SUPPLEMENT

A. Supplementary Methods

A.1. Maximum Entropy Scoring of Splice Sites.

In order to determine whether a splice site candidate is likely to be functional or not we employ well-tested maximum entropy models (MEM) introduced by (Yeo and Burge 2004). These are based on the maximum entropy distributions (MED) of sequence motifs of the donor and acceptor sites. These distributions are specified with constraints that are exclusively estimated from known transcript data. Thus MEDs, consistent with a set of constraints, are the most unbiased approximation for splice site sequence motif modeling. The `MaxEntScan` scores were computed using the perl wrappers provided for download by Burge Lab at `http://genes.mit.edu/burgelab/maxent/download`.

A.2. False Positive Rate Estimation

We sampled random non-exonic positions from the human genome with the additional requirement of a present canonical motif (GT/AG). About 31 % of these sites were alignable to mouse. In order to make an estimation on the false discovery rate on ortholog sites, we scored the aligned sequences with `MaxEntScan`. Only 1.2 % and 3.0 % of all GT and AG sites, respectively, had a score > 3 in the aligned mouse sequence. Therefore the cut-off of > 3 is used throughout this study. Supplemental Fig. S1 shows the distribution of the described `MaxEntScan` scores. It is expected that more distant species, have an even lower false discovery rate.

A.3. Estimation of conservation on transcript level

Throughout the main text a transcript is considered as conserved if it contains at least one conserved splice site. One could argue that this threshold might be too low. To check the effect of the choice of threshold we repeated our analysis requiring that at least 40 % of all splice sites in a transcript must be conserved between two species. A comparison of Supplemental Fig. S2 with Fig. C4 (B) shows that the results change surprisingly little when employing a much more stringent cutoff. Although the absolute number of conserved lncRNAs drops, the relative conservation (disregarding non-aligned sites) still covers more than one third for mammals.

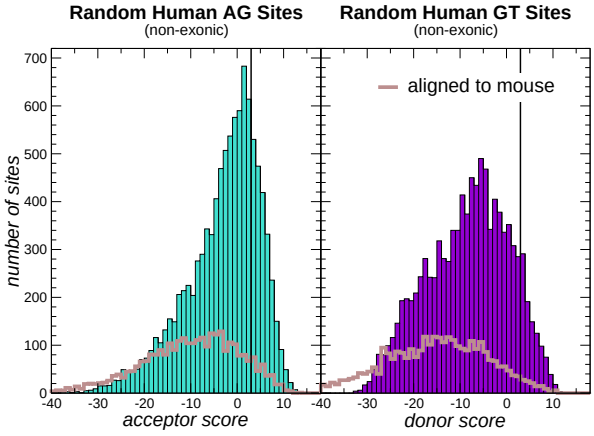

Figure S1: **Distribution of `MaxEntScan` scores for random human non-exonic GT-AG sites.** In brown the distribution of the `MaxEntScan` scores for the ortholog mouse sequence is displayed. Only 1.2 % and 3.0 % of the sampled GT and AG sites, respectively, have ortholog mouse sequence with a score > 3.

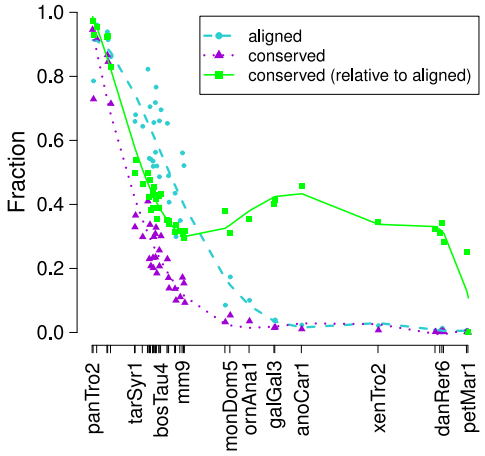

Figure S2: **Conservation of lncRNA transcripts across 46 vertebrates.** Here the effect on the conservation rate of lncRNA transcripts is demonstrated, if a rate of 40 % conserved splice sites per transcript is required for a transcript to be considered as conserved. This figure should be compared to Fig. C4 (B).

