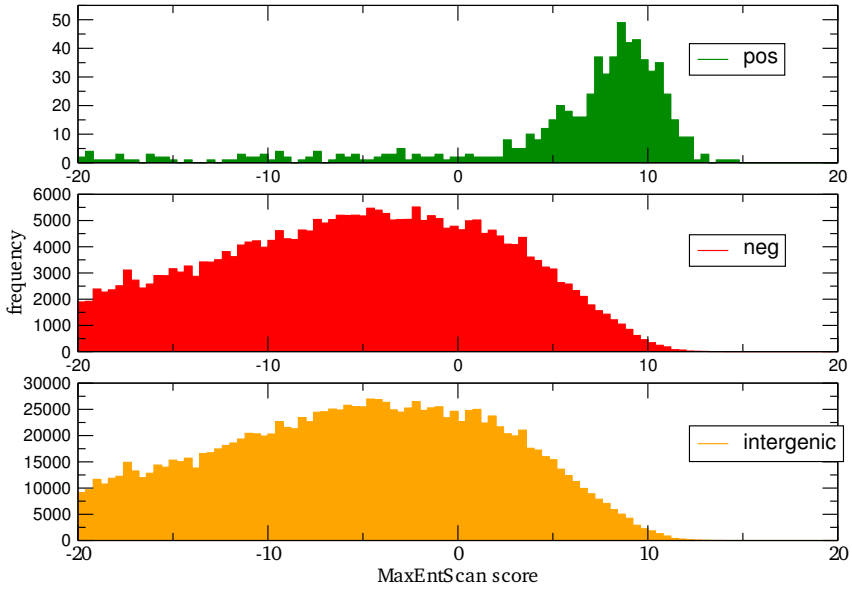

Fig. S4. Distribution of `MaxEntScan` scores. The histograms are based upon scores for real *(pos)*, false *(neg)*, and intergenic splice sites from chr21.

Table S2. **Feature selection of the donor model.** The table outlines the SVM performance of several feature combinations. Each column represents an independent training event based on all marked *('x')* features. Here, we only present a small subset of combinations. Overall, we have tested many more. Ideally, an SVM model is trained with few features and only a few support vectors *(total_sv)* are necessary to yield high AUC values. As expected, the performance improves with more features until we reach a 'stable' AUC of 0.96. Simple features, like GC-content or MPI, and even Shannon entropy only marginally affect the donor model and thus have been neglected. The highlighted column (dark grey) and the associated features represent an optimal tradeoff between effort necessary to train the SVM and resulting predictive power of the model.

`MaxEntScan`-score	x		x		x		x		x		x		x		x		x		x	
Tree LogOdds	x				x		x		x		x		x		x		x		x	
Fly LogOdds			x		x		x		x		x		x		x		x		x	
Median LogOdds							x		x		x		x		x		x		x	
No. sequences									x		x		x		x		x		x	
No. sequences (filtered)											x		x		x		x		x	
`PhastCons` slope & avg													x		x		x		x	
avg GC-content															x		x		x	
MPI																	x		x	
Shannon Entropy																			x	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
p >0.5	86.72	13.96	86.70	14.00	87.42	12.92	86.60	12.40	86.80	10.10	88.96	4.92	89.04	4.22	89.22	4.40	89.38	4.68	89.44	4.76
0.8	73.22	6.66	73.76	6.82	73.62	5.14	72.06	4.58	75.12	3.96	84.10	1.86	83.86	1.36	83.92	1.34	83.98	1.42	83.86	1.48
0.9	45.18	1.20	40.04	1.10	53.52	1.36	59.40	1.94	65.26	1.78	80.72	0.88	80.92	0.72	80.86	0.72	80.82	0.70	80.84	0.72
0.93	35.84	0.40	29.20	0.56	43.80	0.56	51.70	1.04	61.02	1.34	78.76	0.58	79.34	0.60	79.46	0.64	79.48	0.60	79.36	0.62
0.95	30.24	0.18	23.52	0.44	38.14	0.32	44.28	0.54	56.78	1.00	76.24	0.44	78.00	0.46	78.20	0.54	78.10	0.50	78.02	0.54
0.98	20.42	0.04	14.94	0.16	27.86	0.12	32.28	0.16	45.06	0.38	68.16	0.18	74.40	0.18	74.32	0.16	73.76	0.14	73.56	0.14
AUC	0.9244		0.9239		0.9341		0.9374		0.9477		0.9614		**0.9639**		0.9651		0.9660		0.9661	
total_sv	31,159		32,314		29,914		29,910		27,073		19,724		**19,040**		18,924		18,954		19,028	

