BIOINFORMATICS

Vol. 00 no. 00 2011
Pages 1–7

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

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

[1]Bioinformatics Group, Department of Computer Science, University of Leipzig, Germany. [2]LIFE – Leipzig Research Center for Civilization Diseases, University of Leipzig, Germany. [3]Fraunhofer Institut for Cell Therapy and Immunology, AG RNomics, Leipzig, Germany. [4]Department of Molecular Immunology, University of Leipzig, Germany. [5]Interdisciplinary Center of Bioinformatics, University of Leipzig, Germany. [6]Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany. [7]Centre for Biological Signalling Studies (BIOSS), University of Freiburg, Germany. [8]Centre for Biological Systems Analysis (ZBSA), University of Freiburg, Germany. [9]Institute for Theoretical Chemistry, University of Vienna, Austria. [10]Department of Developmental Biology, Stanford University, USA. [11]Sante Fe Institute, Santa Fe, USA.

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Associate Editor: XXXXXXX

ABSTRACT

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.

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. We individually predict splice sites, assemble compatible sites into exon candidates, and obtain multi-exon transcript predictions. 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. Applying our approach to an alignment of 44 vertebrate genomes, we validate many of our predicted exons. Furthermore, predicted exons have a significant tendency to form multi-exon transcript parts and we experimentally validate a novel multi-exon gene.

Our results indicate the existence of novel human transcripts with conserved splice sites. Computational lncRNA gene predictions contribute to the completion of the human transcript catalog.

1 INTRODUCTION

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

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 ncRNA classes including common families like rRNAs, tRNAs and miRNAs evolutionarily preserve their characteristic secondary structure, which can be used to computationally predict them (Washietl et al., 2005; Nawrocki et al., 2009). 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 (Mercer et al., 2009; Ponting et al., 2009; Wilusz et al., 2009; Huarte and Rinn, 2010). 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).

*corresponding author

We recently presented the first computational approach to detect lncRNAs with conserved intron positions in insect genomes (Hiller et al., 2009). This approach is purely 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 (Rodríguez-Trelles et al., 2006; Ponjavic et al., 2007; Chodroff et al., 2010). Since insect genomes feature very short introns (most are $<$100 nt in *Drosophila melanogaster* (Lim and Burge, 2001)) this approach crucially relied on predicting introns as a single unit. Pairs of splice donor (5') and acceptor (3') sites thus are predicted in a single step. Vertebrate introns, in contrast, are substantially longer and more variable in their length, precluding the application of this intron-based method.

Here, we introduce a novel approach to predict novel spliced transcripts from intergenic regions of vertebrate genomes. Using a combination of comparative genomics and machine learning, we first predict novel 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 novel coding and non-coding transcripts with conserved exon/intron structures in the well characterized human genome and validate our predictions with available transcript data and own experiments.

2 RESULTS

De novo splice site prediction. The splice site prediction part of our approach consists of the following steps: (i) detect donor and acceptor splice site candidates in multiple sequence alignments, (ii) train a support vector machine (SVM) with novel features capturing patterns of splice site evolution, (iii) use the trained SVM to score candidate splice sites (Fig. 1A).

To distinguish real from false splice sites by machine learning, we first focused on genomic regions of known coding genes for training. We screened the genome alignment of these genic regions for donor and acceptor candidates and divided them into real splice sites that are annotated (208,282 true positives) and false positives that are not supported by available transcript data ($\sim$12.6 million). Albeit it is clear that these are not experimentally proven false positive splice sites, we reason most real splice sites are annotated by the wealth of mRNA and EST data for coding genes. Thus, the rest of candidates that are not supported by annotation or transcript data are predominantly false positives. The tiny fraction of the false positives that are real splice sites should not interfere with our machine learning approach.

