BIOINFORMATICS ORIGINAL PAPER
Vol. 27 no. 14 2011, pages 1894–1900
doi:10.1093/bioinformatics/btr314

Genome analysis

Advance Access publication May 26, 2011

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

Dominic Rose[1,2,*], Michael Hiller[3], Katharina Schutt[4,5,6,12], Jörg Hackermüller[1,5,12], Rolf Backofen[2,7,8] and Peter F. Stadler[1,5,9,10,11]

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Leipzig, [2]Bioinformatics Group, Department of Computer Science, University of Freiburg, Freiburg, Germany, [3]Department of Developmental Biology, Stanford University, Stanford, CA 94305, USA, [4]LIFE - Leipzig Research Center for Civilization Diseases, University of Leipzig, [5]Fraunhofer Institute for Cell Therapy and Immunology, AG RNomics, [6]Department of Molecular Immunology, University of Leipzig, Leipzig, Germany, [7]Centre for Biological Signalling Studies (BIOSS), [8]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Freiburg, Germany, [9]Interdisciplinary Center of Bioinformatics, University of Leipzig, Leipzig, Germany, [10]Institute for Theoretical Chemistry, University of Vienna, Vienna, Austria, [11]Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe NM 87501, USA and [12]Young Investigators Group Bioinformatics and Transcriptomics, Department of Proteomics, Helmholtz Centre for Environmental Research (UFZ), Leipzig, Germany

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Long non-coding RNAs (lncRNAs) resemble protein-coding mRNAs but do not encode proteins. Most lncRNAs are under lower sequence constraints than protein-coding genes and lack conserved secondary structures, making it hard to predict them computationally.

Results: We introduce an approach to predict spliced lncRNAs in vertebrate genomes combining comparative genomics and machine learning. It is based on detecting signatures of characteristic splice site evolution in vertebrate whole genome alignments. First, we predict individual splice sites, then assemble compatible sites into exon candidates, and finally predict multi-exon transcripts. Using a novel method to evaluate typical splice site substitution patterns that explicitly takes the species phylogeny into account, we show that individual splice sites can be accurately predicted. Since our approach relies only on predicted splice sites, it can uncover both coding and non-coding exons. We show that our predicted exons and partial transcripts are mostly non-coding and lack conserved secondary structures. These exons are of particular interest, since existing computational approaches cannot detect them. Transcriptome sequencing data indicate tissue-specific expression patterns of predicted exons and there is evidence that increasing sequencing depth and breadth will validate additional predictions. We also found a significant enrichment of predicted exons that form multi-exon transcript parts, and we experimentally validate such a novel multi-exon gene. Overall, we obtain 336 novel multi-exon transcript predictions from human intergenic regions. Our results indicate the existence of novel human transcripts that are conserved in evolution and our approach contributes to the completion of the human transcript catalog.

Availability and Implementation: Predicted human splice sites, exons and gene structures together with a Perl implementation of the tree-based log-odds scoring and a supplementary PDF file containing additional figures and tables are available at: http://www.bioinf.uni-leipzig.de/publications/supplements/10-010. The five experimentally confirmed partial transcript isoforms have been deposited in GenBank under accession numbers HM587422-HM587426.

Contact: dominic@bioinf.uni-leipzig.de

Supplementary Information: Supplementary data are available at *Bioinformatics* online.

Received on February 9, 2011; revised on April 28, 2011; accepted on May 19, 2011

1 INTRODUCTION

A series of high-throughput transcriptomics studies utilizing a variety of different technologies revealed that mammalian genomes are pervasively transcribed into a complex mosaic of transcripts (Carninci *et al.*, 2005; ENCODE Project Consortium, 2007; Kapranov *et al.*, 2007a, b). A large extent of these transcripts consists of small and long non-protein-coding RNAs (ncRNAs). Due to the diverse nature of these transcripts, our catalog of genes is still incomplete.

Computational prediction of protein-coding genes is based on characteristic features of coding regions that distinguish them from non-coding DNA (Burge and Karlin, 1997; Cruveiller *et al.*, 2003). Coding genes exhibit a clear evolutionary signature, since mutations are often synonymous and preserve the reading frame. These signals can be exploited to find coding genes by comparative genomics methods (Solovyev *et al.*, 2006; Stark *et al.*, 2007). Also machine learning-based approaches that recognize splice sites, transcriptional and translational start and stop signals can successfully detect coding genes (Gross *et al.*, 2007; Schweikert *et al.*, 2009; Stanke and Waack, 2003).

In contrast to protein-coding genes, ncRNAs form a heterogeneous class of transcripts that lacks common sequence patterns, complicating their detection in genomic DNA. Some

*To whom correspondence should be addressed.

ncRNA classes, including common families like rRNAs, tRNAs and miRNAs, preserve their characteristic secondary structure during evolution, guiding a computational prediction (Nawrocki *et al.*, 2009; Washietl *et al.*, 2005). However, many other ncRNAs exhibit neither conserved secondary structures nor sequence conservation levels as high as coding exons (Pang *et al.*, 2006; Ponjavic *et al.*, 2007), making it hard to find them computationally. Many long non-coding RNAs (lncRNAs) resemble protein-coding mRNAs in that they are often capped, spliced and polyadenylated. They can exhibit cell type-specific expression, are known to be involved in transcriptional regulation, epigenetics, gene silencing, imprinting and are known to play a major role in some human diseases (Huarte and Rinn, 2010; Mercer *et al.*, 2009; Ponting *et al.*, 2009; Wilusz *et al.*, 2009). Examples include *XIST* which is involved in mammalian female X chromosome inactivation and dosage compensation (Senner and Brockdorff, 2009), *MALAT1* which affects the expression of genes controlling synapse formation (Bernard *et al.*, 2010) and *NRON* which regulates nuclear trafficking by repressing the nuclear factor of activated T cells (Willingham *et al.*, 2005).