Table S3. **Feature selection of the acceptor model.** The table outlines the SVM performance of several feature combinations. Each column represents an independent training event based on all marked *('x')* features. Here, we only present a small subset of combinations. Overall, we have tested many more. Ideally, an SVM model is trained with few features and only a few support vectors *(total_sv)* are necessary to yield high AUC values. As expected, the performance improves with more features until we reach a 'stable' AUC of 0.94. Simple features, like GC-content or MPI, and even Shannon entropy only marginally affect the donor model and thus have been neglected. The highlighted column (dark grey) and the associated features represent an optimal tradeoff between effort necessary to train the SVM and resulting predictive power of the model.

`MaxEntScan`-score	x		x		x		x		x		x		x		x		x		x	
Tree LogOdds	x				x		x		x		x		x		x		x		x	
Fly LogOdds			x		x		x		x		x		x		x		x		x	
Median LogOdds							x		x		x		x		x		x		x	
No. sequences									x		x		x		x		x		x	
No. sequences (filtered)											x		x		x		x		x	
`PhastCons` slope & avg													x		x		x		x	
avg GC-content															x		x		x	
MPI																	x		x	
Shannon Entropy																			x	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
p >0.5	81.78	9.74	77.46	10.20	82.00	9.48	82.04	9.60	83.16	10.08	83.60	9.76	84.44	9.44	84.42	9.30	84.48	9.34	84.48	9.28
0.8	74.12	4.14	66.50	3.78	74.20	4.32	74.28	4.14	74.72	4.20	75.90	3.80	77.40	3.56	77.42	3.52	77.56	3.50	77.60	3.42
0.9	69.04	2.38	58.24	1.56	69.32	2.34	69.40	2.30	69.94	2.36	71.96	1.98	73.20	1.80	73.04	1.74	73.14	1.84	73.08	1.84
0.93	67.38	1.78	52.68	0.98	67.72	1.86	67.76	1.76	67.78	1.68	69.62	1.48	70.94	1.34	71.10	1.30	71.12	1.38	71.12	1.36
0.95	65.60	1.28	48.14	0.70	65.74	1.46	66.02	1.42	66.24	1.42	67.62	1.18	69.26	1.10	69.36	1.12	69.44	1.10	69.40	1.10
0.98	61.28	0.86	34.74	0.32	61.20	0.84	61.40	0.86	61.04	0.80	62.88	0.66	64.28	0.72	64.28	0.70	64.50	0.70	64.64	0.70
AUC	0.9245		0.9030		0.9249		0.9255		0.9309		0.9358		**0.9406**		0.9413		0.9423		0.9425	
total_sv	25,233		35,403		25,247		25,455		25,368		24,758		**24,061**		24,217		24,183		24,256	

Table S4
Donor SVM model performance comparison. We randomly generated five sample sets of equal size and trained donor SVMs. The model with the best performance according to its AUC value (set 1) was used for classification. The comparison provides evidence that random sampling only marginally affects the performance.

set	1		2		3		4		5	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
p>0.5	89.04	4.22	87.80	4.78	88.14	5.00	87.80	4.54	87.38	4.18
0.8	83.86	1.36	82.86	1.76	83.52	1.66	82.66	1.58	82.80	1.46
0.9	80.92	0.72	80.30	0.94	80.80	0.86	80.04	0.82	80.14	0.62
0.93	79.34	0.60	78.82	0.58	78.96	0.72	78.78	0.68	78.80	0.44
0.95	78.00	0.46	77.46	0.48	77.20	0.40	77.44	0.58	77.44	0.26
0.98	74.40	0.18	72.98	0.26	72.68	0.08	73.22	0.26	72.80	0.14
AUC	0.9638		0.9605		0.9627		0.9607		0.9616	
total_sv	18,719		18,639		19,006		18,989		19,063	

Table S5
Acceptor SVM model performance comparison. We randomly generated five sample sets of equal size and trained acceptor SVMs. The model with the best performance according to its AUC value (set 1) was used for classification. The comparison provides evidence that random sampling only marginally affects the performance.