Next, we determined characteristic evolutionary features solely extracted from genomic sequence data to distinguish true from false positive splice sites. The first feature captures intrinsic sequence evolution at splice sites. Nucleotide substitutions in splice sites are highly biased to certain substitution patterns that follow the splice site consensus sequence (Fig. S3A). 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. Previous attempts to capture this information relied on pairwise log-odds substitution scores (Hiller et al., 2009) which made use of a reference species but yielded biased results if a substitution had happened in the reference. Therefore, we improved the log-odds scoring scheme by developing a method that explicitly evaluates species- and site-specific substitution patterns along the phylogeny of the aligned species (detailed below). The second feature uses the `MaxEntScan` program (Yeo and Burge, 2004) to score splice site sequences for similarity to typical splice sites. Overall, real splice sites have higher scores than false positives (Fig. S4). A third set of features captures typical sequence conservation around splice sites. Real splice sites are usually highly conserved at the sequence level. In particular the donor and acceptor dinucleotides (GT and AG, respectively) are, on average, even better conserved than the adjacent exons (Fig. S3B). This holds for protein-coding as well as non-coding genes. Therefore, we included the total number of species per alignment block and the number of species with a conserved GT (for donors) and AG (for acceptors) as features. Furthermore, the average sequence conservation significantly decreases at the exon-intron boundary and increases at the intron-exon boundary, see also (Chodroff et al., 2010). The slope of a regression line fitted to the `PhastCons` (Siepel et al., 2005) profile for each splice site captures this information. Overall, we considered eight features (Methods). Figure S4 shows the score distributions and discriminative power of the individual features.

To combine these features into a single prediction value, we trained an SVM classifier on the genic training set. The best donor and acceptor models for the SVM classification yielded a high area under the receiver operating curve (AUC) with 0.96 for donors and 0.94 for acceptors (Fig. 1A), indicating substantial discriminative power. On an independent test-set with 5,000 sites which were not used for training, we correctly detected 89 % of all true donors at a false positive rate of 4 % and 84 % of all true acceptors at a false positive rate of 9 % (SVM classification confidence $p > 0.5$). To reduce the false positive rate 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 demonstrates that our approach is capable of identifying splice sites at high specificity.

Improved log-odds substitution scores. 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 explicitly takes the phylogenetic tree into account (Fig. 2C). We reconstruct 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 leads to a noticeable performance increase compared to the pairwise method, in particular for low false positive rates. Measuring the predictive power of either method alone, the AUC improves from 0.68 to 0.72 for donor and from 0.85 to 0.93 for acceptor sites (Fig. 2D). Acceptors, which usually have more substitutions in their longer poly-pyrimidine tract, particularly benefit from this novel scoring scheme.

Prediction of exons based on individual splice sites. Searching for the short splice sites signature in the huge intergenic space is expected to yield false positives, even at high classification confidence values. However, exons as biological meaningful units

