Tracing the evolution of long non-coding RNAs

Principles of comparative transcriptomics for splice site conservation and biological applications

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

Dissertation

zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)

im Fachgebiet
Informatik

vorgelegt von

Diplom-Informatikerin Anne Nitsche,
geboren am 17. Juni 1984 in Hoyerswerda

Die Annahme der Dissertation wurde empfohlen von:
1. Prof. Dr. Peter F. Stadler, Universität Leipzig
2. Prof. Dr. Dmitrij Frishman, Technische Universität München

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 15.12.2017 mit dem Gesamtprädikat *magna cum laude*.

Abstract

Eukaryotic cells exhibit an extensive transcriptional diversity. Only about a quarter of the total RNA in the human cell can be accounted for by messenger RNA (mRNA), which convey genetic code for protein generation. The remaining part of the transcriptome consists of rather heterogenous molecules. While some classes are well defined and have been shown to carry out distinct functions, ranging from housekeeping to complex regulatory tasks, a big fraction of the transcriptional output is categorized solely based on the lack of protein-coding capacity and transcript length. Several studies have shown, that as a group, mRNA-like long non-coding RNAs (lncRNAs), are under stabilizing selection, however at much weaker levels than mRNAs. The conservation at the level of primary sequence is even lower, blurring the contrast between exonic and intronics parts, which impedes traditional methods of genome-wide homology search. As a consequence their evolutionary history is a fairly unexplored field and apart from a few experimentally studied cases, the vast majority of them is reported to be poorly conserved. However, the pervasive transcription and the highly spatio-temporal specific expression patterns of lncRNAs suggests their functional importance and makes their evolutionary age and conservation patterns a topic of interest. By employing diverse computational methods, recent studies shed light on the common conservation of lncRNA's secondary and gene structures, highlighting the significance of structural features on functionality. Splice sites, in particular, are frequently retained over very large evolutionary time scales, as they maintain the intron-exon-structure of the transcript.

Consequently, the conservation of splice sites can be utilized in a comparative genomics approach to establish homology and predict evolutionarily well-conserved transcripts, regardless of their coding capacity. Since splice site conservation cannot be directly inferred from experimental evidence, in the course of this thesis a computational pipeline was established to generate comparative maps of splice sites

based on multiple sequence alignments together with transcriptomics data. Scoring schemes for splice site motifs are employed to assess the conservation of orthologs. This resource can then be used to systemically study the conservation patterns of RNAs and their gene structures. This thesis will demonstrate the versatility of this method by showcasing biological applications of three distinct studies.

First, a comprehensive annotation of the human transcriptome, from RefSeq, ESTs and GENCODE, was used to trace the evolution of human lncRNAs. A large majority of human lncRNAs is found to be conserved across Eutheria, and many hundreds originated before the divergence of marsupials and placental mammals. However, they exhibit a rapid turnover of their transcript structures, indicating that they are actual ancient components of the vertebrate genome with outstanding evolutionary plasticity. Additionally, a public web server was setup, which allows the user to retrieve sets of orthologous splice sites from pre-computed comparative splice site maps and inspect visualizations of their conservation in the respective species.

Second, a more specific data set of non-colinearly spliced latimerian RNAs is studied to fathom the origins of atypical transcripts. RNA-seq data from two coelacanth species are analyzed, yielding thousands of circular and trans-spliced products, with a surprising exclusivity of the majority of their splice junctions to atypically spliced forms, that is they are not used in linear isoforms. The conservation analysis with comparative splice site maps yielded high conservation levels for both circularizing and trans-connecting splice sites. This fact in combination with their abundance strongly suggests that atypical RNAs are evolutionarily old and of functional importance.

Lastly, comparative splice site maps are used to investigate the role of lncRNAs in the evolution of the Alzheimer's disease (AD). The human specificity of AD clearly points out a phylogenetic aspect of the disease, which makes the evolutionary analysis a very promising field of research. Protein-coding and non-protein-coding regions, that have been identified to be differentially expressed in AD patients, are analyzed for conservation of their splice site and evolution of their exon-intron-structure. Both non-coding and protein-coding AD-associated genes are shown to have evolved more rapidly in their gene structure than the genome at large. This supports the view of AD as a consequence of the recent rapid adaptive evolution of the human brain. This phylogenetic trait might have far reaching consequences with respect to the appropriateness of animal models and the development of disease-modifying strategies.

Zusammenfassung

Eukaryotische Zellen legen eine umfangreiche transkriptionelle Vielfalt an den Tag. Nur etwa ein Viertel der in der menschlichen Zelle enthaltenen RNA ist messenger RNA (mRNA), welche den genetischen Code für die Proteingenerierung übermittelt. Der verbleibende Anteil des Transkriptoms besteht aus eher heterogenen Molekülen. Während einigen wohldefinierten Klassen spezifische Funktionen zugeordnet werden können, welche von Zellhaushalt bis zu komplexen regulatorischen Aufgaben reichen, wird ein großer Teil der transkriptionellen Produktion ausschließlich auf Grundlage der fehlenden Kodierungskapazität und der Transkriptlänge kategorisiert. Einige Studien zeigten, dass mRNA-ähnliche lange nicht-kodierende RNA (lncRNA) als Gruppe unter stabilisierender Selektion stehen, wenn auch in einem weitaus geringeren Ausmaß als mRNAs. Die Konservierung auf Ebene der primären Sequenz ist sogar noch niedriger, wodurch der Kontrast zwischen exonischen und intronischen Elementen verschwimmt und Methoden der traditionellen Homologiesuche erschwert werden. Infolgedessen ist die evolutionäre Geschichte der lncRNAs ein recht unerforschtes Gebiet und abgesehen von ein paar vereinzelten Fallstudien wird die große Mehrheit als schwach konserviert vermeldet. Die tiefgreifende Transkription und die in Raum und Zeit hochspezifischen Expressionsmuster von lncRNA deuten jedoch auf deren funktionelle Bedeutung hin und machen ihr evolutionäres Alter und ihre Konservierungsmuster zu einem Thema von Interesse. Durch die Verwendung von computergestützten Methoden konnten jüngste Studien die verbreitete Konservierung von Sekundär- und Genstruktur von lncRNAs aufzeigen, was die Signifikanz von strukturellen Merkmalen in Bezug auf deren Funktionalität unterstreicht. Spleißstellen im besonderen werden oft über lange evolutionäre Zeitspannen erhalten, da sie die Intron-Exon-Struktur des Transkripts bewahren.

Folglich, kann die Konservierung von Spleißstellen durch einen Ansatz der ver-

gleichenden Genomik benutzt werden, um Homologie herzuleiten und evolutionär gut konservierte Transkripte unabhängig von deren Kodierungskapazität zu prognostizieren. Da es nicht möglich ist die Spleißstellenkonservierung direkt anhand von experimentellen Indikatoren abzulesen, wurde im Zuge dieser These eine computergestützte Methode entwickelt, welche, basierend auf multiplen Sequenzalignments und Transkriptomikdaten, "Vergleichskarten" von Spleißstellen erstellt. Ein Punktebewertungssystem für Spleißstellenmotive wird benutzt um die Konservierung der Orthologen zu beurteilen. Diese Resource kann anschließend verwendet werden um systematisch die Konservierungsmuster von RNAs und deren Genstrukturen zu untersuchen. Diese Arbeit wird die Vielseitigkeit dieser Methode demonstrieren, indem die biologische Anwendung in drei verschiedenen Studien präsentiert wird.

Zuerst wird eine umfassende Annotation des menschlichen Transkriptoms, basierend auf RefSeq, EST und GENCODE, benutzt, um die Evolution von humanen lncRNAs nachzuvollziehen. Es konnte festgestellt werden, dass eine große Mehrheit der menschlichen lncRNAs innerhalb der Eutheria konserviert ist und mehrere hundert bereits vor der Auseinanderentwicklung von Beuteltieren und höheren Säugetieren entstanden. Dennoch zeigen sie eine rasante Veränderung in ihren Transkriptstrukturen, welche darauf hindeutet, dass sie tatsächlich alte Bestandteile von Vertebratengenomen mit bemerkenswerter evolutionärer Formbarkeit sind. Zusätzlich wurde ein öffentlicher Webserver aufgesetzt, der dem Nutzer ermöglicht Datensätze orthologer Spleißstellen aus vorgenerierten Vergleichskarten zu extrahieren und Visualisierungen der Konservierung in den jeweiligen Spezies zu betrachten.

Als zweites wird ein spezifischerer Datensatz von nicht-linear gespleißten Latimeria-RNA untersucht um die Ursprünge untypischer Transkripte zu ergründen. Die Analyse der RNA-seq Daten zweier Exemplare des Quastenflossers ergab tausende zirkulärer und Transspleiß-Produkte, wobei die Mehrheit der Spleißverbindungen eine überraschende Exklusivität für untypisch gespleißte Formen aufzeigt, d.h. diese werden nicht für lineare Isoformen genutzt. Die Konservierungsanalyse mit Spleißstellen-Vergleichskarten ergibt hohe Konservierungsniveaus sowohl für zirkulärisierende als auch für trans-verbindende Spleißstellen. Diese Tatsache in Kombination mit ihrem häufigen Vorkommen, deutet stark darauf hin, dass untypische RNAs evolutionär alt und von funktioneller Bedeutung sind.

Zuletzt werden Spleißstellen-Vergleichskarten benutzt um die Rolle von lncRNAs in der Evolution der Alzheimer-Krankheit (AK) zu untersuchen. Die Spezifität der AK auf den Menschen weist klar auf einen phylogenetischen Aspekt der Krankheit hin, was deren evolutionäre Analyse zu einem vielversprechenden Forschungsgebiet macht. Proteinkodierende und nicht-proteinkodierende Regionen, bei denen eine differentielle Expression in AK-Patienten erkannt wurde, werden auf die Konservierung

ihrer Spleißstellen und Evolution ihrer Exon-Intron-Strukturen hin analysiert. Es kann nachgewiesen werden, dass sich die Genstruktur von sowohl nicht-kodierenden als auch von proteinkodierenden AK-assoziierten Genen schneller entwickelt als das Genom im Allgemeinen. Das unterstützt die Auffassung, dass AK die Folge einer kürzlichen rasanten adaptiven Evolution des menschlichen Gehirns ist. Diese phylogenetische Eigenschaft könnte weitreichende Konsequenzen in Bezug auf die Angemessenheit von Tiermodellen und die Entwicklung von krankheitsmodifizierenden Strategien haben.

Danksagung

Zuerst möchte ich mich ganz besonders bei meinem Supervisor Peter bedanken. Danke, dass du mir diese Arbeit ermöglicht hast. Danke für das Beantworten unzähliger E-Mails mit einer spektakulären Reaktionszeit, sogar spät in der Nacht.

Vielen Dank, Steve, dass ich Teil deiner Gruppe sein durfte und für all den Kaffee.

Ich möchte mich bei allen Kollegen aus Leipzig, Strasbourg und Berkeley bedanken, dass sie ihre wissenschaftlichen Kenntnisse und Geburtstagkuchen mit mir geteilt haben. Besonderer Dank geht an meine "Langzeit-Raumteiler", Christian, Gero und Helene, sowie an Berni für die gute Laune und die Unterhaltung in den Kaffeepausen. Danke an alle Mensagänger, die gemeinsam mit mir während alberner Gespräche die außergewöhnlichen Mahlzeiten ertragen haben.

Vielen Dank an Petra, die gute Seele des Büros, für die Unterstützung in allen bürokratischen Dingen und fürs Plaudern. Danke, Jens, für die Erfüllung aller IT-Wünsche in einem Wimpernschlag.

Von ganzem Herzen möchte ich mich bei meiner Familie bedanken. Danke, dass ihr da seid und an mich geglaubt habt. Danke, Mutti und Papa. Danke, Dirki, Oma Toni, Opa Josef, Oma Heidi und Opa Eberhard. Ohne eure anhaltende Unterstützung wäre es um so vieles schwerer gewesen. An dieser Stelle möchte ich mich auch bei meiner Krötenfamilie bedanken! Stellvertretend für alle: Danke, Surki, fürs Glückbringen.

Mein besonderer Dank geht an dich, Christoph. Danke für deine Unterstützung, fürs Zuhören, Mutmachen, Motivieren, Mahnen und Erinnern. Ohne dich gäbe es die folgenden Seiten nicht. DDFJIQ.

Contents

Contents

Chapter 1

Motivation

The discovery of the DNA double helix in 1953 by Watson and Crick [1] was a starting shot of the race for its decoding. Since then, our knowledge of genomes has grown immensely and our conception of their functional principles has kept changing. Just until recently, the central dogma of biology was that the genetic code on the DNA is transcribed into RNA and subsequently translated into proteins. Those were assumed to be the essential building blocks of life responsible for basic structural, regulatory or catalytic cell functions in all species.

1.1. Overthrow of a dogma

Biosynthesis of proteins was considered to be the main purpose of the cryptic genetic code. In 1977 the finding that protein-coding genes of mammals are interspersed with seemingly arbitrarily long segments of intervening non-protein-coding sequence, now called introns, which are not included in the mature product, triggered the first earthquake on the ground of protein-centric genetic research [2, 3]. Since studies showed, that those elements were simply spliced out of the transcript and subsequently degraded, it was concluded that introns are non-functional evolutionary relics.

