Universität Leipzig

Fakultät für Mathematik und Informatik
Institut für Informatik

Hidden Treasures in unspliced EST data

Masterarbeit

Betreuer:
Prof. Peter F. Stadler

vorgelegt von:
Jan Engelhardt

Studiengang:
Informatik

Studienrichtung:
Bioinformatik

Leipzig, January 2014

Acknowledgment

I would like to thank everybody who helped me during my study and writing of this thesis.
Special thanks goes to my supervisor Peter for being the most knowledgeable person I ever met and also for always supporting my funny ideas. Another great help with everything I had to do in the last years was Sonja. My family has been greatly supportive and patient. My working room mates, Sebastian and Axel, kept me motivated all the time and have been very helpful with spell checking. Petra and Jens provided an atmosphere nearly free of adminstrative or computer trouble.
Furthermore, I am very grateful to nature for evolving Coffea arabica and Coffea canephora.

Contents

Chapter 1

Introduction

Non-coding RNA (ncRNA) constitutes a significant portion of the mammalian transcriptome [1, 2]. These transcripts form by no means a homogeneous class, however. A large sub-class of long ncRNAs (lncRNAs) that includes mRNA-like RNAs [3] differs from their protein-coding siblings only in coding capacity: these transcripts are capped, spliced, and polyadenylated. At least some of them, furthermore, are conserved over long evolutionary time-scales [4, 5]. Nuclear retained ncRNAs, on the other hand, are often spliced transcripts but not polyadenylated. These "*dark matter* RNAs", which have remained largely un-annotated so far, can in fact be the dominating non-ribosomal RNA component in a mammalian cell [6]. To-date, most examples of long ncRNAs for which detailed functional information is available are spliced, see e.g. [7] for a recent review.

About 4.5% of the human protein-coding genes are intron-less [8]. Until very recently, they have not received much attention in the literature even though they share special, distinctive features that set them apart from their spliced cousins: Efficient mRNA export to the cytoplasm is normally linked to splicing [9]. Most intron-less mRNAs instead require polyadenylation and specific sequence elements to facilitate efficient packaging into the TREX mRNA export complex [10]. As a group, they are expressed at lower levels, tend to be more tissue specific, evolve at faster rates, and are relatively recent origins [11]. Another distinctive class of intron-less genes is formed by the extremely well-conserved replication-dependent histone genes featuring a unique 3' end structure [12].

In this contribution we focus on the even less-well unspliced non-coding transcripts and characterize their genomic distribution and their organization relative to the much better characterized spliced transcriptional output of the human genome. A survey of the literature, briefly summarized below, suggests that there are at least three major groups of unspliced non-coding transcripts: intronic transcripts typically associated with protein-coding genes, transcripts associated with long 3'-UTRs, and independent

unspliced RNAs found in intergenic regions.

Tens of thousands of totally and partially intronic transcripts (TINs and PINs) have been reported in the human and mouse transcriptomes [13, 14, 15], many of which are unspliced. A large fraction of these comprises unspliced long anti-sense intronic RNAs [16, 17]. Intronic transcripts have been implicated in gene regulation, presumably employing a variety of different mechanisms [18, 19]. Although a detailed analysis of nearly 40,000 putative ncRNAs from RIKEN's FANTOM3 transcript data set showed that a large fraction of the intronic and intergenic transcripts are potentially internally primed from even longer transcripts [20], there are nevertheless thousands of transcripts for which there is no indication that they might be technical artifacts. It is unclear at present how many of them are functional.

The 3' untranslated regions (3'UTRs) of eukaryotic genes not only regulate mRNA stability, localization and translation. In addition, a large number of 3'UTRs in animals can be decoupled from the protein-coding sequences to which they are normally linked. This independent expression of 3'UTR-derived RNAs (uaRNAs) is regulated and conserved. They appear to function as non-coding RNAs in trans [21]. In the form of independent uaRNAs, they are often detectable as unspliced ESTs. A recent study, furthermore, found parallels in sequence composition between lncRNAs and 3' UTRs that sets both groups apart from 5'UTRs and coding regions [22].

The best-known examples of independently located unspliced ncRNAs are MALAT-1 and MENβ. These long ($\sim$ 8.7kb and $\sim$ 20kb, resp.) ncRNAs organize nuclear structures known as SC35 speckles and paraspeckles, respectively [23, 24, 25]. Both transcripts are spliced only infrequently [26] and are rather well-conserved [27]. This class also contains important disease-associated RNAs such as PRNCR1 [28]. In contrast, the even longer ($\sim$ 100kb) transcripts involved in the regulation of imprinted loci (e.g. Airn [29, 30], and KCNQ1OT1 [31]) are very poorly conserved. Similar macroRNAs have been observed in tumor cells [6].

Recent evidence demonstrates that non-coding RNAs can affect gene expression both *in cis* and *in trans* by modulating the chromatin structure [32]. In fact, reports that RNA is an integral component of chromatin have been published already in the 1970s [33]. Chromatin-associated RNAs (CARs) are predominantly non-polyadenylated [34]. Recently, deep sequencing was used to characterize 141 intronic regions and 74 intergenic regions harboring CARs [35]. Many promoter-associated long ncRNAs [36] might also function via chromatin modification [37].

Although a large body of unspliced EST data is available in public databases, they have received little attention as a source of information on non-coding RNAs apart from a seminal work [13]. Here we analyze these data in detail, focusing on their relationships with recently described types on unspliced and rarely spliced transcripts, such as nuclear retained species and independent UTR transcripts.

Chapter 2

Background

2.1 Biological background

2.1.1 What is a gene?

The fact that there must be a mechanism for transferring information from one generation to the next has already been obvious when Darwin published his famous book "On the origin of species" in 1859. However he was lacking sufficient methods to get a grasp on how this mechanism could work. It was not before 1909 that the danish botanist Wilhelm Johannsen coined a word for the Mendelian units of heredity. He called them: "genes". That century was filled with spectacular discoveries, e.g. the double helix structure of DNA in 1953 by Watson and Crick [38], the smei-conservative replication of DNA [39] and the genetic code - deciphering the translation of RNA sequence into amino acids sequence [40]. The technical progress was not less astonishing. A newly developed method for rapid DNA sequencing allowed Friederick Sanger and his colleagues to determine the whole sequence of a bacteriophage genome [41]. The method, later called Sanger sequencing, ultimately resulted in the first draft of the human genome in 2001 [42, 43].
In science it often happens that answering one question gives rise to multiple others. It was the same in genomic research. At the beginning researchers had a very simplicstic view of gene expression it was believed that a certain region on the DNA is transcribed to mRNA and subsequently translated into a protein. These proteins were thought to be the key molecules of life. Being responsible for nearly all processes in the cells of living organisms. Other biomolecules in the cell, namely RNAs not coding for proteins, have been neglected and underestimated in quantity. They have been regarded as a mere transmitter of information and fulfiller of supportive tasks during protein biosynthesis. At the latest since the publication of a study thoroughly analysing 1% of the human genome it is evident that gene regulation in humans (and other eukaryotes) is

much more complex. Non-coding RNAs play important roles in silencing, activating or otherwise manipulating protein-coding genes. Discovery of more and more parts of the gene regulation machinery makes it increasingly difficult to find a clear definition of genes. Earlier, it was sufficient to define a region on the genome and call it a gene, when it gave rise to a specific protein. But nowadays there are examples known when one locus contains many different overlapping transcripts giving rise to functionally different protein isoforms and functional non-protein coding RNAs. Introns can be the host of functional active non-coding RNAs, e.g. snoRNAs. Trans-splicing can combine proteins originating from different chromosomes.
Nevertheless, until a better definition is found and generally accepted we have to work with a variety of different "pseudo"-definitions that are implemented in databases and that are present in the minds of researchers.

Simplistic gene structure

Putting all possible obstacles like intronic genes, overlapping reading frames, trans-splicing and others [44] aside a traditional gene has a reading direction from 5' to 3' and consists of alternating segments of exons and introns. Exons are expressed regions of a gene which are present on the transcript after splicing. Introns (intragenic regions) are regions on the genome, located between exons, they are removed during splicing and are therefore not present on the final RNA transcript. A gene always starts and ends with an exonic region. The beginning of the gene encodes the 5' Untranslated Region (UTR). It can stretch over just a part of the first exon up to several complete exons. UTRs become transcribed and are present on mRNAs but they are not translated to amino acid sequence. They are important for the regulation of translation. Similarly, there is a 3'UTR at the end of the gene of variable length. Figure 2.1 depicts an example of a simple gene structure.

2.1.2 Transcription

An ribonucleic acid (RNA) molecule is created by an RNA polymerase. It is an enzyme that binds to the deoxyribonucleic acid (DNA) and reads the information on the template strand. Simultaneously a stretch of ribonucleotides composed of the complementary bases is generated. The new molecule contains the same information as the coding strand of the DNA. Adenine (A) and thymine (T) are complementary as well as

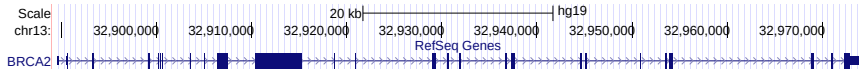


Figure 2.1: *Gene structure of BRCA2 (Breast Cancer 2). It is located on chromosome 13 and shown in forward direction. It consists of 26 exons, depicted by blue bars. The first and a part of the second exon contain the 5'UTR. The remaining part of the second exon and the other 24 exons contain coding sequence (illustrated by a larger bar). A part of the last exon contains the 3'UTR. The exons are of variable size and there are in total 25 introns in between. According to RefSeq gene annotation there is just this splice variant. Source: Genomic location chr13:32889100-32974301 of Homo Sapiens (hg19) in the UCSC genome browser.*

guanine (G) and cytosine (C). There are two important chemical differences between DNA and RNA. In RNA molecules the complementary base of adenine is uracil instead of thymine. Furthermore there is also a difference in the ribose-phospat backbone, since RNA contains ribose instead of deoxyribose. Meaning that the 2'-carbon of ribose has not been reduced to deoxy ribose making RNA less stable.

Simultaneous to RNA transcription a so called 5'-cap is added at the 5'-end of the newly synthesized RNA. The cap is a modified guanosine. This happens just to products of RNA polymerase II, which is responsible for transcription of mRNAs and certain non-coding RNAs. Products of RNA polyermase I and III do not undergo the capping process. The cap's function is to protect the 5'-end from exonucleases and to regulate the protein synthesis process.

The initial transcript of the whole genomic region is called pre-mRNA. Often this transcript is further processed by an RNA-protein complex called spliceosome. The spliceosome cuts out parts of the RNA, so called introns. The adjacent regions, called exons, are spliced together, subsequently, see Fig. 2.2. Introns have been thought to be negligible parts of a gene without function. Nowadays, they are know to play an important role in gene regulation, e.g. for creating different splice variants, or as hosts for small RNAs. While capping normally happens during transcription the timing of splicing is less clear. It can happen co-transcriptionally as well as after termination of transcription.

Not all genes contain introns and therefore undergo splicing. An even more general post-processing mechanism is polyadenylation. It concerns almost all RNA polyermase II products, with histone mRNAs as a notable exception. After transcription is finished a poly-A tail is added to the 3' end of the precursor transcript. The tail is a stretch of

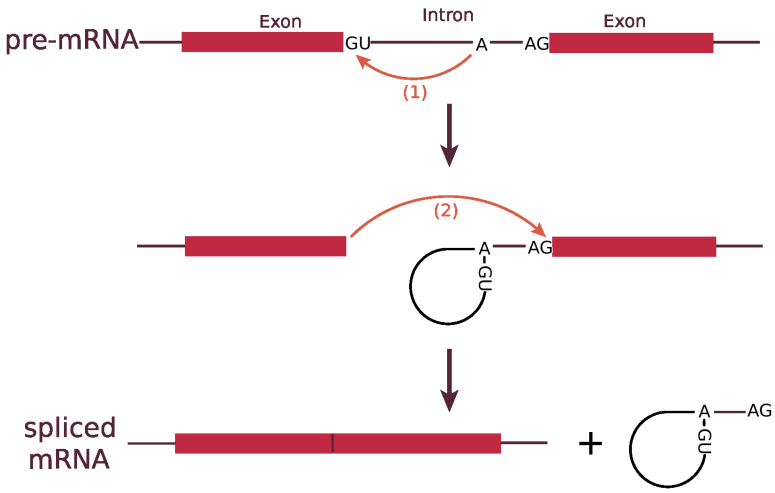


Figure 2.2: *Removal of an intron happens in two steps. In the first step the 5'-end of the intron is cliped from the 3'-end of the upstream exon. The released end is temporarily connected to a site inside the intron. In the second step the 3'-end of the intron is clipped from the 5'-end of the downstream exon and both exons are joined together. Note that the picture just depicts a part of the full RNA. Source:* `http://en.wikipedia.org/wiki/File:RNA_splicing_reaction.svg`

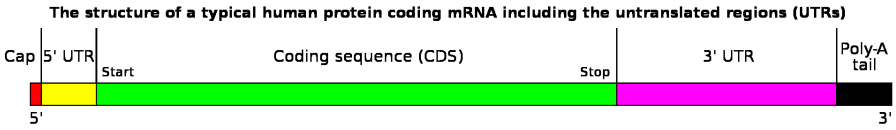


Figure 2.3: *The final form of a typical mRNA befor it leaves the nucleus. 5'UTR, coding sequence and 3'UTR are equivalents of genomic DNA generated by the polymerase. The cap and poly-A tail are added to the transcript through additional post-processing mechanisms. If there have been introns in the pre-mRNA, the are removed by now. Source:* `http://en.wikipedia.org/wiki/File:MRNA_structure.svg`

nucleotides that is build exclusively from adenosine nucleotides. The length of it is normally approximately 250 nucleotides. Sometimes there are several sites for, a possible addition of a poly-A tail which results in alternative transcripts. These differences are known to influence the function of these RNA molecules. Alternative polyadenylation can therefore have a similar effect on alternative splicing. While the 5-cap protects the 5-end, the poly-A tail prevents degradation of the RNA from the 3'-end and is involved in the regulation of protein translation, see Fig. 2.3 for an overview of the final mRNA product.

2.1.3 Transcription regulation

The fact that a gene exists in the DNA does not mean that it is transcribed all the team. Actually, if a transcript is needed highly depends on the current circumstances of a cell, e.g. cell type, age, developmental stage and more. The mechanisms how regulation of transcriptions works are subject to current research. Transcription factors, for example, are proteins binding to the DNA and herby either activating or inactivating the transcription of nearby genes. Even modifications directly on the DNA can regulate the expression of genes. A prominent example is DNA methlyation, which is a modification of the base cytosine. It normally inhibits transcription. However, methylation can be involved in more complex regulation pathways like inhibiting the binding of a repressing factor and therefore activating gene transcription.
Important factors can not just be located directly on the DNA. In the nucleus DNA is winded around protein complexes. These proteins are called histones they are one of

the most conserved protein family in eukarya. Histones have long tails that can be post-translationally modified. Some of these modifications are known to have an effect on the regulation of the DNA wrapped around the respective protein. Mono-methylation and acetylation are in general lead to activation of transcription. Di-methylation and tri-methylation can be a repressing as well as an activating mark depending on their location. There are much more modifications described but for most of the the function is not understood very well. Nevertheless, knowledge of the current histone modification in a certain region of the DNA can help to understand its transcriptional status.