B. ENSEMBL *versus* UCSC Alignments

B.1. Estimates of splice site conservation are limited by alignment coverage and quality

The multiple sequence alignment underlying the splice site map has a major influence on the estimates of splice site conservation. Even though the total coverage of the two alignments is quite similar: in the UCSC

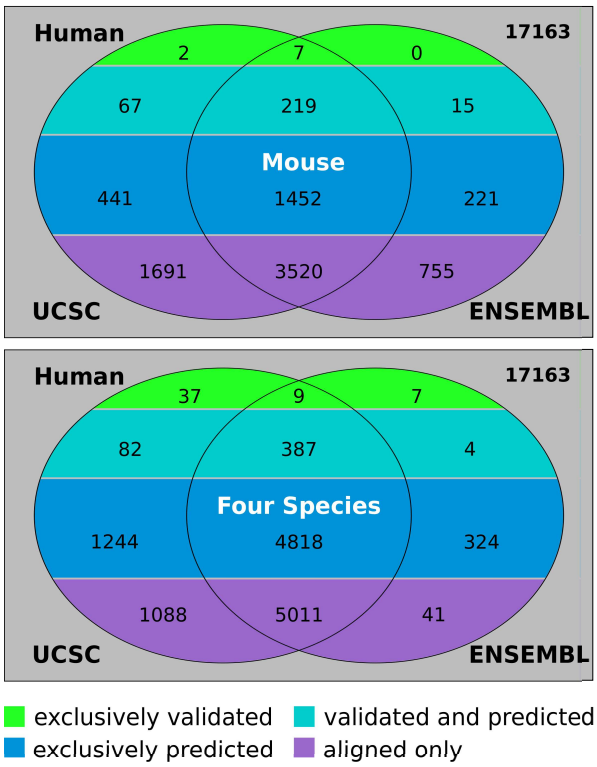

Figure S3: **Comparison of UCSC and ENSEMBL alignment** regarding influence on the estimates of splice site conservation. All splice sites of 17, 163 human lncRNAs, aligned to the considered species are shown distinguished in four groups within the venn diagram. Top: number of human splice sites found in mouse. Bottom: number of human splice sites present in at least one out of mouse, rat, dog, and cow.

alignments, about 31 % of the whole human genome is aligned to a mouse sequence, in the ENSEMBL alignments, the fraction is 27 %. This small difference cannot explain the discrepancy of about one fifth in the coverage of splice sites.

B.2. Differences in lncRNA sets

The observed splice site conservation differs significantly between the two genome-wide alignments used here.

For the majority of the human GENCODE lncRNA splice sites, no aligned mouse sequence is reported in either alignment. Supplemental Fig. S3 shows the overlaps between the two alignments. Surprisingly, the alignable sequence fragments differ quite a bit between the two different alignments. Although the coverage of the UCSC alignment is larger ($\sim$ 4 %), there are still nearly one thousand human splice sites for which the ENSEMBL alignment proposes homologous sequence while no sequence at all is aligned in the UCSC alignment. Integrating over the four eutherian species, however, increases the overlap by 16 % to more than 78 %.

As expected from the larger coverage of UCSC alignment, we also observe more aligned lncRNA splice sites. The fraction of conserved ones among those that are alignable is also comparable. Interestingly, most (89 %) of the loci that are alignable in the ENSEMBL alignment *only*, correspond to conserved splice sites in at least one the four non-primate mammals, Supplemental Fig. S3. Combining the results of the two alignments, we obtain a lower bound estimate of 40 % for the fraction of splice sites in lncRNAs that originated early in the evolution of placental mammals. Although the two alignments show a substantial overlap, the fact that we can find hundreds of splice sites whose conservation is visible only in the more stringent ENSEMBL alignment strongly suggests that the actual numbers might still be higher.