The universal conception for a long time remained “the more protein-coding genes,

the more complex the organism." While the non-coding rRNA and tRNA and their "housekeeping" functions in the cell's translational machinery were already unraveled in the 1970s, they were deemed an exception to the rule and it was not until the late 1990s that studying the field of non-coding RNA shifted into the focus of scientists. More functional RNA molecules were discovered, such as the famous Xist RNA or first microRNAs.

Recent advances in the technology of large-scale genome sequencing revealed even more surprising insights. Among those were intriguing facts like: The number of protein-coding genes in the roundworm *Caenorhabiditis elegans* is almost the same as in humans [4, 5], we share about 99 % of our DNA with chimps and bonobos [6, 7] and less than 2 % of the human genome encodes for proteins. Extensive transcriptomic studies using high throughput sequencing showed that nonetheless the mammalian genome is pervasively transcribed in a well regulated manner, that is highly specific to certain developmental stages or cell tissues in the case of non-coding RNAs [8–11]. All of these findings point us towards the importance of non-protein-coding parts of the genome, once neglected as "junk DNA".

To date, more and more functional non-coding transcripts and classes of non-coding transcripts have been discovered amongst the huge transcriptional output. It became increasingly evident that the complexity of an organism is in fact correlated to the proportion of the genome that is non-protein-coding rather than its sheer number of protein-coding genes [12]. Therefore the protein-centric view of molecular biology gave way to the era of non-coding RNAs, which hold the key to understanding human cognition, development and evolution.

1.2. The era of long non-coding RNAs

With tens of thousands of transcripts expressed from the mammalian genomes, long non-coding RNAs (lncRNAs) make up the largest and most peculiar and at the same time the least explored class of non-coding RNAs. Transcripts from this group often resemble protein-coding messenger RNA and undergo capping, polyadenylation and splicing. These particular transcripts are classified as mRNA-like lncRNAs (mlncRNAs). Recent studies identify up to almost 60,000 well defined lncRNAs produced from the human genome [13]. Although they usually have a very low expression rate compared to protein-coding RNAs, they are expressed highly spatio-temporal specific [14–16]. Some have been shown to be involved in gene regulation processes associated with essential roles during development, organ growth and diverse disease pathogenesis [17, 18].

Apart from a few detailed case studies, global statistical analyses have demonstrated that, as a group, mlncRNAs are under stabilizing selection. However, their evolutionary history is poorly understood. While their primary sequence is better conserved than putative neutrally evolving stretches of the genome, the average sequence conservation across species is weak [19–21]. This provides only very limited contrast between intronic and exonic parts, so that it is difficult at best to infer complete gene structures for orthologs. Not only the level of sequence conservation is low compared to other functional transcripts [20, 22], but characteristic secondary structures, like in rRNA or tRNA, are also missing. This absence of typical evolutionary patterns makes it hard in practice to computationally predict and identify homologs in genome-wide searches based on sequence similarity. As a consequence $> 95\,\%$ have been reported as poorly conserved, and suggested to be transcriptional noise [19].

The rapid development of sequencing technology has made it feasible to obtain high coverage transcriptome data sets for a wide variety of cell and tissue types. In addition to the systematic efforts to exhaustively catalog the human transcriptome in the ENCODE project and large cDNA resources amassed by the FANTOM project [23], rapidly growing resources are also becoming available for a diversity of model organisms. As a consequence, comparative transcriptomics approaches become feasible, see e.g. [24, 25] and the review [26].

More recent studies were able to detect higher percentages of conserved lncRNAs. Washietl *et al.* [27] demonstrated that $30 - 40\,\%$ of nearly $2,000$ human lncRNAs show conserved expression in rodents or ungulates based on direct comparison of transcriptome sequencing data for six mammalian species. In a similar approach Necsulea *et al.* [28] investigated 11 tetrapod species and reported $11,000$ primate-specific lncRNAs contrasted by $2,500$ highly conserved ones. These numbers are somewhat lower ($19\,\%$ of lncRNAs are older than primates), presumably because only one non-primate mammal was included and a direct `blast`-based homology search was used in this study. A maximum likelihood approach from Managadze *et al.* [29] to estimate the number of lncRNAs from publicly available data resulted in an estimate of $40,000 - 50,000$ lncRNAs of which about $60 - 70\,\%$ are conserved between man and mouse. In 2015 Hezroni *et al.* [30] used a method of direct transcriptome comparison from RNA-seq data sets and identified thousands of human lncRNAs that have homologs with similar expression patters in other species. But still, $> 70\,\%$ of lncRNAs had no sequence-similar orthologs in species that diverged > 50 million years ago.

An alleged lack of evolutionary conservation, however, does not imply absence of functionality. Beyond global sequence conservation, it is possible to utilize the conservation of gene structures to establish homology. Splice sites, in particular, are

retained over very large evolutionary time scales in many cases. Indeed, conserved splice site patterns in combination with multiple genome alignments can be used to successfully predict novel evolutionarily well conserved non-coding transcripts [31, 32]. While in flies the procedure is conveniently based on intron predictions, one has to resort to predicting internal exons in mammals. A considerable fraction of the transcripts detected in this manner shows very little sequence conservation and resembles lncRNAs. Probably they would not have been detected based on sequence homology alone.

1.3. Scope and outline

The scope of this thesis is to shed light on the realm of lncRNAs and their evolutionary history to help understanding their biological role in present-day humans. In the course of this contribution a method was developed to systematically study the conservation patterns of spliced RNAs, particularly lncRNAs and the evolution of their gene structures. Therefore comparative maps of splice sites, constructed from genome-wide multiple sequence alignments together with transcriptomic data, were employed. Building on the work and results of my diploma thesis [33], the method was refined and extended and applied in three different biological contexts that yielded new insights in the field of evolutionary history of human lncRNA, atypically spliced RNA transcripts of coelacanth and Alzheimer's disease associated genes.

This thesis is divided into three major parts. Part I elucidates the biological background that inspired this thesis. The molecular-biological concept and evolutionary relevance of RNA splicing are highlighted in Chapter 2.

Chapter 3 elaborates on the definition of non-coding RNAs and their functional mechanisms. Furthermore, the current knowledge about non-coding RNA evolution and the challenges of their systematic analysis in regards of conservation are discussed. These sections are based on the review:

> Nitsche A, and Stadler PF (2017). Evolutionary clues in lncRNAs.
> *Wiley Interdisciplinary Reviews: RNA* 8. doi: 10.1002/wrna.1376

Part II expounds the technical component of the developed method, regarding its underlying mathematical principles as well as the framework of its computational pipeline. Since multiple sequence alignments, provided through online databases, are prerequisite input files for the developed program, algorithmic concepts of alignment methods are explained in Chapter 4, specifying in particular the two programs that have been used to generate the employed files. Furthermore the mathematical model of `MaxEntScan` scoring is explained, which is an essential tool of the developed

method. Chapter 5 explains in detail the pipeline of the designed computational method, including the collection of necessary data, the utilization of multiple sequence alignments to establish homology of splice sites and how the integration of `MaxEntScan` scoring amplifies the power to infer the conservation of transcripts.

Part III showcases three biological applications and presents the conclusions that can be drawn from their results. First the results of a broad genome-wide approach to investigate the conservation of human lncRNAs across 46 vertebrates via the introduced splice site maps are unrolled in Chapter 6 based on the publication:

> Nitsche A, Rose D, Fasold M, Reiche K, and Stadler PF (2015). Comparison of splice sites reveals that long non-coding RNAs are evolutionarily well conserved. *RNA* 21:801–812. doi: 10.1261/rna.046342.114

This publication comprises a substantially extended and revised reanalysis of preliminary results of splice site conservation in lncRNAs originally described in my diploma thesis [33].

Chapter 7 takes a more differentiated turn by particularly analyzing the evolution of atypically spliced transcripts, such as circular and trans-spliced RNA found in the RNA-seq data of two coelacanth species. The results, indicating they are of an evolutionary old age, are presented based on the publication of

> Nitsche A, Doose G, Tafer H, Robinson M, Saha NR, Gerdol M, Canapa A, Hoffmann S, Amemiya CT, and Stadler PF (2014). Atypical RNAs in the coelacanth transcriptome. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 322:342–351

Chapter 8 focuses on the evolution of genes associated with Alzheimer's disease. A genome-wide RNA-profile was established comprising protein-coding an non-coding transcripts that are differentially expressed in Alzheimer's disease patients. The systematic study of those genes with comparative splice site maps revealed their accelerated evolution. This chapter is based on the following publication.

> Nitsche A, Reiche K, Ueberham U, Arnold C, Hackermüller J, Horn F, Stadler PF, and Arendt T (2017). Alzheimer related genes show accelerated evolution. bioRxiv: 10.1101/114108. submitted

The thesis is concluded in Chapter 9, where the findings are discussed in the light of lncRNA evolution and with respect to their relevance for future research. The three distinct appendices provide supplementary information of the respective studies for the interested reader, including further results and methods as well as research that has been done preliminary.

1.4. Author contributions and use of personal pronoun

In scientific writing the impersonal style used to be expected and even required. This convention changed and in a diverse range of scientific publications the use of the personal pronoun “we” is common to account for the collective work of a group, even if specific parts have been contributed by a single individual. Since the deliberations and results of multiple collaborative projects are presented throughout this thesis, the personal pronoun “we” will be used as well. This does not invalidate the statement made in the declaration of independence.

Part I.

Biological Background

Contents

Chapter 2

RNA splicing

In eukaryotic cells protein-coding genes are interrupted by non-coding stretches, called introns, which are transcribed but later removed from the transcript in the process of RNA maturation. This process is called splicing – a reaction catalyzed by a ribonucleoprotein (RNP) complex, whose components recognize particular intronic elements. The high accuracy of splicing is complemented by spatio-temporal regulatory mechanisms which make the process highly specific. The vast majority of human genes produces alternatively spliced transcripts and therefore contribute not only to the proteomic variety but also to that of the non-coding RNAome. Mutations in elements of this sensitive splicing machinery can have far reaching effects on the functional transcriptome.

The information in this chapter is based on the textbooks of Elliott and Ladomery [38] and Hertel [39], if not stated otherwise.

2.1. Split genes

When Richard Roberts and Phillip Sharp independently discovered "split genes" of the adenovirus in 1977 [2, 3], the perception of the gene organization changed dramatically and led to further research about the origin of introns. It also sparked debates

about their potential beneficial role in evolution.

Around 94% of all mammalian genes are interrupted by at least one intron. In the brief time period after transcription and before transcript processing, the immature precursor mRNA (pre-mRNA) in the nucleus corresponds in length and content with the DNA sequence on the gene. These long RNA molecules are also called heterogeneous nuclear RNA (hnRNA). The process of splicing, which in fact may already occur during transcription, will excise intronic sequences and ligate the retained segments, which are called exons.

In the case of protein-coding genes, exons encode the amino acid sequence of a protein in an ORF with an average length of $\sim 150\,\text{bp}$ and are therefore relatively short compared to the average intron, which is $\sim 6,000\,\text{bp}$ long but can exceed extremes of $> 400,000\,\text{bp}$. Not only the intron length, but also the number of introns per gene varies greatly. The median number of exons in human protein-coding genes is 7, but there are numerous extreme cases with > 100 exons per gene. The longest human gene is that of the dystrophin protein. While the 79 exons only comprise a coding sequence of 14 kb, the large amount of long introns inflates the gene with seemingly futile sequences to a total length of 2.5 Mb.

2.1.1. Splice sites

Each intron has important sequence elements that play an essential role in the splicing process (Figure 2.1). Located at the exon-intron-boundaries are the splice sites with a highly conserved but very short consensus sequence. The 5' splice site, the **donor** site, has the consensus sequence `AG|GURAGU` (exonic|intronic; `R` = `A` or `G` nucleotide) and the 3' splice site, the **acceptor** site, has an intronic `AG` dinucleotide preceding the downstream exon. The polypyrimidine tract, located directly upstream of the acceptor site, is a stretch of $10-20$ pyrimidines (`Y` = `U` or `C` nucleotide), predominantly uridine. Located further upstream ($\sim 100\,\text{nt}$) of the 3' splice site region is the branch point site (BPS), which is the location of a single `A` nucleotide that, besides the donor and acceptor site, is the only sequence that participates in the chemical splicing reaction. It is surrounded by a poorly conserved sequence, which makes it hard to identify BPS in introns. In human introns the consensus is `YNYURAC` (`N` = any nucleotide).

While there are slight differences in the intronic splicing elements between species, e.g. yeast has no polypyrimidine tract but a much more conserved BPS sequence (`UACUAAC`), the vast majority of introns conforms to the "GT-AG rule", which describes the first two and the last two nucleotides of an intron. This circumstance induces functional equality between splice sites, meaning that any donor can be spliced

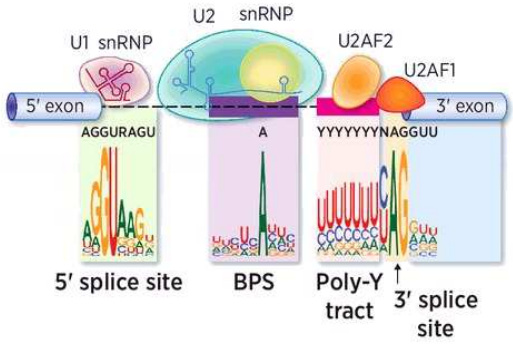

Figure 2.1.: Sequence elements of major spliceosomal introns in human are highly degenerate but follow certain consensus motifs, especially the intron-exon boundaries, which conform to the typical GT-AG motif, at the first and last two intronic positions, respectively. Donor (5' splice site), acceptor (3' splice site), branch point site (BPS) and polypyrimidine (poly-Y) tract interact with components of the major spliceosome to mediate the splicing reaction (see Section 2.2.1). `R` = `A` or `G`; `Y` = `C` or `U`; `N` = `A`, `C`, `G` or `U`. Figure adopted from [40].

to any acceptor.