2.1.4 Protein-coding RNAs

The "standard" genes in molecular biology research have been protein-coding genes for quite some time. They give rise to one of the most important classes of molecules in living organisms: proteins. As described in the previous section the information on the DNA is transcribed and processed into mRNA. However there is still one transformation missing till the final product, this step is called "translation". It is a more complex conversion of information compared to the one from DNA to RNA. While DNA and RNA are biochemically very similar and consist of nucleotides, proteins are build from amino acids.

Every nucleotide triplet in the coding part of an mRNA codes for one amino acid. These triplets are also called codons. An exception are stop codons. One of these is located at the end of the coding part of an mRNA and results in termination of the translation process. The three letter code is redundant, meaning there are several codons leading to the same amino acid, see Fig. 2.4. Twenty-two amino acids are naturally occuring in the protein sequence of organisms. 20 of these are encoded by the universal genetic code. The remaining two, selenocysteine and pyrrolysine, require special mechanisms to be incorporated.

The process of translation is similar to transcription. It requires a large complex of molecules, the ribosome, to bind to an mRNA. Transfer RNAs (tRNAs) are going to bind to this complex and translate the information of the codons one by one into amino acids. The first amino acid is usually methionine, encoded by the so called start codon (AUG). Subsequently, tRNAs loaded with an amino acid specific to the respective codon bind to the ribosome-mRNA complex. While the RNA-binding part of the tRNA, the anti-codon, is bound to the mRNA its amino acid is added to the

growing protein chain. This process is repeated until a stop codon is reached. It signals the end of the translation process. The complex is disassembled and while the protein can go on to fullfill its duty the mRNA either serves as a template for another round of translation or is degraded.

2.1.5 Non-coding RNAs

As mentioned before not all RNAs go down the path of translation. The classic examples are transfer RNAs (tRNAs) and ribosomal RNAs (rRNA) which play important parts during translation. However, nowadays there is a plethora of non-coding RNAs (ncRNAs) known. They fullfill various functions and are involved in many gene regulatory processes, e.g. micro (mi)RNAs and small interfering (si)RNA in gene silencing, small nuclear (sn)RNA in splicing, small nucleolar (sno)RNA in nucleotide modification, telomerase RNA in telomere synthesis and many more. Non-coding RNAs are a rather diverse group of molecules. Some of them undergo complex post-processing steps after transcription, like tRNAs and miRNAs.
A plethora of new non-coding RNAs have been discovered in the last decade. However, a certain class has gained attention recently, so called long noncoding RNAs (lncRNA). They are defined as ncRNAs longer than 200 nucleotides. This cutoff is arbitrary and mainly caused by experimental constraints. There are just a few well characterised lncRNAs like Airn [45] and Xist [46]. Their analysis is much more difficult with regards to experimental as well as computational procedures, mostly due to their enormous length of up to several 100 kb. Nevertheless, they are thought to be involved in many genetic and epigenetic mechanisms and can be important factors in developping severe diseases like cancer. New experimental methods and additional studies are needed before one will be able to tell how they interact with other components in the cellular environment and how they evolved to do so.

2.2 Experimental background

2.2.1 EST - Expresed Sequence Tag

In the early 1990's the official Human Genome Project (HGP) was just slowly starting the systematic sequencing of the whole genome. Craig Venter was pondering options to get a glimpse on important parts of the genome in a faster way. Therefore he

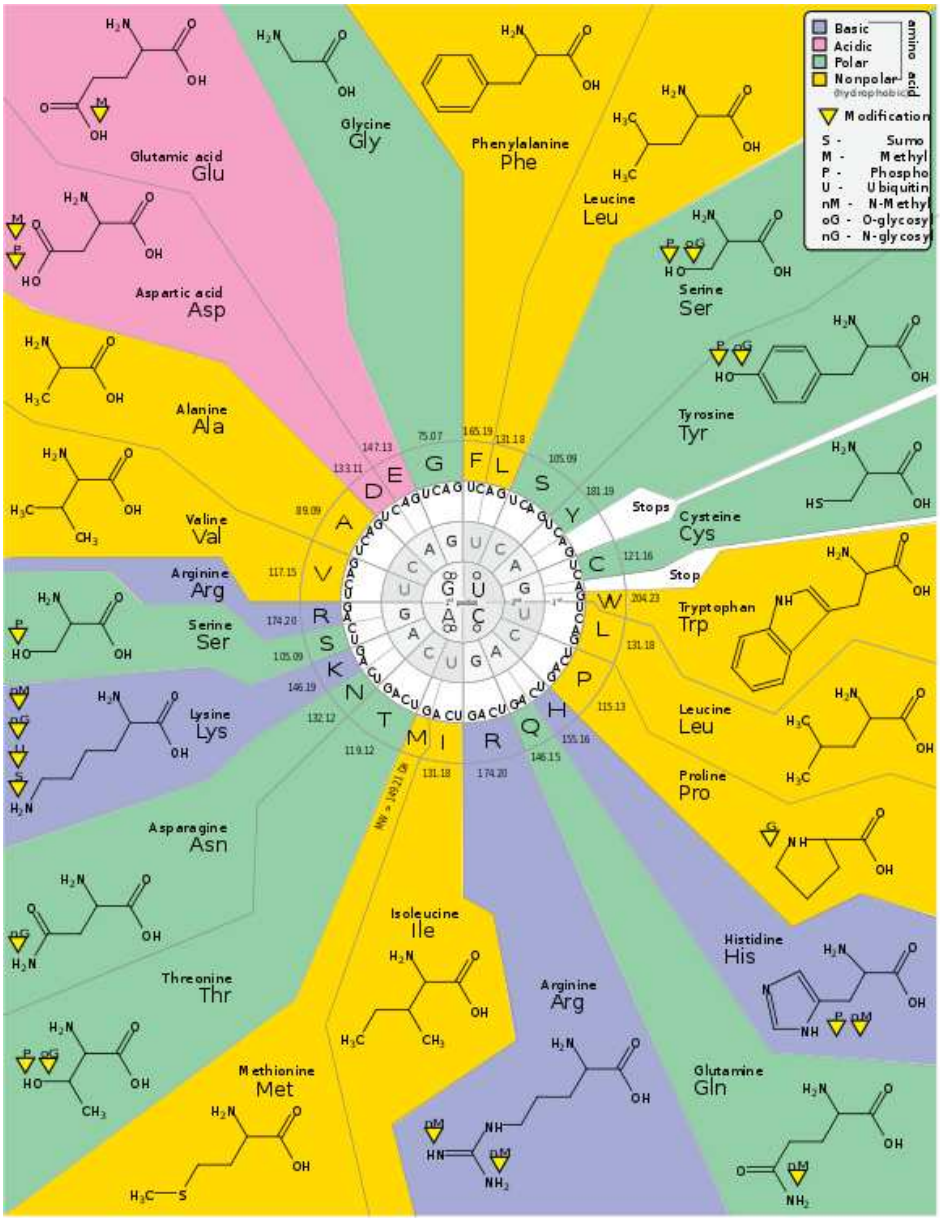


Figure 2.4: *The figure shows codons of the twenty universally used amino acids. The codon has to be read from the inner circle to the outside. The letters in the 4th circle are the one-letter code of the respective amino acid. The number next to it is the molar weight. Methionine, for example, is encoded by "AUG" and the one-letter abbreviation is "M". Its molar weight is 149.21 Dalton. On the outside is the structure of the amino acid and some of the potential modifications known to date. See the legend for further information. Source:* `http://upload.wikimedia.org/wikipedia/commons/d/d6/GeneticCode21-version-2.svg`

proposed to do an initial analysis of just the approximately 3% of the human genome that code for proteins. The way to achieve this was not trying to sequence genomic DNA but expressed RNAs. Venter and colleagues developed the method of "Expressed Sequence Tags" [47].

Library preparation

As a first step one has to isolate RNA from single cells or tissue. There are several standard protocols for RNA extraction. Filtering or enrichment for a special type of RNA makes a great difference for the result. A commonly used method is isolation of poly-A containing RNAs which concerns mostly mRNAs. It is also possible to filter for short RNAs. Without additional purification the whole RNA content is used. RNA can not be cloned directly. It has to be converted into double stranded DNA first. This can be done using the enzyme reverse transcriptase. Like normal DNA polymerase it needs primers to bind to the ribonucleic acid. The selection of these primers is highly dependent on the purpos of the experiment. One can use polyT primers which bind to the poly-A tail of the RNA sequence to start reverse transcription from the 3'-end. Alternatively, there are random primer of length eight which consist of all possible combinations of nucleotides. Random primers can bind to any position of the RNA. A benefit is that the resulting cDNA is shorter and the likelihood that the 5-end of the RNA is reached is higher. One should note that even polyT primers can bind inside the RNA molecule if there are motifs similar to the poly-A tail. Irrespective of the used primer the reverse transcription results in double stranded DNA equivalent to the original RNA. It is called complementary DNA (cDNA). The cDNA fragments can be included into a vector. The vector itself can be inserted into a living organism, often a bacterium. During the normal replication of the organism the vector is also replicated and the amount of cDNA is increased considerably. In case of a bacterial host like `Escherichia coli` the cells containing the cDNA can easily be frozen and stored for later usage. This is refered to as cDNA library.

EST sequencing

The fragment inserted into a vector can easily be extracted again since the vector is artificially constructed and the whole sequence is know. Normally it contains sequencing

primers upstream and downstream of the inserted fragment. The host cell is destroyed and purified until just the vector remains. They are sequenced from both directions in a single run resulting in 5'- and 3'-ESTs. The length of an EST is usually between 200-800 nucleotides [48, 49]. "A typical EST sequence is only a very short copy of the mRNA itself and is highly error prone, especially at the ends." [49]. Sequencing errors are hard to discover because of the single sequencing run on random fragments. Also sequence amplification in bacterial vectors causes artifacts resulting in alterations in the final sequence output compared to the original. Sequence artifacts from the vector can be detected and removed, in theory. However, this was not always done for ESTs in public databases. If one uses these databases, e.g. dbEST, there is not much information about the generation of the ESTs available. Therefore one has to treat the data as a mixture of all possible experimental procedures.

2.2.2 CAGE - Cap Analysis Gene Expression

Sequencing of parts of a genome, e.g. protein-coding genes, often involves a reverse transcription of the original RNA sequence. In case of mRNAs oligo (dT) primers can be used conveniently. But starting the reverse transcription from the end of the DNA fragment has the drawback that the sequencing coverage is the worst at the beginning of the fragment. Some issues render it crucial to know the exact transcription start site. For example detecting alternative start positions or folding the secondary structure of an mRNA.
This issue was adressed by *Shiraki et. al* [50] by developing the so called "Cap analysis gene expression" (CAGE). It is an experimental technique specialized on sequencing the first 20 base pairs of a transcript. The technique is based on the abilities of *MmeI*, a restriction enzyme that cuts cDNAs at a sequence 20 and 18 bp downstream from its recognition site, creating a two-base overhang. After generation of full-length cDNA a linker containing a *MmeI* recognition site is ligated to the single stranded cDNA. This primes the syntesis of the second DNA strand. Subsequently, the double stranded cDNA is digested with *MmeI* to cut of the first 20/18 bp of the mRNA. At the cutting point on the end of the small mRNA fragment a second linker is ligated. These two linkers provide primers for PCR amplification and allow convenient sequencing. In distinction to other tag cloning methods [51] the CAGE protocol uses a second linker to mark the back end of the CAGE tag. This is advantageous because sometimes the

cut is 19/21 base pairs or even farther away from the recognition site. For an overview of the method see Fig. 2.5.

2.3 Computational background

2.3.1 Public data sources

Most of the data was gained from the "UCSC Genome browser" [52]. Originally it was created to "provide a rapid and reliable display of any requested portion of genomes at any scale, together with dozens of aligned annotation tracks" because "it is not helpful to have 3 billion letters of genomic DNA shown as plain text!" (`http://genome.ucsc.edu/goldenPath/help/hgTracksHelp.html`). Nowadays it's also a major hub of different data tracks from collaborators all over the world [53]. For example all data of the ENCODE project is available for visualization in the UCSC genome browser [54]. Another prominent example is the European BLUEPRINT Epigenome project [55] which provides access to its data via a "Public Hub" [54] at the genome browser. All visualization tracks can be downloaded and further processed like we did. This allowed use to get access to different data sets without accessing the original database like dbEST [56] for EST data.
Another track widely used by us was the data provided by the Reference Sequence (RefSeq) database. It is "a collection of taxonomically diverse, non-redundant and richly annotated sequences representing naturally occurring molecules of DNA, RNA, and protein" ("The NCBI Handbook", `http://www.ncbi.nlm.nih.gov/books/NBK21091/`). It is an attempt to obtain a reference set of genes in a large number of organisms. The data is derived from publicly available sequence data and analysed by a combination of automated methods and manual curation. However, due to the large amount of data even just for the human genome it is hard for the curators to keep up. Therefore it can occur that single entries are not yet reviewed and may be erroneous, in particular concerning annotation like splice variants, UTR length or contaminations. Nevertheless its a valuable ressource especially for evolutionary studies since there are not many alternative gene anotation databases that keep manually curated genomes other than human.

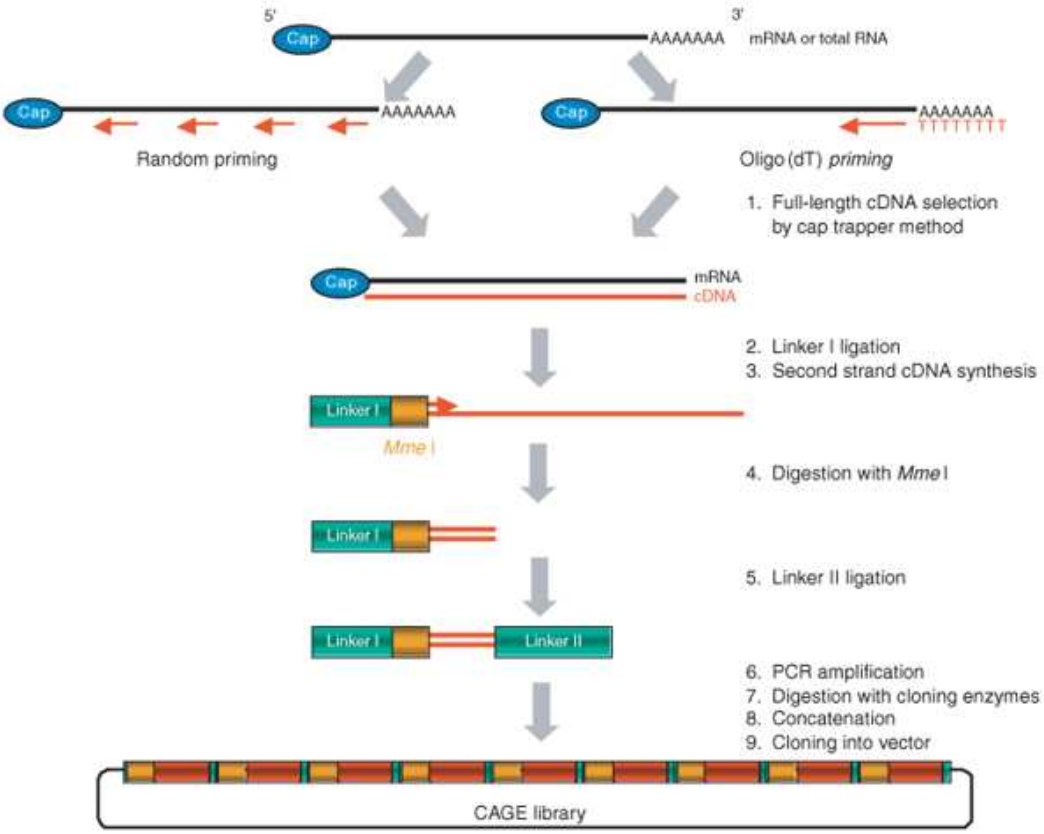


Figure 2.5: *Complementary DNA can be generated by different methods but it is filtered for full-length cDNA by the cap trapper method. A experimental procedure which enriches for caped RNAs. A linker containing a MmeI recognition site is ligated to the single stranded cDNA and the second strand is synthesized. Subsequently the double-stranded cDNA with the linker is digested by MmeI, a restriction enzyme, leaving just the 5-end of the RNA. The strands have slightly different lengths, in general 20 and 18 base pairs. Afer ligating a second linker to the end of the mRNA fragment it can be easily amplified and sequenced. The linker provide binding sites for primer. Picture by Harbers and Carninci, 2005 [51].*

2.3.2 RNAz 2.1

`RNAz` is the implementation of an efficient method for detecting functional RNAs [57, 58]. Given a multiple alignment `RNAz` can predict the possibility that a conserved secondary structure is present in the complete alignment or a subset of the sequences. The prediction is based on two independend criteria, thermodynamic stability of the RNA structure and structural conservation. The thermodynamic stability is calculated by estimating the standard and mean deviation using a support vector regression (SVR). According to the authors the SVR is necessary because of the large number of random sequences for which a energy evaluation would be necessary [58]. Structural conservation is determined using the structure conservation index (SCI), $SCI = \frac{E_{consensus}}{\frac{1}{N}\sum_{x \in A} E_x}$. $E_{consensus}$ is the consensus secondary structure of the input alignment predicted by *RNAalifold*. $\frac{1}{N}\sum_{x \in A} E_x$ is the average minimum free energy of the sequences contained in the alignment, individual minimum free energies are predicted by *RNAfold*.
Both criteria are evaluated by a second support vector machine which classifies the input alignment as "structural RNA" or "other". The workflow of `RNAz` is graphically depicted in Fig. 2.6.