We recently presented a computational approach to detect lncRNAs with conserved intron positions in insect genomes (Hiller *et al.*, 2009). This approach is solely based on a genomic screen for regions that evolve like introns, exploiting that the intron boundaries (splice sites) are well conserved and under purifying selection in both coding and non-coding genes (Chodroff *et al.*, 2010; Ponjavic *et al.*, 2007; Rodríguez-Trelles *et al.*, 2006). The approach crucially relied on predicting introns as a single unit, which means pairs of splice donor (5′) and acceptor (3′) sites are predicted in a single step. This strategy works in insect genomes where introns are very short [according to Lim and Burge (2001) most are <100 nt in *Drosophila melanogaster*]. In contrast, vertebrate introns are substantially longer and more variable in their length, preventing the application of this intron-based method.

Here, we introduce a novel approach to *de novo* predict spliced transcripts in intergenic regions of vertebrate genomes. Using a combination of comparative genomics and machine learning, we first predict new splice donor and acceptor sites in whole genomes and subsequently assemble them into exon predictions. Applying this approach to an alignment of 44 vertebrate genomes, we predict previously unreported coding and non-coding transcripts with conserved exon/intron structures in the well-characterized human genome and validate these predictions with available transcript data and own experiments.

2 METHODS

Input data: our analysis is based on the multiple alignment of 44 vertebrate genomes with the human hg18 assembly as the reference downloaded from the UCSC Genome Browser (Rhead *et al.*, 2010).

De novo splice site prediction: we trained support vector machines (SVMs) to solve the binary classification problem of *de novo* splice site prediction comprising the following steps: (i) detect donor and acceptor splice site candidates in multiple sequence alignments; (ii) train splice site SVMs with features capturing patterns of splice site evolution; and (iii) use the SVM to score candidate splice sites.

To this end, we screened the genome alignment of genic regions for donor and acceptor candidates and divided them into real splice sites that are annotated (>208 000 true positives) and false positives that are not supported by available transcript data (∼12.6 million). To perform supervised machine learning, we compiled three disjoint sets: (i) positive and (ii) negative samples to train and test individual donor and acceptor SVM models and (iii) a set of intergenic candidate sites forming the search space for putative novel splice sites. The positive set contained the splice sites of the `UCSC`, `RefSeq` and the `Human mRNA` gene tracks. Negative training data were the remaining genic sites [unannotated sites within introns, exons or untranslated regions (UTRs)]. We considered GT (donor) and AG (acceptor) dinucleotides (both strands) which were conserved in at least five species. Alignment blocks had to contain the intervals $[-3,6]$ for donors and $[-19,2]$ for acceptors (Supplementary Fig. S3). To avoid obvious false positives, sites with a `MaxEntScan` (Yeo and Burge, 2004) score <0 were discarded (Supplementary Fig. S4). We only considered canonical (GT/AG) splice sites.

We generated five representative sample sets to efficiently train/test donor-/acceptor SVMs [`LIBSVM` (Chang and Lin, 2001), rbf-kernel, default parameters]. Each training set consisted of 100 000 randomly chosen sites (50 000 positives and 50 000 negatives) and 10 000 independent instances to test the resulting models (5000 positives and 5000 negatives). We evaluated SVM performances by comparing receiver operating characteristics (ROC) and the area under the ROC curve (AUC), see Figure 1. Observed AUC values were nearly identical among all sets, demonstrating that random sampling did not bias our data. We kept the best performing 5′ and 3′ splice site models and classified the broad set of intergenic candidates (∼54 million) to identify novel splice sites.

The splice site classification was based upon the following eight features: (1) human `MaxEntScan` splice site score; (2–4) log-odds substitution scores s_{tree}, s_{pair}, s_{median}; (5) number of species in the alignment; (6) number of species with conserved GT/AG dinucleotides and a positive `MaxEntScan` score; (7) slope of a regression line fitted to the `PhastCons` conservation profile of the splice site; and (8) average `PhastCons` score.

The `MaxEntScan` program (Yeo and Burge, 2004) scores splice site sequences for similarity to typical splice sites. Overall, real splice sites have higher scores than false positives (Supplementary Fig. S4). Next, we computed log-odds scores to capture intrinsic sequence evolution of splice sites (see also next paragraph). Splice sites are usually highly conserved at the sequence level. Therefore, we included the total number of species per alignment and the number of species with a conserved GT (for donors) and AG (for acceptors) as SVM features. The average sequence conservation significantly decreases at the exon–intron boundary and increases at the intron–exon boundary, see also Chodroff *et al.* (2010). On average, the splice sites are even more conserved than the adjacent exons (Supplementary Fig. S3B). This holds for protein-coding as well as non-coding genes. The slope of a linear regression line fitted to the `PhastCons` (Siepel *et al.*, 2005) profile of the region $[-20,+20]$ for each splice site captures this information. Supplementary Figure S5 shows the score distributions and discriminative power of the individual features.

Log-odds substitution scores: we computed three variants of species- and site-specific substitution scores (s_{tree}, $s_{\text{pair}}$and s_{median}) based on the substitution frequencies in real and false splice sites. These log-odds capture splice site evolution among species (Fig. 2) and the score is >0 if the region of interest conforms to real splice site evolution and <0 otherwise. The more substitutions are consistent with splice site evolution, the higher is the total score. We evaluated the donor region $[-3,6]$ and the acceptor region $[-19,2]$ for all three score variants.