2.1.2. The chemical reaction of splicing

The splicing process is one of three post-transcriptional modifications in the course of RNA maturation. The other two are: capping of the 5' end and polyadenylation of the 3' end. Both of those processing steps primarily serve to enhance the stability of the nascent mRNA.

Splicing as the third post-transcriptional modification in which introns are precisely removed from the pre-mRNA transcript and remaining exons are joined together. Its basic biochemical mechanism is a well characterized process. During the reaction phosphodiester bonds are split and reformed via hydroxyl groups (—OH), which is called transesterification. The splicing reaction can be described as a double transesterification (Figure 2.2A).

Step 1 The 2'—OH of the BPS adenosine attacks the 5'-phosphate residue at the donor site, cleaving the 3'—5' phosphodiester bond and forming a 2'—5' bond at the BPS, resulting in an intron lariat (loop) intermediate and a disconnected upstream exon.

Step 2 The 3'—OH, at the end of the free upstream exon attacks the 3'—5' phosphodiester bond at the acceptor site, forming a new bond between both exons and releasing the lariat intermediate. The intron lariat debranches and gets degraded.

Each splicing process occurs for each intron individually. The order of spliced out introns does not necessarily comply with the order in which they are present on the transcript.

2.2. Classes of introns

Introns can be distinguished into three classes that perform splicing in different ways: (1) spliceosomal introns (2) self-splicing introns (3) tRNA introns.

In eukaryotic cells the splicing process happens predominantly with the help of the major spliceosome, a complex of RNA and proteins, which assembles directly on the pre-mRNA. However, in a wide range of organisms (including prokaryotes) the introns of diverse transcripts (mRNA, rRNA) catalyze the splicing reaction themselves. The splicing of tRNAs is an exception, since it does not occur via transesterifications.

2.2.1. Spliceosomal introns

Unlike in the transcription or translation process, where the RNA/DNA is scanned and processed from 5' to 3' end, an independent spliceosome complex assembles for each intron removal and gets degraded after the completed splicing reaction. This makes seven "spliceosome cycles" for the average human protein-coding pre-mRNA transcript.

Major spliceosome

A full major spliceosome consists of five small nuclear ribonucleoproteins (snRNP): U1, U2, U4, U5 and U6. The units are named after their corresponding small nuclear RNA (snRNA) component, which is rich in uridine. RNA–RNA base pairing interactions of these snRNAs with conserved sequence elements of the pre-mRNA transcript and other snRNAs are essential for the spliceosome assembly and ensure an efficient and precise splicing process. For each intron to be spliced out a spliceosome complex assembles directly on the primary transcript from its subunits (Figure 2.2B).

Complex E The step that commences the spliceosome pathway. U1 binds the donor site in an RNA–RNA interaction. U2 Auxiliary Factor (U2AF) proteins interact with the 3' splice site region elements for stability and protein SF1 binds the BPS.

Complex A The pre-splicing complex. SF1 is replaced with U2, which imperfectly base pairs with the BPS sequence. This causes an adenosine bulge, exposing the

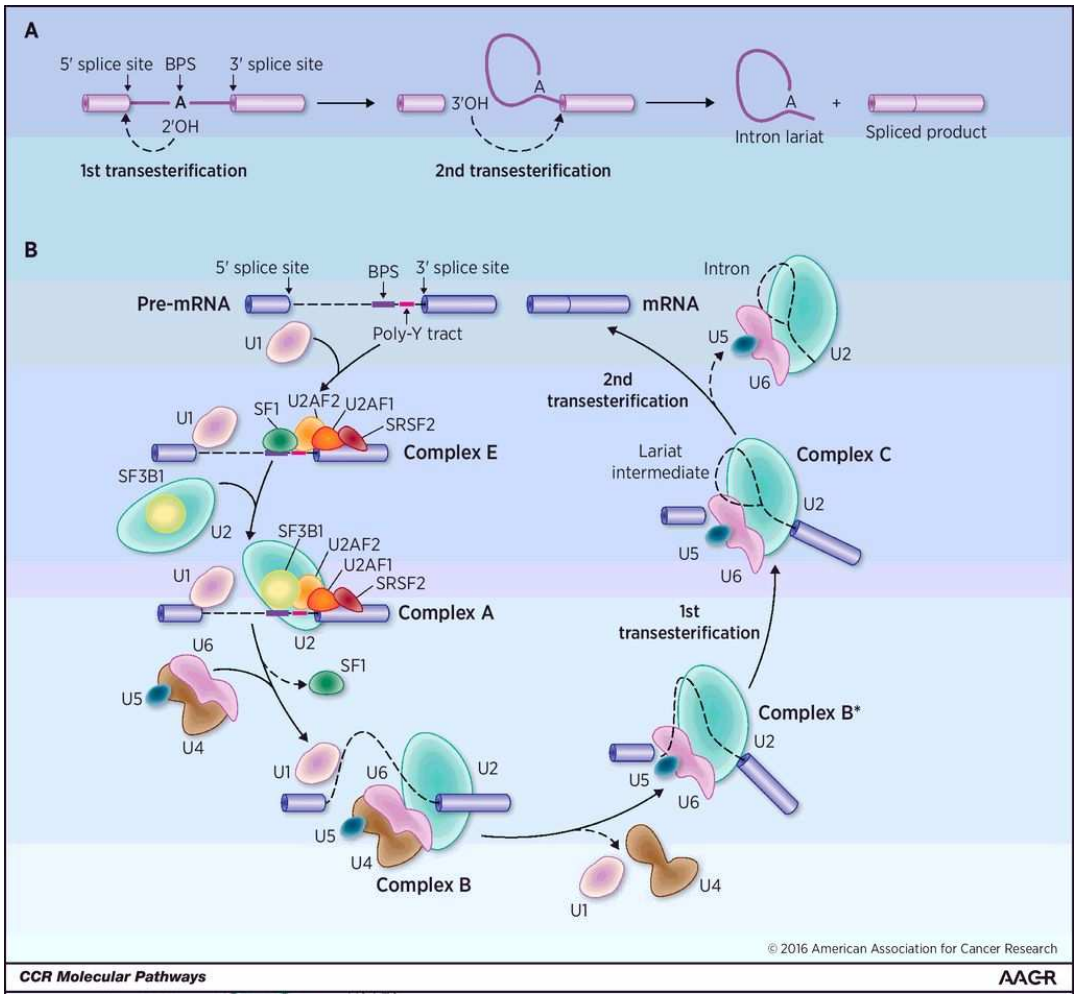

Figure 2.2.: Splicing mechanism. (**A**) The double transesterification of the chemical splicing reaction. Re-formation of phosphodiester bonds between the branch point site (BPS) and 5' and 3' splice site of the intron ligate both adjacent exons by forming an intron lariat intermediate. (**B**) Illustration of the spliceosome cycle. The spliceosome assembles stepwise from five snRNP subunits directly on the primary transcript strongly supported via RNA–RNA base pair interactions. Additional proteins drive the re-arrangement of the complex to eventually form a catalytic core that promotes the splicing reaction. Afterward the completed splicing the complex is disassembled and recycled in the next spliceosome assembly. Figure adopted and rearranged from [40]

2'—OH of the adenosine BPS and preparing it for the upcoming nucleophilic attack. The U2AF binding with the poly-Y tract stabilizes the base pairing interaction of U2.

Complex B The pre-catalytic complex. A trimer of snRNPs U4/U6 and U5 is added to the spliceosome body, which now contains all snRNP subunits. U5 base pairs with both exons holding the spliceosome in place.

Complex B* The catalytically active complex. The spliceosome undergoes structural changes to initiate the catalytic splicing process. Subunit U1 and U4 are dissociated from the complex. This enables U6 to base pair with the now vacant donor site and U2, which brings the BPS physically closer to the 5' splice site and generates the catalytic core of the spliceosome. The first transesterification takes place.

Complex C The spliceosome only consists of three remaining subunits (U2, U5 and U6), the 5' exon and the lariat intermediate bound to the 3' exon. While U5 still holds both transcript parts together, the second transesterification occurs.

Disassembly All components disassemble. The final spliced product is transported to the cytoplasm, the intron lariat is degraded and the subunits are re-used for the next spliceosome assembly.

At each stage additional proteins contribute to the progress of the spliceosome cycle. Proteins like U1C, splicing factors (U2AF) and serine-rich (SR) proteins stabilize the complex at diverse steps of the assembly. A key role is performed by RNA helicases, which regulate the re-arrangement of the complex and ensure a correct timing of events under the consumption of ATP or GTP.

Major spliceosomal introns are the most common introns in eukaryotes. We will employ their specific canonical GT-AG splice site motif to assess the conservation of splice sites with our method, which will be introduced in Chapter 5.

Minor spliceosome

A small fraction of spliceosomal introns ($1:300-1:670$) belong to the minor class of U12-dependent introns, which occur in metazoans and plants. They are distinct from canonical introns of the major spliceosomal class, and usually follow the AT-AC rule instead of the GT-AG rule. The BPS consensus sequence differs as well. All of them are spliced by an alternative spliceosomal complex - the minor spliceosome, which uses snRNP U11, U12 and U4atac/U6atac, instead of U1, U2 and U4/U6. These are functionally equivalent to the subunits of the major spliceosome, but do not show a

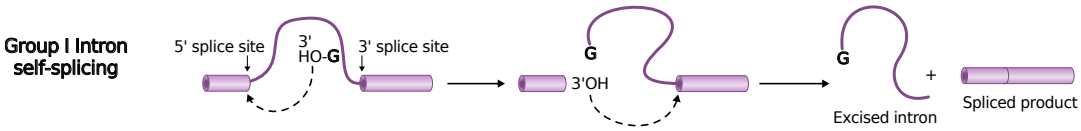

Figure 2.3.: Group I introns. The splicing proceeds as two-step transesterification. Opposed to Group II introns, however, the first nucleophilic is not performed by an intronic BPS but by the 3'—OH of an external guanosine. Hence no lariat structure is formed during the process. Graphic style inspired by [40].

big sequence similarity. Both spliceosomes use the U5 subunit and a similar set of proteins is involved to control the chemically identical splicing process. The minor spliceosome consensus motif will not be considered in our method.

2.2.2. Autocatalytic introns

Some introns are able to excise themselves from the primary transcript without the help of a spliceosome. They are also referred to as ribozymes since they perform the splicing in an autocatalytic reaction by folding into a secondary structure that resembles the catalytic core of the spliceosome complex. Therefore their secondary structure is highly conserved. According to distinct secondary structures and the actual splicing reaction, the class of self-splicing introns can be distinguished into two groups. Although the catalytic reaction proceeds for both groups as a double transesterification, there are slight differences.

Group I These introns do not develop a lariat intermediate during their splicing process. The nucleophilic attack is performed by the 3'—OH of a free exogenous guanosine, that was previously bound to a specific G-binding site on the transcript (Figure 2.3).

Group II The chemical splicing reaction of Group II introns is analogous to those of spliceosomal introns, meaning they form a lariat with a adenosine BPS.

While spliceosomal introns solely occur in eukaryotes, autocatalytic introns are present in prokaryotes as well. Group II introns are found in bacteria and in subcellular organelles like mitochondria and chloroplasts. Due to the strong resemblance of their splicing pathway with those of spliceosomal introns and their self-splicing ability, it is assumed, that spliceosomal pre-mRNA splicing actually evolved from those formerly "parasitic DNA elements".

2.2.3. tRNA introns

The introns of tRNAs are unusually short and get removed in a different splicing pathway than the autocatalytic or spliceosomal introns. Opposed to the other splicing reactions, tRNA splicing is a process of successive cleavage and ligating reactions catalyzed by several enzymes, that occurs in three stages. Here the process is described on the example of yeast.

Stage 1 Cleavage. An endonuclease enzyme cleaves the intron on both splice sites, producing two tRNA half-molecules and a linear intron. This leaves unusual 2'—3' cyclic phosphate (P) and a 5'—OH ends on the half-molecules.

Stage 2 Ligation. An RNA ligase joins both exon-molecules in a multistep reaction. First catalyzed by phosphodiesterase and kinase the unusual ends are altered into a 2'—P and 5'—P under the consumption of GTP. Then both ends are joined by synthetase and ligase forms a 5'—3' phosphodiester bond. This reaction requires another nucleoside triphosphate, this time ATP.

Stage 3 Removal of 2'—P. The extra 2'—P group from the original donor site that remained at the splice junction after ligation is transfered to a nicotinamide adenine dinucleotide (NAD) by a phosphotransferase. The splicing process is completed and the mature tRNA is present.

2.3. Alternative splicing

A primary transcript of a single gene can be processed into various isoforms or different gene products by alternative splicing. This increases the coding capacity of the genome without increasing the number of genes. About 95% of multi-exonic genes in the human genome are differentially spliced [41]. This explains the non-proportionate relation between gene count and complexity of organisms. Through alternative splicing $\sim 20,000$ human genes produce a proteomic diversity of hundreds of thousands of proteins.