2.3.3 RNAcode

A large problem in genome-wide screens for ncRNAs is the rate of false positives. To support `RNAz` a second program has been developed: `RNAcode` [59]. Its purpose is to predict the probability of a region to be protein coding. Combining the predictions of `RNAz` and `RNAcode` leads to a decrease in the rate of false positive ncRNA predictions. Similar to compensatory mutations in the structure of an RNA most mutations in functional important parts of protein coding genes are synonymous. Meaning that the alteration of one nucleotide does not chance the resulting amino acid because of the redundant genetic code, see Fig. 2.4. This provides valuable information for deciding if a region is protein coding.
`RNAcode` uses multiple sequence alignments with a reference sequence as input. This alignment is used to estimate a phylogenetic tree assuming a noncoding nucleotid model. Mutations in the alignment are subsequently evaluated for beeing under negative selection. This score combined with a gap scoring scheme are the basis for a dynamic programming (DP) algorithm that detects local high-scoring coding segments.

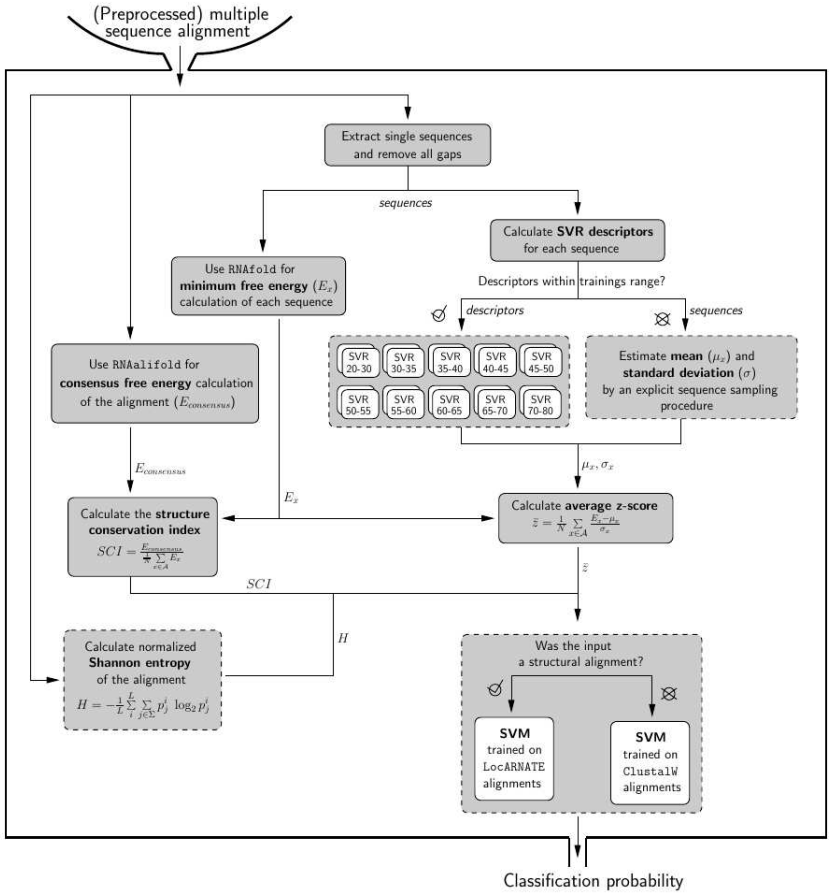


Figure 2.6: *Overview of the RNAz algorithm: A multiple sequence alignment is used as input. Using RNAalifold the consensus free energy of the alignment is calculated. Together with the individual minimum free energies, predicted by RNAfold, the structure conservation index (SCI) is calculated.*
Simultaneously, it is estimated if the mean free energy and the standard deviation are withing the training boundaries. If yes, the are passed to the SVR based on the G+C content. Otherwise they have to be evaluated explicitly to clalculate the z-score.
The SCI, the z-score and the "normalized Shannon entropy" of the alignment are passed to the classification SVM, which returns a probability estimate that the given alignment contains a evolutionary conserved RNA secondary structure. For a more detailed description, see the source of the picture: Gruber et al., 2010 [58].

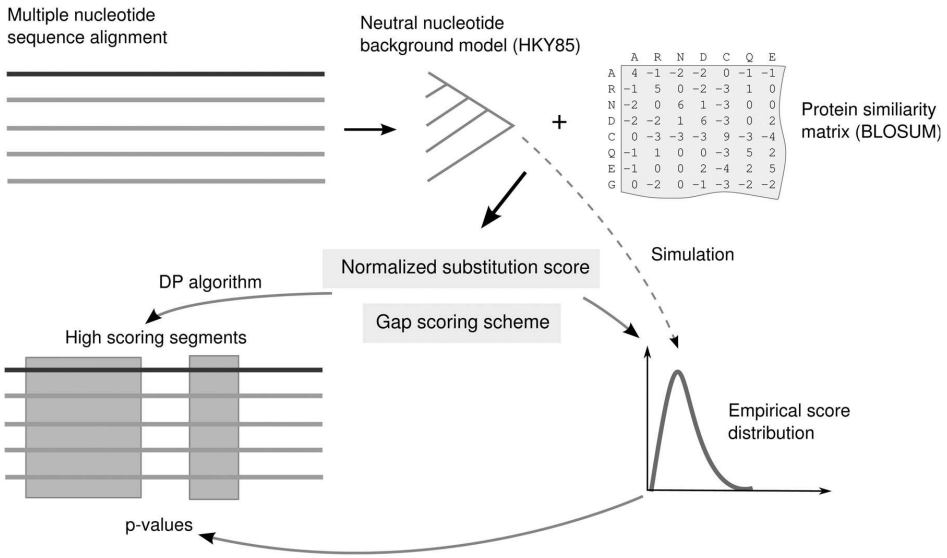


Figure 2.7: *"Overview of the RNAcode algorithm. First, a phylogenetic tree is estimated from the input alignment including a reference sequence (darker line) under a noncoding (neutral) nucleotide model. From this background model and a protein similarity matrix, a normalized substitution score is derived to evaluate observed mutations for evidence of negative selection. This substitution score and a gap scoring scheme are the basis for a dynamic programming (DP) algorithm to find local high-scoring coding segments. To estimate the statistical significance of these segments, a background score distribution is estimated from randomized alignments that are simulated along the same phylogenetic tree. The parameters of the extreme value distributed random scores are estimated and used to assign P-values to the observed segments in the native alignment." [59]*

The statistical significance of the segments is estimated by comparision to a background score distribution estimated from randomized alignments along the same phylogenetic tree. An overview of the algorithm can bee seen in Fig. 2.7.

Chapter 3

Materials and Methods

3.1 Data

The annotation track 'all_est' for human genome assembly hg19 was downloaded from the UCSC genome browser (30th of April 2013) [60]. Starting from 'ESTs including unspliced' we removed all ESTs with more than 30 deleted nucleotides compared to the reference genome. This gets rid of all the annotated spliced ESTs contained in this track. It also discards sequences with sometimes long, intron-like, gaps that are not annotated as spliced ESTs because they map without canonical splice sites. The cutoff of 30 nucleotides was choosen since even smaller introns are extremly rare [61, 62]. We furthermore discarded ESTs mapped with more than 5% mismatches. The hg19 release of the human genome contains 9 alternative haplotypes for highly variable regions on chromosome 6 [63]. They were not included in the further analysis to prevent counting identical ESTs several times. We obtained 3,425,788 unspliced ESTs (Tab. 3.1). The same procedure was done for the 'all_est' track for mouse genome assembly mm10 also downloaded from the UCSC genome browser (30th of April 2013). We obtained 2,177,648 unspliced ESTs in mouse (Tab. 3.1).
A possible source of contamination is nuclear encoded mitochondrial DNA (NUMT) [64]. Since the mitochondrial transcripts remain unspliced at least in mammals [65], it is impossible to reliably distinguish unspliced ESTs mapping to recent NUMTs from fragments of mitochondrial transcripts. An annotation track for human and mouse NUMTs was recently provided in [66].
The gene locations and structure were taken from the 'RefSeq Genes' track [67] downloaded from the UCSC genome browser. Multiple genome alignments in maf format for human (46way) and mouse (60way) have also been downloaded using the UCSC table browser [60]. Input alignments for `RNAcode` have to be filtered for alignments with a lenght of less than 21 nucleotides for computational reasons. These alignments were discarded. CAGE data for hg18 and mm9 assembly were downloaded from 'Fantom

Table 3.1: *Summary of human EST data. Analysis of the EST annotation track of hg19 and mm10 downloaded from the UCSC genome browser in April 2013. The numbers indicate successive filters. Which means just the ESTs which do not have introns larger than 30 nucleotides are included in the number of introns with more than 5% mismatches and so on.*

Type	Human	Mouse
all ESTs	8,675,182	4,370,322
ESTs on Haplotypes	534,970	n/a
$> 30nt$ gaps	4,514,773	2,107,660
$> 5\%$ mismatches	169,967	68,239
overlap with NUMTs	29,684	16,775
unspliced ESTs	3,425,788	2,177,648
unspliced EST cluster	104,980	66,109

Table 3.2: *Summary of human EST data. Analysis of the mRNA annotation track of hg19 and mm10 downloaded from the UCSC genome browser in November 2013. The numbers indicate successive filters. Which means just the mRNAs which do not have more than one block are included in the number of mRNAs with more than 60nt in length.*

Type	Human	Mouse
all mRNAs	399,812	329,819
> 1 block	247,389	188,945
$< 60nt$ lenth	93,301	91,182
unspliced mRNAs	59,122	49,665

Web Resource' [68, 69]. Their coordinates were converted to hg19/mm10 using the UCSC `liftOver` tool [52]. The genome segmentation based on ENCODE data for all six cell lines (GM12878, H1-hESC, K562, HeLa-S3, HepG2, HUVEC) was downloaded from the ENCODE public hub at the UCSC genome browser (8th of January 2014). The combinded segmentation was used. Coordinates of chromatin-associated RNAs (CARs) were taken from the supplemental material of [35] and lifted over to hg19.

3.1.1 Unspliced mRNAs

The track "all_mrna" was downloaded for Human and Mouse from UCSC genome browser on 15th of November 2013 and 18th of November, respectively. We removed

all mRNAs with more than one block, since we were interested in unspliced ones. A large part of the remaining data consisted of rather short mRNAs. As far as we know the shortest mini-protein known is an artificially designed 20-residue construct [70]. It is therefore safe to say that mRNAs not even able to get translated to 20 amino acids are unlikely to be naturally occuring protein-coding transcripts. Consequently, we removed all mRNAs with a length shorter than 60 nucleotides, see Tab. 3.2.

3.2 Analysis pipeline

While the reading direction of spliced ESTs can be determined with high accuracy from the asymmetric structure of the splice sites, unspliced EST data in practice have to be regarded as undirected. The only exception are ESTs arising from polyT-primed libraries, provided the polyA-tail is part of the sequenced fragment. However, there are also internally-primed ESTs containing A-rich regions [71], so that even such a sequence-based detection of the reading direction is not reliable in all cases. Unspliced ESTs are therefore treated as undirected annotation items.
We also have to bear in mind that ESTs are, by definition, not full-length mRNAs. In particular they may cover an unspliced stretch of a spliced transcript only and thus appear unspliced. In particular we are interested in genomic loci where unspliced ESTs cluster together. To this end we define unspliced ESTs that overlap each other or that are separated by at most 30nt as members of the cluster. In order to avoid artifacts of both the experimental and the computation procedures we only consider unspliced EST cluster comprising at least three individual ESTs, leaving 104,980 human and 66,109 mouse unspliced EST cluster for further analysis.
In order to computer overlaps with existing annotation we used the functionality of the Table Browser integrated in the UCSC genome browser and the “Operate on Genomic Intervals”-tools of the Galaxy Browser [72] as well as *BEDTools* [73].

3.3 Classification

In order to gain insights about the location of unspliced EST (uEST) cluster relative to RefSeq gene components we classified the uEST cluster in different classes. To a lesser extent this was also done by *Nakaya et al.* [13] who introduced the terms TIN (totally intronic) and TEX (totally exonic).

For every uEST cluster the class for every overlapping RefSeq gene and a 5kb upstream/downstream region was determined to be one of the following classes:

- TEX - Overlap exclusively with a RefSeq exons including UnTranslated Regions (UTR) and non-coding RNA
- TIN - Overlap exclusively with a RefSeq intron
- 5'PIN - Overlap exclusively with an adjoining exon-intron pair, where the exon is located upstream of the intron
- 3'PIN - Overlap exclusively with an adjoining exon-intron pair, where the exon is located downstream of the intron
- rI - Overlap exclusively with one intron and both adjacent exons
- 5'R - Overlap with a RefSeq exon and the 5kb upstream region
- 3'R - Overlap with a RefSeq exon and the 5kb downstream region
- UT - Overlap with 5kb upstream region
- DT - Overlap with 5kb downstream region
- IGR - No overlap with any RefSeq gene or the 5k upstream/downstream region
- NO_CLASS - None of the other classes could be applied

If the uEST cluster was determined to have the same class for all overlapping RefSeq genes, it was assigned to this class. In case of conflicting data it was assigned to the class "NO_CLASS". An exception was made if the conflict was between one of the classes TEX, TIN, 5' PIN, 3' PIN, rI, 5'R and 3'R and one of the classes UT or DT. In this case the overlap with an additional upstream/downstream region was ignorred. Otherwise a class assignment would be impossible in regions with a high gene density. To avoid misclassifications due to small mistakes in the borders of RefSeq gene components the overlap between the uEST cluster and the gene component had to be more than 20 nucleotides to be considered.