To compute score s_{tree}, we reconstructed ancestral sequences for each splice site region using `prequel` (Siepel *et al.*, 2005). It computes marginal probability distributions for bases at ancestral nodes in a phylogenetic tree. For each edge e of the reconstructed binary tree and for each site i of each two related sequences, we computed the frequency f^i of substitutions of nucleotide x_i to nucleotide y_i for each position in the splice site region ($x \neq y$ and $x,y \in \Sigma$, $\Sigma = \{A,C,G,T\}$). We tabulated the log-odds ratio of the total number of pairwise substitutions observed between all positive and negative training samples. Given a set of sequences, the sum of all log-odds of all observed substitution events along each edge of the reconstructed phylogenetic tree is

$$s_{\text{tree}} = \sum_{e}\sum_{i} \log_2 \left(\frac{f^{i}_{\text{pos}}(x \rightarrow y) / \sum_{n \in \Sigma} f^{i}_{\text{pos}}(x \rightarrow n)}{f^{i}_{\text{neg}}(x \rightarrow y) / \sum_{n \in \Sigma} f^{i}_{\text{neg}}(x \rightarrow n)} \right) \quad . \qquad (1)$$

The log-odds substitution score s_{pair} was previously used in Hiller *et al.* (2009). We counted substitution frequencies of each splice site position of human against each other species, learned the log-odds ratio of positive and negative samples and scored intergenic candidates with the sum of observed log-odds.

Score s_{median} was inspired by the CSF metric for codon substitution frequencies (Stark *et al.*, 2007). Similarly to s_{pair}, we summed up log-odds for each splice site position, but took the median instead of totaling the position-specific scores. Since SVM training- and test sets have to be independent (disjoint), log-odds substitution scores were always learned on training sets, never on test sets.

Exon prediction: searching for short splice site signatures in the huge intergenic space is expected to yield false positives, even at high classification confidence values. However, exons as biologically meaningful units consist of an acceptor–donor pair in close proximity. To find parts of novel transcripts (exons) and to reduce false positives, we derived candidate exons from individually predicted splice sites by searching for acceptor–donor pairs on the same DNA strand separated by not >300 nt. This is a natural cut-off since 85% of all `RefSeq` exons are <300 nt (and still 80% of all non-coding exons).

To predict novel exons in intergenic regions, we evaluated all candidate exons using a second SVM (exon-SVM) that was trained on characteristic signatures of transcript-confirmed exons.

Each internal exon of a multi-exon transcript is flanked by an upstream acceptor and a downstream donor. Conservation of such an exon implies compatible acceptor–donor pairs present in many species. To capture the conservation and the compatibility of the particular acceptor–donor pair, we considered the absolute number and the fraction of species having both a conserved acceptor and donor as two SVM features. Other features were the previously assigned class probabilities of the splice site SVM and the distance between particular splice site pairs. The exon-SVM was trained with six features: (1–2) acceptor and donor SVM classification probability; (3) exon length; (4) number of species that have conservation for both splice sites; (5) fraction of (4) and the number of species with conserved AG in the acceptor alignment; (6) fraction of (4) and the number of species with conserved GT in the donor alignment.

To train this exon-SVM, we obtained a set of real exons by requiring that both splice sites are (i) not annotated as part of a pseudogene (according to the `Yale` and the `UCSC` browser tracks); (ii) evolutionary conserved in the same species (at least five); and (iii) confirmed by ≥ 2 spliced expressed sequence tags (ESTs) as well as $\geq 20\%$ of all spliced ESTs present at the particular locus. This was fulfilled for 334 (22% of 1521) EST-confirmed exons. The stringent filtering assures that we exclude cases of transcriptional noise and use only high-quality true positives during training. We randomly selected 284 of the 334 exons for training and used the remaining 50 to evaluate the SVM. Then, 1000 EST-unconfirmed exons were randomly selected and 900 of these were used as negative training examples and the remaining 100 for evaluation. We repeated this procedure 10 times, kept the best-performing model with respect to sensitivity and specificity and classified the whole exon candidate pool to detect exons that exhibit signatures specific to EST-confirmed loci.

Candidate gene structures: most coding genes and lncRNAs consist of several exons and introns. If the predicted exons are real and belong to multi-exon genes, we expect that they have a tendency to cluster spatially within the genome. Human introns have a mean length of 6 kb which we used as a cut-off to define genomic clusters of predicted exons.

We performed a simulation test to determine whether predicted clusters (defined here as ≥ 2 exons separated by at most 6 kb on the same strand) occur more often than expected. To generate a background distribution, we selected as many rejected exons ($p \leq 0.5$) as we observed positively classified clustered exons ($p > 0.5$) and counted the number of (random) clusters. Repeating this sampling procedure 10 000 times yields empirical *P*-values which indicate the statistical significance of predicted exon clusters. In case of overlapping exons, one representative according to the highest SVM probability was selected to generate non-overlapping gene structures.

Coding versus non-coding exons: exons without protein homology (using `BLASTX` against the NCBI nr database with -e 1e-5, -F F, -S 1) and no protein-coding potential as predicted by `RNAcode` (Washietl *et al.*, 2011) (-b -r -s -p 0.01) were classified as 'non-coding'. In addition, exons containing stop codons in all three reading frames were classified as 'non-coding'.

Secondary structures: we applied `RNAz` (Washietl *et al.*, 2005) with default parameters (window size 120 nt, step width 40 nt) to search for signatures of conserved and stable RNA secondary structures in alignments of predicted exons and alignments covering possible exon–exon junctions. For the latter, we concatenated 60 nt up- and downstream of the predicted intron using the Galaxy Browser (Goecks *et al.*, 2010). These 120 nt regions mimic alignments of the mature transcript and were scored by `RNAz` to identify possible structures formed by long-range base pairs.