For the alternative data set of microRNA and snoRNA host genes, the data for UCSC and ENSEMBL alignments are also quite similar, Supplemental Tab. S2. Here again the coverage is a bit smaller for the ENSEMBL alignments.

B.3. Differences in RefSeq annotated sets

Supplemental Tab. S1 outlines major differences in the observed conservation of splice sites compared to Tab. 1 in the main text. For splice sites in coding regions it makes a difference of nearly 12 %, for UTRs even up to 15 % change of estimated conservation rate. By disregarding the non-aligned sites, the resulting upper bounds on conservation rate are almost the same for both alignments. Interestingly, the upper bounds in ENSEMBL are slightly higher (up to 0.3 %) than in UCSC alignments.

B.4. Differences in upper bound estimation

The ENSEMBL alignments contain relatively more conserved splice sites than the UCSC data, this difference is even enhanced when data aggregated to the level of transcripts, Supplemental Tab. S3. While in coding sequences the gap between the estimated upper bounds on the level of single splice sites is only 0.3 %, the difference increases by 10-fold on the transcript level of lncRNAs.

C. Conservation of Protein-Coding Splice Sites

We used RefSeq annotated transcripts in order to investigate the conservation of protein-coding sequences.

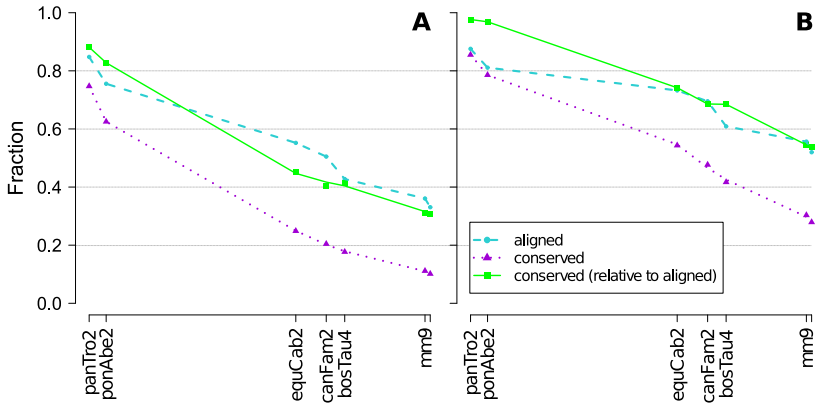

Figure S4: **Conservation of lncRNAs across eight mammals according to ENSEMBL alignment.** The estimated conservation on the level of **(A)** 17,163 single splice sites and **(B)** 5,413 transcripts is similar to the estimation resulting from the UCSC alignment. In (B) a transcript is considered conserved if a single splice site is conserved in that transcript.

Table S1: **Conservation of RefSeq splice sites between human and mouse based upon ENSEMBL alignment.** This table is to compared with Tab. 1 from the main text, in order to see differences of estimations based upon ENSEMBL or UCSC alignment.

	human	mouse			
Data Set	*N*	aligned	predicted	validated	conserved
RefSeq coding	355,573	260,507	249,588	251,385	256,045
RefSeq 5'-UTR	16,035	8,622	6,022	5,024	6,120
RefSeq 3'-UTR	1,124	608	501	445	511

Table S2: **Conservation of miRNA and snoRNA host genes based upon ENSEMBL alignment.** We tabulate the number of conserved lncRNAs in selected species and in at least one of five Eutheria (human, mouse, rat, cow, dog).

	aligned	predicted	validated
128 human transcripts hosting microRNAs			
mouse	82	53	12
dog	106	81	1
5 Eutheria	109	99	17
73 human transcripts hosting snoRNAs			
mouse	47	42	26
dog	62	54	19
5 Eutheria	63	57	34

Table S3: **Upper bounds on the percentage of conserved splice sites and transcripts in lncRNAs.** The numbers are estimated from the fraction of conserved splice sites within aligned sequence blocks only.