3.4 Conservation

For detecting conserved regions we use already existing pairwise alignments between human and mouse generated by the UCSC/Penn State Bioinformatics comparative genomics alignment pipline [74]. The human-mouse alignment was created by aligning both genomes using blastz [75]. The coordinates were afterwards corrected using in-house scripts (`blastz-normalizeLav` by Scott Schwartz of Penn State). The mouse-human alignment was done by using `chainSwap` "to translate hg19-reference blastz alignment to mm10 into mm10-referenced chains aligned to hg19." [52]. Interestingly the mm10-based alignment leads to more ortholog clusters than the original one.
The chain-files were used as input to `liftOver` [52], a tool original developed to switch between genome assemblies. The human unspliced EST cluster were lifted over to mm10 by using the hg19-reference alignment, mouse unspliced EST cluster were lifted over to hg19 by using the mm10-referenced alignment, respectively. We combined all ortholog pairs detected by one of the `liftOver` runs.

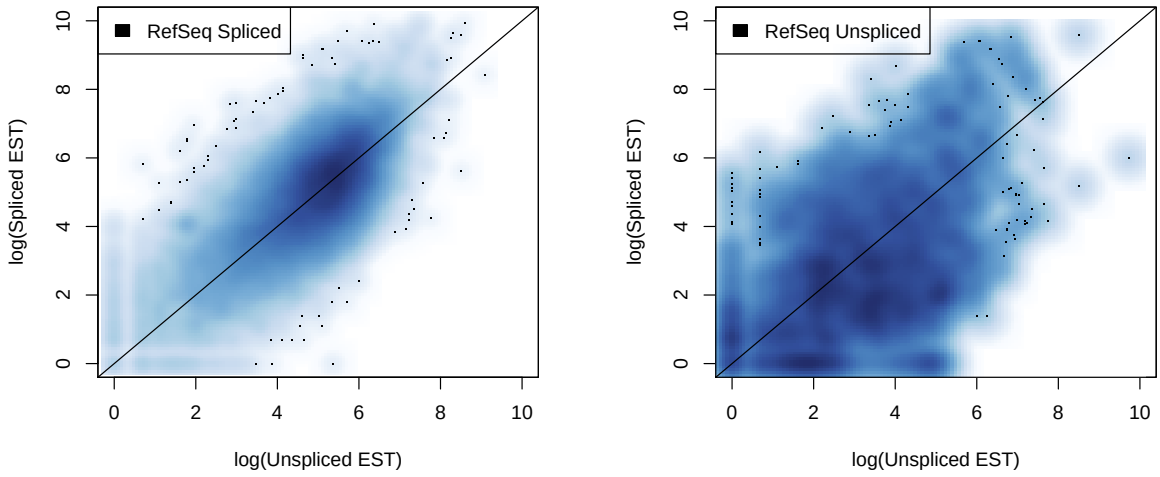
Chapter 4

Results

4.1 ESTs "within range" of RefSeq genes

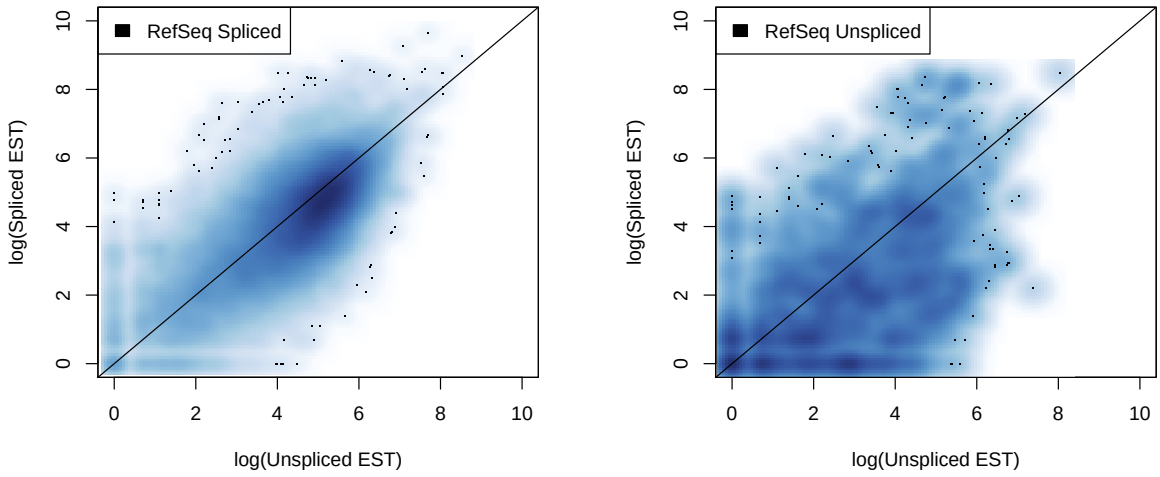
The majority of unspliced EST cluster overlaps, or at least lies in close proximity of, already annotated RefSeq genes. This begs the question whether unspliced ESTs are just a by-product of "normal" spliced transcripts. To address this point we compare the abundance of spliced and unspliced ESTs with the range of annotated RefSeq genes (including a 5kb flanking region), see Fig. 4.1 for human and Fig. 4.2 for mouse. Although there is a correlation ($r = 0.72$) within the range of spliced RefSeq genes, we observe a surprising variability in the numbers of unspliced ESTs, with relative abundances of spliced and unspliced sequences typically varying by a factor of ten or more. Conversely, there is a surprisingly large number of spliced ESTs overlapping with annotated unspliced RefSeq genes. In these loci, the correlation between spliced and unspliced output is even less pronounced. The 'Pearson's product moment correlation coefficient' (r) are computed from the logarithms of EST counts ignorring all genes that did not have an overlap with a spliced/unspliced EST. Taken together, these quantitative data strongly suggest that at least a sizeable fraction of unspliced transcripts is independent of overlapping spliced forms either already at the transcriptional level or by subsequently being processed in a different way.
We therefore analyzed in detail the relative location of unspliced EST cluster and components of RefSeq genes, Fig. 4.3.

Compared to *Nakaya et al.* (Table 1) [13], we can work with a much larger data set of *Totally INtronic* (TIN) unspliced EST cluster, comprising 38,803 (vs. 5,678) but nearly the same amount of *Partially INtronic* (PIN), 10,025 (vs. 9,132). Interestingly, the number of detected TINs has increased by a factor of 6.8, while PINs increased by only a factor of 1.1, suggesting that the coverage of TINs is much farther from saturation than that of the PINs. See Fig. 4.4 for an example of a TIN unspliced EST cluster. Additionally we can utilize a comparable, but slightly smaller, amount of data



spliced ($r = 0.72$) unspliced ($r = 0.50$)

Figure 4.1: *Scatter plots of the relative abundance of human spliced and unspliced ESTs within the range of both spliced and unspliced RefSeq genes show only a relatively poor correlation.*



spliced ($r = 0.72$) unspliced ($r = 0.48$)

Figure 4.2: *Scatter plots of the relative abundance of mouse spliced and unspliced ESTs within the range of both spliced and unspliced RefSeq genes show only a relatively poor correlation.*

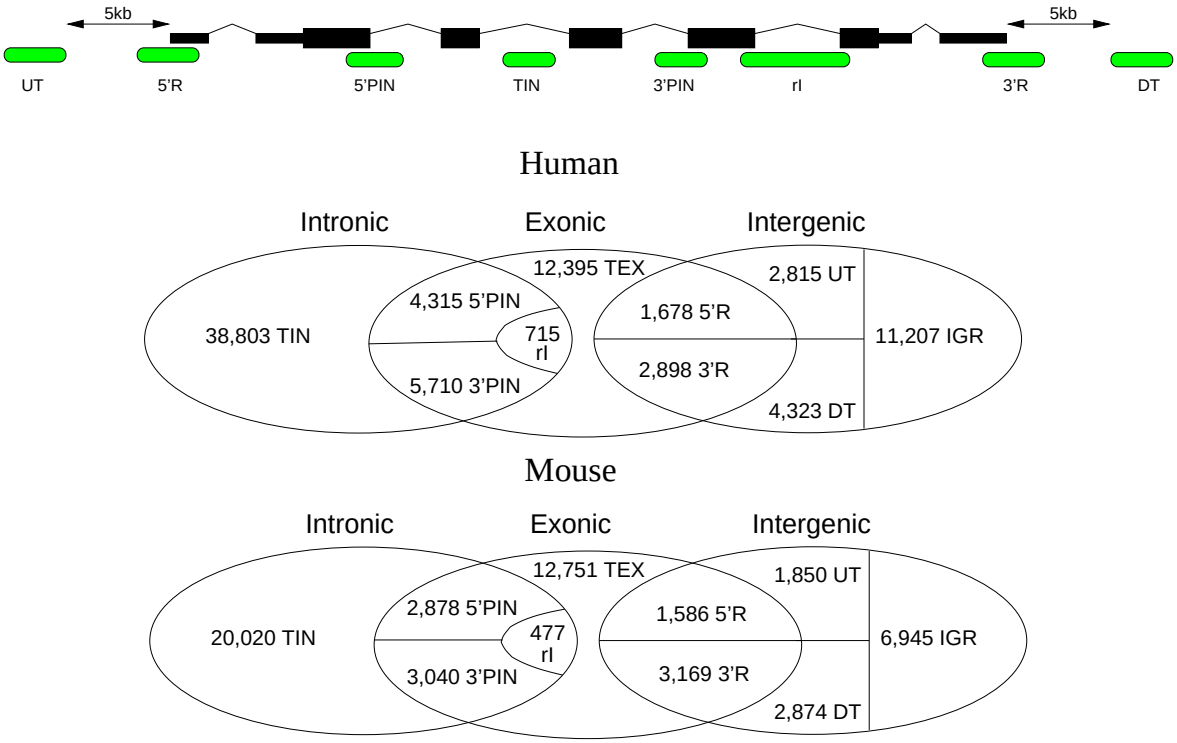


Figure 4.3: *Classification of unspliced EST cluster w.r.t. their location relative to RefSeq genes. With the exception of totally intronic RNAs (TINs) and cluster in the upstream (UT) and downstream (DT) region within 5kb, all other classes partially overlap RefSeq exons: 5' and 3' partially intronic RNAs (5'PIN, 3'PIN), EST cluster overlapping 3'UTR and downstream region (3'R) or 5'UTR and upstream region (5'R), resp., and cluster covering complete introns indicating retained introns (rI) are distinguished in the statistical analysis. Furthermore, we record totally exonic cluster (TEX). Below, the data are summarized as a Venn diagram. Some cluster cannot be classified unambiguously, mostly because two or more RefSeq genes may overlap the same locus. 20,121 (19%) of the unspliced EST cluster in human and 10,519 (16%) in mouse can not be classified. For more details see section 3.3.*

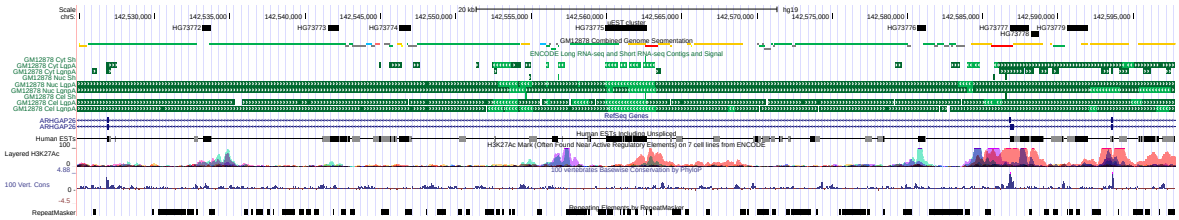


Figure 4.4: *ARHGAP26 (Rho GTPase activating protein) is a long protein-coding gene on chromosome 5 in human, stretching over appr. 46kb (blue line with arrows). Between the 20th and 21st exon we find a unspliced EST (uEST) cluster (HG73775) with a length of 2,700 nt (black box). Additional evidence for functional importance is given by ENCODE data. Long RNAseq transcripts antisense to the protein-coding gene have been reported (green bar, white arrows depict the reading direction). Additionally the region upstream of the uEST cluster is classified as transcription start [76] (the tiny red bar enclosed by yellow and green bars below the EST track), assuming the reading direction predicted by RNAseq. There is also an enrichment of the histone modification H3K27Ac around the uEST cluster (the colored transparent overlays near the bottom). This modification is often found near active regulatory elements.*

for the rodent *Mus musculus*. The protein-coding part of RefSeq genes overlaps 39,791 unspliced EST cluster (27,897 in mouse). Less than 4% of these cluster in each species, however, are located completely in the coding region.

In order to obtain a more fine-grained view of distribution of TINs and PINs, we distinguish PINs depending on whether they overlap the exon/intron boundary at the donor or the acceptor side, see Fig. 4.3. Furthermore, we treat coding regions and UTRs of coding RefSeq genes separately.

In addition to the unspliced ESTs overlapping the body of the RefSeq genes we also consider EST cluster in the immediate proximity, here defined as within 5kb of the annotated ends of the RefSeq entry. There are 4,868 (3,682) unspliced EST cluster in upstream region of human (mouse) RefSeq genes and 7,035 (5,201) downstream, see Fig. 4.3.

4.2 Independent UTR-derived RNAs

Unspliced EST cluster show a bias towards the 3' end of the RefSeq genes. There are more than 7,000 (6,000 in mouse) cluster of unspliced ESTs overlapping, or located within 5000 nt downstream of, the 3'UTR of human (mouse) RefSeq genes. The fre-

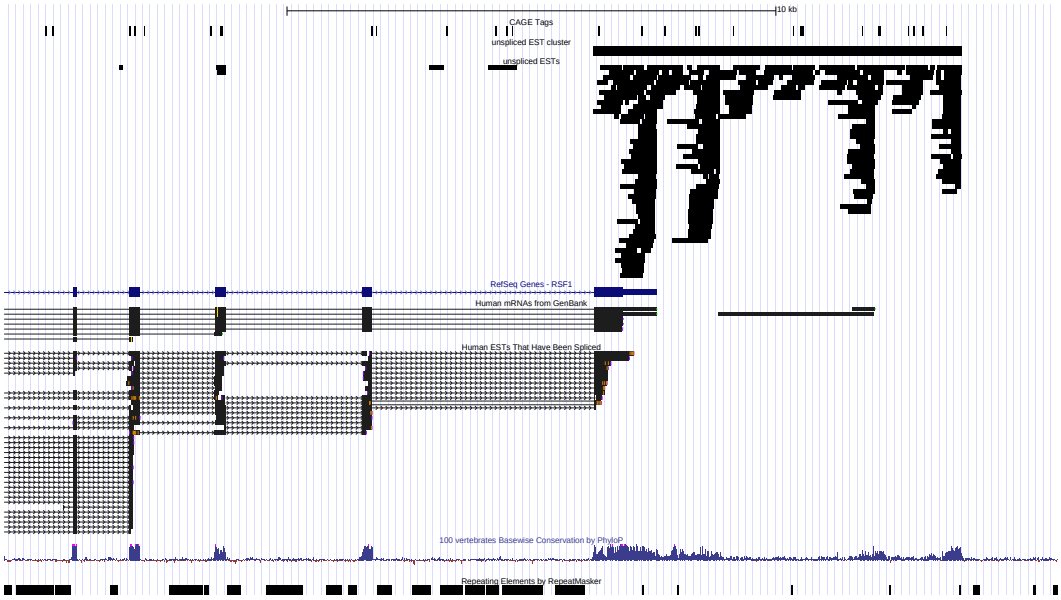


Figure 4.5: *The 3'UTR of the RSF gene is extended by a single cluster of unspliced ESTs covering about 7 kb. Compared to the large number of unspliced ESTs, only a few spliced ESTs cover the inner exons of RSF. The presence of several CAGE tags in the extended UTR region suggests that independent uaRNAs are produced from this locus [hg19 Chr11:77369073-77390595, reverse strand]. This is further supported by two unspliced GenBank mRNAs that map to the extended UTR. It is interesting to note that a large fraction of the extended UTR is very well conserved and nearly devoid of repetitive sequence.*

quent observation of downstream signals is consistent with the recent observation that the length of UTRs is often underestimated in current gene annotations [77]. An impressive example is shown in Fig. 4.5. We frequently find that CAGE tags, i.e., markers for transcription start sites, are located within long unspliced 3'UTRs. This supports the observation found by *Mercer et al.* [21] that these 3'UTRs are also transcribed in an independent mode, or are at least processed to become independent gene products [78]. 5.9 million (2.7 million in mouse) of the approximately 13.4 million (5.5 million) CAGE tags overlap with a total of 71,941 (38,805) unspliced EST cluster.