RNA-seq data: transcription of predicted exons was validated by RNA-seq data from Wang *et al.* (2008). We basically relied to their mappings, but re-mapped their short read data to confirm additional exon–exon junctions. For each predicted intron, we concatenated 26 nt of the predicted trailing exons and built 52 nt long putative mature mRNA fragments. We performed a `BLASTN` search of all 32 nt long reads against this database. Short reads producing nearly perfect BLAST hits ($\geq$30 nt in length, $\leq$2 mismatches, $\geq$3 read coverage) spanning the 52 nt exon–exon junction support our predicted splice junctions whenever they lack a better hit to other loci in the genome. Apart from 37 exon–exon junctions that were directly verified by the annotations of Wang *et al.* (2008), this procedure additionally confirmed nine previously unreported exon–exon junctions.

Experimental validation: to experimentally verify a prime multi-exon transcript candidate, we designed primers to the predicted exonic regions in the human genome using `Primer3` (v0.4.0, default parameters). Primer sequences: fwd 5′-gcagtgcagaatggcaagt-3′; rev 5′-gcctcagcatattcatctcca-3′. Total RNA from LNCaP and RWPE-1 cells (human prostate cancer cells) was extracted using TRIZOL$^{\text{TM}}$ reagent according to the manufacturers' instructions (Invitrogen). To eliminate genomic DNA, a DNase digestion was performed using the TURBO DNA-free$^{\text{TM}}$ Kit (Applied Biosystems/Ambion, manufacturers instructions). Next, 1 µg of total RNA was reverse transcribed with SuperScript$^{\text{TM}}$ III Reverse Transcriptase (Invitrogen). Genomic DNA was isolated using DNeasy Blood & Tissue Kit (Qiagen). PCRs were performed using Taq-DNA-Polymerase (NEB) in a 30 µl reaction containing 1 µl cDNA or genomic DNA. PCR products were analyzed on 1.5% agarose gels, extracted from the gel using the MinElute Gel extraction Kit (Qiagen), cloned using TOPO TA Cloning^{R} (Invitrogen) and sent out for sequencing (Seqlab).

3 RESULTS

De novo splice site prediction: as illustrated in Figure 1, we trained SVM classifier using the evolutionary signatures of vertebrate splice sites to distinguish real from false splice sites. Indicating a good discriminative power, our models achieved high AUC values of 0.96 for donors and 0.94 for acceptors (Fig. 1A). On an independent test set with 5000 sites which were not used for training, we correctly detected 89% of all true donors at a false positive rate (FPR) of 4 and 84% of all true acceptors at an FPR of 9% (SVM classification confidence $p > 0.5$). To reduce the FPR to less than 2%, we used a more stringent SVM classification confidence of $p > 0.9$, which still correctly identifies 81% (73%) of real donor (acceptor) sites. This