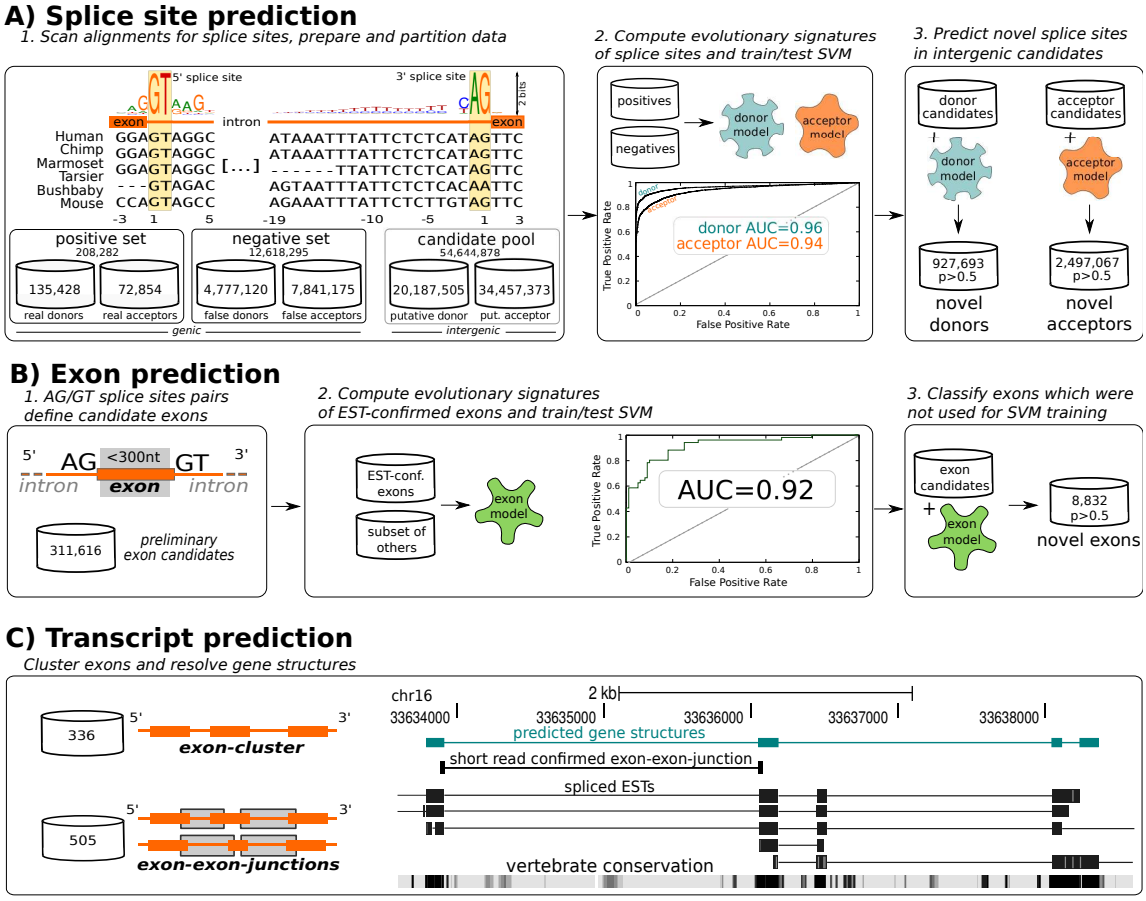

Fig. 1. Overview on the computational procedure to identify novel spliced transcripts in vertebrate genomes. (A) First, we extracted splice site candidates from genome-wide alignments and set up a training set (left panel). We distinguished between (i) real annotated splice sites in genic regions, (ii) 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. Second, we compiled a set of evolutionary signatures that are characteristic for vertebrate splice sites and trained SVM models of donor and acceptor splice sites (middle panel, receiver operating curves (ROC) are shown for both models). Thirdly, we used the SVMs 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) to rank 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.

consist of an acceptor-donor pair in relatively close proximity. To find parts of novel transcripts (exons) and to reduce false positives, we derived from our individual splice site predictions potential exon candidates by searching for acceptor-donor pairs on the same DNA strand separated by not more than 300 nt. We chose 300 nt as the cut-off since 85 % of all `RefSeq` exons are shorter than 300 nt (and still 80 % of all non-coding exons).

Using all splice sites ($p > 0.5$) that we predicted in intergenic regions, we obtained 311,616 exon candidates. Of those, only 1,521 (0.5 %) exons are confirmed by ESTs. To predict novel exons in intergenic regions, we evaluated all candidate exons using a second SVM that was trained on characteristic signatures of transcript-confirmed exons. Conservation of a real exon implies that both acceptor and donor sites are conserved in a species. To capture the compatibility of the particular acceptor-donor pair, we used the absolute number and the fraction of species having both a conserved acceptor and donor as two features. Other features were the previously assigned class-probabilities of the splice site SVM and the distance between particular splice site pairs.

To train this exon-SVM we obtained a set of real exons by requiring that both splice sites are (i) not annotated as pseudogene, (ii) evolutionary conserved in the same species (at least five), and (iii) confirmed by ≥ 2 spliced ESTs as well as ≥ 20 % of all spliced ESTs present at the particular locus. This is fulfilled for 334 (22 % of 1,521) EST-confirmed exons. We randomly selected 284 of the 334 exons for training and used the remaining 50 to evaluate the SVM. Then, 1,000 EST-unconfirmed exons were randomly selected and 900 of these were used as negative training examples and the remaining 100 for evaluation. The exon-SVM achieved an AUC of 0.92 (Fig. 1B shows the ROC curve).