Figures 4.6 and 4.7 show that CAGE tags are predominantly located at the ends of unspliced EST cluster. The symmetry of the diagram is expected since we treat unspliced ESTs as undirected. In fact, we can use the CAGE tags to determine the reading direction of those cluster in which the tag distribution unambiguously concentrates on one end, which is the 5' end.

Here, we require that the density of CAGE tags in the first 84 nt is at least threefold

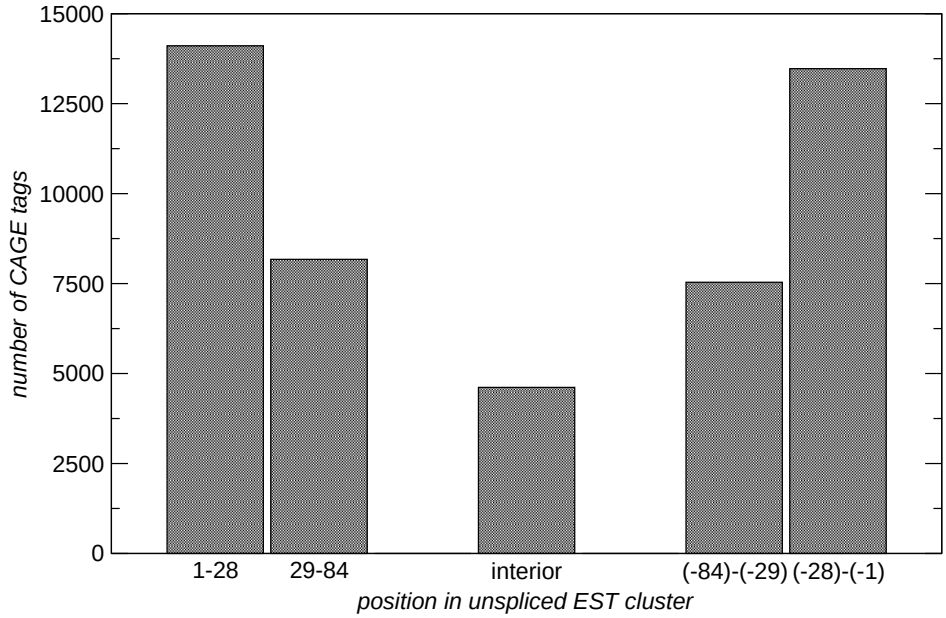


Figure 4.6: *Distribution of CAGE tags on human unspliced EST cluster (with length* ≥ 170*). The number of tags is divided by the length of the region.*

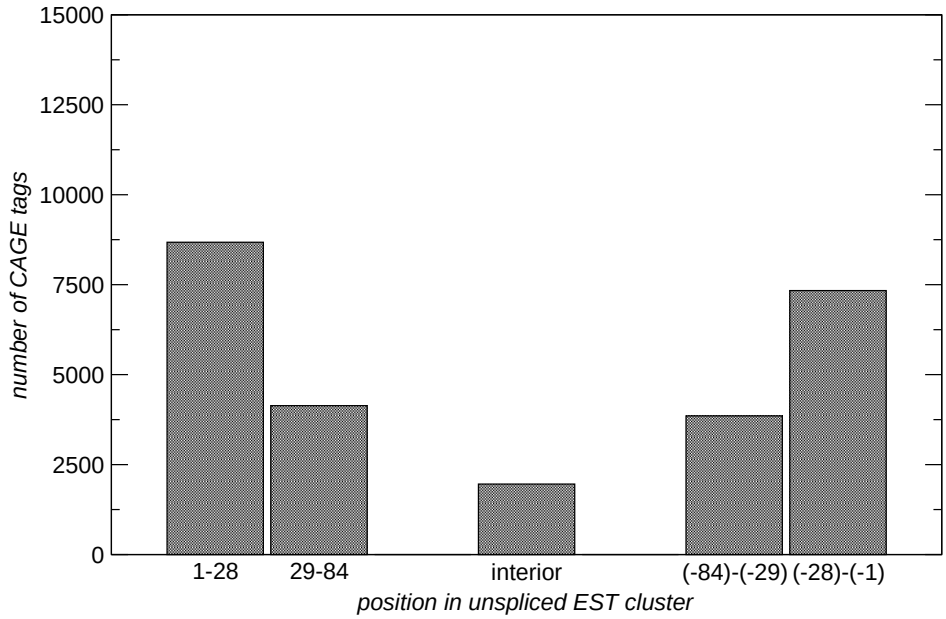


Figure 4.7: *Distribution of CAGE tags on mouse unspliced EST cluster (with length* ≥ 170*). The number of tags is divided by the length of the region.*

higher than in the remainder of the cluster. We furthermore required at least 10 CAGE tags at the 5' end. By utilizing the strand information of the CAGE tags further reliability can be gained. Tags in the beginning of an unspliced cluster have to be on the sense strand while in the end they have to be on the antisense strand.
We found at total of 2,872 (2,989 in mouse) such cluster with a CAGE-supported transcription start and an unambiguous reading direction. A subset of 1,847 (1,612 in mouse) overlap at least 100 CAGE tags at the 5' end, each.
Not surprisingly, many cluster of unspliced ESTs with overlaping CAGE tags correspond to the 5' ends or 5' extensions of annotated RefSeq genes. These were excluded in the detailed analysis since we can not distinguish the transcription start site (TSS) of the corresponding RefSeq gene and a independent one of an unspliced EST cluster, see Tab. 4.1. This is also the reason why we are just considering 3'UTR-associated RNAs. As expected, most of these cluster share the reading direction of the RefSeq gene. This is also the case for PINs. For the relatively small number of human TINs with a strong CAGE signal we find a 3:1 ratio of sense *versus* antisense orientation, consistent with previous observations [13]. At least some of these unspliced cluster may correspond to a class of functional promoter-associated ncRNAs similar to the CCND1-pncRNAs [37], which negatively regulate CCND1 transcription by recruiting TLS to the promoter.

A particular class of most unspliced transcripts are the 215 chromatin-associated RNAs (CARs) [35]. Of these, 143 (67%) overlap unspliced EST cluster. 191 are located within the range of RefSeq genes ±5 kb of flanking region. We therefore evaluated in more detail how these CARs are distributed relative to the organization of their RefSeq gene: 97 (51%) are located totally in intronic region. 30 (15%) are located in the 3' area (5 are exclusively in the 3' UTR, 16 overlapping the 3' UTR and 9 in the 3' flanking region), while 20 are associated with the 5'UTR or 5' flanking region. 14 CARs are totally in exonic regions.

4.2.1 Chromatin architecture

CAGE is a method developed for exactly determining the 5'-end of RNAs. A more indirect way to detect the transcription start site (TSS) of a gene is analysing the chromatin structure, see Section 2.1.3 for more information.

We used a genome-segmentation track based on ENCODE data by *Hoffmann et*

Table 4.1: *Unspliced EST cluster whose reading direction is determined by CAGE tags and their orientation relative to the surrounding RefSeq gene. Only cluster overlapping with at least 10 individual CAGE tags in their 5' region are considered. RefSeq TSS are all unspliced EST cluster that overlap any 5'UTR or that are located completely within the upstream region of a RefSeq gene. TINs and PINs are defined in Fig. 4.3. All unspliced EST cluster that have an overlap with the 3'UTR but not with the 5' UTR are interpreted as uaRNAs. "Sense" and "Antisense" are relative to the reading direction of the RefSeq gene. "Ambiguous" cluster could not be assigned to a orientiation because of conflicting directions. The column "All" gives the equivalent numbers for all unspliced EST clusters without considering an overlap with CAGE-tags.*
** - unspliced EST cluster overlapping 5' UTR have been excluded for this analysis*

Type	Sense		Antisense		Ambiguous		All	
Species	Human	Mouse	Human	Mouse	Human	Mouse	Human	Mouse
RefSeq TSS	1,638	1,802	94	32	586	659	12,693	9,805
TINs*	53	45	17	7	16	15	38,803	20,200
5' PINs*	73	61	0	0	37	33	2,926	1,527
3' PINs*	9	8	4	0	1	1	5,439	2902
uaRNAs	21	24	3	1	16	18	17,331	17,126

Table 4.2: *Unspliced EST (uEST) cluster overlapping with chromatin elements predicted from ENCODE data [76]. 'TSS' represents "Predicted promoter region including TSS". 'Transcribed' refers to "Predicted transcribed region". 'All' is the number of all uEST cluster of this class with the relative amount of cluster that have an overlap with 'TSS' or 'Transcribed' in brackets. The chromatin elements have been computed for six different cell lines independently by Hoffmann et al. [76]. We took all cell lines together for this analysis.*

TYPE	TSS	Transcribed	All (%)
TIN	3,406	34,709	38,803 (92.6%)
UT	1,172	1,550	2,815 (83.8%)
DT	320	3,670	4,323 (87.7%)
3R	259	7,299	3,169 (85.5%)
IGR	733	2,664	11,207 (68.8%)

al., 2013 [76]. They developed computational methods for unsupervised chromatin state annotation. "The input datasets consist of ChIP-seq assays for multiple histone modifications, general transcription factors and chromatin accessibility assays." [76] Annotation for six different cell lines (GM12878, H1-hESC, K562, HeLa-S3, HepG2, HUVEC) are available. We took all cell lines together for this analysis. Every position in the genome was assigned to one of the following classes: Predicted promoter region including TSS (TSS), Predicted promoter flanking region (PF), Predicted enhancer (E), Predicted weak enhancer or open chromatin cis regulatory element (WE), CTCF enriched element (CTCF), Predicted transcribed region (T) and Predicted Repressed or Low Activity region (R).
We analysed several classes of unspliced EST cluster which should not be transcribed according to the current RefSeq annotation. More than 80% of the cluster close to or within annotated genes but outside of exons (TIN,UT,DT) are overlapping a segment predicted to be transcriptionally active ('TSS' or 'Transcribed') in at least one cell line. In theory this could be due to incorrect annotated exon borders or splice variants. Small mistakes in gene annotation are not relevant for unspliced EST cluster in intergenic region (IGR) which are at least 5,000nt away from the next annotated gene. In total 2,664 intergenic cluster are overlapping regions annotated to be transcriptionally active. 733 of them intersect predicted 'TSS'. Regions associated to TSS can also be found in unspliced EST cluster downstream of annotated genes. Similar to the previous section, these could be non-coding RNAs transcribed from within the 3'-UTR of protein-coding genes, so called uaRNAs, see Tab. 4.2 for all numbers.
Although we already knew that unspliced EST cluster are transcribed regions chromatin data can help to provide additional evidence for the significance of our findings.

4.3 Unspliced mRNAs

A track of unspliced mRNAs was created as described in section 3.1.1. 50,498 of 59,122 (35,430 of 49,665 in mouse) annotated 'unspliced mRNAs' are located within the extended RefSeq regions, of which 41,361 (16,303 in mouse) overlap RefSeq exons (including annotated unspliced RefSeq genes). We find that 51% (72% in mouse) of the unspliced mRNAs overlap with about 11% (23% in mouse) of the unspliced ESTs,

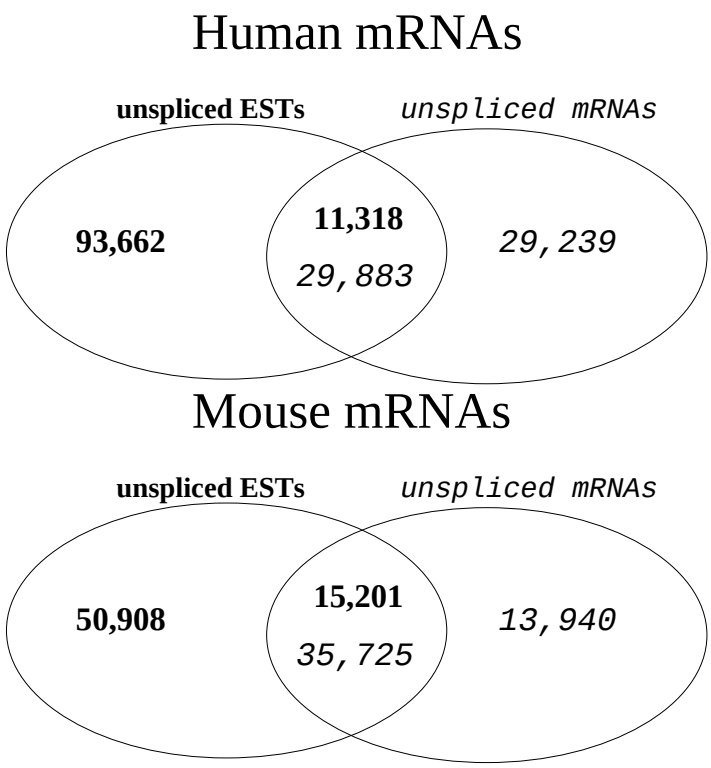


Figure 4.8: *Overlap of unspliced EST (uEST) cluster with annotated unspliced mRNAs in human and mouse. Bold numbers indicate the amount of uEST cluster while non-bold numbers indicate the amount of unspliced mRNAs. Numbers depicted in the overlap of both entities refer to uEST overlapping unspliced mRNAs and vice versa.*

Table 4.3: *Coverage of well-known long ncRNAs by unclustered unspliced ESTs (# uESTs). The first group is annotated as predominantly unspliced. The second group has annotated spliced isoforms. Coordinates taken by* `lncrna db`*[79].*

Gene	Chr.		approx.loc.		# uESTs	
Species	Human	Mouse	Human	Mouse	Human	Mouse
MALAT1	11	19	65.27mb	5.80mb	16,917	864
NEAT1	11	19	65.19mb	5.82mb	2,848	876
KCNQ1OT1	11	7	2.66mb	143.21mb	96	192
PTCSC3	14	-	36.60mb	-	10	-
XIST	X	X	73.04mb	103.46mb	1,341	314
TUG1	22	11	31.37mb	3.64mb	907	227
HOTAIR	12	15	54.36mb	102.94mb	22	3
AIR	-	17	-	12.74mb	-	218
ANRIL	9	-	21.99mb	-	120	-

see Fig. 4.8. Among these unspliced mRNAs there are several famous transcripts, cf. Tab. 4.3.