2.3.1. Forms of alternative splicing

While constitutive exons are always included in the spliced product, alternative exons are elements of the primary transcript, that are variably spliced to be included or excluded in the mature RNA. The order of the exons, however, is always maintained. It can be distinguished between five major forms of alternative splicing (Figure 2.4).

Exon skipping. The simplest and most common form of alternative splicing, where

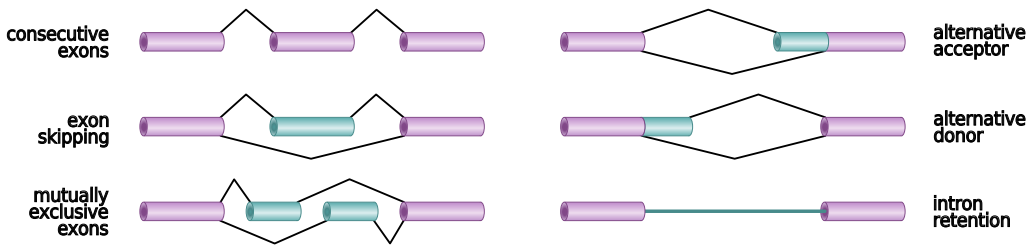

Figure 2.4.: Five basic forms of alternative splicing. Transcript isoforms can be produced by variably including exons or using alternative splice site on the 5' or 3' end, which splices exons of altered length. Black lines indicate splicing events. Graphic style inspired by [40].

an alternative exon is removed entirely together with its adjacent introns.

Mutually exclusive exons. One of two alternative exons is skipped so that only one of them is present in the mature RNA.

Alternative acceptor. An exon has two or more possible acceptor sites that can be selected for splicing. This choice will influence the length of the exon (start point of the exon).

Alternative donor. An exon has two or more possible donor sites, that can be chosen for splicing. This choice will influence the length of the exon (end point of the exon).

Intron retention. An intron is not spliced out of the transcript and thus becoming an "exon" itself. In the case of protein-coding RNA these introns do have an open reading frame (ORF) corresponding to that of the neighboring exons. In humans this is the rarest form of alternative splicing.

The biological reality of alternative splicing is more often than not a combination of these basic forms. As another way to achieve more variety of transcript isoforms, the transcription machinery can employ alternative transcription start or polyadenylation sites.

2.3.2. Regulation of alternative splicing

Some genes express all of their isoforms in any cell, while others only produce certain variants under distinct spatio-temporal conditions, like tissue type, developmental stage or gender. This is relevant for the regulation of gene expression levels.

The splicing code

Certain sequence elements within exons and introns function as *cis*-active sites that regulate gene expression. They encipher whether the current sequence belongs to an exon or intron and therefore serve to distinguish between both during the spliceosomal splicing process. Hence, these sequences are called the splicing code.

They can lead to silencing or enhancing effects on adjacent splice sites, by recruiting *trans*-regulatory splicing factors, e.g. SR-proteins, which are essential for the spliceosome assembly. According to location and function these auxiliary regulatory elements are referred to as exonic splicing enhancers (ESE) or silencers (ESS) and intronic splicing enhancers (ISE) and silencers (ISS). Enhancers and silencers "strengthen" or "weaken" the associated splice sites, respectively.

Large introns can contain numerous sequences similar to the consensus of functional splice sites. Resulting pseudoexons pose a high risk of erroneous splicing. The splicing code is crucial for correctly splicing authentic alternative or constitutive exons.

Exon and intron definition

The spliceosome composition described in Section 2.2.1 follows the so-called intron definition, in which the spliceosome recognizes the intron and assembles directly on it. Due to the fact that introns are usually huge stretches of sequence, whereas exons are rather short, the spliceosome composition in higher eukaryotes happens through exon definition. In this case the spliceosome recognizes the exons first by binding early spliceosome factors. U1 subunits bind the donors of consecutive introns, while U2AFs bind the acceptors. This basically marks the beginning and end of the enclosed exon, as a signal for the spliceosome. An SR-protein chain now connects U2AF of the upstream intron and U1 of the downstream intron, spanning the complete exon. The ESEs of this exon are required for this process. When U2 binds the 3' splice site region, a rearrangement occurs and the interactions between subunits are now across the introns. Their subunits form Complex A and follow along the intron definition pathway.

The mode of spliceosome assembly has a major influence on the form of alternative splicing that is used. In organisms (e.g. human) that use exon definition, exon skipping is most common, while intron retention is more common in species that use the intron definition pathway (e.g. yeast). This makes sense since exons as well as introns are recognized by the spliceosome and therefore can specifically be included in or excluded from the mature RNA.

Other regulating factors

There are additional factors that are able to regulate alternative splicing.

Concentration of regulatory proteins. Proteins that *trans*-actively bind to auxiliary elements of exons (ESS, ESE) in the primary transcript can block or activate splicing, e.g. SR-proteins, heterogeneous nuclear RNPs (hnRNPs).

Epigenetic modifications. RNA-guided modifications of histones affect the splicing of alternative exons. It has been shown that RNA transcripts recruit histone-modifying complexes to induce DNA methylation, changes in chromatin structure and other histone modifications. The involved RNA molecules can be small or long non-coding RNAs, and even the primary transcript itself [42].

Transcription. Splicing often occurs co-transcriptionally. It has been demonstrated that the elongation rate of RNAP II has an impact on the inclusion of alternative exons. The kinetic model proposes that a decelerated elongation rate gives more time to a weak alternative exon to recruit the splicing machinery, before a stronger competing exon emerges during the transcription process.

Secondary structure. A single-strand motif can be masked by secondary structure and blocked from recognition by the splicing machinery. The effect can be positive or negative, according to the function of the masked element. Another way of regulating splicing through secondary structure is the ability to bring two distant elements in close proximity. Alternative secondary structures can also influence which of multiple mutually exclusive exons is integrated into the final transcript.

2.4. Atypical splicing

Some splicing events do not follow the described canonical splicing mechanism, where a 5' donor and a 3' acceptor are spliced so that two consecutive exons from a single preliminary transcript are ligated into a linearly mature RNA product. The advances in computational and experimental techniques that study the transcriptome unraveled that unconventional splicing events that produce non-colinear transcripts are biological reality and much more abundant than previously assumed. This section highlights two types of atypical splicing events.

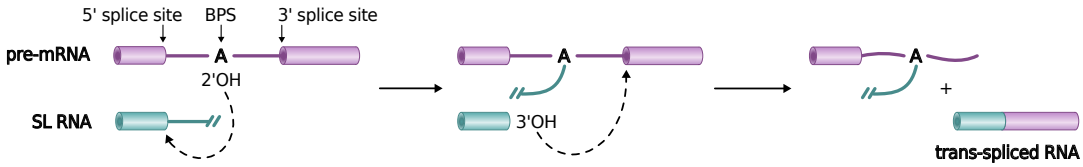

Figure 2.5.: Trans-splicing in *Trypanosoma brucei* follows the double transesterification pathway. A spliced leader (SL) RNA, independently transcribed, serves as both donor and U1 subunit in this spliceosomally catalyzed reaction and provides the 5' cap for the mature RNA transcript. The BPS is a conserved adenosine within the intron of a long polycistronic transcript. Graphic style inspired by [40].

2.4.1. Trans-splicing

It is well known, that some species, such as psychosomatics or nematodes, frequently produce inter-molecular transcripts, that are *trans*-spliced products from independently transcribed parts, which originate from distant genomic regions or even distant chromosomes [43]. While this is a rare event in humans and most other organisms, where splicing usually occurs in *cis* – within the same molecule – all RNAs are spliced in *trans* in *Trypanosoma brucei*. In this case the splicing involves a "spliced leader" RNA (SL RNA), see Figure 2.5.

Trans-splicing is crucial for species like *Trypanosoma brucei* and *C. elegans*. It is required to split long polycistronic RNA molecules, which contain coding information of multiple genes, into shorter individually translatable transcripts.

The chemical reaction of *trans*-splicing is a similar double transesterification as in *cis*-splicing. It is equivalently catalyzed by a spliceosome, where the SL RNA takes over the role of the U1 subunit. The SL RNA, a mini-exon of 39 nt length, provides the donor site for the nucleophilic attack by a BPS of a second molecule. This results in a free SL RNA molecule and a Y-shaped intermediate molecule. In the second reaction the 3'—OH attacks the acceptor site of the downstream exon, joining both molecules and releasing the branched intermediate. In the resulting RNA molecule the SL RNA is now the leading sequence of the spliced exon. This is important for the stability and efficient translation of the mature transcript, since the mini-exon features the 5' cap, which is not present on the RNAP I-transcribed polycistronic transcripts.

Trans-splicing does also occur in human, but has been generally thought to be a rare phenomenon playing only a subordinate role in vertebrates [44]. Although large-scale transcriptome sequencing showed, that non-colinear or chimeric RNAs are abundant in a variety of species, a considerable fraction of them might be "RTfacts", like short homologous sequences (SHS) at the junctions, that are generated by reverse

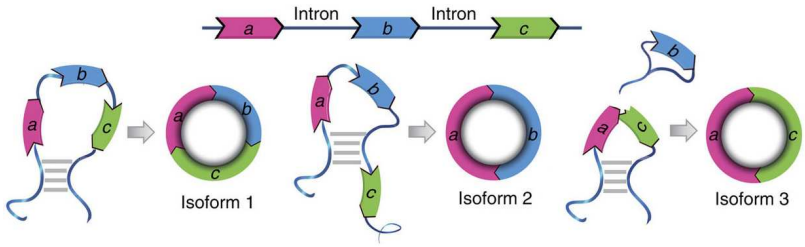

Figure 2.6.: Alternative Circularization. Back-splicing can be promoted and regulated by specific RNA–RNA interactions of intronic elements, e.g. inverted *Alu* repeats, which bring the ends of exons in close proximity. When different hybridization is possible the exons can be circularized alternatively. Figure adopted and rearranged from [54]

.

transcription [45–49].

On the other hand reports of high tissue specific expression of chimeric transcripts, together with evidence that they contain complete protein domains and detectably produce a multitude of proteins, emphasize the biological origin and functionality of these RNAs [10, 50–52].

Strand-switching

In a special case of *trans*-splicing the transcribed units are derived from opposite strands. This is called a strand-switch and has been observed in *Trypanosoma brucei* and in *Drosophila*. The best characterized case is that of the local strand-switch at the mod(mdg4) gene in *Drosophila* [49, 53].

2.4.2. Back-splicing

Back-splicing events occur when the BPS of an upstream intron attacks the donor site of a downstream exon, and thus produce circular RNA (circRNA) molecules. This is also called head-to-tail splicing. The resulting circRNAs are mono- or multi-exonic, but consist usually of two to three exons, which can be alternatively spliced (Figure 2.6) [55]. A minimum exon length, however, is required in the case of a single exon circRNA, such as ciRS-7. The back-splicing event itself employs regular canonical splice sites and is guided by the spliceosome, however at a lower efficiency than that of their linear counterparts. Exon skipping events are positively correlated with circRNA biogenesis, as they commonly derive from spliced alternative middle exons of pre-mRNAs [55].

The circularization is regulated in *cis* and *trans*. The exons of circular RNAs are often flanked by long intronic sequences, containing reverse complementary elements

(e.g. *Alu* repeats) that can base pair with each other to bring both exon ends in close proximity [55] and therefore promoting the back-splicing event in *cis*. Some RBPs can regulate the circRNA genesis in *trans*, with repressive or enhancing effects [56, 57].

Due to their relatively low expression level and their non-polyadenylated nature, they got under the radar of typical poly(A) enriched library sequencing. Just recently the sequencing of selective libraries of non-polyadenylated RNAs, that were treated to have all linear RNAs degraded and only retain circular RNA, revealed that a substantial fraction of spliced human transcripts produce circRNA [58, 59]. Among those identified are long known prominent examples [60–62], but also several new interesting circular isoforms, that are conserved between human and mouse [59].

As they had no function assigned originally, they were first thought to be by-products of defective irregular splicing and received little attention from the scientific community. While more and more circular transcripts have been validated in several studies, some of them highly and spatio-temporal specifically expressed, it has become evident that they emerge from purposeful and well regulated splicing events. Some circular transcripts are found to be expressed even more frequently than their linear isoforms [63].

A proposed function of circRNAs is the regulatory role as microRNA sponges as a crucial component of gene expression. The best studied case is that of ciRS-7, which harbors over 60 conserved binding sites for miR-7 [64, 65]. However, there are only a few other known cases in mammals, suggesting that this might not be their primary role [66].

2.5. Relevance

Even though it takes significantly longer to transcribe genes prolonged with intronic sequences, and their increased risk of impactful mutations that adds up along the transcript length, the majority of eukaryotic genes are interrupted by introns.

2.5.1. Enhancement of eukaryotic gene expression

It is well known that introns in fact are beneficial for an efficient expression of eukaryotic genes [67]. Besides experiments of intron insertion/removal, which showed the enhancing effect on transcription [68], this is also visible when comparing the yeast genome and its transcriptomic outcome. While only 4 % of its genes contain introns, those genes contribute to more than a third of the overall mRNA transcripts in the

cell. While a decelerated maturation of RNA due to an extended gene length can be profitable to the organism in certain contexts (e.g. tissue patterning), this "intron delay" [69] can be canceled out on several levels of gene expression when needed.