We applied the exon-SVM to the remaining exon candidates that were not used for training. 8,832 candidates were predicted to be

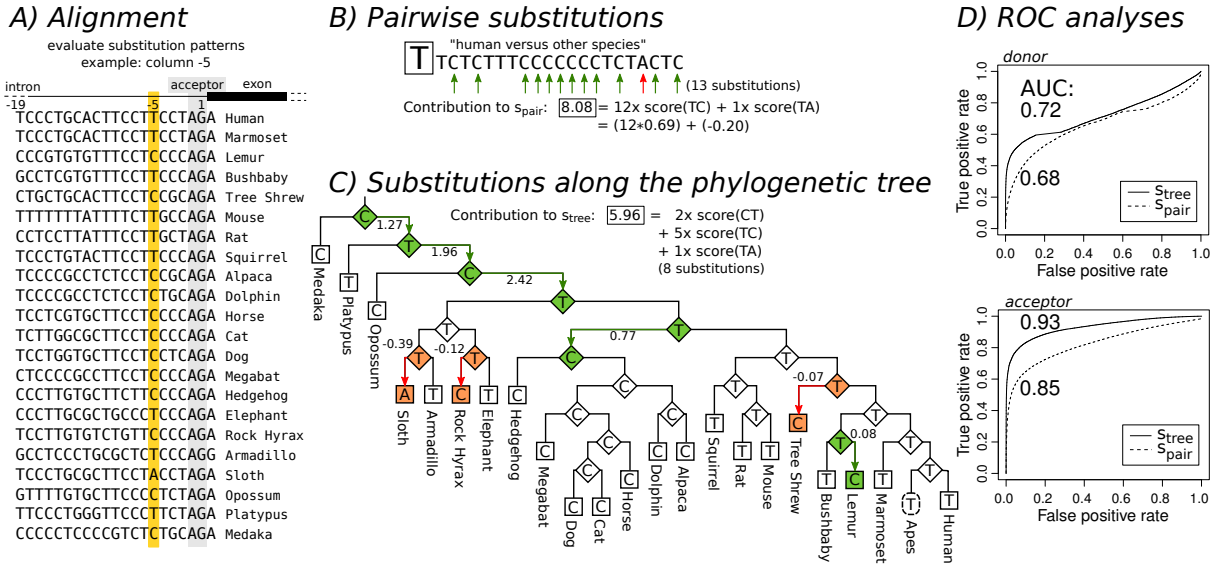

Fig. 2. Log-odds substitution scores. Given the sequence alignment of a 3' splice site (A), we compare the pairwise (B) and the tree-based (C) approach to score splice site substitution patterns, here focusing on substitutions at alignment column -5. The pairwise method would sum 13 substitutions (B), however only the eight substitutions that likely happened along the phylogenetic tree are considered by the tree-based method (C). The leaves of the tree are the nucleotides from extant species in the alignment, inner nodes are reconstructed ancestral nucleotides. 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.

real exons at confidence $p > 0.5$. At confidence $p > 0.9$ we obtained 898 predicted exons.

Confirmation by RNA-seq data and recent `RefSeq` *annotations.* Wang et al. (2008) used deep sequencing to characterize the transcriptomes of 15 human tissues and cell lines. We used the RNA-seq reads mapped to the genome to validate our exons predictions and found evidence for transcription of 5 % $(469/8,832)$, see Tab. S9. Next, we used only a fraction of the RNA-seq data for confirmation to test if deeper sequencing might confirm more exon predictions. We observed that the number of confirmed exons increases linearly with the fraction of RNA-seq reads without saturation (Fig. 3, S8), which suggests that additional data is likely to verify more 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. 281 exons are only supported by reads from a single tissue. This clearly indicates tissue-specific transcription of these genes.

The human gene catalogue 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, 50 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* (nebulin) gene (Fig. S7).

Predicted exons are mostly non-protein-coding and unstructured. Only 8 % (674 of 8,832) of exons have homology to protein-coding genes, 40 % (3,508 of 8,832) have stop-codons in all three reading frames, and 92 % (8,124 of 8,832) are classified as non-coding by `RNAcode` (Washietl et al., 2010). Considering all exons without protein homology and no predicted `RNAcode` coding potential as non-coding, the great majority (89 % or 7,894) of our exon predictions is likely non-coding.