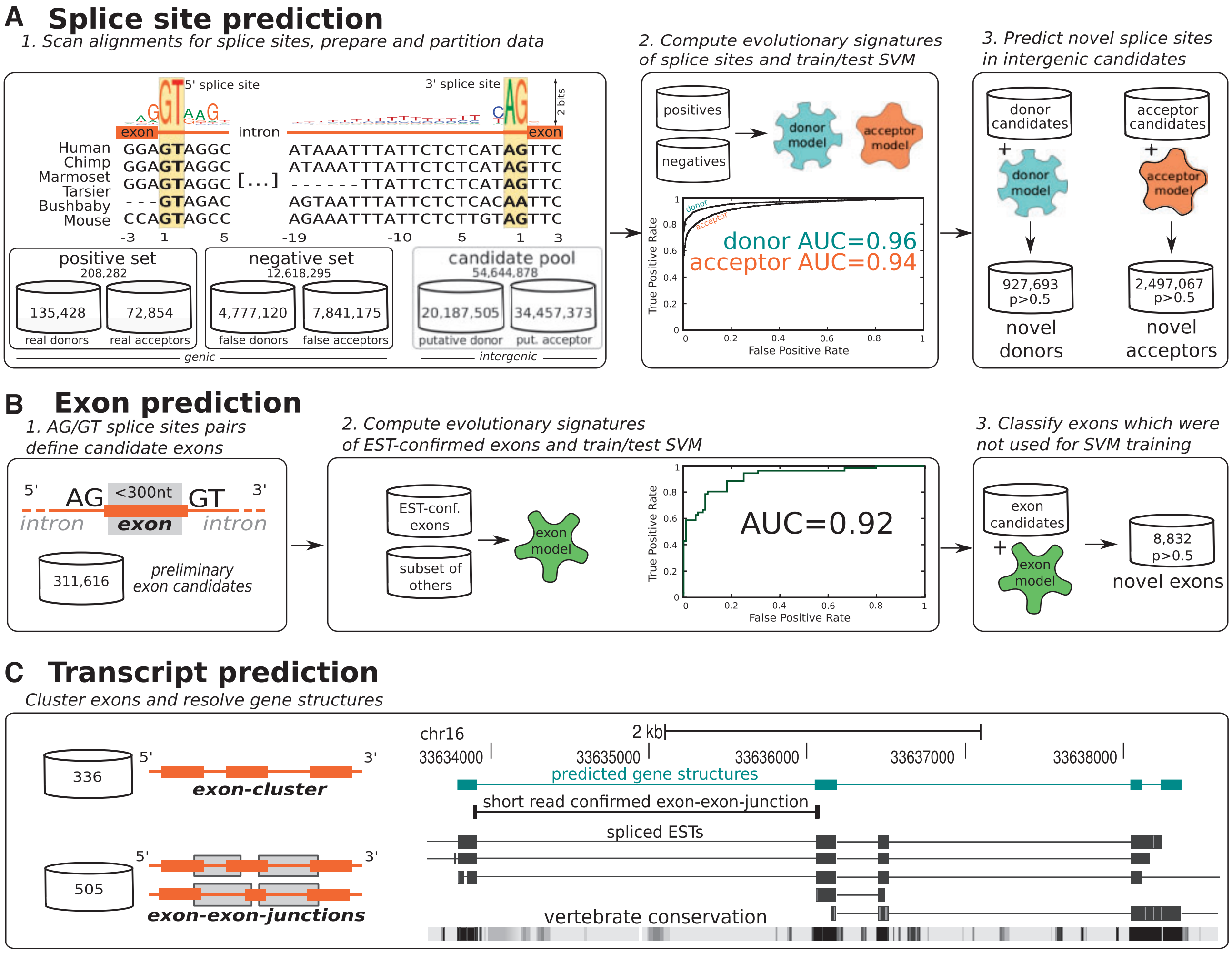

Fig. 1. Overview on the computational procedure to identify novel spliced transcripts in vertebrate genomes. (**A**) First, we extracted splice site candidates from genome-wide alignments and set up a training set (left panel). We distinguished between (i) real annotated splice sites in genic regions; (ii) false splice sites in genic regions (GT/AG dinucleotides that were similar to real splice sites but were not supported by transcript data); and (iii) the remaining set of intergenic splice site candidates. Secondly, we compiled a set of evolutionary signatures that are characteristic for vertebrate splice sites and trained donor and acceptor SVM models (middle panel, ROC curves are shown for both models). Thirdly, these SVMs were used to classify intergenic candidates as either real or false splice sites (right panel). (**B**) To obtain exon predictions, we searched for pairs of splice sites with a maximal distance of 300 nt (left), trained a second SVM that considered features of EST-confirmed exons (middle) and ranked predicted exons. Finally, predicted exons were then clustered into partial multi-exon transcripts. (**C**) Several splice sites and exons/introns of the shown example are confirmed by ESTs and short RNA-seq reads.

demonstrates that our approach is capable of identifying splice sites at high specificity. Out of 54 million intergenic candidates, about 3.4 million sites were predicted to be real ($p > 0.5$).

Improved log-odds substitution scores: nucleotide substitutions in splice sites are highly biased to certain substitution patterns that follow the splice site consensus sequence (Supplementary Fig. S2 and S3A). This holds for protein-coding as well as non-coding genes (Supplementary Fig. S2). For example, A and G are the preferred nucleotides at the donor consensus position +3 and A/G substitutions are the most frequent substitution at this position in real donors. The pairwise approach used in Hiller *et al.* (2009) considers substitutions between a reference and orthologous sequences for each alignment column (Fig. 2B). This can over- or underestimate the real number of substitutions that happened in evolution. In particular, if a strictly conserved base has changed in the reference sequence, the pairwise method will sum the log-odd scores for all pairs reference ortholog, although only a single change has happened. To avoid these biases, we developed a method that evaluates species- and site-specific substitution patterns along the phylogeny of the aligned species (Fig. 2C). We reconstructed the likely ancestral bases at each internal node in the phylogenetic tree. This allowed us to compute log-odd scores that only consider real substitutions. Our tree-based approach let to a noticeable performance increase compared with the pairwise method, in particular for low FPRs. Measuring the predictive power of either method alone, the AUC improved from 0.68 to 0.72 for donor and from 0.85 to 0.93 for acceptor sites (Fig. 2D). Acceptors, for which