Transcription. Some introns contain *cis*-regulatory elements that enhance transcription of genes. Secondly, by nucleosome positioning introns can make the DNA more accessible for transcription [70]. A third possible way to influence the transcription is the promotion of RNA polymerase II activity, including transcriptional initiation and elongation.

Transcript Processing. The splicing of the last (most downstream) intron positively affects the polyadenylation of the transcript [71].

Export. The splicing machinery actively promotes the export of the mature RNA from the nucleus to the cytoplasm, while retaining unspliced transcripts, ensuring that translation only occurs on mature RNA products.

Translation. The splicing process leaves protein marks on the splice junctions, known as the exon junction complex (EJC), which influence the efficiency of translation [72].

2.5.2. Evolutionary role

The frequency of discontinuous genes rises with the evolutionary stage of the organism, indicating an important evolutionary role for eukaryotic organisms.

The feature of splicing ability provides a path to accelerated evolution of new proteins. Considering exons as modules, that decode functional units of proteins or structural elements of ncRNA, the addition of new exons as mobile genetic elements, can easily create new more sophisticated and complex proteins or ncRNAs. As long as the exon is flanked by introns, the new transcript will be correctly spliced and translated. The option of alternative splicing further expands the coding capacity of the genome.

A precise splicing process is essential to achieve the correct and functional gene product. Hence the position of an intron within a gene and its splice sites are usually highly conserved. The positions of splice sites therefore pose suitable reference points to analyze the evolutionary history of non-coding genes in particular, since their exonic sequences on the other hand show little sequence conservation. By tracing the evolution of splice sites, we can trace the evolution of gene structure, which means the evolution of the transcript itself.

Contents

Chapter 3

The curious case of non-coding RNA

Over the last decade the technology of high-throughput sequencing helped revealing that an overwhelmingly large fraction of the mammalian genome is transcribed in some cell type, tissue, or developmental stage [8, 73–75]. Less than 3 % of the genome encode all of the $\sim 19,800$ protein coding genes [76, 77] and the coding sequence itself barely exceeds 1 %. However, they account for a disproportionately large fraction of the mass of the human transcriptome (disregarding rRNAs), due to their high expression rate [10, 13, 78]. Nearly a quarter of the total RNA in the cell can be attributed to UCSC-annotated exons and thus to mature mRNAs [79]. The vastly diverse remainder of the genome harbors tens of thousands of non-protein-coding transcripts, many of which are expressed only at very low levels or under very specific circumstances [80].

3.1. Classes of ncRNA

The complex and extensive non-coding transcriptome comprises a diverse array of RNA molecules, varying in size, function, abundance, and genomic location and orientation of transcription. Therefore a classification of ncRNAs can be made based on various aspects. The most common and coarsest distinction is made by length, where a rather arbitrary length cutoff of 200 nt is dividing the huge diversity of ncRNAs into

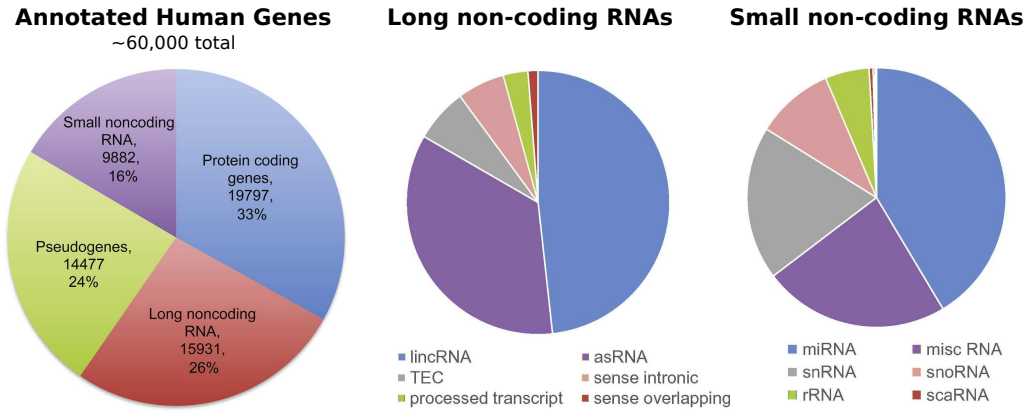

Figure 3.1.: Distribution of GENCODE annotated genes. The pie chart on the left (adapted from [81]) represents the statistic of annotated human transcripts from GENCODE v23. The two rightmost pie charts (adapted and redrawn from [82]) show the distribution of annotated non-coding genes in GENCODE v22, further classified into long and small non-coding RNAs. The fractions of mitochondrial tRNA and other small non-coding RNAs were too small to display. Label "TEC" stands for "to be experimentally confirmed."

two classes: small and long non-coding RNAs (lncRNAs), Figure 3.1 and Table 3.1.

Among the group of small ncRNA are well studied classes of ncRNAs, which fulfill known and well defined tasks in the cell. This category is therefore further distinguished by the established function of the RNA molecule. Famous representatives belong to the long-known group of infrastructural "housekeeping" ncRNAs, comprising small nuclear (snRNA), small nucleolar (snoRNA), ribosomal (rRNA) and transfer RNA (tRNA), which are involved in transcription, post-transcriptional modifications (e.g. splicing) and translation of mRNA.

The fairly heterogeneous class of lncRNAs is far more unexplored. The broad definition of a lncRNA is currently only based on size ($> 200\,\mathrm{nt}$), and the feature they do not have, namely protein-coding capacity, since there are no (known) further characteristics they all exhibit. Only for a tiny fraction exist detailed case studies that elucidate their functionality. Hence the majority is further distinguished by genomic location and orientation of transcription, rather than by function. It is discriminated between lncRNAs that originate from intergenic regions (lincRNAs) or from already known annotated genes. In the latter case, their source may be exonic as well as intronic (PINs and TINs) in both sense or antisense (asRNA) direction.

Identifying these long non-coding transcripts is not trivial. They may contain potential ORFs of moderate length, are usually transcribed by RNA polymerase (RNAP) II and frequently spliced, polyadenylated and capped in the same way as mRNAs [83–85]. About 30% of non-coding RNA even produce alternatively spliced transcripts [86–88]. The strong resemblance to mRNA transcripts might be one of

Table 3.1.: Brief classification of ncRNA. Non-coding RNA molecules are classified mostly based on length and genomic origin, since little is known about their functionality. It is likely that they share a number of qualities, instead of belonging to a single RNA category.

Class	RNA type
"Housekeeping" non-coding RNA	transfer (tRNA), ribosomal (rRNA), small nuclear (snRNA), small nucleolar (snoRNA), precursor microRNA (pre-miRNA)
Regulatory small non-coding RNA	microRNA (miRNA), small RNA processed from structured RNA, small interfering (siRNA), piwi-interacting (piRNA), promoter associated (paRNA), termini associated (taRNA), splice site associated RNA
mRNA-like non-coding RNA (mlncRNA)	long intergenic non-coding (lincRNA), snoRNA host genes, primary microRNA (pri-miRNA), antisense (asRNA), intronic RNA (TINs, PINs), UTR-derived (uaRNA), chromatin associated RNA
Other	DualRNA, mRNA with IRES, macroRNA, circular RNA (circRNA), circular intronic RNAs (ciRNAs)

the reasons that some, now known as non-coding transcripts, have been previously annotated as protein-coding genes for years [89]. The following features can indicate that a transcript is non-coding, even if a potential ORF is present: The transcript is predominantly present in the nucleus; The ORF is not substantially longer than expected by chance considering the length of the transcript; Codon frequencies are not random; The nucleotide substitution rates are not biased towards the third codon position; The aminoacid sequence of the ORF is not similar to known proteins or protein domains. However, experimental tests of *in vitro* translation are often critical to undoubtedly exclude coding capacity.

With the current speed of lncRNA discovery by far outpacing their functional annotation, the question has become what fraction of the detectable lncRNAs actually convey biological functions, as opposed to being coherently transcribed and processed byproducts without biological relevance. Only a small minority of the nearly genome-spanning primary transcriptional output of mammalian genomes have been detected as stable processed RNA products such as protein-coding mRNAs, mRNA-like ncRNAs (mlncRNA), or a plethora of short RNA products [90, 91]. Nonetheless, the discussion whether these "dark matter transcripts" are real or merely technical artefacts seems to have been (largely) settled in favor of the reality of pervasive transcription [8, 9, 14, 78, 92, 93].

3.2. Transcriptional noise vs. functionality

Pervasive transcription, which seems to be prevalent since the last common ancestor of eukaryotes a billion years ago, produces a widely diverse repertoire of thousands of long non-coding RNAs from mammalian genomes. However, the classification for a group of heterogeneous molecules with a tremendous expected functional diversity which could rival the proteome's is based merely on size and missing protein-coding capacity.

Most lncRNAs lack levels of sequence conservation comparable to their protein-coding sisters. Recent studies estimate that less than 10 % of the human genome are evolutionarily constrained at the sequence level [94]. Such estimates are based on a comparison with 4-fold degenerate codon positions or ancient repetitive sequences that are taken as neutrally evolving. As such they are lower bounds limited by the power of statistical tests and the assumption that the reference really evolves without constraint. In the light of pervasive transcription this is not necessarily true [95]. This low level of sequence conservation impedes "traditional" genome-wide searches for homologs both between species and within the same genome. It has led some researchers to conclude that most lncRNAs convey no important biological functions [96, 97]. Even if much of the transcriptome evolves (nearly) neutrally at the sequence level, substantial selection pressures may still act e.g. on gene structure or RNA structure, as it will be discussed below.

3.2.1. Mechanisms of lncRNA action

A growing body of detailed functional studies about lncRNAs demonstrates that they can have strong cellular effects and exert non-trivial influence on the organismal level. LncRNAs affect gene expression through diverse mechanisms and in a wide variety of genomic contexts. They may exercise positive and negative regulation, act in *cis* or in *trans*, impact transcription, post-transcriptional maturation, or translation, and function through interaction with RNA, DNA, or proteins. The transcript itself can serve as a scaffold for binding sites, as molecular decoy, as a guide to target elements, or as a recruiter or inducer for building molecule complexes (reviewed e.g. in [98–100]). The best studied cases are those associated with human diseases (reviewed e.g. in [17, 18, 81]).

There are five established roles for long non-coding RNAs (Figure 3.2), according to their mechanism of action. One lncRNA can exert multiple of these roles:

Signal lncRNAs (e.g. HOTAIR, HOTTIP, Xist) play essential roles in signal regula-

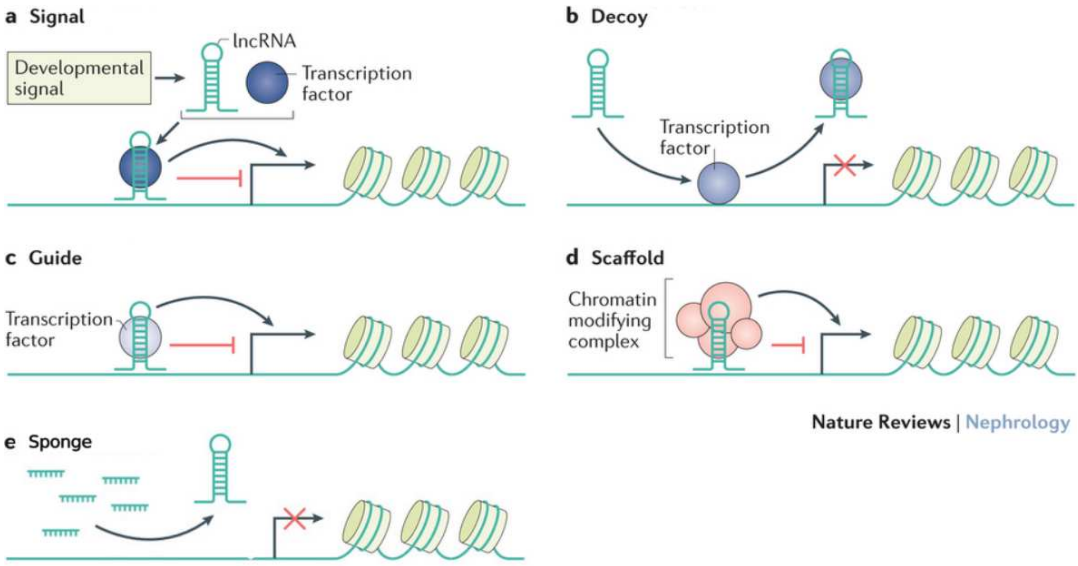

Figure 3.2.: Functional mechanisms of lncRNA. (**a**) Precisely expressed lncRNA over space and time as response to developmental cues serve as signal regulators. (**b**) By titrating away DNA-binding proteins (e.g. TF) lncRNA act as decoys. (**c**) They may also guide protein complexes to specific target sites and/or (**d**) act as scaffolds to link those proteins and form a complex. Figure adapted from Lorenzen and Thum [17] and extended.

tion when expressed in a highly spatio-temporal specific manner as a response to various stimuli.

Decoy lncRNAs (e.g. MALAT1, MHRT, GAS5) can regulate transcription by binding to and detracting transcription factors or other regulatory proteins away from chromatin.

Guide lncRNAs (e.g. Xist, HOTTIP, HOTAIR, Fendrr) help chromatin modifying complexes, typically PRC2, to localize certain target sites on the genome in *cis* as well as in *trans*.

Scaffold lncRNAs (e.g. ANRIL, HOTAIR, Kcnq1ot1) help assemble, link and hold together aggregations of proteins to form ribonucleoprotein (RNP) complexes which can affect histone modifications on chromatin.