Next, we used `RNAz` (Washietl et al., 2005) to search for signatures of conserved and stable RNA secondary structures. We found `RNAz` hits in only 245 non-coding exons (3 % of 7,894), indicating that our exon candidates are mostly unstructured or do not contain conserved secondary structures and consequently cannot be detected by secondary structure based ncRNA gene-finders.

Predicted exons form potential multi-exon transcripts. 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. Human introns have a mean length of 6 kb which we used as a cut-off. We found that 8 % $(734/8,832)$ of the exon-SVM predictions ($p > 0.5$) are separated by less than 6 kb from the nearest adjacent prediction on the same strand, leading to 336 exon clusters which represent parts of potential multi-exon transcripts. The largest cluster contains seven adjacent exons. Again, the majority of these clusters $(241/336)$ has no evidence of coding for proteins. 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 < 0.0001. 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,

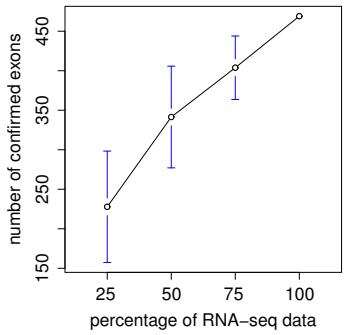

Fig. 3. Deeper sequencing will likely confirm further exon predictions. We split the data of Wang et al. (2008) to subsets containing 25%, 50%, and 75% of all reads and computed the corresponding number of confirmed exons. We repeated this procedure five times to avoid biases of a single random split. 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.

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) directly verified 46 (9%).

Experimental validation of a predicted multi-exon transcript. There are two top-scoring intergenic exon clusters, each consisting of seven adjacent exons (Fig. 4A,B). The first cluster is already confirmed by ESTs and, according to BLASTX, likely protein-coding. The second cluster has RNA-seq support for one exon-exon junction. Therefore, we used RT-PCR and subsequent cDNA sequencing to experimentally verify the remaining predicted transcript structure. 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, which contain even additional novel exons.

3 DISCUSSION

The value of conserved introns for the prediction of conserved, and hence likely functional, ncRNAs has previously been demonstrated in insect genomes (Hiller et al., 2009). Here, we 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 of vertebrate introns. Therefore, we developed a two step procedure that first uses a novel method to predict individual splice sites, which are in a second step combined to predicted exons. A key improvement is a log-odds score for splice site substitutions that explicitly takes the phylogenetic tree into account, avoiding biases of previous approaches. This tree-based method substantially improves the power of splice site detection.

In contrast to gene-finders designed to predict only coding genes, our approach is solely based on detecting typical splice site evolution by combining comparative genomics and machine learning. Specifically, our method does not rely on the characteristic evolutionary signatures of coding regions, which are a dominating signal used by coding gene-finders. As conserved splice sites are a hallmark of both non-coding and protein-coding transcripts, this enables us to predict both classes of transcripts from comparative genomics data.

Although transcript discovery recently have become a domain of high throughput sequencing methods, *ab initio* computational predictions of conserved transcripts nevertheless complement experimental approaches. Independent of low expression levels and specificity for rare cell types, our method points at evolutionarily conserved, and hence likely functional, transcripts that still remain hidden in mammalian genomes. Our data show that the currently available sequencing data by far do not saturate all existing transcripts. On the one hand, we found that a significant fraction of the predicted exons is confirmed by already available transcript data. Despite the fact that detecting novel exons in the already well-characterized human genome is challenging, our results suggest the existence of further evolutionary conserved multi-exon transcripts, one of which we directly validated experimentally. High scoring exon and transcript predictions can be included in ongoing large-scale RT-PCR based efforts to further validate gene predictions (Harrow et al., 2009; RGASP, 2010). Our approach complements other gene prediction approaches and contributes to completing the catalog of human transcripts.

4 MATERIALS AND METHODS

Alignments, genomes, annotations. We downloaded the genome-wide 44-way vertebrate multiple alignment with the human hg18 assembly as the reference and the following annotation tracks from the UCSC Genome Browser (Kent et al., 2002): `UCSC Genes`, `RefSeq Genes`, `Human mRNAs`, `Human ESTs`, site-specific `PhastCons` scores, and the phylogenetic tree of the 44-way alignment. Pseudogenes were annotated according to the `Yale` and the `UCSC` tracks.