Alignment	mouse	rat	cow	dog	*union*
splice sites					
UCSC	29.6	29.7	40.4	39.5	51.6
ENSEMBL	30.9	30.7	41.4	40.5	52.3
transcripts					
UCSC	50.7	50.5	66.3	67.1	79.6
ENSEMBL	54.5	53.7	68.5	68.6	79.5

Supplemental Tab. S4 shows conservation of coding splice sites in numbers in four chosen mammals, namely mouse, rat, dog, and cow. The chart in Supplemental Fig. S5 illustrates that the predicted level of conservation (blue and cyan colored) is similar in all of these species, while only in mouse it is almost identical with the number of validated splice sites (cyan). This suggests that there still might be several thousands of unannotated protein-coding splice sites, in this example especially in dog.

Table S4: **Conservation of RefSeq splice sites.** RefSeq annotated transcripts were used for estimation of coding transcripts only, since the majority of the non-coding RefSeq transcripts are still associated with coding loci.

	all	coding	3'-UTR	5'-UTR
Human		355,573	1,124	16,035
Mouse				
aligned		340,327	828	11,737
predicted		325,323	680	8,200
validated		326,401	607	6,908
conserved		333,661	693	8,339
Rat				
aligned		324,604	770	10,954
predicted		310,135	627	7,669
validated		276,676	522	5,090
conserved		317,055	635	7,753
Dog				
aligned		343,042	915	12,111
predicted		327,591	761	8,485
validated		149,575	455	2,614
conserved		331,434	768	8,527
Cow				
aligned		337,301	880	12,711
predicted		322,453	747	9,109
validated		269,218	543	5,404
conserved		329,448	753	9,217

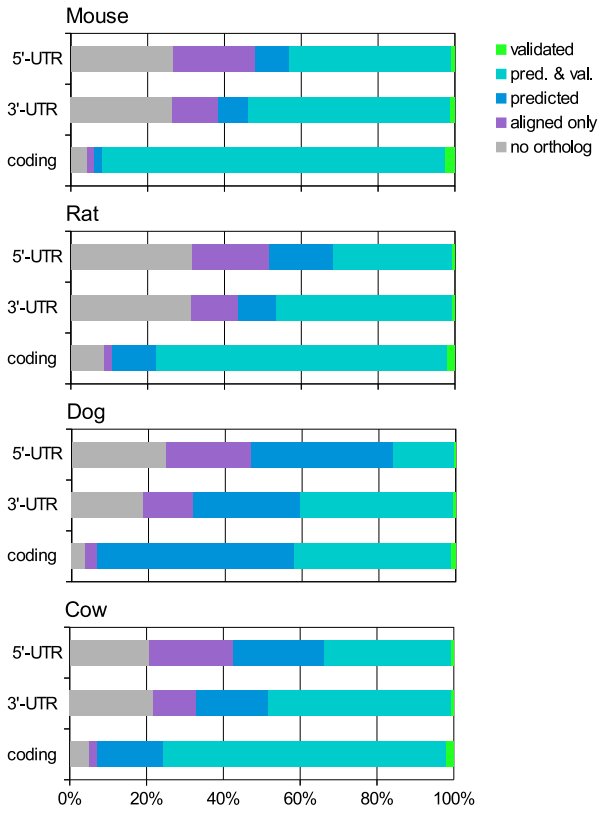

Figure S5: **Conservation of RefSeq splice sites in mouse, rat, dog and cow.** Graphical illustration of numbers displayed in Supplemental Tab. S4.