set	1		2		3		4		5	
	TP	FP	TP	FP	TP	FP	TP	FP	TP	FP
p>0.5	84.44	9.44	84.38	8.84	84.14	8.38	84.36	7.92	84.84	8.16
0.8	77.40	3.56	77.02	3.36	76.28	2.94	77.20	3.28	76.84	2.78
0.9	73.20	1.80	72.86	2.00	71.58	1.64	72.78	1.94	72.82	1.52
0.93	70.94	1.34	70.74	1.44	69.30	1.16	70.98	1.34	70.62	1.14
0.95	69.26	1.10	68.32	1.10	67.50	0.84	69.08	0.94	68.62	0.86
0.98	64.28	0.72	63.44	0.54	62.70	0.40	64.46	0.30	63.78	0.46
AUC	0.9411		0.9387		0.9390		0.9407		0.9394	
total_sv	23,394		23,530		23,492		23,410		23,452	

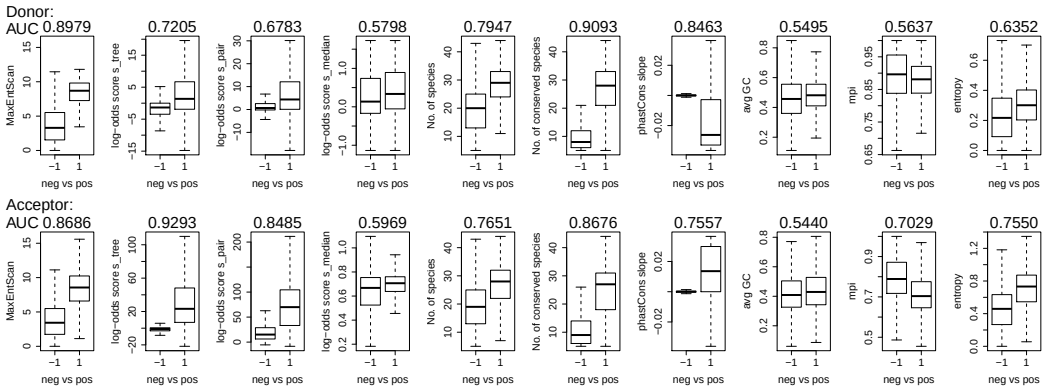

Fig. S5. **Score distributions of splice site features.** For both donors (top row) and acceptors (bottom row) the figure depicts the score distributions for all positive and negative SVM training samples. AUC values indicate the ability of each feature to distinguish between real and false splice sites (1 would be the optimum). The `MaxEntScan` and the improved tree-based log-odds scores are key features of our approach. Some ordinary features, like GC-content, badly separate the training data on their own. The AUC values reveal that the novel tree-based log-odds substitution scores (second boxplot from the left) are superior to the pairwise method of previous studies (third boxplot).

Table S6

Prediction of splice sites in human intergenic regions. The table gives the number of predictions made for different SVM classification confidence cut-offs. Confirmed splice sites are supported by human ESTs and/or have been experimentally validated by available RNA-seq data (Wang et al., 2008).

	novel donors			novel acceptors		
p	*total*	*confirmed*	%	*total*	*confirmed*	%
0.5	927,693	17,250	1.86	2,497,067	18,213	0.73
0.8	309,192	9,565	3.09	886,180	10,474	1.18
0.9	155,890	6,898	4.42	468,934	7,280	1.55
0.93	111,992	5,911	5.28	344,063	6,012	1.75
0.95	83,323	5,235	6.28	258,302	5,059	1.96
0.98	38,877	3,777	9.72	120,670	3,202	2.65
0.99	22,225	2,982	13.42	68,583	2,353	3.43

Table S7
Number of predicted splice sites that coincide with other gene-finding approaches. ∪ denotes the number of splice sites that are found by us and the union of the listed methods.

approach	*donor*		*acceptor*	
	$p > 0.5$	$p > 0.9$	$p > 0.5$	$p > 0.9$
`AceView`	6,215	2,552	5,719	2,321
`CONTRAST`	2,201	1,688	1,652	1,195
`Ensembl`	740	554	583	381
`Gencode` (auto)	121	90	80	60
`Gencode` (manual)	327	191	364	176
`GeneID`	3,046	1,362	3,803	1,784
`Genscan`	7,038	2,456	8,836	3,331
`N-SCAN`	2,863	1,703	2,297	1,258
`SGP`	2,695	1,515	3,109	1,714
`VEGA`	1,909	873	1,768	797
∪	15,661	5,644	18,846	7,318

Table S8
Evolutionary conservation of predicted exons. The table lists the number of exon candidates with respect to lower bounds of the number of conserved species in both splice sites at different SVM class-probabilities. Here, while restricted to splice sites with $p > 0.5$, p denotes the average of both splice site SVM probabilities. See Fig. S6 for a heatmap of these data and further interpretation.