Sponges are a special form of decoys that act on microRNA (miRNA) targets. They bind a multitude of microRNA molecules and titrate them away from the target site, causing a downregulation. Hence lncRNAs act as competing endogenous RNAs (ceRNAs) to regulate miRNA networks.

3.2.2. Regulation modes of gene expression

A critical layer of genetic regulation is formed by lncRNAs. They largely contribute to the high organizational complexity of organisms, while maintaining a steady number of protein coding genes. Since RNAs are molecules that can be synthesized and degraded quickly and without high energy costs, they are a perfect match to act as regulators. It is known that they regulate the expression of neighboring (*cis*-active) or distant (*trans*-active) protein-coding genes, with **enhancing** or **repressive** effects on the expression of a gene.

Transcriptional regulation. Non-coding RNAs transcribed from enhancers or promoters in close proximity to the protein-coding gene can act as co-factors in ribonucleic complexes to affect transcription factors and indirectly regulate the transcriptional activity *cis*. lncRNA genes are in fact enriched in these locations.

Post-transcriptional regulation. Post-transcriptional processing is a multistep event, which provides regulation targets on a variety of different levels: mRNA maturation, including (alternative) splicing, capping and other editing; transport; translation; and degradation, including molecule stability. The natural ability of RNA to base pair with and thus recognize a complementary sequence makes RNA, especially transcribed from antisense, a highly specific potential post-transcriptional regulator.

Epigenetic regulation. The modification of chromatin states seems to be the most common way of lncRNA mediated gene regulation. They recruit chromatin remodeling complexes to specific genomic loci, which affect the methylation of histones and therefore regulate transcription. The Polycomb repressive complex 2 (PRC2) induces the repressive chromatin mark H3K27me3. lncRNAs can contribute to this silencing pathway by binding to its subunits Ezh2 and JARID2 and guiding them to a specific target chromatin, or mediating the PRC2 assembly and/or stabilizing it. The mixed-lineage Leukemia (MLL) complex, which induces the activating chromatin mark H3K4me3, can also be recruited by lncRNAs. Another way for epigenetic activation via lncRNAs is to titrate silencing factors like PRC2 away from chromatin.

See Table 3.2 for a more detailed description of the named lncRNA examples. Nevertheless, these are only a tiny fraction of all reported lncRNAs, that have been explored with respect to their function and mechanism of action, which leaves a great number of transcripts with unknown function.

Table 3.2.: Examples of important functional lncRNAs. It has been shown that lncRNAs play key roles in cellular pluripotency, differentiation and developmental patterning, dosage compensation, and genomic imprinting (reviewed in [101]). Amongst the best studied lncRNAs regarding their molecular function are Xist, HOTAIR and ANRIL.

lncRNA	Target	Level of regulation	Acts in	Effect on expression	Mechanism	Biological context
ANRIL [102, 103] alias CDKN2B-AS1; located on chr 9p21; expressed antisense of INK4A	p15 (INK4B); CDK6	epigenetic	*cis*	repressive activating	repressive: binds PRC1 and PRC2, inducing chromatin modification (scaffold, guide); activating: molecular sponge for miR-99a	targets a large number of genes throughout the genome regulation of cell proliferation and senescence; associated with cardiac diseases, diabetes and various cancers
Fendrr [104, 105]; expressed adjacent of TF gene FOXF1	FOXF1, Pitx2	epigenetic	both	dual	binds to PRC2 and/or MLL complex to induce chromatin marks	crucial for heart and body wall development in mouse
GAS5 [106]; located on chr 1q25	glucocorticoid-responsive genes	transcript.		repressive	acts as decoy for glucocorticoid receptor and stops it from binding to glucocorticoid response elements	tumor suppressor; key role in cell apoptosis and growth
H19 [107–109]; located on chr 11q15.5; maternally expressed	IGF2; miR-7 targets	multiple	both	silencing	exact mode of action is still elusive; has been reported to bind to PRC2, recruit MBD1 and act as molecular sponge for miRNA let-7	imprinting control and regulation; muscle differentiation; growth control; implicated as tumor suppressor
HOTAIR [110–112]; expressed from the HOXC locus	HOXD locus	epigenetic	*trans*	repressive	mediating chromatin modifying complexes PRC2 and LSD1-CoREST to the target (scaffold, guide, signal)	involved in distal limb development
HOTTIP [105, 113]; expressed antisense of INK4A	HOXA locus	epigenetic	*cis*	activating	binds WDR5 of MLL complex forming chromatin loops and catalyzing the H3K4me3 chromatin mark	limb morphology, including muscular and skeletal tissue
Kcnq1ot1 [114, 115]; located on chr 11p15; paternally expressed	Kcnq1	epigenetic	*cis*	silencing	recruits both PRC2 and G9a to the target gene to induce two silencing histone marks: H3K27me3 and H3K9me3 (scaffold, guide, signal)	imprinting control of paternal allele; expressed in placental tissue
MALAT1 [116] alias NEAT2; located on chr 6p24.3	motility-related, growth control genes	multiple	*trans*	activating repressive	activating: binds to the unmethylated PRC2 inducing activating acetylation marks of histone 2; repressive: sequesters SR splicing factors affecting alternative splicing (decoy)	regulates synaptogenesis, endothelial proliferation; associated with a multitude of cancer types (metastasis, cell growth, apoptosis) and other diseases, e.g. diabetes
MEG3 [117]; located on chr 14q32.3	p53; Dlk1	epigenetic	*trans*; *cis*	activating	recruitment of PRC2 and chromatin modification	putative tumor suppressor; imprinting; crucial for growth and development
PANDA(R) [118, 119]; located on chr 6p21.2 at the CDKN1A locus	Bcl-2	transcript.	*trans*	repressive	activated by p53 binding to the CDKN1A locus; titrates the nuclear TF NF-YA away from the chromatin of target genes (decoy)	induced through DNA damage; affects p53-mediated cell apoptosis and cell-cycle arrests; associated with cancer
TUNA(R) [120] alias megamind (zebrafish) [16]; located on chr 14q32.2	Nanog, Sox2, Fgf4	transcript.	*trans*	activating	binds three RBPs, guiding the complex to and occupying the promoters of the target genes	neural differentiation; maintenance of pluripotency; associated with Huntington's disease
Xist [121]; located on chr Xq13.2; expressed from the future inactive chrX	chrX	epigenetic	*cis*	silencing	it physically coats chrX and recruits PRC2, which induces the repressive chromatin marks (scaffold, guide, signal)	inactivation of the second X-chromosome in mammalian females; dosage compensation

3.3. Conservation of lncRNAs

For a large fraction of mRNAs the only evidence to deem them functional is the presence of an evolutionarily conserved ORF or the presence of an ORF that would translate into an amino acid sequence similar to known protein or at least containing known protein domains. In other words, homology and conservation are accepted as indicators of biological function and relevance in the world of protein coding genes and for the highly conserved families of ncRNAs such as miRNA, snoRNAs, snRNAs, tRNAs, and rRNAs.

Extending this reasonable standard to the entire transcriptome, this section reviews the available evidence for the evolutionary age and conservation patterns of lncRNAs

3.3.1. Primary sequence

Naturally, ncRNAs do not obey the same evolutionary constraints as protein-coding transcripts. Instead of featuring high sequence conservation, the majority of non-coding genes has highly variable intronic and exonic sequences, gene length and structure, as well as transcriptional start sites. In view of this fact, established ncRNA orthologs among amniotes are rare. Only a small subset of lncRNAs shows levels of sequence conservation comparable to protein-coding genes or some of the evolutionarily old, well-conserved families of ncRNAs.

MALAT1 is one of the best-conserved lncRNAs and regulates alternative splicing as well as gene expression [116]. It is conserved throughout the jawed vertebrates, but may have been lost in birds [122]. MALAT1 shares several characteristics, including nuclear retention [123] and a non-standard processing of its 3' end [124], with the longer, but less well conserved, eutheria-specific MENβ RNA. The Xist RNA, which is the key player in X chromosome inactivation [125] in eutheria, originated from the pseudogenization of the ancestral LNX3 protein-coding gene under inclusion of several transposable elements [126, 127].

The lncRNA TUNA was discovered in zebrafish as *megamind* [16, 120]. TUNA is involved in brain development but also expressed in spinal chord and eye tissues. The exonic regions of TUNA feature atypically strong sequence conservation across vertebrates. In particular, it contains a sequence element of $\sim 200\,\text{bp}$ length with $> 80\,\%$ sequence similarity between human and zebrafish. This level of conservation exceeds that of most coding regions. Well-studied functionally important ncRNAs with orthologs over a wide phylogenetic range of species include genes such as Fendrr, Braveheart, cyrano, and Evf-2 (reviewed in [101]) as well as H19X [28]. A computationally generated high quality set of 233 constrained lncRNAs was recently reported

in [128].

Ultraconserved regions (UCRs) are genomic segments that are highly conserved across almost all vertebrates, with a remarkable 100 % sequence identity between human, rat and mouse [129]. The majority of non-coding UCRs is transcribed into RNA (T-UCRs) [130] and gives rise to lncRNAs. It remains unclear, however, whether the extreme conservation of UCRs is caused by direct selection pressure on their RNA products.

The majority of lncRNAs, exhibits very little measurable sequence conservation. This does not imply that they are lacking function. Indeed, there are good examples of lncRNAs without substantial sequence conservation but unambiguous biological functions, like ANRIL [102] and GAS5. The latter harbors about ten distinct snoRNAs in its introns [131], which makes it possible to track the gene throughout the vertebrates. The very poorly conserved exonic product acts as a riborepressor blocking the DNA-binding domain of the glucocorticoid receptor [106, 132].

In extreme cases, rates of sequence evolution are even faster in functional RNA genes than in neutrally evolving genomic background. In [133] segments in the human genome were identified that show an atypically strong sequence divergence between human and chimp that at the same time are highly conserved between chimp and non-primates. Of these "human accelerated regions", 96% are located in non-coding regions [133, 134]. The most famous case, HAR1, might be of importance in the evolution of the human brain [135]. We refer to [136] for a recent review on HARs. Although it remains a matter of debate whether the accelerated rate of evolution is caused by positive selection or is the consequence of compensatory substitutions [137], examples like this emphasize that a lack of nucleotide-wise sequence conservation cannot be used as proof for the absence of function.

A series of global statistical analyses [19–22] showed that a large fraction of lncRNAs is under stabilizing selection. The measured levels of conservation, however, are much smaller than for protein-coding genes. A comparison of human lncRNAs with 18 mammals [80] used `exonerate` [138] to map human lncRNAs to genomic regions identified by `blast` as candidate orthologs and counted a human lncRNA as conserved when 70 % were recovered by `exonerate`. An estimated 44 % of the GENCODE 7 lncRNA set was found to be conserved across the major groups of placental mammals [80].

Estimates based on the average sequence conservation of a lncRNA locus may be criticized, however, because there is no guarantee that the observed selection pressure really acts on the RNA. Instead, "phylogenetic footprints", that is well-conserved local elements, might also function as transcription factor binding sites at the DNA

level. The HOX clusters serve as a particularly impressive example. On the one hand, many well-characterized mRNA-like lncRNAs [110, 139], including HOTAIR [110, 140], HOTTIP [113] and several microRNA precursors [141], are transcribed from the intergenic regions. On the other hand, the same region is also packed with conserved functional DNA elements [142–144]. Hence, the observable conservation of genomic sequence does not in itself provide sufficient information to disentangle the evolutionary history of lncRNAs. In other words, the fact that the genomic sequence of a lncRNA exon is sufficiently conserved to be identifiable by `blast` or `infernal` cannot be taken as adequate evidence that the exon is conserved. For instance, [145, 146] report the conservation of all but one exon of HOTAIR between man and mouse. Detailed sequencing data [112], on the other hand, indicate that mouse HOTAIR, like kangaroo HOTAIR [146], completely lacks the first three exons. Later we will show that the method developed in the course of this thesis recapitulates this observation by considering the conservation of splice sites only (Figure 6.7B).

Stabilizing selection on lncRNAs also reveals itself as a significant, albeit sometimes weak, contrast of conservation levels between exonic and intronic sequence. A detailed analysis of, for instance, mammalian lncRNAs uncovered an increased GC content in exons compared to introns [147]. Several studies reported strong negative selection on the promoters of lncRNAs [21, 22]. Based on `PhastCons` scores [148], lncRNA promoters match the conservation levels of protein-coding genes [80]. A good example is the lncRNA *RMST*, which plays an important role in neuronal development [149], and is highly conserved at least among tetrapods [150]. We later show, that this can also be confirmed by tracing the conservation of its splice sites (Figure 6.7D).

3.3.2. Secondary structure

Many well-studied ncRNAs exhibit well-conserved RNA secondary structures. Well known examples are the many families of structured RNAs compiled in the `Rfam` database. It comprises both independent ncRNA genes and a large collection of structured RNA elements that function as part of larger transcripts. Examples of the latter are internal ribosomal entry sites (IRES), selenocystein insertion elements (SECIS), aptamer domains of riboswitches, or the autoregulatory domains of many of the mRNAs that encode ribosomal proteins [151]. The non-coding RNA MALAT1 has a conserved cloverleaf structure at the 3' end of the transcript [152]. Importantly, the mere presence of stable secondary structures cannot be taken as an indication of biological or molecular function: Random RNA sequences may also fold into highly complex and stable structures that statistically are no different from known functional secondary structures [153, 154]. It is necessary therefore to assess the evolutionary

conservation of the secondary structures.