MALAT-1, for instance, appears among the loci with the highest coverage of unspliced ESTs. Together with NEAT1, which is located in the same genomic region, it belongs to a class of long nuclear retained transcripts involved in the organization of nuclear speckles [26, 80], see also [27] and the references therein. Other examples, such as KCNQ1OT1 [31] and PTCSC3 [81] are clearly visible in our data, albeit with moderate coverage.

Unspliced ESTs are also reported as parts of known spliced non-coding transcripts, in particular those with very long exons such as XIST [82]. In other cases, such as TUG1 [83], we observe predominantly unspliced ESTs that cover (nearly) the entire primary transcript, even though the genomic location is annotated by the spliced forms.

4.4 Unannotated intergenic EST clusters

11,207 unspliced EST cluster (6,945 in mouse) are located outside of spliced and unspliced RefSeq genes and their 5kb area. 4,985 of these overlap with already annotated mRNA, leaving 6,222 candidates (3,412 in mouse) for novel, predominantly unspliced genes.

Figure 4.9: *HNRNPR 3'UTR on chromosome 1 shown in reverse direction. Examples of "intergenic" unspliced EST cluster that are probably part of very long extensions of the 3'UTR are shown. Cluster of CAGE tags indicate transcription start sites of uaRNAs. The expression much beyond the end of the annotated 3'UTR is supported by RNAseq data from a plethora of cell lines (green bars with white arrows). Genome segmentation based on chromatin architecture annotates the region mostly as transcribed (green bar) as well with just smaller fragments in some cell lines being annotated as "Repressed or Low Activity region" (grey bar) or enhancer (orange bar).*

A manual inspection showed, however, that many of them are conspicuously well-conserved and can be identified as recent retropseudogenes [84] of known spliced genes. This concerns in particular the EST cluster with the largest numbers of ESTs. We therefore compared the sequences of the cluster against the "Retroposed Genes, Including Pseudogenes"-track from UCSC genome browser to determine the fraction of cluster that correspond to retropseudogenes. 760 loci (12%) of the candidates derive from protein-coding genes. In mouse the relative number is higher: 864 (25%) of the candidates. These unspliced EST cluster might be mapping artifacts and cannot be reliably distinguished from revived retrogenes [85].

The largest class among the remaining loci are probably long extensions of 3'UTRs of both coding and non-coding spliced transcripts. We suspect that in many of these cases we might in fact also see uaRNAs. In the examples shown in Fig. 4.9 this hypothesis is supported by large CAGE tag cluster at the 5' end of the 3'UTR. A more complex transcript organisation can be found in Fig. 4.10.

In order to further investigate the 5,462 intergenic unspliced EST cluster that were not recognized as likely retrogenes, we used `RNAz 2.1` [57, 86] to detect signatures of

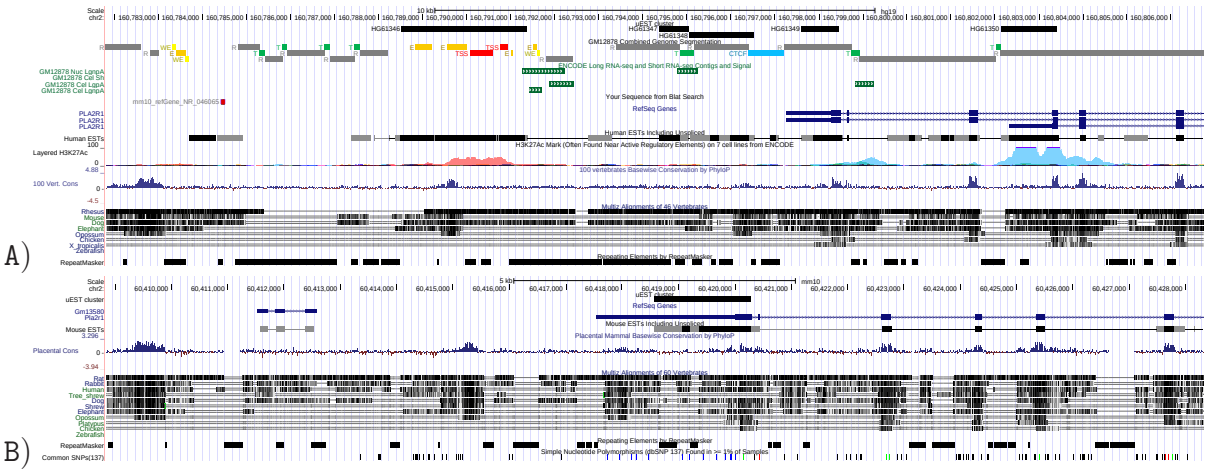


Figure 4.10: *A)PLA2R1 and downstream region on chromosome 2 in human is an example of a candidate for a new lncRNA. The importance of the unspliced EST (uEST) cluster HG61346 is supported by various evidence. Transcripts antinsense to the neighboring gene PLA2R1 have been detected by long RNAseq (green bars, white arrows depict the reading direction). The uEST overlaps with a region annotated as transcription start site by analysing the chromatin structure (red bar named TSS). In the middle of the uEST cluster there is a siginificant peak of evolutionary conservation visible (blue histogram). B) PLA2R1 and downstream region with lncRNA Gm13880 on chromosome 2 is the homologous region in mouse. In contrast to human there is an antisense and spliced long non-coding RNA annotated downstream of PLA2R1. A small part of its sequence (76nt) can be found in the human genome using BLAT. To know the exact structure of the lncRNA additional investigation, most likely by experimental methods is necessary.*

stabilizing selection on RNA secondary structure and `RNAcode` [59] to find evidence for conserved protein-coding regions. We used multiple genome alignments taken from the UCSC genome browser as input. This data set covers 4,059,185 nt and yields 1,104 RNAz hits with a classification probability $P_{RNAz} > 0.5$ and 371 with $P_{RNAz} > 0.9$. In mouse the 1,550,680 nt are covering 249 RNAz hits with a classification probability $P_{RNAz} > 0.5$ and 77 with $P_{RNAz} > 0.9$. Compared to the predicted 6880 low confidence and 2259 high confidence hits in the 30 Mb long ENCODE regions [86], the human result amounts to a very moderate enrichment by a factor of 1.2. It remains unclear whether this enrichment is confounded by the restriction of unspliced EST cluster to genomic regions with relatively high expression. It could hint a function of long unspliced regions in specific binding with proteins. An example for a conserved secondary structure element is shown in Fig. 4.11. Using `RNAcode`, we detected 632 regions with a p-value $p_{RNAcode} < 0.01$ in human. 232 of these regions have a length of at least 60 nucleotides. In mouse there are 652 regions with a p-value $p_{RNAcode} < 0.01$. 265 have a length of at least 60 nt. Shorter regions mostly appear to be artifacts caused by short segments of coding region in retrogenes.

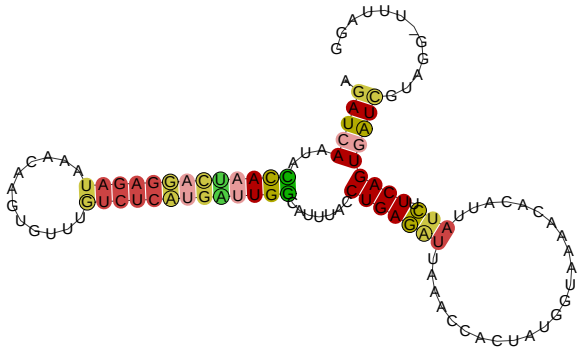
4.5 Conservation

The first publication on unspliced EST analysis [87] focused just on the human genome. The motivation for including mouse data is to get an evolutionary perspective. After detecting unspliced EST cluster in both, human and mouse, we are able to predict evolutionary conserved ones.

Unfortunately, for most parts of the data we could not find a homologous cluster. This is either the case if there is no homologous region detected by `liftOver`, see section 3.4 for more information, or if there is no unspliced EST cluster in the homolog region. Both cases sum up to 87% of human and 80% of the mouse unspliced EST cluster, see Fig. 4.12. There are 14,396 unspliced EST cluster conserved between human and mouse. Of the conserved cluster, 5,325 are classified in the same class, according to Fig. 4.3, 9,071 belong to a different class in both species, see Fig. 4.13.

Unsurprisingly the class Totally EXonic (TEX) comprises the most conserved cluster (2,464). One reason for that is the fact that detection of conservation between already known genes is much easier. The same goes for Totally INtronic (TIN) cluster, see

Conserved secondary structure on forward strand



Conserved secondary structure on reverse strand

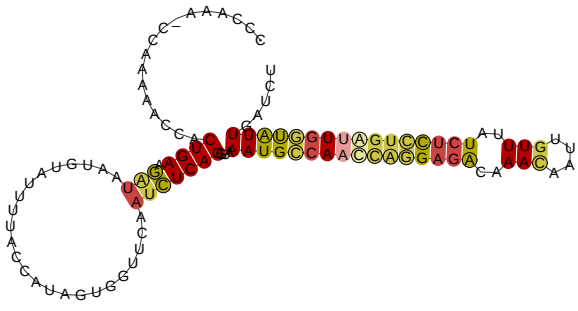


Figure 4.11: *Conserved secondary structure elements prioviding one of the best* `RNAz` *scores. The multiple genome alignment of the region hg19.chr9:66,766,331 66,766,440 contained sequences from diverse mammals, including Pteropus vampyrus and Loxodonta africana. The upper picture is the structure on the forward strand while the lower one is on the reverse strand. Both have a p-value better than* 0.99*. The color code, from red to ochre and green indicates that 1, 2, or 3 different types of basepairs are observed in the corresponding alignment columns, unsaturated colors indicate basepairs that cannot be formed by 1 or 2 of the 6 sequences in the alignment. Substitutions in stem regions are indicated by circles.*

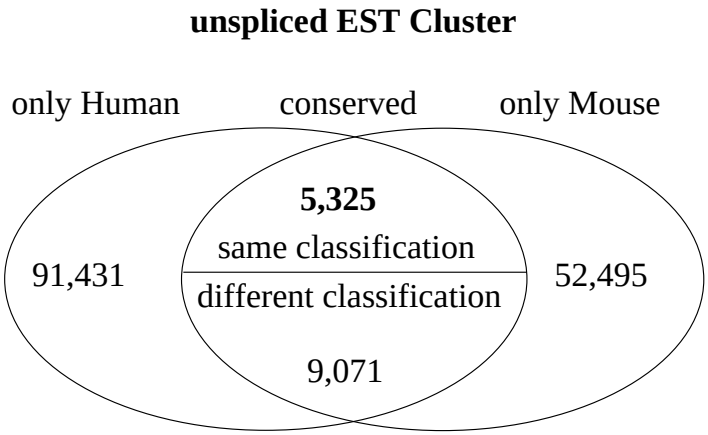


Figure 4.12: *Using the pairwise alignments of human and mouse we could detect 14,396 pairs of unspliced EST cluster which are conserved. The pairs consist of 13,278 different cluster from human and 13,277 from mouse. 5,325 pairs are between cluster which are classified in the same class, see Fig. 4.3 for details about classification.*
91,431 (87%) of 104,980 unspliced EST cluster in human and 52,495 (80%) of 66,109 in mouse can not be associated with an homologous unspliced EST cluster in the other species.

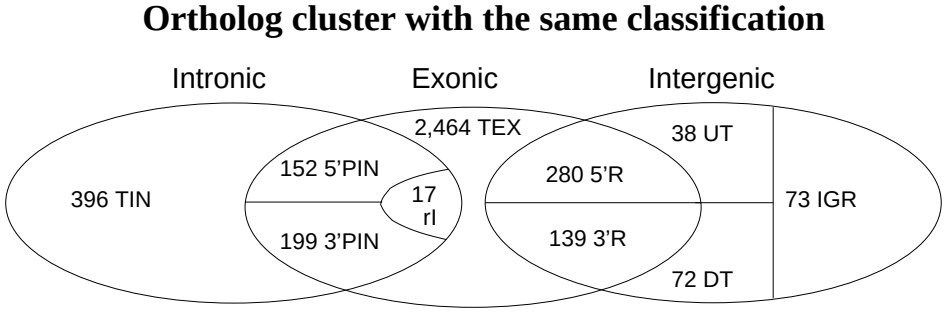


Figure 4.13: *Distribution of the classes in pairs between homologous cluster which are classified in the same class in human and mouse. 1,495 cluster are assigned to "NO_CLASS".*

Fig. 4.14 for an example. However, we also found 73 conserved intergenic cluster. One example can be seen in Fig. 4.15. The conservation between human and mouse implicitly hints to functional transcripts which have to be protected from harmful mutations.

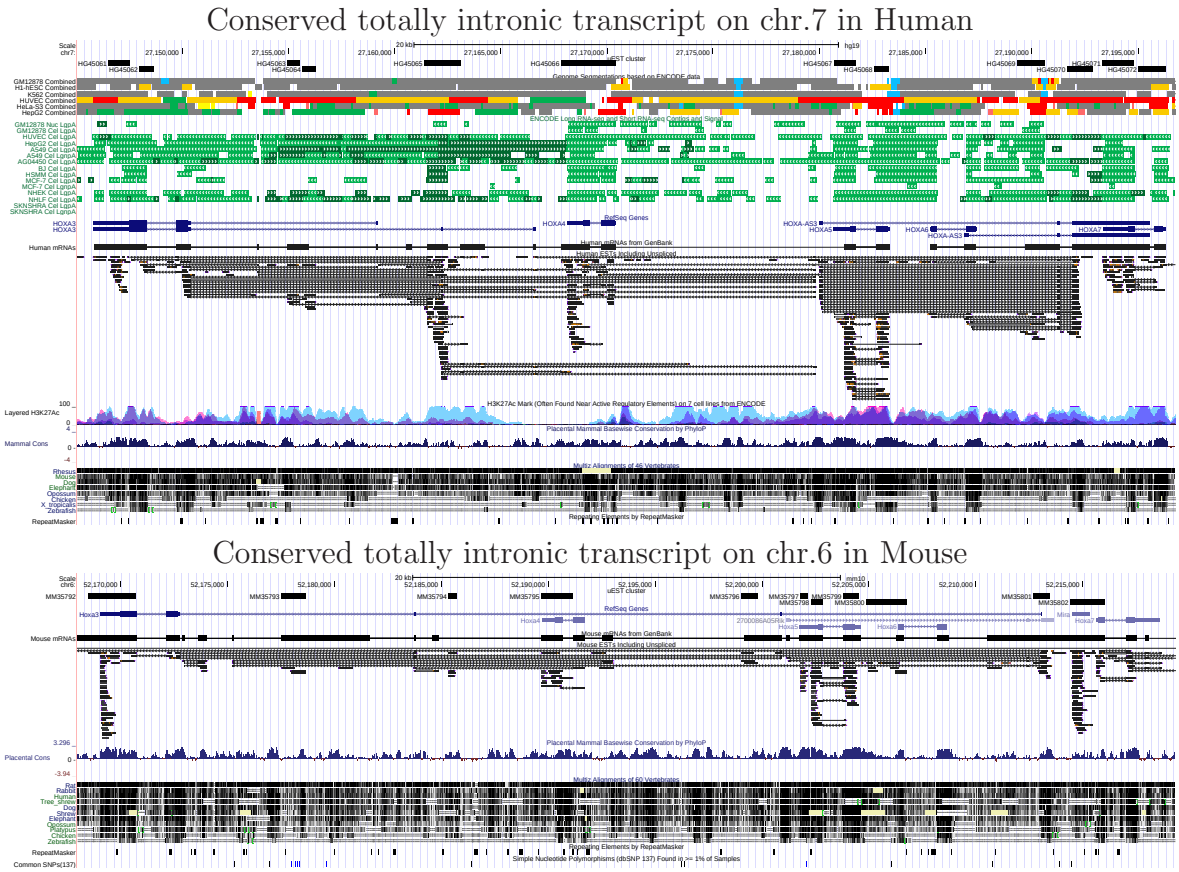


Figure 4.14: *Example of conserved totally intronic unspliced EST (uEST) cluster. The cluster HG45063 and HG45064 in human are homolog to MM35793 in mouse. As far as we know there is no independent transcript described for that region. The overlapping gene is HOXA3. The genomic loci is rather complex as one can see by the large number of differently connected ESTs and mRNAs. The uEST cluster HG45065 has been recently described as beeing an apoptosis repressor in certan cell lines and was called HOXA-AS2 [88]. The independence of HG45063/HG45064 can not be determined without additional experiments. It migt be a splice variant of HOXA-AS2 or even of HOXA-AS3. In fact they could turn out to be a single large antisense ncRNA, maybe depending on the used gene definition. Nevertheless, transcription is supported by chromatin marks (cell lines HeLa-S3 and HepG2) and antisense RNAseq data (see caption of Fig. 4.10).*