No. of species	exons	
	$p > 0.5$	$p > 0.9$
>10	216,442	31,300
>15	157,828	26,580
>20	90,341	18,007
>25	38,051	8,444
>30	11,425	2,499
>35	1,507	325
>40	87	13

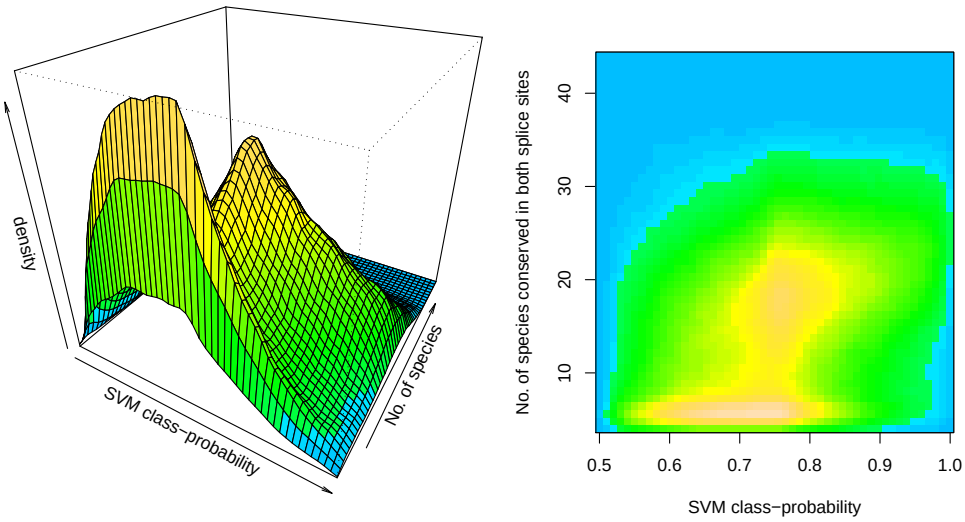

Fig. S6. Density heatmaps. The figure illustrates the number of species that are conserved in sequence alignments of two particular splice sites enclosing predicted exons in dependence of their average SVM class-probability. Requiring at least ten conserved species in both splice sites seems to be appropriate to filter for reliable exons. Below ten, we assume to observe either false positives, or more intuitively explaining the second peak at five species, genes that are specific to certain subgroups, e.g. primates or teleosts.

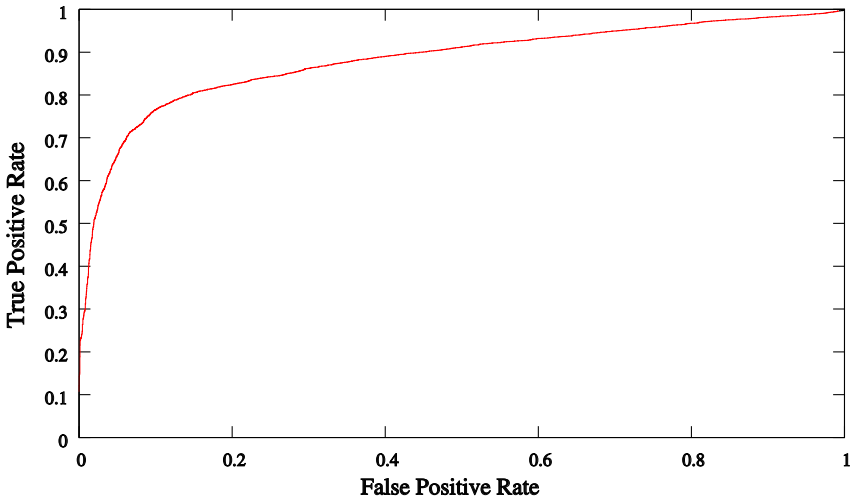

Fig. S7. Performance of the exon-SVM on an independent set of `RefSeq` exons. Similarly to our procedure of assembling novel exons by searching for pairs of compatible splice sites at intergenic regions, we assembled a set of 9,333 real and 4,722 false exons from the `RefSeq` database. On this independent test-set which was not used in any previous training step (neither the splice site SVM nor the exon-SVM), we obtained an AUC of 0.88. At $p > 0.9$ we obtained a true positive rate of 38% (3,528/9,333) and a false positive rate of 1% (57/4,722).