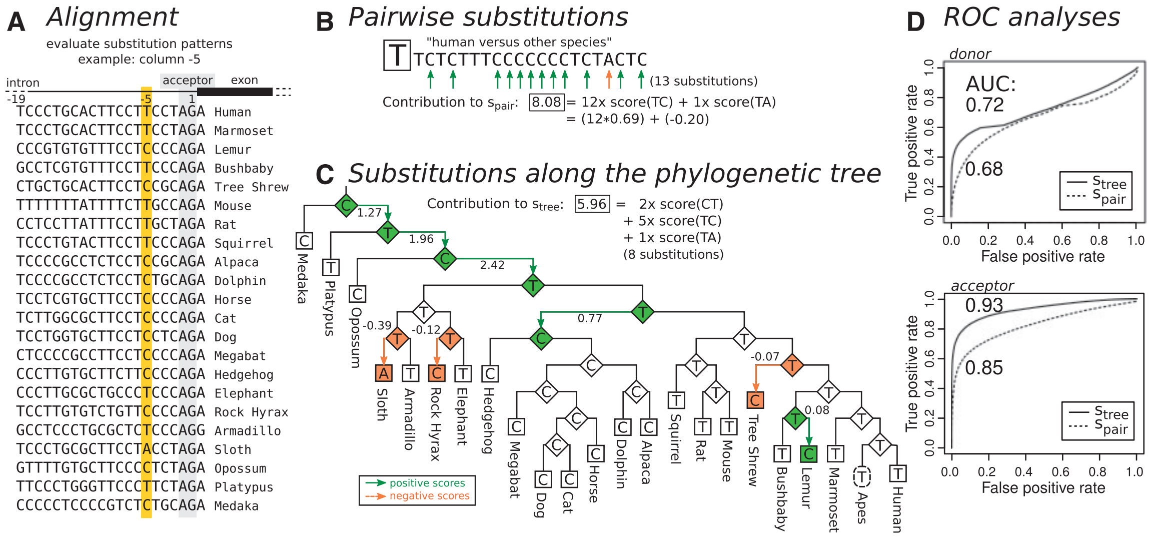

Fig. 2. Log-odds substitution scores. Given an exemplary acceptor alignment (**A**), we compare the pairwise (**B**) and the tree-based (**C**) approach to score splice site substitutions, here focusing on substitutions at alignment column -5. While the pairwise method evaluates 13 substitutions (B), the tree-based method considers only eight that likely happened along the phylogenetic tree (C). The leaves of the tree represent the sequences of extant species, and inner nodes reflect the reconstructed ancestral state. Substitutions with positive log-odd scores happen more frequently in the evolution of real than false splice sites. (**D**) ROC curves demonstrate that the tree-based method (s_{tree}) significantly outperforms the pairwise method (s_{pair}) for both donor and acceptor sites.

we score a longer region that comprises the poly-pyrimidine tract, particularly benefit from this novel scoring scheme.

Prediction of exons based on individual splice sites: we obtained 311 616 candidate exons from *de novo* predicted splice sites ($p > 0.5$) of which 1521 (0.5%) exons were confirmed by ESTs. The purpose of the second SVM (exon-SVM) is to rank these preliminary candidates. Trained on a subset of most reliably EST-confirmed exons (334/1521), the SVM achieved an AUC of 0.92 (Fig. 1B). In addition, we validated the performance on a different test set consisting of 9333 real and 4722 false RefSeq exons whose splice sites were correctly classified by the splice site SVM and have not been used for training. At $p > 0.5$, we obtained an AUC of 0.88 (Supplementary Fig. S7) with a TPR of 78% (7262/9333) and a FPR of 11% (537/4722). At $p > 0.9$, we still obtained 38% (3528/9333) TPR and 1% (57/4722) FPR. We applied the exon-SVM to the remaining exon candidates that were not used for training. Finally, 8832 candidate exons were predicted to be real at confidence $p > 0.5$ (898 exons at $p > 0.9$).

Confirmation by RNA-seq data and recent RefSeq annotations: The RNA-seq data published by Wang *et al.* (2008) provide evidence for transcription of 5% (469/8832) of predicted exons (Supplementary Table S10). To test if deeper sequencing might confirm more exon predictions, we used only a fraction of their data for validation. We observed that the number of confirmed exons increases linearly with the fraction of RNA-seq reads without saturation (Fig. 3, Supplementary Fig. S9), which suggests that additional data are likely to verify additional predictions.

To evaluate tissue-specific expression, we found that only 14 of the 469 exons confirmed by RNA-seq are supported by reads from at least 10 of the 15 tissues/cell lines. The 281 exons are only supported by reads from a single tissue. This clearly indicates tissue-specific transcription of these genes, in agreement with previous findings from Wang *et al.* (2008).

The human gene catalog is continuously updated and refined. Therefore, we expect that some of our predictions, unknown at the time we made them, are now validated by new annotations. Indeed, 44 of our predicted and previously unknown exons have meanwhile been included in the RefSeq transcript annotation. For example, a complete predicted cluster consisting of five exons is now part of the official consensus gene structure of the *NEB* gene (Supplementary Fig. S8).

Predicted exons form potential multi-exon transcripts: Of total, 8% (734 of 8832) of the predicted exons form 336 clusters (≥ 2 exons separated by at most 6 kb). With up to seven adjacent exons, these clusters are parts of potential multi-exon transcripts.

The remaining 8098 exons that are not in clusters might still belong to multi-exon transcripts for the following reasons. First, our approach can only detect internal exons that are flanked by a donor and acceptor site. In particular, the first exon (lacking an acceptor) and the last exon (lacking a donor) cannot be detected, which means that only genes with at least four exons (containing at least two internal exons) can form clusters. Second, our method is optimized for specificity not sensitivity and is likely to miss other exons belonging to the same transcript. Third, lncRNAs have fewer exons than coding genes, decreasing the chance to detect clusters. Finally, our cut-off for the maximum intron length of 6 kb prevents the detection of clusters for transcripts having longer introns.

To assess if the number of 336 clusters is higher than expected by chance, we used a simulation that builds exon clusters from an equal number of exons receiving low SVM confidence scores ($p < 0.5$). Running the simulation 10 000 times, we never obtained 336 or more clusters, yielding an empirical P-value $< 10^{-5}$. Remarkably, this remains true when empirical P-values are computed separately for clusters with cardinalities between two and seven exons. The predicted exons thus have a strong tendency to form potential multi-exon transcripts, which makes them good candidates for novel protein-coding genes and lncRNAs.

The 336 clusters contain 505 exon–exon junctions of which RNA-seq reads (Wang *et al.*, 2008) verified 46 (9%). Interestingly, 29% of the predicted exons in clusters are independently predicted by coding gene finders, indicating that they are part of multi-exon coding genes (Supplementary Table S9). For all predicted exons that do not cluster, 7.5% are predicted by coding gene finders, which can be due to a higher false positive rate in these exons but also the fact that lncRNAs have fewer exons than coding genes (Guttman *et al.*, 2010). The latter, in particular, complicates the prediction of lincRNAs which on average have only 1.7 internal exons.

Predicted exons are mostly non-protein-coding and unstructured: only 8% (674 of 8832) of predicted exons have homology to protein-coding genes, 40% (3508 of 8832) have stop codons in all three reading frames, and 92% (8124 of 8832) are classified as non-coding by RNAcode (Washietl *et al.*, 2011). This shows that the great majority (89% or 7894) of predicted exons and exon-clusters (241/336) is likely non-coding.