Figure 4.15: *A) Conserved intergenic unspliced EST cluster on chromosome 8 in human. The cluster HG102499 and HG102500 in human are homolog to MM42420 in mouse. The distrubtion of uESTs in the flanking region indicates a larger spliced transcript with at least three exons. In human there is additional evidence by chromatin marks and RNAseq data (see caption of Fig. 4.10).* B) *Conserved intergenic unspliced EST cluster on chromosome 4 in mouse is homologous to the ones in* A

Chapter 5

Discussion

5.1 Conclusion

Unspliced EST data, although typically disregarded in transcriptome analysis, can provide interesting insights into the structure of human transcriptomes. They outline a major component of transcriptional output that totally or at least partially escapes splicing. Nevertheless, this valuable resource has not been mined comprehensively in the past. So far, only the partially and totally intronic transcripts (PINs and TINs), which constitute more than half of the clusters of unspliced ESTs, have received attention in systematic studies [13, 14, 15]. Our analysis confirmed the conclusions drawn from the literature on these abundant class of transcripts in human and mouse. In addition, we find direct evidence of the independent transcription of PINs and TINs based on CAGE tag and chromatin data. Several promising candidates for previously unknown long RNAs could be found.

A large class of unspliced EST clusters forms extensions of the UTRs of well-annotated genes. On the 5'-side they provide additional information about the transcriptional start sites (TSS) of the annotated RefSeq genes themselves. Not surprisingly, the majority of clusters with clear support for a TSS from CAGE data falls into this class. More interestingly, however, the relative majority of UTR extensions is located at the 3'-end of RefSeq genes and forms typically extensive, several kb-long extensions. In line with the finding of [21], a lot of these 3'UTR extensions contain a TSS for an independently transcribed uaRNA. We identify at least 24 likely uaRNAs in human and 25 in mouse, with strong support from the CAGE data. Although TINs and PINs have been reported to be transcribed mostly independently of their surrounding gene, a small fraction of them has a sufficient concentration of CAGE tags for a recognizable TSS. This suggest that hundreds or even thousands of the 3'UTR extensions are EST clusters overlapping the 3'UTR of a RefSeq gene but originate rather from independent transcripts than a single 3'UTR.

The remaining set of unspliced EST clusters outside a 5kb range around RefSeq genes comprises 11% of the data. About a tenth (a forth in mouse) of these clusters overlaps retrogenes and retropseudogenes. For these cases, it is often impossible to distinguish between truly expressed loci and mapping artifacts of ESTs arising from the spliced original of the gene. The manual inspection of a random sample of the remaining cases shows that a large fraction of these EST clusters constitute even larger extensions of 3'UTRs. In many cases they also appear to be long 3'UTRs/uaRNAs belonging to previously unannotated, typically non-coding, spliced transcripts.
In summary, the analysis of unspliced ESTs uncovers a largely unexplored realm of long transcripts. The frequently postulated connection between lack of splicing and nuclear retention, see e.g. [89], and the surprising overlap of chromatin-associated transcripts suggests that this class of transcripts might be involved in chromatin organization and possibly other mechanisms of epigenetic control.
One might argue that the insights gained about the human genome could have been retrieved by using RNAseq transcriptome data as well. This might be true, indeed. However, there are already studies claiming that for complex genomic loci it works best to combine different transcriptome analysing methods like ESTs and RNAseq [90]. Another aspect is that whole transcriptome sequencing is still an expensive method and especially for large genomes not practical. Even for the very standard model organism Mus musculus there is less RNAseq data available compared to human. Especially plant genomes can reach a very large size. The Norway spruce for example is seven times larger than the human genome, bread wheat is still five times larger. This makes whole-genome sequencing much more expensive. In these cases EST sequencing can still provide an easer to reach first glimpse on expressed genes like it once did for the human genome.

5.2 Outlook

We compiled two genome-wide screens of unspliced ESTs. Analysing the whole genome necessarily means that one hast to focus on specific parts of the data just because of the sheer amount. Nevertheless, there is still much left to explore. Intronic RNAs have gained much attention recently [91, 92]. Using our data it is possible to find promising candidates for currently unknown functional RNA transcripts, like Fig. 4.10 and Fig. 4.15. This leads to the problem of validating our candidates. Additional experi-

mental verification would add striking evidence for the significance of our predictions. It would be a natural proceeding to hand on some of the most promising candidates for intronic or intergenic transcripts to experimental partners for validation.
In addition the computational possibilities are not exhausted yet. One could further analyse additional subsets of the data. Especially unspliced EST cluster conserved between human and mouse already contain implicit additional evidence for functional importance.
There are still possibilities left how to further analyse the created data sets. Furthermore using the introduced methods one could also redo the study for other organisms. As mentioned before whole-genome sequencing for plants is not an easy task. According to dbEST there are several plants for which more than one million ESTs are available: Zea mays (maize), Arabidopsis thaliana (thale cress), Glycine max (soybean), Triticum aestivum (wheat), Oryza sativa (rice). Long non-coding RNAs are studied mostly from a medical point of view. Therefore much less is known about this class of RNAs in plant. A study examining the mentioned genomes could help to get a completly new insight into unspliced transcripts of plants and their evolution in Eukarya in general.

Bibliography

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

[2] Michael B. Clark, Paulo P. Amaral, Felix J. Schlesinger, Marcel E. Dinger, Ryan J. Taft, John L. Rinn, Chris P. Ponting, Peter F. Stadler, Kevin J. Morris, Antonin Morillon, Joel S. Rozowsky, Mark Gerstein, Claes Wahlestedt, Yoshihide Hayashizaki, Piero Carninci, Thomas R. Gingeras, and John S. Mattick. The reality of pervasive transcription. *PLoS Biology*, 9:e1000625, 2011.

[3] M Guttman, M Garber, J Z Levin, J Donaghey, J Robinson, X Adiconis, L Fan, M J Koziol, A Gnirke, C Nusbaum, J L Rinn, E S Lander, and A Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat Biotechnol.*, 28:503–510, 2010.

[4] Michael Hiller, Sven Findeiß, Sandro Lein, Manja Marz, Claudia Nickel, Dominic Rose, Christine Schulz, Rolf Backofen, Sonja J. Prohaska, Gunter Reuter, and Peter F. Stadler. Conserved introns reveal novel transcripts in *Drosophila melanogaster*. *Genome Res.*, 19:1289–1300, 2009.

[5] R A Chodroff, L Goodstadt, T M Sirey, P L Oliver, K E Davies, E D Green, Z Molnár, and C P Ponting. Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol*, 11:R72, 2010.

[6] P Kapranov, G St Laurent, T Raz, F Ozsolak, C P Reynolds, P H Sorensen, G Reaman, P Milos, R J Arceci, J F Thompson, and T J Triche. The majority of total nuclear-encoded non-ribosomal RNA in a human cell is 'dark matter' un-annotated RNA. *BMC Biol.*, 8:149, 2010.

[7] K C Wang and H Y Chang. Molecular mechanisms of long noncoding RNAs. *Mol Cell*, 43:904–914, 2011.

[8] A Louhichi, A Fourati, and A Rebaï. IGD: a resource for intronless genes in the human genome. *Gene*, 488:35–40, 2011.

[9] A Köhler and E Hurt. Exporting RNA from the nucleus to the cytoplasm. *Nat. Rev. Mol. Cell Biol.*, 8:761–773, 2007.

[10] H Lei, A P Dias, and R Reed. Export and stability of naturally intronless mRNAs require specific coding region sequences and the TREX mRNA export complex. *Proc Natl Acad Sci USA*, 108:17985–17990, 2011.

[11] S A Shabalina, A Y Ogurtsov, A N Spiridonov, P S Novichkov, N A Spiridonov, and E V Koonin. Distinct patterns of expression and evolution of intronless and intron-containing mammalian genes. *Mol. Biol. Evol.*, 27:1745–1749, 2010.

[12] W F Marzluff, E J Wagner, and R J Duronio. Metabolism and regulation of canonical histone mRNAs: life without a poly(A) tail. *Nat Rev Genet.*, 9:843–854, 2008.

[13] H I Nakaya, P P Amaral, R Louro, A Lopes, A A Fachel, Y B Moreira, T A El-Jundi, A M da Silva, E M Reis, and S Verjovski-Almeida. Genome mapping and expression analyses of human intronic noncoding RNAs reveal tissue-specific patterns and enrichment in genes related to regulation of transcription. *Genome Biol.*, 8:R43, 2007.

[14] R Louro, H I Nakaya, P P Amaral, F Festa, M C Sogayar, A M da Silva, S Verjovski-Almeida, and E M Reis. Androgen responsive intronic non-coding RNAs. *BMC Biol.*, 5:4, 2007.

[15] R Louro, T El-Jundi, H I Nakaya, E M Reis, and S. Verjovski-Almeida. Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci. *Genomics*, 92:18–25, 2008.

[16] J L Rinn, G Euskirchen, P Bertone, R Martone, N M Luscombe, S Hartman, P M Harrison, F K Nelson, P Miller, M Gerstein, S Weissman, and M Snyder. The transcriptional activity of human chromosome 22. *Genes Dev.*, 17:529–540, 2003.

[17] E M Reis, H I Nakaya, R Louro, F C Canavez, A V Flatschart, G T Almeida, C M Egidio, A C Paquola, A A Machado, F Festa, D Yamamoto, R Alvarenga, C C da Silva, G C Brito, S D Simon, C A Moreira-Filho, K R Leite, L H Camara-Lopes, F S Campos, E Gimba, G M Vignal, H El-Dorry, M C Sogayar, M A Barcinski, A M da Silva, and S Verjovski-Almeida. Antisense intronic non-coding RNA levels correlate to the degree of tumor differentiation in prostate cancer. *Oncogene*, 23:6684–6692, 2004.

[18] R Louro, A S Smirnova, and S. Verjovski-Almeida. Long intronic noncoding RNA transcription: expression noise or expression choice? *Genomics*, 93:291–298, 2009.

[19] A C Tahira, M S Kubrusly, M F Faria, B Dazzani, R S Fonseca, V Maracaja-Coutinho, S Verjovski-Almeida, M C Machado, and E M Reis. Long noncoding intronic RNAs are differentially expressed in primary and metastatic pancreatic cancer. *Mol Cancer*, 10:141, 2011.

[20] K J Nordström, M A Mirza, M S Almén, D E Gloriam, R Fredriksson, and H B Schiöth. Critical evaluation of the FANTOM3 non-coding RNA transcripts. *Genomics*, 94:169–176, 2009.

[21] Tim R. Mercer, Dagmar Wilhelm, Marcel E. Dinger, Giulia Soldà, Darren J. Korbie, Evgeny A. Glazov, Vy Truong, Maren Schwenke, Cas Simons, Klaus I. Matthaei, Robert Saint, Peter Koopman, and John S. Mattick. Expression of distinct RNAs from 3' untranslated regions. *Nucleic Acids Res.*, 2393-2403:39, 2011.

[22] Farshad Niazi and Saba Valadkhan. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3'utrs. *RNA*, 2012. doi: 10.1261/rna.029520.111.

[23] Yasnory T F Sasaki, Takashi Ideue, Miho Sano, Toutai Mituyama, and Tetsuro Hirose. MENϵ/β noncoding RNAs are essential for structural integrity of nuclear paraspeckles. *Proc Natl Acad Sci USA*, 106:2525–2530, 2009.

[24] Hongjae Sunwoo, Marcel E. Dinger, Jeremy E. Wilusz, Paulo P. Amaral, John S. Mattick, and David L. Spector. MEN ϵ/β nuclear-retained non-coding RNAs are up-regulated upon muscle differentiation and are essential components of paraspeckles. *Genome Res.*, 19:347–359, 2009.

[25] Y S Mao, H Sunwoo, B Zhang, and D L Spector. Direct visualization of the co-transcriptional assembly of a nuclear body by noncoding RNAs. *Nat. Cell Biol.*, 13:95–101, 2011.

[26] John Hutchinson, Alexander W Ensminger, Christine M Clemson, Christopher R Lynch, Jeanne B Lawrence, and Andrew Chess. A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics*, 8:39, 2007.

[27] P. F. Stadler. Evolution of the long non-coding RNAs MALAT1 and MENβ/ϵ. In Carlos Eduardo Ferreira, Satoru Miyano, and Peter F. Stadler, editors, *Advances in Bioinformatics and Computational Biology, 5th Brazilian Symposium on Bioinformatics*, volume 6268 of *Lecture Notes in Computer Science*, pages 1–12, Heidelberg, 2010. Springer Verlag.

[28] S Chung, H Nakagawa, M Uemura, L Piao, K Ashikawa, N Hosono, R Takata, S Akamatsu, T Kawaguchi, T Morizono, T Tsunoda, Y Daigo, K Matsuda, N Kamatani, Y Nakamura, and M. Kubo. Association of a novel long non-coding RNA in 8q24 with prostate cancer susceptibility. *Cancer Sci.*, 102:245–252, 2011.

[29] C I M Seidl, S H Stricker, and D P Barlow. The imprinted Air ncRNA is an atypical RNAPII transcript that evades splicing and escapes nuclear export. *EMBO J*, 25:1–11, 2006.

[30] S H Stricker, L Steenpass, F M Pauler, F Santoro, P A Latos, R Huang, M V Koerner, M A Sloane, K E Warczok, and D P Barlow. Silencing and transcriptional properties of the imprinted Airn ncRNA are independent of the endogenous promoter. *EMBO J.*, 27:3116–3128, 2008.

[31] L Redrup, M R Branco, E R Perdeaux, C Krueger, A Lewis, F Santos, T Nagano, B S Cobb, P Fraser, and W. Reik. The long noncoding RNA Kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development*, 136:525–530, 2009.