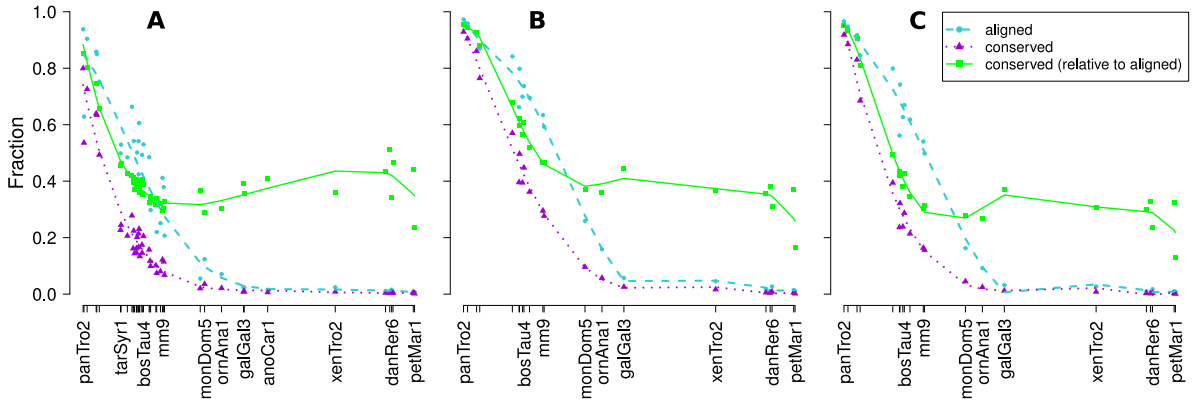

Figure S6: **Conservation of lncRNAs from (Cabili et al. 2011)**. The estimated conservation on the level of **(A)** 32, 515 single splice sites, and 14, 274 transcripts with a required conservation rate of **(B)** at least one splice site per transcript and **(C)** more than 40 % of splice sites per transcript - is similar to the estimation resulting from our filtered lncRNA data set. In Panel (B) and (C) only the results of 22 species of the 46 vertebrates of the UCSC alignment are plotted in the graphs.

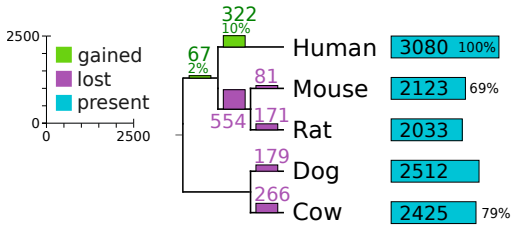

Figure S7: **Turnover of individual lncRNA splice sites.** The figure gives an overview on the number of gained and lost splice sites for 814 lncRNAs that have at least one splice site conserved between human and all of the four depicted mammals.

Table S5: **Multi-exon lncRNAs.** In dependence of the number of exons per transcript, we provide the number of lncRNAs, the underlying number of splice sites and their average human `MaxEntScan` score. For each splice site, we furthermore report the average as well as the maximum number of species in which we found it. The splice site scores only slightly increase with the number of exons per transcript. Furthermore, we observed some "ultra-conserved" splice sites which can be traced in nearly all vertebrate genomes.

exons	2	3	4	≥ 5
lncRNAs	2,493	1,545	791	584
splice sites	4,770	5,665	4,260	4,342
score_{avg}	7.1	7.4	7.6	7.6
species_{avg}	9.8	9.6	10.0	10.1
species_{max}	44	41	39	40
$\text{splice sites}_{\geq 40}$	6	8	0	1

COLOR FIGURES

Figure C1: **Splice site map of the GAS5 locus.** Each line represents a splice site, each column a vertebrate genome arranged in increasing phylogenetic distance from human; `MaxEntScan` scores for splice site quality are color coded; missing data are indicated as gray background. Light green entries indicate validated splice sites present in RefSeq or sites that are experimentally confirmed by more than a single EST.