We found RNAz hits in only 1.2% (94 of 7894) of the non-coding exons. This fraction is not higher than for coding exons (1.7%, 16 of 938). Furthermore, only 7% (35 of 505) of the predicted exon–exon junctions and 14% (46 of 336) of all exon clusters contain RNAz hits. This indicates that our predicted loci are mostly unstructured or do not contain conserved secondary structures and consequently cannot be detected by structure-based ncRNA finders.

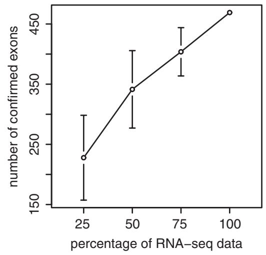

Fig. 3. Deeper sequencing will likely confirm further exon predictions. We split the data of Wang *et al.* (2008) to subsets containing 25, 50 and 75% of all reads and computed the corresponding number of confirmed exons. We repeated this procedure five times to avoid biases of a single random split. Error bars show the minimum and maximum of the five iterations. Since the number of confirmed exons increases almost linearly with the number of reads, it is likely that future sequencing projects will provide further experimental evidence for our approach.

Experimental validation of a predicted multi-exon transcript: clusters of predicted exons with highest cardinality are prime candidates for novel genes. Our data contain two such top-scoring clusters, each consisting of seven adjacent exons (Fig. 4A and B). The first cluster is confirmed by ESTs and, according to BLASTX, likely protein coding. The second cluster has RNA-seq support for only one exon–exon junction. Therefore, we designed primers to positions in the human genome that allow to verify the remaining predicted transcript. We used RT–PCR in a human prostate cancer cell line and subsequent cDNA sequencing. This confirmed eight of the nine predicted splice sites and three of five predicted exons (Fig. 4C). Furthermore, our experiments revealed complex alternative splicing at this locus with five different isoforms in prostate cancer cells. These contain even additional novel exons, neither detected by existing methods nor by our current approach.

4 DISCUSSION

We present a computational procedure to identify novel spliced transcripts in vertebrate genomes using conserved splice sites and signatures of length-constrained exons. We deliberately neglect features of protein-coding genes and outline a first conceptual approach toward the *de novo* prediction of lncRNAs.

We have previously demonstrated the value of conserved introns for the prediction of conserved, and hence likely functional, ncRNAs in insect genomes (Hiller *et al.*, 2009). In this work, we now tackled the problem of applying this conceptual idea to vertebrate genomes, where *ab initio* splice site and intron prediction is challenging due to the drastically increased absolute length and length variability

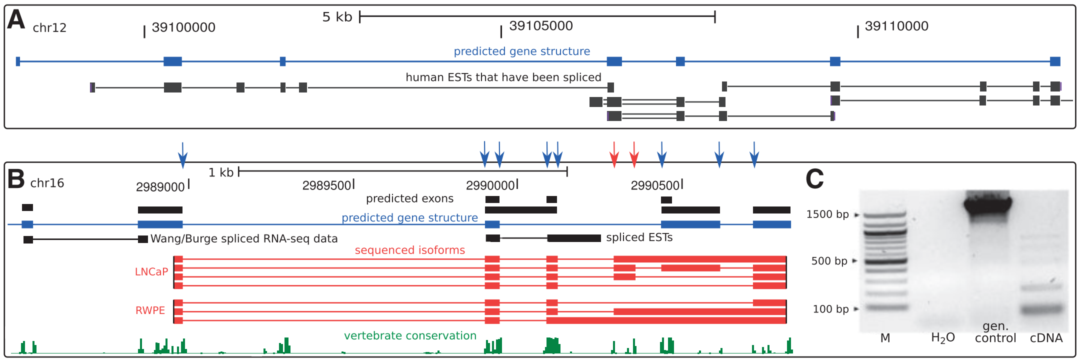

Fig. 4. Examples of multi-exon transcripts. (**A**) A cluster consisting of seven exons supported by ESTs. (**B**) RT–PCR followed by sequencing cloned isoforms verified another predicted transcript (only five of seven exons are shown). We sequenced seven transcripts with a total of 10 splice sites at this locus. Eight of these ten splice sites are predicted (black arrows) and missed two (dashed arrows). Within the range of the RT–PCR primers, eight of nine predicted splice sites and both splice sites for three of five predicted exons are verified. (**C**) Gel electrophoresis shows several bands indicating the spliced isoforms (cDNA) depicted in (B) as well as the genomic DNA (control). The observed transcripts are shorter than the corresponding genomic interval due to splicing.

of vertebrate introns. We developed a two step procedure that first predicts novel splice sites, which are in a second step combined to predicted exons. A key improvement is a novel log-odds score for splice site substitutions that explicitly takes the phylogenetic tree into account, avoiding biases of previous approaches. We show that this tree-based method substantially improves the power of splice site detection and outperforms two other approaches. The general concept behind this tree-based method can be applied to detect other biologically relevant signals and motifs in multiple sequence alignments.

Our predicted exons and exon clusters mostly belong to non-coding transcripts. These transcripts also rarely contain conserved secondary structures. This means that current methods to find coding genes or structured ncRNAs will miss most of our predictions and that our approach complements existing methods.