Splice site prediction. We trained SVMs to solve the binary classification problem of *de novo* splice site prediction. Therefore, we compiled three disjoint sets of (i) positive and (ii) negative samples to train and test individual donor and acceptor SVM models and (iii) a set of candidate sites forming our search space for putative novel splice sites. To this end, we filtered the 44-way alignment for GT (5', donor) and AG (3', acceptor) dinucleotides (both reading directions) which are conserved in at least five species and possess enough informative aligned flanking sequence according to the sequence logos presented in Fig. S3. We rejected all sites with a `MaxEntScan` (Yeo and Burge, 2004) score below 0 to avoid GT/AG dinucleotides that are unlikely to be potential splice sites (see Fig. S4). Alignment blocks had to contain the nucleotides of the region $[-3, 6]$ for donors and $[-19, 2]$ for acceptors (position 1 of each interval corresponds to the first/last G of an intron). We did not consider non-canonical splice sites without a GT/AG. The positive set contained splice sites annotated in the `UCSC`, `RefSeq`, and the `Human mRNA` gene tracks. Negative training examples are the remaining genic sites (unannotated sites within introns, exons, or untranslated regions (UTRs)).

Given all positives ($\sim$200,000) and all negatives ($\sim$12 million), we generated five representative sample sets, each containing 55,000 randomly selected splice sites, in order to compute substitution scores and to efficiently train/test donor-/acceptor SVMs. For each set, we trained SVMs with all except 5,000 randomly chosen positives and 5,000 randomly chosen negatives (rbf-kernel, default c and g). Each training-set thus comprised 100,000 sites (50,000 positives and 50,000 negatives) and the remaining 10,000 were used to test the resulting models. We kept the two best performing 5' and 3' splice site models and classified the broad set of intergenic candidates ($\sim$54 million) to identify novel splice sites.

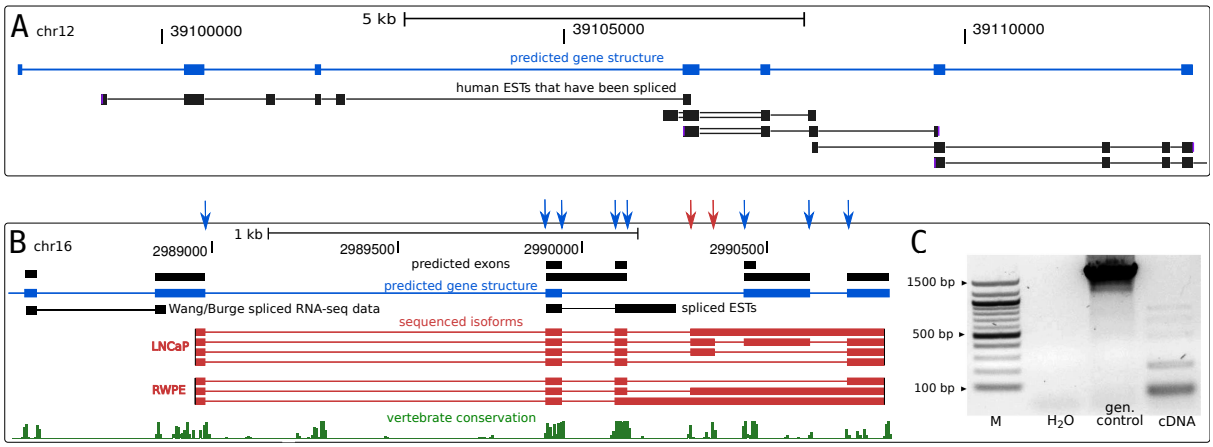

Fig. 4. Examples of multi-exon transcripts. (A) A cluster consisting of seven exons with reliable EST support. (B) RT-PCR followed by sequencing cloned isoforms verified another predicted transcript (only five of seven exons are shown). We sequenced seven transcripts with a total of ten splice sites at this locus. Eight of these ten splice sites are predicted (blue arrows) and two are not predicted (red arrows). Within the range of the RT-PCR, 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.