Over the last two decades a variety of computational methods have been developed to identify negative selection on RNA secondary structure, i.e., the preservation of base pairs, and to distinguish it from selection pressure acting to maintain the sequence. They fall into two broad classes. Most tools are alignment-based. These include `qrna` [155], `AlifoldZ` [156], `EvoFold` [157], `RNAz` [158], or `SISSIz` [159], and alignment-free methods such as `CMfinder` [160]. Although the technical details differ widely [161], the basic idea is the same: characteristic properties of the input alignments are measured and compared to the prediction from a background model that for a given level of sequence conservation assumes that there is no conservation of structure. The discrepancy between background prediction and foreground measurement is then converted into measure of selection on the secondary structure. By construction alignment-based approaches depend on the reliability of the multiple sequence alignment that is used as input. Hence they are limited in practice to regions that are at least moderately conserved also at the sequence level. In a genomic screen, transcript boundaries are usually unknown, hence sliding windows are used, leading to an unavoidable increase of noise in the predictions.

Alignment-free screens start from homologous sequences that have been identified based on synteny [162, 163], i.e., the order-preserving arrangement of closely-spaced homologous genomic elements. Then structure-based alignments are computed as best estimates for conserved RNA structure. For this task, `Foldalign` has been used in [162]. More recent screens [163] used `CMfinder` [160]. As in the alignment-based approaches, the predictions are compared to randomized controls to determine cutoff levels and to estimate false discovery rates (FDR). Alignment-free screens on very poorly conserved regions still reported large numbers of sequence elements that appear to be under stabilizing selection for RNA secondary structures. This emphasizes the point that a lack of observable constraints on individual nucleotides does not necessarily imply a lack of selective pressure on an entire sequence element.

All computational methods that measure selection on secondary structure are sensitive to modification of the background model, and thus are plagued by relatively high FDRs. It should be noted that FDR estimates are not without problems as well as they are obtained from re-running the screen on a computationally randomized control. The choice of the background model therefore influences the reported FDR values. Surveys conducted with different tools, show little overlap, see e.g. [166], where `RNAz` and `EvoFold` is compared on the ENCODE regions (Table 3.3). While this observation appears to speak against the reliability of the available computational methods, a closer inspection shows that the lack of overlap has a simple explanation: the sensitivity of the methods depends strongly on sequence conserva-

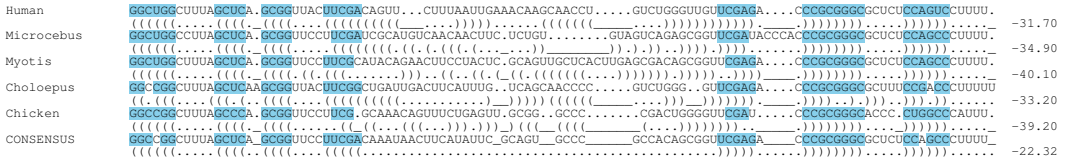

Figure 3.3.: Conservation pattern of amniote vault RNAs [164]. The very uneven conservation pattern, here a well-conserved stem structure at the ends and highly variable interior regions is typical of many evolutionarily conserved RNA elements. Secondary structure predictions for each sequence and the consensus structure of the alignment computed with `RNAalifold` [165] are shown in "`ViennaRNA` notation": matching pairs of parentheses denote base pairs, dots indicate unpaired bases. Bases pairs present in the consensus marked in color. Note that the consensus sequence (defined as the majority vote over an alignment column) does not fold into the consensus structure. The energy of the consensus structure (-22.32 kcal/mol) differs substantially from the average folding energy of the unconstrained sequences (-35.84 kcal/mol). The ratio, here 0.623, serves as a statistically robust measure of structure conservation e.g. in `RNAz` [158].

tion and sequence composition. While `EvoFold` works best on very conserved AU-rich sequences, `RNAz` is most reliable on moderately conserved GC-rich sequences. In a recent, very detailed analysis [167], different tools are therefore combined to a meta-method that selects the best individual tool for given input parameters. This reduced the estimated false discovery rate to only $5-22\,\%$. In [167], more than 4 million structured RNA components were identified that are conserved in mammals. This yields an estimate of $13.6\,\%$ for the fraction of the genome with selective constraints on RNA structure. Of these, $88\,\%$ fall outside the sequence-constrained regions.

`CMfinder` [160] infers covariance models from unaligned sequences as a means of detecting secondary structure conservation. It therefore does not pre-suppose significant sequence conservation. On a test set of 19 known ncRNA families from `Rfam`, with randomly generated flanking sequences (200 nt), `CMfinder` yields more accurate motif predictions than RNA alignment tools such as `RNAalifold` [170], `Pfold` [171], `Foldalign` [172], in particular for very short elements and for conserved RNAs with low sequence similarity. Because of the latter, `CMfinder` is well suited for identifying secondary structure conservation in lncRNAs.

In [163], `CMfinder` was employed for a large-scale screen of the ENCODE regions. Here, `multiz` alignment blocks were used to identify syntenic genomic locations. The screen excluded known coding exons as well as conserved regions as defined by `PhastCons`. They report 4933 candidates in non-repetitive regions detected with `CMfinder`, compared to 3134 and 3267 of `EvoFold` and `RNAz` predictions from [166] (for comparability a posteriori filtered in [163]). $78\,\%$ of the predictions in [163] are complementary to the predictions in [166], adding a total of 3861 `CMfinder` predicted candidates to the set of the filtered `EvoFold` and `RNAz` predictions. When includ-

Table 3.3.: Overview of the latest screens for conserved secondary structures in the human genome. The numbers are directly taken from the publications specified in column "ref". Note that not all inputs/predictions have been filtered for coding regions. The screen in [167] for example was applied on a genome wide input, with a compound of SISSIz, RNAz and other prediction programs. Some numbers (*) have been recalculated or converted from fraction to Mb, or vice versa (based on a human genome size of 3095 Mb and a *D. melanogaster* genome size of 120 Mb) , to fit the measuring units of our table. As for the RNAz results we only show the number of high confidence ($p > 0.9$) loci.

Species	Method	Input	Input (Mb)	Loci	% Input	% Genome	Ref.
hg17	RNAz	PhastCons conserved (excl. coding regions)	82.64	35,985	6.6	1.76*	[168]
hg17	RNAz	ENCODE (excl. repeat regions)	9.76	3707	4.2	0.01*	[166]
hg17	EvoFold	ENCODE (excl. repeat regions)	14.44	4986	2.5	0.01*	[166]
hg18	CMfinder	ENCODE (excl. PhastCons conserved, incl. repeat regions)	8.68	6587	6.1	0.02*	[163]
hg19	SISSIz +	EPO alignment (35 eutherian mammals)	∼ 2600*	> 4 M	18.5	13.6	[167]
Drosophila	RNAz	ENCODE (excl. 5S rRNAs, SRP RNAs)	57.4	16,377	3.8*	1.75*	[169]

ing repeat regions into the screening, 1654 further candidates have been found with `CMfinder`, adding a total of previously uncovered 5515 candidates to the comprehensive `RNAz/EvoFold` set of 17,046 candidates, extending it by 32 %. The total of 6587 predicted candidates spans 0.53 Mb, which equals 6.1 % of the input sequence, where in non-repetitive regions twice as many candidates are detected than in repetitive regions (7.9 % vs. 3.9 %).

The high false discovery rates have promoted several authors to devise postprocessing methods. In the simplest case, stringent filters are used as in [173]. Structure-based re-alignments with `LocARNA-P` [174] and a consistency-based scoring scheme that measures structure-based alignment reliabilities provide much more accurate boundaries of regions with evolutionarily conserved secondary structures and considerably reduce the false positive rate. The `REAPR` tool [175] achieves a substantial increase in computational efficiency for such approaches.

A major issue with the various computational screens for conserved RNA structure (see Table 3.3) is the disappointingly small overlap of the actual predictions. This is readily explained, however, by the very different characteristics of the secondary structure inference tools. In particular their sensitivities strongly depend on GC content and sequence conservation. The predictions therefore have to be expected to be largely complementary. Despite substantial false positive rates, taken together they demonstrate that evolutionarily conserved structured RNA is an abundant genomic

feature.

There is a correlation of conserved structures and their distance to the nearest protein-coding genes. Predominantly, the structurally conserved elements are 50 kb up- or downstream from the next protein-coding element, indicating their potential role as molecular functional *cis*-regulators of those genes.

RNA secondary structures can be modular and very complex. It is likely that the diversity in functionality and modularity influences the conservation pattern of specific regions within the lncRNA. This in turn impacts the sensitivity of computational methods, contributing to the discrepancies between different screens. Finally, all currently available approaches focus on the base paired regions. Linker sequences, in which a depletion of base pairs may be a conserved feature, therefore are likely to remain unnoticed.

A comparison of predicted secondary structures in human and murine transcripts showed little differences between mRNAs and lncRNAs [176]. The survey [167] reported a small but significant (1.4-fold) enrichment of conserved structures in lncRNAs. It appears, however, that this enrichment is not uniform. While functionally important secondary structure elements have been suggested for several lncRNAs, the majority of annotated lncRNA is not enriched in evolutionarily conserved RNA elements [121, 177]. In fact, the relative enrichment of secondary structure elements in protein-coding exons is more than twice as strong, possibly reflecting the importance of structured elements in post-transcriptional regulation.

So far, systematic investigations into the conservation of RNA secondary structure are based only on computational methods. Recently, several experimental techniques to assay RNA secondary structures on a genome-wide scale have become available, see [178, 179] for timely reviews. At the time of writing, however, these have not been employed in a comparative context.

An interesting special case are transcripts with dual functions as both protein-coding and non-coding RNAs. The paradigmatic example is the *steroid receptor RNA activator* (SRA), which produces both a well conserved protein and an elaborately conserved secondary structure [180–183]. As a lncRNA it coactivates steroid nuclear receptors and also participates, like many others, in chromatin bases gene regulation [184], while the protein product SRAP appears to function by stabilizing specific intermolecular interactions in the nucleus [185]. It is unclear at present how wide-spread such cases are. There is, however, statistical evidence that conserved, structured RNA elements are quite frequently superimposed on coding sequences [186, 187]. Experimentally studied examples, such as *oskar* [188] in fruitflies are very rare.

3.3.3. Gene structure

Spliceosomal splice sites constitute highly conserved sequence motifs that can relatively easily be recognized in genome DNA sequences by various statistical pattern search methods [189, 190]. Evolutionarily conserved splice donors and acceptors are therefore identifiable in genome-wide multiple sequence alignments. In combination with additional machine learning techniques that evaluate the sequence between consecutive splice donor–acceptor pairs, either short introns [31] or exons [32] have been used successfully to find evolutionarily well-conserved lncRNAs. With the availability of large sets of transcriptome sequencing data this genome-centered approach has become obsolete as means of genome annotation. It serves as a demonstration, that conservation of splicing patterns can be used to establish orthology of lncRNAs that are otherwise not sufficiently well-conserved at sequence level.

The host genes of snoRNAs and microRNAs form a special class of lncRNAs whose evolution can be studied with relative ease: their payloads, the microRNA precursor hairpins and the snoRNAs, respectively, are typically very well conserved and, despite their small size, can be traced at least at phylum level, see e.g. [191, 192] and the references therein. Although snoRNAs and miRNAs are known to be mobile to a certain extent, their associations with coding and non-coding host genes are evolutionarily stable at long time-scales [193–195]. Consequently it is possible to identify putative orthologs in distantly related species. It is not surprising, that many of the non-coding host genes such as UHG (SNHG1), U87HG [196], or GAS5 [131] also exhibit deeply conserved gene structures.

Among more distant species orthology of lncRNAs cannot be established unambiguously due to rapid sequence divergence. Several authors noted that lncRNAs can often be found at syntenic positions [16, 30, 197]. These also seem to have significantly correlated expression patterns which may hint at analogous functions. Due to poor sequence similarity `blastn` fails to identify conservation between the orthologs of Miat/Gomafu/Rncr2 in human, mouse, frog and chicken. However, the syntenic position of the locus strongly suggests that they are indeed homologs [99]. This is supported by the presence of multiple copies of a short motif within the last exon of the Miat transcripts in all species.

3.4. Evolution of lncRNAs

Taken together, the various threads of evidence outlined in Section 3.3 show that many lncRNAs indeed convey selectable functions whether or not these selective constraints result in levels of nucleotide-wise sequence conservation that is detectable

with classical measures. It is of key interest, to obtain an overview of the actual numbers of lncRNAs that are under measurable evolutionary constraints and to estimate their evolutionary age. At present, estimates from various studies still differ quite a bit, but there appears to be an emerging consensus that conserved lncRNAs are numerous and much older than their poor sequence conservation might suggest.

In an RNA-seq based study covering eleven tetrapods with a total of 185 samples of eight tissues $> 13,500$ multi-exonic homologous families of lncRNAs were identified [28]. Of these, about ~ 2500 families are highly conserved, dating back at least to the eutherian ancestor some 90 million years ago. More than 400 lncRNAs could even be traced back 300 million years ago. However, the majority (81 %) were reported as primate-specific, classifying only the remaining 19 % as conserved beyond primates. Orthology assignments in this study were based on sequence similarity recognizable by pairwise `blastn` comparisons with an additional assessment of synteny. The numbers therefore have to be regarded as lower bounds on the conserved part of the mammalian lncRNA system.