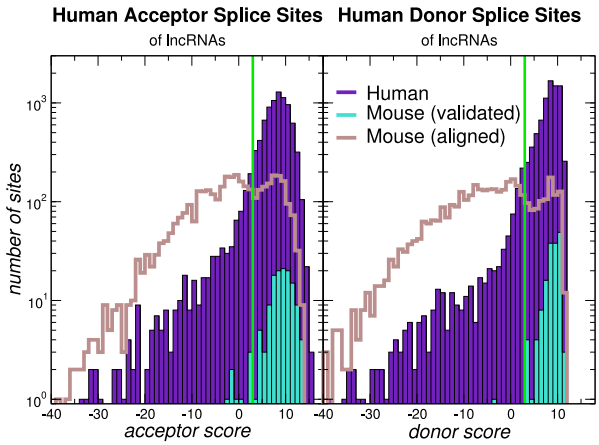

Figure C2: **Conservation of splice sites of human lncRNAs in the mouse.** Filled curves designate the distributions of `MaxEntScan` scores for human splice sites (purple) and orthologous positions that are known to be splice sites in mouse (cyan). The score distribution of all orthologous positions in mouse (brown) is a superposition of conserved functional splice sites and positions that have been destroyed by substitutions. The cut-off value of 3.0 is indicated by a green line.

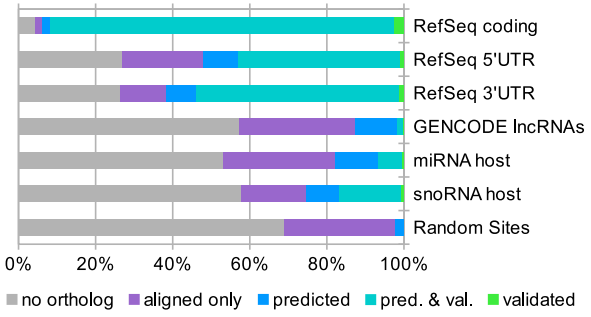

Figure C3: **Conservation of splice sites between human and mouse in different contexts.** In non-gray colors the fraction of all alignable splice sites is shown. Colors from green to blue display the estimated conservation rate. Consequently, the fraction of alignable but likely non-conserved splice sites is shown in purple. The overlap of our predicted and the validated splice sites is displayed in turquoise. In protein-coding RNAs 95 % of the splice sites are at least alignable to mouse, and of those almost all are conserved. While in lncRNAs the rate of alignable sites drops to around 40 %. The fraction of validated splice sites amongst the predicted ones turned from nearly 98 % to only 13 %, indicating that there are a lot of unannotated splice sites.

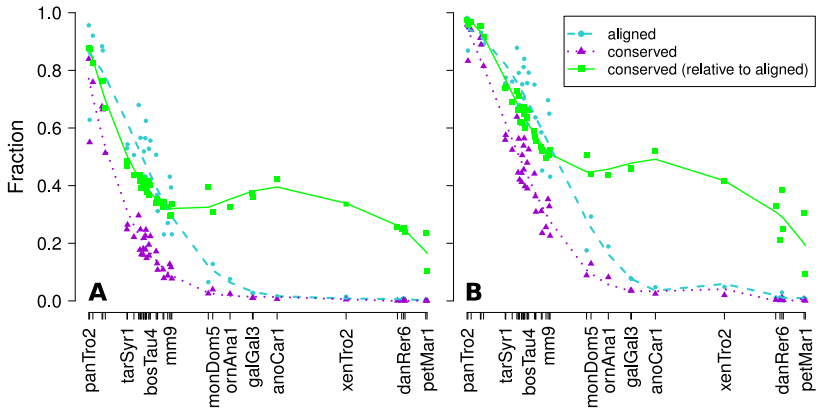

Figure C4: **Conservation of lncRNAs across 46 vertebrates.** Indicated in blue is the fraction of aligned splice sites, in purple the fraction of splice sites that are validated and/or predicted to be a functional splice site in the regarding species. In green the upper bounds on the fraction of conserved splice sites are shown. The numbers are estimated from the fraction of conserved splice sites within aligned sequence blocks only. Panel **(A)** shows the conservation rate of 17, 163 single splice sites, while panel **(B)** illustrates the conservation on the level of transcripts for 5, 413 lncRNAs.