We evaluated SVM performances by receiver operating characteristics (ROC), expressed in a single number, the area under the ROC curve (Fig. 1). Observed AUC values were nearly identical among all sets, demonstrating that random sampling did not bias our data. Sequence conservation was measured by analyzing the `PhastCons` profile of splice site regions. In particular, we performed a linear regression based on `PhastCons` scores of the region [-20,+20] and used the slope of the resulting regression line and the average `PhastCons` score of this region as SVM features. We used gap-free subsequences to compute sequence-based scores (no significant differences were found using aligned splice site regions). In summary, splice site classification was based upon the following 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. (8) Average `PhastCons` score.

Log-odds substitution scores. We computed three variants of species- and site-specific substitution scores (s_{tree}, s_{pair}, and s_{median}) based on the substitution frequencies in real and false splice sites. We evaluated the donor region [-3,6] and the acceptor region [-19,2].

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 columnwisely counted the frequency f^i of substitutions of nucleotide x_i to nucleotide y_i ($x \neq y$ 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. These log-odds are designed to model splice site evolution (Fig. 2). Given a set of sequences, the sum of all log-odds of all observed substitution events along each edge of the reconstructed phylogenetic tree, formally written as

$$s_{tree} = \sum_e \sum_i \log_2 \left(\frac{f^i_{pos}(x \to y) / \sum_{n \in \Sigma} f^i_{pos}(x \to n)}{f^i_{neg}(x \to y) / \sum_{n \in \Sigma} f^i_{neg}(x \to n)} \right) \quad (1)$$

expresses whether the region of interest conforms to real splice sites ($s_{tree} > 0$) or not ($s_{tree} < 0$). The more substitutions are consistent with splice site evolution, the higher the total score.

The log-odds substitution score s_{pair} previously was successfully applied to detect novel transcripts in insect genomes (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.

Inspired by the CNF-metric for codon substitution frequencies (supplemental material of (Stark et al., 2007)), we trained the SVM with an additional log-odds score s_{median}. Similarly to s_{pair}, we summed up log-odds for each splice site position but, instead of totaling the position-specific scores, took the median of all intermediate totals. Since SVM training- and test-set have to be independent (disjoint), log-odds substitution scores were always learned on training-sets, never on test-sets.

Exon prediction. We trained a second SVM to identify a subset of meaningful candidates that resemble characteristics of transcript-confirmed loci. Out of all 334 positives (acceptor/donor pairs with a distance $\leq$300 nt, not annotated as pseudogenes, with splice sites conserved in $\geq$ 5 species, and which are confirmed by $\geq$ 2 spliced ESTs as well as $\geq$ 20 % of all spliced ESTs at the particular locus), we randomly selected 50 samples to test and kept the remaining 284 to train the model. Out of >300,000 negatives (the remaining preliminary acceptor/donor splice site pairs which are not in the positive set), we randomly selected a subset of 1,000 samples to reduce computational complexity and split it to a negative test-set (training-set) of size 100 (900). We repeated this procedure ten 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. The EST-model was trained with six features: (1) Acceptor SVM classification probability. (2) Donor SVM classification probability. (3) Exon length. (4) Number of common 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.

Candidate gene structures. We used a simulation test to determine if exons have the tendency to occur clustered (defined here as $\geq$ 2 exons separated by at most 6 kb on the same strand). Exon clusters only reliably indicate novel genes if the number of observed clusters differs significantly from the background. To generate a background distribution, we selected as many rejected exons ($p_{exon-SVM} \leq 0.5$) as we observed positively classified clustered exons ($p_{exon-SVM} > 0.5$) and counted the number of (random)

clusters. Repetition of this sampling procedure (10,000 times) yields empirical *P*-values which indicate the statistical significance of predicted exon clusters. In case of overlapping exon predictions, one representative according to the highest SVM class-probability was selected to generate non-overlapping gene structures.

Coding vs. non-coding candidates. 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., 2010) (-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".