Table S9
Number of predicted exons that are already annotated by existing gene-finding approaches. We distinguish between all exons that passed the exon-SVM (8,832 exons) and the fraction of predicted exons that are part of genomic clusters (734 exons). ∪ denotes the number of predicted exons that are found by us and the union of the listed methods. Overall, 10% of our predicted exons are already listed by competing gene-finding approaches. Remarkably, this increases to 29% if we consider clustered exons.

approach	*no. of predicted exons*		
	all	clustered	non-clustered
`AceView`	346	59	287
`CONTRAST`	381	147	234
`Ensembl`	92	25	67
`Gencode` (auto)	89	30	59
`Gencode` (manual)	207	57	150
`GeneID`	250	90	160
`Genscan`	415	127	288
`N-SCAN`	374	130	244
`SGP`	271	100	171
`VEGA`	163	46	117
∪	877	216	661

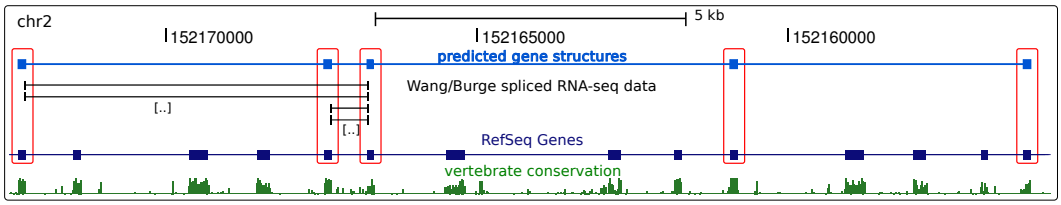

Fig. S8. ***De novo* prediction of exons in the *NEB* gene.** Due to excessive alternative splicing at this locus, `RefSeq` does not provide a complete list of all isoforms but, by the time of SVM training, released a novel consensus gene structure containing additional exons which were not part of our prior training set. This update validated all five predicted exons at the depicted locus. RNA-seq data confirms some exon-exon junctions and indicates the existence of exemplary exon skipping isoforms.

Table S10
Overview of the number of exon-SVM-derived exons with short read support. Here, we only consider reads of high confidence which Wang *et al.* could uniquely map to the genome ($\geq$ 2 mismatches). To be counted, at least two reads have to fit to our exon over their full length (32 nt). We distinguish between reads of different tissues or cell-lines but also list the total number of read-supported exons as the union over all reads. (A) Total number of exons with intersecting reads. (B) Number of exons without a known exon at the opposite strand according to the UCSC `RefSeq` and `UCSC Genes` track. This gives a more reliable count that corrects for ambiguous cases: the reading direction of randomly primed reads is unknown and in cases of two opposing exons in sense and antisense direction it is unclear which of them is actually indicated by the particular reads. (C) Read-indicated exons without any read at their 200 nt up-/downstream flanking region (this ensures that our predicted exon is not part of a larger exon but may also wrongly discard some true positive exons, especially in cases of intronically hosted transcripts (e.g. miRNAs) in close proximity to the exon. The intersection of (A-C) constitutes a reliable but very conservative estimate of the number of short read confirmed exons. As outlined, we applied stringent filters to resolve ambiguities resulting from randomly primed short read data, and most likely considerably underestimate the true number of read-confirmed exons.

		A		B		C		A+B+C	
		SVM class-probability threshold							
GEO accession	tissue/cell-line	0.5	0.9	0.5	0.9	0.5	0.9	0.5	0.9
GSM325476	brain	621	48	137	18	109	13	58	7
GSM325477	liver	370	27	78	9	77	6	34	3
GSM325478	heart	479	40	111	16	113	13	60	7
GSM325479	skeletal muscle	513	46	117	24	97	14	53	11
GSM325480	colon	615	56	147	27	102	18	61	15
GSM325481	adipose	615	52	163	25	103	16	59	13
GSM325482	testes	872	98	301	61	187	40	132	35
GSM325483	lymph node	685	63	201	37	111	16	70	15
GSM325484	BT474	593	62	203	38	118	14	71	11
GSM325485	HME	590	59	177	33	120	17	69	13
GSM325486	breast	586	53	158	29	112	16	59	13
GSM325487	MCF-7	583	60	184	35	88	11	49	8
GSM325488	MB435	570	56	166	34	88	14	54	12
GSM325489	T47D	575	56	204	41	98	14	64	13
GSM325490	MAQC human	606	61	170	26	122	18	70	14
GSM325491	MAQC UHR	495	40	122	22	100	8	48	7
	$\cup$	1,613	193	782	140	736	103	469	87