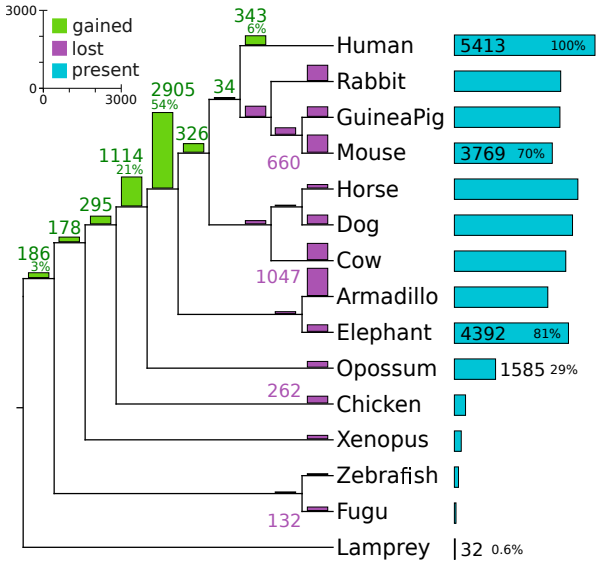

Figure C5: **Gains and losses of human GENCODE lncRNAs across the vertebrates.** Events are inferred by the parsimony criterion: a gene is deemed lost along the edge leading to a maximal subtree for which it is not observed at any leaf; a gain event is placed on the edge leading to the last common ancestor of all observed occurrences. The vertebrate phylogeny is the `phyloFit` tree provided by the UCSC browser. The primate subtree is omitted.

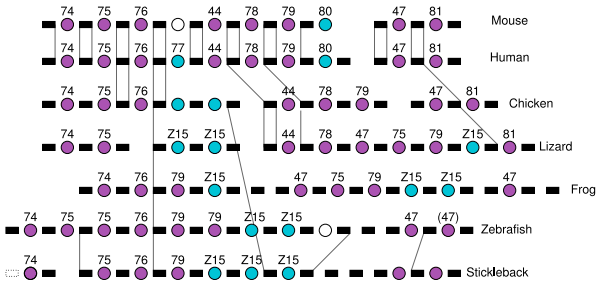

Figure C6: **Conserved splice sites of the GAS5 lncRNA.** The GAS5 snoRNA host gene is among the most highly conserved lncRNAs. Its homologs are easily identifiable via the well-conserved snoRNAs (circles) located within its introns. Members of the SNORD80/Z15 family are shown in blue. Black boxes indicate the major exons supported by RefSeq and/or EST data. Gray lines indicate splice sites that can be traced manually in at least one of the genome-wide alignments available in the UCSC browser. Note that only a subset of these is represented in any individual alignment, *c.f.* Fig. C1. The transcript structure as well as its snoRNA payload has changed also by means of duplications and deletions.