RNA-seq data. We performed a `BLASTN` search of all 32 nt long reads published by (Wang et al., 2008) against a database consisting of concatenated exon-ends and adjacent exon-starts of clustered exons, each 26 nt in length. These 52 nt long sequences constitute putative mature mRNA fragments. 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. Beyond 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. For the experimental validation of predicted transcripts we designed primers for the region 100 nt up- and downstream of predicted introns 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 was extracted using TRIZOL^{TM} reagent according to the manufacturers instructions (Invitrogen). To eliminate genomic DNA a DNase digestion was performed using the TURBO DNA-free^{TM} Kit (Applied Biosystems/Ambion, manufacturers instructions). After DNase digestion $1\mu g$ of total RNA was reverse transcribed with SuperScript^{TM} III Reverse Transcriptase (Invitrogen). Genomic DNA was isolated using DNeasy Blood & Tissue Kit (Qiagen). PCR reactions were performed using Taq-DNA-Polymerase (NEB) in a $30\mu l$ reaction containing $1\mu l$ 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).

DATA AVAILABILITY

Predicted human splice sites, exons, and gene structures together with a supplemental 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`.

ACKNOWLEDGMENT

We thank Thomas Derrien and Roderic Guigó for fruitful discussions. 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 data sets used in this study. This work is supported by the SAB grant no. 14494, the FP-7 project QUANTOMICS (no. 222664), the German Research Foundation (Hi 1423/2-1), the Excellence Initiative of the German Federal and State Governments (EXC 294 to R.B.), and the Human Frontier Science Program (Fellowship LT000896/2009-L).

REFERENCES

Bernard, D., et al., 2010. A long nuclear-retained non-coding RNA regulates synaptogenesis by modulating gene expression. *EMBO J* **29**: 3082–93.

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.

Chodroff, R., Goodstadt, L., Sirey, T., Oliver, P., Davies, K., Green, E., Molnár, Z., and Ponting, C., 2010. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol* **11**: R72. doi:10.1186/gb-2010-11-7-r72.

Cruveiller, S., Jabbari, K., Clay, O., and Bernardi, G., 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.

Harrow, J., Nagy, A., Reymond, A., Alioto, T., Patthy, L., Antonarakis, S. E., and Guigó, R., 2009. Identifying protein-coding genes in genomic sequences. *Genome Biol* **10**: 201.

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**: R152–61.

Kapranov, P., Willingham, A. T., and Gingeras, T. R., 2007a. Genome-wide transcription and the implications for genomic organization. *Nat Rev Genet* **8**: 413–423.

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

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D., 2002. The human genome browser at UCSC. *Genome Res* **12**: 996–1006.

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

Mercer, T. R., Dinger, M. E., and Mattick, J. S., 2009. Long non-coding RNAs: insights into functions. *Nat Rev Genet* **10**: 155–159.

Nawrocki, E. P., Kolbe, D. L., and Eddy, S. R., 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**: 1335–1337.

Pang, K. C., Frith, M. C., and Mattick, J. S., 2006. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet* **22**: 1–5.

Ponjavic, J., Ponting, C. P., and Lunter, G., 2007. Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* **17**: 556–565.

Ponting, C. P., Oliver, P. L., and Reik, W., 2009. Evolution and functions of long noncoding RNAs. *Cell* **136**: 629–641.

RGASP, 2010. RGASP - The RNAseq Genome Annotation Assessment Project: http://www.sanger.ac.uk/PostGenomics/encode/RGASP.html.

Rodríguez-Trelles, F., Tarrío, R., and Ayala, F. J., 2006. Origins and evolution of spliceosomal introns. *Annu Rev Genet* **40**: 47–76.

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., Kosarev, P., Seledsov, I., and Vorobyev, D., 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol* **7 Suppl 1**: S10.1–S1012.

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

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.

Washietl, S., Hofacker, I. L., and Stadler, P. F., 2005. Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci U S A* **102**: 2454–2459.

Washietl, S., Hofacker, I. L., Stadler, P. F., Findeiß, S., and Goldman, N., 2010. RNAcode: robust prediction of protein coding regions in comparative genomics data. *Submitted* .

Willingham, A. T., Orth, A. P., Batalov, S., Peters, E. C., Wen, B. G., Aza-Blanc, P., Hogenesch, J. B., and Schultz, P. G., 2005. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. *Science* **309**: 1570–1573.

Wilusz, J. E., Sunwoo, H., and Spector, D. L., 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.