High-throughput transcriptome sequencing has led to the discovery of many unknown exons and transcripts. However, for the following reasons we believe that *ab initio* computational predictions of conserved transcripts complement experimental approaches. First, tissue-specific transcripts can only be detected when a large variety of conditions such as cell types, tissues, time points is sampled. Second, detecting transcripts with low expression levels requires sufficiently deep sequencing. These limitations are particularly relevant for ncRNAs, which often have low and highly specific expression patterns (Mercer *et al.*, 2008; Ravasi *et al.*, 2006). Consistent with this and the fact that our predictions are mostly non-coding, we observed that more transcript data lead to the confirmation of more predictions without any observable saturation. Furthermore, our predictions confirmed by transcriptome sequencing data mostly have tissue-specific expression patterns. While our predictions will inevitably contain false positives, these observations suggest the existence of further evolutionarily conserved, and hence likely functional, multi-exon transcripts that still remain hidden in the human genome. Future increases in transcriptome sequencing depth and breadth will confirm additional predictions. Also, our exon and transcript predictions can be included in ongoing large-scale RT–PCR based efforts to further validate human gene predictions (Harrow *et al.*, 2006). Our approach complements other gene prediction approaches and contributes to completing the catalog of human transcripts.

ACKNOWLEDGEMENTS

We thank Thomas Derrien and Roderic Guigó for fruitful discussions and Sita Lange for proof reading. We are grateful to the UCSC Genome Browser team, the Mammalian Genome Project and the sequencing centers for providing many of the available genomes and related datasets used in this study.

Funding: SAB (grant no. 14494); FP-7 project QUANTOMICS (no. 222664); German Research Foundation (Hi 1423/2-1); Excellence Initiative of the German Federal and State Governments (EXC 294 to R.B.); Human Frontier Science Program (Fellowship LT000896/2009-L).

Conflict of Interest: none declared.

REFERENCES

Bernard,D. *et al.* (2010) A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J.*, **29**, 3082–3093.

Burge,C. and Karlin,S. (1997) Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, 78–94.

Carninci,P. *et al.* (2005) The transcriptional landscape of the mammalian genome. *Science*, **309**, 1559–1563.

Chang,C. and Lin,C. (2001) LIBSVM: a library for support vector machines. *ACM Trans. Int. Syst. Technol.*, **2**, 27:1–27:27.

Chodroff,R. *et al.* (2010) Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol.*, **11**, R72.

Cruveiller,S. *et al.* (2003) Compositional features of eukaryotic genomes for checking predicted genes. *Brief. Bioinform.*, **4**, 43–52.

ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.

Goecks,J. *et al.*; Galaxy Team (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, **11**, R86.

Gross,S.S. *et al.* (2007) CONTRAST: a discriminative, phylogeny-free approach to multiple informant *de novo* gene prediction. *Genome Biol.*, **8**, R269.

Guttman,M. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.

Harrow,J. *et al.* (2006) GENCODE: producing a reference annotation for ENCODE. *Genome Biol.*, **7** (Suppl. 1), S4.1–S4.9.

Hiller,M. *et al.* (2009) Conserved introns reveal novel transcripts in *Drosophila melanogaster*. *Genome Res.*, **19**, 1289–1300.

Huarte,M. and Rinn,J. (2010) Large non-coding RNAs: missing links in cancer? *Hum. Mol. Genet.*, **19**, R152–R161.

Kapranov,P. *et al.* (2007a) RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, **316**, 1484–1488.

Kapranov,P. *et al.* (2007b) Genome-wide transcription and the implications for genomic organization. *Nat. Rev. Genet.*, **8**, 413–423.

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

Mercer,T.R. *et al.* (2008) Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl Acad. Sci. USA*, **105**, 716–721.

Mercer,T.R. *et al.* (2009) Long non-coding RNAs: insights into functions. *Nat. Rev. Genet.*, **10**, 155–159.

Nawrocki,E.P. *et al.* (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.

Pang,K.C. *et al.* (2006) Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.*, **22**, 1–5.

Ponjavic,J. *et al.* (2007) Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res.*, **17**, 556–565.

Ponting,C.P. *et al.* (2009) Evolution and functions of long noncoding RNAs. *Cell*, **136**, 629–641.

Ravasi,T. *et al.* (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.*, **16**, 11–19.

Rhead,B. *et al.* (2010) The UCSC Genome Browser Database: update 2010. *Nucleic Acids Res.*, **38**, D613–D619.

Rodríguez-Trelles,F. *et al.* (2006) Origins and evolution of spliceosomal introns. *Annu. Rev. Genet.*, **40**, 47–76.

Schweikert,G. *et al.* (2009) mGene: accurate SVM-based gene finding with an application to nematode genomes. *Genome Res.*, **19**, 2133–2143.

Senner,C.E. and Brockdorff,N. (2009) Xist gene regulation at the onset of X inactivation. *Curr. Opin. Genet. Dev.*, **19**, 122–126.

Siepel,A. *et al.* (2005) Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, **15**, 1034–1050.

Solovyev,V. *et al.* (2006) Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.*, **7** (Suppl. 1), S10.1–S1012.

Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19** (Suppl. 2), i215–i225.

Stark,A. *et al.* (2007) Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures. *Nature*, **450**, 219–232.

Wang,E.T. *et al.* (2008) Alternative isoform regulation in human tissue transcriptomes. *Nature*, **456**, 470–476.

Washietl,S. *et al.* (2005) Fast and reliable prediction of noncoding RNAs. *Proc. Natl Acad. Sci. USA*, **102**, 2454–2459.

Washietl,S. *et al.* (2011) RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA*, **17**, 578–594.

Willingham,A.T. *et al.* (2005) A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science*, **309**, 1570–1573.

Wilusz,J.E. *et al.* (2009) Long noncoding RNAs: functional surprises from the RNA world. *Genes Dev.*, **23**, 1494–1504.

Yeo,G. and Burge,C.B. (2004) Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals. *J. Comput. Biol.*, **11**, 377–394.