[32] J Whitehead, G K Pandey, and C Kanduri. Regulation of the mammalian epigenome by long noncoding RNAs. *Biochim Biophys Acta*, 1790:936–947, 2009.

[33] Izhak J. Paul and Jacob D. Duerksen. Chromatin-associated RNA content of heterochromatin and euchromatin. *Mol. Cell. Biochem.*, 9:9–16, 1975.

[34] Antonio Rodríguez-Campos and Fernando Azorín. RNA is an integral component of chromatin that contributes to its structural organization. *PLoS ONE*, 2:e1182, 2007.

[35] T Mondal, M Rasmussen, G K Pandey, A Isaksson, and C. Kanduri. Characterization of the RNA content of chromatin. *Genome Res.*, 20:899–907, 2010.

[36] P Kapranov, J Cheng, S. Dike, D Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermüller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, G. Madhavan, A Piccolboni, V Sementchenko, H. Tammana, and T. R. Gingeras. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488, 2007.

[37] Xiangting Wang, Shigeki Arai, Xiaoyuan Song, Donna Reichart, Kun Du, Gabriel Pascual, Paul Tempst, Michael G. Rosenfeld, Christopher K. Glass, and Riki Kurokawa. Induced ncRNAs allosterically modify RNA-binding proteins *in cis* to inhibit transcription. *Nature*, 454:126–130, 2008.

[38] JD Watson and FHC Crick. Molecular structure of nucleic acids: A structure for deoxyribose nucleic acid. *Nature*, 171:737–738, 1953.

[39] M. Meselson and F.W. Stahl. The replication of dna in e. coli. *Proc Nat Aca Sci USA*, 44:671–682, 1958.

[40] M Nirenberg, P Leder, M Bernfield, R Brimacombe, J Trupin, F Rottman, and C O'Neal. Rna codewords and protein synthesis, vii. on the general nature of the rna code. *Proc Nat Aca Sci USA*, 53:1161–1168, 1965.

[41] F Sanger, GM Air, BG Barrell, NL Brown, AR Coulson, CA Fiddes, CA Hutchison, PM Slocombe, and M Smith. Nucleotide sequence of bacteriophage phi x174 dna. *Nature*, 265:687–95, 1977.

[42] JC et al Venter. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[43] International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001.

[44] Mark B. Gerstein, Can Bruce, Joel S. Rozowsky, Deyou Zheng, Jiang Du, Jan O. Korbel, Olof Emanuelsson, Zhengdong D. Zhang, Sherman Weissman, and Michael Snyder. What is a gene, post-encode? history and updated definition. *Genome Research*, 17(6):669–681, 2007.

[45] Wutz A, Smrzka OW, Schweifer N, Schellander K, Wagner EF, and Barlow DP. Imprinted expression of the igf2r gene depends on an intronic cpg island. *Nature*, 389:745–9, 1997.

[46] CJ Brown, A Ballabio, JL Rupert, RG Lafreniere, M Grompe, R Tonlorenzi, and HF Willard. A gene from the region of the human x inactivation centre is expressed exclusively from the inactive x chromosome. *Nature*, 349:38–44, 1991.

[47] MD Adams, JM Kelley, JD Gocayne, M Dubnick, MH Polymeropoulos, H Xiao, CR Merril, A Wu, B Olde, RF Moreno, and al. et. Complementary dna sequencing: expressed sequence tags and human genome project. *Science*, 252(5013):1651–1656, 1991.

[48] Shivashankar H. Nagaraj, Robin B. Gasser, and Shoba Ranganathan. A hitchhiker's guide to expressed sequence tag (est) analysis. *Briefings in Bioinformatics*, 8(1):6–21, 2007.

[49] J S Aaronson, B Eckman, R A Blevins, J A Borkowski, J Myerson, S Imran, and K O Elliston. Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput est sequence data. *Genome Research*, 6(9):829–845, 1996.

[50] T Shiraki, S Kondo, S Katayama, K Waki, T Kasukawa, H Kawaji, R Kodzius, A Watahiki, M Nakamura, T Arakawa, S Fukuda, D Sasaki, A Podhajska, M Harbers, J Kawai, P Carninci, and Y Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc Natl Acad Sci U S A*, 100(26):15776–81, Dec 2003.

[51] M Harbers and P Carninci. Tag-based approaches for transcriptome research and genome annotation. *Nat Methods*, 2(7):495–502, Jul 2005.

[52] W. James Kent, Charles W. Sugnet, Terrence S. Furey, Krishna M. Roskin, Tom H. Pringle, Alan M. Zahler, and David Haussler. The human genome browser at ucsc. *Genome Research*, 12(6):996–1006, 2002.

[53] Donna Karolchik, Galt P. Barber, Jonathan Casper, Hiram Clawson, Melissa S. Cline, Mark Diekhans, Timothy R. Dreszer, Pauline A. Fujita, Luvina Guruvadoo, Maximilian Haeussler, Rachel A. Harte, Steve Heitner, Angie S. Hinrichs, Katrina Learned, Brian T. Lee, Chin H. Li, Brian J. Raney, Brooke Rhead, Kate R. Rosenbloom, Cricket A. Sloan, Matthew L. Speir, Ann S. Zweig, David Haussler, Robert M. Kuhn, and W. James Kent. The ucsc genome browser database: 2014 update. *Nucleic Acids Research*, 42(D1):D764–D770, 2014.

[54] Brian J. Raney, Timothy R. Dreszer, Galt P. Barber, Hiram Clawson, Pauline A. Fujita, Ting Wang, Ngan Nguyen, Benedict Paten, Ann S. Zweig, Donna Karolchik, and W. James Kent. Track data hubs enable visualization of user-defined genome-wide annotations on the ucsc genome browser. *Bioinformatics*, 2013.

[55] D et al Adams. Blueprint to decode the epigenetic signature written in blood. *Nature Biotechnology*, 30:224–226, 2012.

[56] MS Boguski, TM Lowe, and CM Tolstoshev. dbest database for "expressed sequence tags". *Nature Genetics*, 4:332–333, 1993.

[57] Stefan Washietl, Ivo L. Hofacker, and Peter F. Stadler. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA*, 102:2454–2459, 2005.

[58] A R Gruber, R Neuböck, I L Hofacker, and S. Washietl. The RNAz web server: prediction of thermodynamically stable and evolutionarily conserved RNA structures. *Nucleic Acids Res.*, 35:W335–W338, 2007.

[59] Stefan Washietl, Sven Findeiß, Stephan Müller, Stefan Kalkhof, Martin von Bergen, Ivo L. Hofacker, Peter F. Stadler, and Nick Goldman. RNAcode: robust prediction of protein coding regions in comparative genomics data. *RNA*, 17:578–594, 2011.

[60] Donna Karolchik, Angela S. Hinrichs, Terrence S. Furey, Krishna M. Roskin, Charles W. Sugnet, David Haussler, and W. James Kent. The ucsc table browser data retrieval tool. *Nucleic Acids Research*, 32(suppl 1):D493–D496, 2004.

[61] J D Hawkins. A survey on intron and exon lengths. *Nucleic Acids Res.*, 16:9893–9908, 1988.

[62] Xin Hong, Douglas G. Scofield, and Michael Lynch. Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol*, 23:2392–2404, 2006.

[63] R Horton, R Gibson, P Coggill, M Miretti, R J Allcock, J Almeida, S Forbes, J G Gilbert, K Halls, J L Harrow, E Hart, K Howe, D K Jackson, S Palmer, A N Roberts, S Sims, C A Stewart, J A Traherne, S Trevanion, L Wilming, J Rogers, P J de Jong, J F Elliott, S Sawcer, J A Todd, J Trowsdale, and S Beck. Variation analysis and gene annotation of eight mhc haplotypes: the mhc haplotype project. *Immunogenetics*, 60(1):1–18, Jan 2008.

[64] Einat Hazkani-Covo, Raymond Zeller, and William Martin. Molecular poltergeists: Mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS Genet.*, 6:e1000834, 2010.

[65] J Asin-Cayuela and C M Gustafsson. Mitochondrial transcription and its regulation in mammalian cells. *Trends Biochem Sci.*, 32:111–117, 2007.

[66] J Tsuji, M C Frith, K Tomii, and P Horton. Mammalian numt insertion is non-random. *Nucleic Acids Res*, 40(18):9073–88, Oct 2012.

[67] K. D. Pruitt, T. Tatusova, and D. R. Maglott. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, 35:D61–D65, 2007.

[68] H Kawaji, J Severin, M Lizio, A Waterhouse, S Katayama, K M Irvine, D A Hume, A R Forrest, H Suzuki, P Carninci, Y Hayashizaki, and C O Daub. The fantom web resource: from mammalian transcriptional landscape to its dynamic regulation. *Genome Biol*, 10(4):R40, 2009.

[69] H Kawaji, J Severin, M Lizio, A R Forrest, E van Nimwegen, M Rehli, K Schroder, K Irvine, H Suzuki, P Carninci, Y Hayashizaki, and C O Daub. Update of the fantom web resource: from mammalian transcriptional landscape to its dynamic regulation. *Nucleic Acids Res*, 39(Database issue):D856–60, Jan 2011.

[70] JW Neidigh, RM Fesinmeyer, and Andersen NH. Designing a 20-residue protein. *Nature Structural Biology*, 9:425–430, 2002.

[71] Masaaki Furuno, Ken C. Pang, Noriko Ninomiya, Shiro Fukuda, Martin C. Frith, Carol Bult, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, John S. Mattick, and Harukazu Suzuki. Clusters of internally primed transcripts reveal novel long noncoding RNAs. *PLoS Genetics*, 2:e37, 2006.

[72] J Goecks, A Nekrutenko, J Taylor, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11:R86, 2010.

[73] Aaron R. Quinlan and Ira M. Hall. Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, 26(6):841–842, 2010.

[74] F Chiaromonte, VB Yap, and W Miller. Scoring pairwise genomic sequence alignments, 2002. In *Pac. Symp. Biocomput*, pages 115–26, 2002.

[75] Scott Schwartz, W. James Kent, Arian Smit, Zheng Zhang, Robert Baertsch, Ross C. Hardison, David Haussler, and Webb Miller. Humanmouse alignments with blastz. *Genome Research*, 13(1):103–107, 2003.

[76] Michael M. Hoffman, Jason Ernst, Steven P. Wilder, Anshul Kundaje, Robert S. Harris, Max Libbrecht, Belinda Giardine, Paul M. Ellenbogen, Jeffrey A. Bilmes, Ewan Birney, Ross C. Hardison, Ian Dunham, Manolis Kellis, and William Stafford Noble. Integrative annotation of chromatin elements from encode data. *Nucleic Acids Research*, 41(2):827–841, 2013.

[77] Lieven Thorrez, Leon-Charles Tranchevent, Hui Ju Chan, Yves Moreau, and Frans Schuit. Detection of novel 3' untranslated region extensions with 3' expression microarrays. *BMC Genomics*, 11:205, 2010.

[78] T R Mercer, M E Dinger, C P Bracken, G Kolle, J M Szubert, D J Korbie, M E Askarian-Amiri, B B Gardiner, G J Goodall, S M Grimmond, and J S Mattick. Regulated post-transcriptional RNA cleavage diversifies the eukaryotic transcriptome. *Genome Res.*, 20:1639–1650, 2010.

[79] P P Amaral, M B Clark, D K Gascoigne, M E Dinger, and J S Mattick. lncrnadb: a reference database for long noncoding rnas. *Nucleic Acids Res*, 39(Database issue):146–151, Jan 2011.

[80] R Lin, M Roychowdhury-Saha, C Black, A T Watt, E G Marcusson, S M Freier, and T S Edgington. Control of RNA processing by a large noncoding RNA over-expressed in carcinomas. *FEBS Lett.*, 585:671–671, 2011.

[81] H He, R Nagy, S Liyanarachchi, H Jiao, W Li, S Suster, J Kere, and A de la Chapelle. A susceptibility locus for papillary thyroid carcinoma on chromosome 8q24. *Cancer Res.*, 69:625–631, 2009.

[82] E A Elisaphenko, N N Kolesnikov, A I Shevchenko, I B Rogozin, T B Nesterova, N Brockdorff, and S M Zakian. A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. *PLoS One*, 3:e2521, 2008.

[83] T. L. Young, T. Matsuda, and C. L. Cepko. The noncoding RNA taurine upregulated gene 1 is required for differentiation of the murine retina. *Current Biology*, 15:501–512, 2005.

[84] E J Devor and K A Moffat-Wilson. Molecular and temporal characteristics of human retropseudogenes. *Hum. Biol.*, 75:661–672, 2003.

[85] Z Yu, M Morais, D Ivanga, and P M Harrison. Analysis of the role of retrotransposition in gene evolution in vertebrates. *BMC Bioinformatics*, 8:308, 2007.

[86] Andreas R. Gruber, Sven Findeiß, Stefan Washietl, Ivo L. Hofacker, and Peter F. Stadler. `RNAz 2.0`: improved noncoding RNA detection. *Pac. Symp. Biocomput.*, 15:69–79, 2010.

[87] J Engelhardt and P F Stadler. Hidden treasures in unspliced est data. *Theory Biosci*, 131(1):49–57, May 2012.

[88] Hang Zhao, Xueqing Zhang, Josias Brito Frazo, Antonio Condino-Neto, and Peter E. Newburger. Hox antisense lincrna hoxa-as2 is an apoptosis repressor in all trans retinoic acid treated nb4 promyelocytic leukemia cells. *Journal of Cellular Biochemistry*, 114(10):2375–2383, 2013.

[89] A B Eberle, V Hessle, R Helbig, W Dantoft, N Gimber, and N. Visa. Splice-site mutations cause rrp6-mediated nuclear retention of the unspliced RNAs and transcriptional down-regulation of the splicing-defective genes. *PLoS ONE*, 5:e11540, 2010.

[90] N. Schurch, C. Cole, A. Sherstnev, J. Song, C. Duc, K. G. Storey, W. H. Irwin McLean, S. J. Brown, G. G. Simpson, and G. J. Barton. Improved annotation of 3-prime untranslated regions and complex loci by combination of strand-specific Direct RNA Sequencing, RNA-seq and ESTs. *ArXiv e-prints*, November 2013.

[91] Felipe C. Beckedorff, Ana C. Ayupe, Renan Crocci-Souza, Murilo S. Amaral, Helder I. Nakaya, Daniela T. Soltys, Carlos F. M. Menck, Eduardo M. Reis, and Sergio Verjovski-Almeida. The intronic long noncoding rna ¡italic¿anrassf1¡/italic¿ recruits prc2 to the ¡italic¿rassf1a¡/italic¿ promoter, reducing the expression of ¡italic¿rassf1a¡/italic¿ and increasing cell proliferation. *PLoS Genet*, 9(8):e1003705, 08 2013.

[92] Angela Fachel, Ana Tahira, Santiago Vilella-Arias, Vinicius Maracaja-Coutinho, Etel Gimba, Giselle Vignal, Franz Campos, Eduardo Reis, and Sergio Verjovski-Almeida. Expression analysis and in silico characterization of intronic long noncoding rnas in renal cell carcinoma: emerging functional associations. *Molecular Cancer*, 12(1):140, 2013.

Statement of Authorship

I hereby affirm that I have written this thesis independently and only used the sources and aids stated in the bibliography.

Leipzig, January 24, 2014

Signature