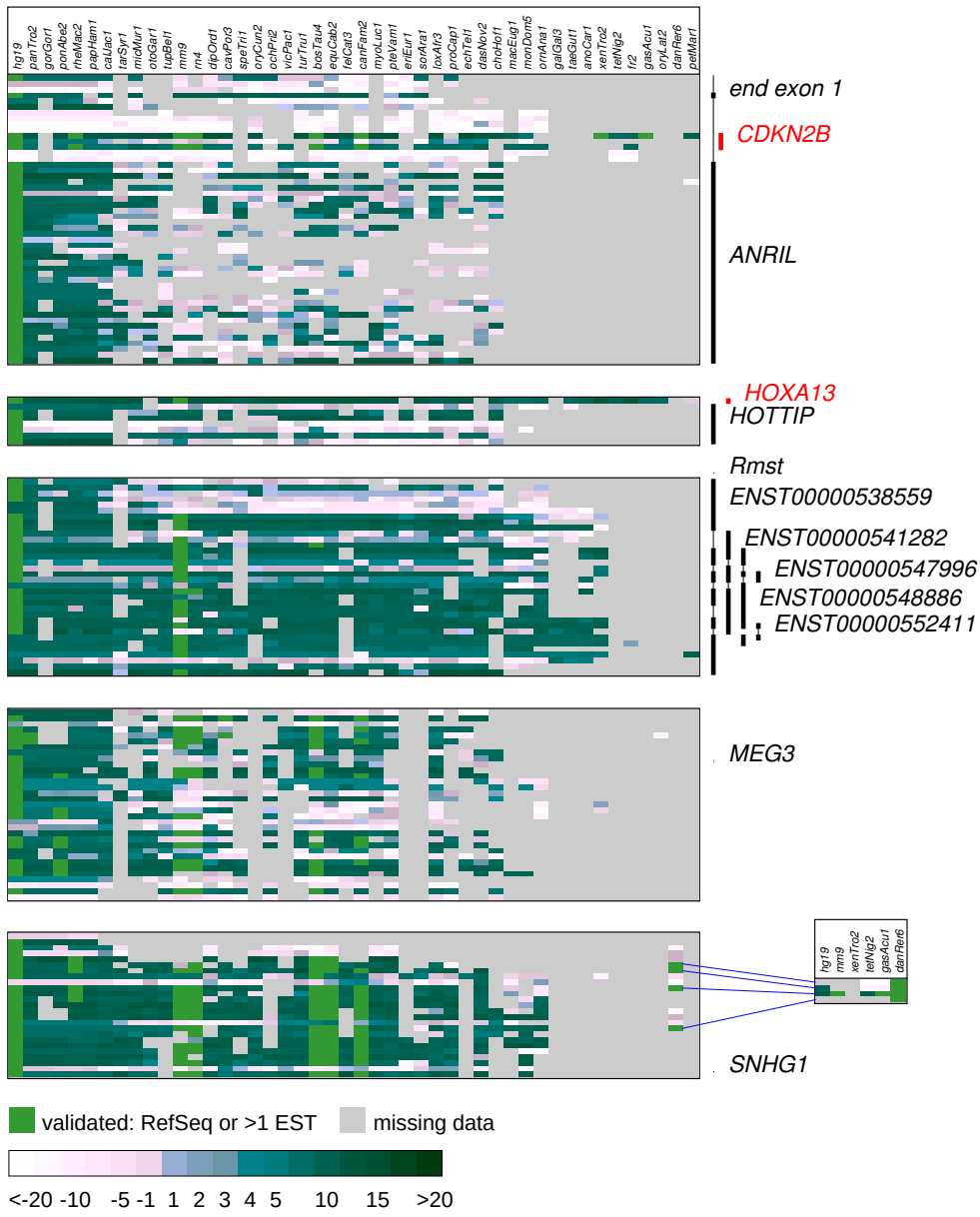

Figure C7: **Variation of splice site conservation.** The patterns of splice site conservation vary substantially between different lncRNAs, even when their evolutionary age is comparable. The main panel refers to the UCSC 46-way alignment. In the case of ANRIL, only a few splice sites are conserved outside the primates. Although the mouse ortholog shares at least some functions with human ANRIL (Pasmant et. al 2010), there are only four shared conserved splice sites. HOTTIP, with few exons that are partially conserved, is also a rather typical chromatin-related lincRNA. In contrast, the overwhelming majority of splice sites is conserved in Rmst. MEG3 shows an intermediate pattern, with more lineage-specific losses. The snoRNA host gene SNHG1 contains several splice sites that are deeply conserved among vertebrates. Some are even found in teleosts. Experimentally known splice sites from zebrafish SNHG1 were searched also in the 6-way zebrafish `multiz` alignment (inset). Additional homologous splice sites in two teleosts demonstrate once more the limitations arising from alignment quality. The color scheme is explained in Fig. C1. Thick vertical bars on the right mark splice sites that belong to a specific transcript (black: plus strand, red: minus strand). Thin lines between these bars indicate conserved splice sites, that are not part of the annotated transcripts.