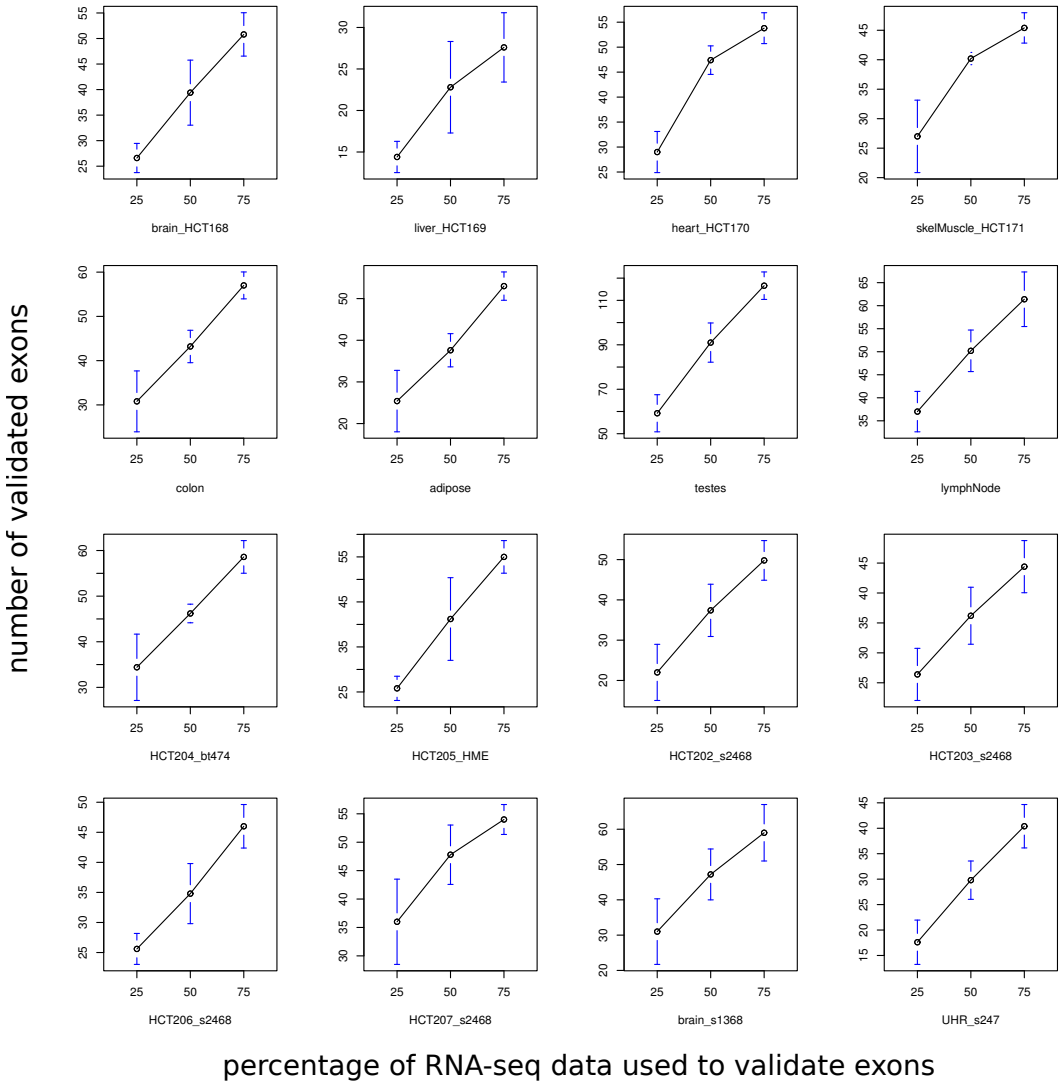

Fig. S9. The number of confirmed exons increases almost linearly with the number of reads. We split the data of Wang *et al.* to subsets containing 25%, 50%, and 75% of all reads and computed the corresponding number of confirmed exons. We repeated this procedure five times to avoid biases of a single random split. Since the number of confirmed exons increases almost linearly with the number of reads, it is likely that future, probably deeper, sequencing projects provide further experimental evidence for our approach.

References

Wang, E. T., Sandberg, R., Luo, S., Khrebtukova, I., Zhang, L., Mayr, C., Kingsmore, S. F., Schroth, G. P., and Burge, C. B., 2008. Alternative isoform regulation in human tissue transcriptomes. *Nature* **456**: 470–476.