In a direct comparison of transcriptome sequencing data for six mammalian species comprising nine tissues, [27] showed that $30 - 40$ % of nearly 2000 lncRNAs exhibit conservational expression patterns between human and rodents and/or ungulates. In accordance with [28] 80 % of the human lncRNAs had orthologs in chimpanzee. The identification of orthologs between human and each other species, was accomplished by employing genome-wide pairwise alignments from the UCSC genome browser. When looking at the level of splice sites, the rate of primary sequence conservation was significantly higher than for complete lncRNA transcripts (in rhesus 90 % of splice sites vs. 63 % of lncRNAs, in rat 62 % vs. 35 %). By comparing a `cufflinks` [198] generated transcript annotation made from the RNA-seq data sets of rhesus, cow, mouse and rat, with the human GENCODE annotation, $40 - 73$ % of all cufflinks-constructed non-coding exons were found to be expressed in human.

The comparison of publicly available data of > 4000 lncRNAs of human and mouse to estimate the size of the mammalian lncRNome via a maximum likelihood approach, resulted in a prediction of $40,000$ to $50,000$ lncRNAs of which about $30,000$ ($60 - 70$ %) are conserved between man and mouse [29].

Due to their close proximity and relevance for protein-coding genes, individual splice sites within untranslated regions (UTRs), as expected, have significantly higher levels of conservation ($\sim$ 52 % and $\sim$ 62 % for 5'UTRs and 3'UTRs respectively) [35]. Like other estimates of conservation, estimates based on conserved splicing patterns suffer from the uneven quality of genome-wide multiple sequence alignments. Since coding exons provide a dense set of high quality anchors, they are better and

more complete in regions containing much coding sequence. The conservation of "intergenic" transcripts is therefore systematically underestimated.

Although lncRNAs typically have a relatively low expression rate compared to protein-coding genes [28, 80], their expression shows very distinctive spatio-temporal patterns, featuring a substantially higher tissue-specificity than mRNAs [15]. The tissue-specificity is conserved in all primates in 47 % of the cases, and in 28 % throughout the eutherian clade [28]. The extent of spatial conservation, is significantly lower than in mRNAs. Changes in the tissue-specificity seem to be common. The lncRNA H19X, for example, is predominantly expressed in placental tissue in human and mouse, while in opossum it is highly expressed in testis [28].

Not only the expression patterns of lncRNAs are subjects to rapid evolutionary turnover. This effect can also be observed in the evolution of non-coding gene structure. When tracing back the conservation of splice sites it becomes obvious that non-coding gene structures evolve rapidly. While in about 35 % of the cases at least one splice site of a non-coding transcript can be traced back to mouse, less than 13 % of the entirety of all non-coding splice sites -present in human- can still be found in mouse [35]. This is also visible in the example of HOTAIR, see Figure 6.7B.

3.5. Perspective

Diverse patterns in lncRNAs evolution match the observation that lncRNAs are by no means a homogeneous group but apparently comprise transcripts with very different fates, interactions, and biological functions. Nevertheless, there are some commonalities that make it meaningful to study them as a group. There is mounting evidence that thousands and maybe tens of thousands of lncRNAs are subject to some selective constraints on their gene structure, including promoters, and their splicing patterns in addition to commonly very weak or even undetectable selection pressures on their sequences. Expression patterns are frequently very specific to tissues, cell types, and developmental stages, and are often conserved across species. It is worth keeping in mind that quantitative estimates of lncRNA conservation suffer from multiple sources of errors and biases that limit their accuracy. These include (i) the uneven quality of reference genomes, (ii) the decreasing sensitivity of computational homology assignments with phylogenetic distance, (iii) ascertainment biases in certain model organisms such as human, mouse, or fruitfly for which much more data are available, and (iv) limits to the detectability of transcripts specific to rare cell types in complex tissues such as brain.

Available data suggest an extensive turnover of entire lncRNA gene structures and

at a rapid evolution of individual transcripts. This view, however, may to a certain extent also be confounded by biases in the data. The well-documented specificity of spatio-temporal expression patterns makes it difficult in practice to find perfectly matching RNA-seq data sets for cross-species comparison. The relatively low expression levels of many lncRNAs in most samples presumably reduce the congruence even further because transcript reconstruction pipelines frequently retrieve only fragments rather than complete lncRNAs. Our current inability to adequately infer lncRNA functions from sequence features, furthermore, does not allow us to zoom in on important parts of a transcript to reduce the noise in evolutionary comparisons. It is not unlikely, that the current view on evolutionary age and turnover is biased towards inferring ages that are too young and conservation levels that are too low.

The mounting evidence for selection beyond nucleotide-wise conservation [199] implies that there is need for statistical and computational methods to assess constraints on secondary structure, distances between recognizable anchor points, and similar features not only for deep phylogenetic conservation but also at a population level. Methods to estimate the effects of SNPs on RNA secondary structure [200–202] are a promising first step, but by no means provide a satisfactory way of measuring selection pressures on non-coding regions.

Part II.

Methodology

Contents

Chapter 4

Technical background

This chapter will give an overview on the technical principles underlying the developed method. In particular, Section 4.1 elucidates the algorithmic concepts of multiple sequence alignments. The section is based on the review of Chatzou *et al.* [203] and the textbook of Böckenhauer and Bongartz [204], which are referred to for further reading.

In addition, we will give a more detailed description of the `MULTIZ` [205] and `EPO` [206, 207] software employed by the UCSC Genome Browser[1] [208] and the Ensembl Project[2] [209, 210] to generate the two major file formats that we use as a basic building block of our pipeline. Hereafter we will refer to those online databases as UCSC and ENSEMBL, respectively.

In Section 4.2 we will explain the mathematical model of maximum entropy behind the `MaxEntScan` software by Yeo and Burge [211] as an essential tool that we employ to assess the conservation of splice sites.

[1] http://genome.ucsc.edu/
[2] http://www.ensembl.org/

4.1. Mulitple sequence alignment

Multiple sequence alignments (MSAs) are a fundamental tool for comparative genomics analyses. A robust accurate alignment is crucial for the correctness of the predicitions made in the course of this work.

The purpose of MSAs in general is to determine the similarity between sequences of RNA, DNA or amino acids. Their potential scope ranges from phylogenetic tree reconstruction, RNA structure predicitions and identifying functional features of proteins, such as catalytic sites, target signals or specific domains. We will use MSAs to draw conclusions about the conservation of splice sites between certain organisms.

4.1.1. Sequence alignment methods

A sequence alignment is a rectangular arrangement of two or more sequences so that similar features are aligned in one column to reflect their evolutionary relationship. The goal is to maximize the sum of similarities by inserting gaps into the sequences so that homologous positions are aligned with each other. The resulting pattern is a combination of four possible operations per aligned column.

Insertions. A gap is present in the upper sequence, but not in the lower one.

Deletions. A gap is present in the lower sequence, but not in the upper one.

Matches. Both sequences share the same nucleotide.

Substitutions. The sequences have mismatching nucleotides in this colum.

Example 4.1 (Pairwise Alignment)**.**
This is a possible alignment of sequence `GACTAGGTCACAG` and `GTAGATCATCA` with the respective operations for each aligned pair (represented by the leading letter of the operation).

```
Sequence 1: GACTAGGTCA-CAG
Sequence 2: G--TAGATCATCA-
  (Pattern: MDDMMMSMMMIMMD)
```

These operations hypothesize the events that occured during the evolution from a common ancestor. To find the mathematically optimal MSA, the possible operations are dynamically computed based on a scoring model, which assigns gap penalties and substitution costs. This scoring function can be rather simple by giving a single penalty score per gap column, or more sophisticated by introducing higher penalties

for gap opening than for gap extensions, or by favoring/penalizing certain substitutions based upon biological probabilities (e.g. `PAM` or `BLOSUM` matrices).

To this day a multitude of heuristics, adapting and extending the basic global or local dynamic programming algorithms (Needleman-Wunsch [212], Smith-Waterman [213] or Viterbi [214]), have been developed to gradually built up MSAs from pairwise sequence alignments. Commonly, these heuristics are all based on the progressive alignment approach [215, 216]. These algorithms follow a certain phylogenetic order when adding new sequences to the alignment. The order is based on a guide tree, whose computation via an estimated distance matrix is an essential step of the algorithm itself. The most famous representative of MSA methods is probably `ClustalW` [217].

Developing and improving MSA methods remain to be a very active field of research. Its biggest challenge is to balance out evolutionary accuracy with computability. There is an increasing amount of available data such as structural dependencies, phylogenetic relationships, substitution matrices or other data of biological context that can be incorporated in the MSA computation. This, however, in turn increases speed and memory usage exponentially. One approach to improve the accuracy of MSAs is to deal with local minima by introducing consistency to the alignment method. Consistency-based algorithms reestimate the costs of all possible sets of pairwise alignments to be in the best agreement with the optimal multiple alignment. The archetype of these tools is `T-Coffee` [218] and its probabilistic variation `ProbCons` [219].

4.1.2. Multiple whole genome alignment methods

When dealing with whole genomes, the task of alignment computation is stocked with a variety of new challenges. The extreme long sequences are highly heterogenous in function and conservation rate and including structure data on a genome scale is not possible. To account for the more complex evolutionary events on a larger scale, at least three additional operations have to be considered in the alignment:

Duplication. The repetition of a sequence segment (e.g. gene, exon) occured.

Inversion. A sequence segment has been reversed.

Translocation. Sequence segments have been exchanged between distant parts.

These operations are incorporated into a segmentation step previous to the alignment procedure in which the genomes are split into bins of homologous fragments.

Since our method relies greatly on multiple whole genome alignments provided

through UCSC and ENSEMBL, this section discusses the basic principles of the algorithms applied by both online databases to create the respective multiple genome alignments (MGA).

MULTIZ

The MGAs provided by UCSC are generated with an independently operating component of the threaded blockset aligner (`TBA`) software by Blanchette *et al.* [205]. `TBA` is one of the first programs that splits a whole genome alignment problem into a set individual distinct local alignment blocks. The `MULTIZ` program performs the dynamic programming step on those sets. To describe the method of `MULTIZ` adequatly, we first have to introduce the vocabulary.

Block and blockset. An optimal local alignment between two or more sequences is a block. Not necessarily all given have to be included in every block. A group of two or more blocks is a called blockset.

Ref-blockset. A designated "reference" sequence, is present in each block of the blockset, always in the first row with a positive orientation. Each position of the reference sequence appears exactly once throughout the blockset, averting overlapping regions between blocks.

Thread. Sequence S "threads" a blockset, if each position of S appears exactly once throughout the blockset. A ref-blockset is always threaded by its reference sequence.

Threaded blockset. A blockset is threaded by all of its sequences. A ref-blockset for any chosen reference sequence S from the threaded blockset, can be achieved by projection onto S (`maf_project`).

`TBA` generates a threaded blockset for a given set of sequences. While `TBA` does not take into account inversion or duplication events, `MULTIZ` can deal with this special events. It dynamically computes the alignment for three or more sequences, based on pairwise alignments generated by `BLASTZ` [220].

The biggest difference to a common alignment program is that `MULTIZ` is able to merge two existing MSAs, present as ref-blocksets, into one bigger MSA rather than just aligning single sequences. This process is guided by a `BLASTZ` pairwise alignment between the reference sequences of both ref-blocksets (Figure 4.1). The two MSAs are treated as sequences (of columns) for which a pairwise alignment is generated algorithmically similiar to the progressive methods described in Section 4.1.1.

To generate the union U of an S-ref blockset M and a T-ref blockset N, `MULTIZ`

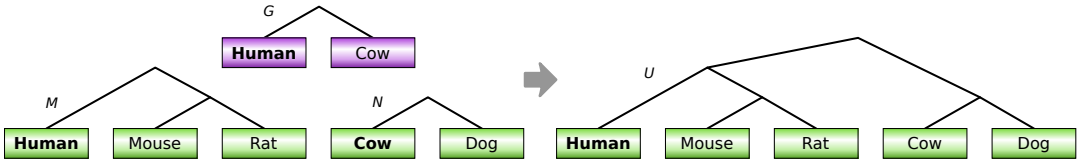

Figure 4.1.: Generation of a `MULTIZ` alignment. Human-ref blockset M and cow-ref blockset N are merged into a new MSA, guided by human-ref blockset G (a pairwise alignment generated with `BLASTZ`). The resulting human-ref blockset U contains all sequences (species) from both M and N. The reference sequence for each ref-blockset is written in bold letters. Figure adapted and redrawn from [205]

uses the guiding pairwise S-ref blockset to find a combination of blocks (g, m, n) and positions (w, x, y, z) in a maximal segment of $[w, x] \in S$ and $[y, z] \in T$, so that:

1. w and x are in the same block $m \in M$
2. y and z are in the same block $n \in N$
3. (w, y) and (x, z) are aligned pairs in the same block $g \in G$

When such a combination is found the respective columns, that are aligned to $[w, x] \in m$ and $[y, z] \in n$ get aligned in a new block $u \in U$. This method makes realigning unnecessary.

The algorithm proceeds in order of the positions in S, where the alignment G is used to translate the current position of S to the aligned positions of T and therefore to the corresponding segments in N. In the example of Figure 4.1 human is the reference sequence S. An algorithm called `stageMULTIZ` uses the alignment G between human and cow to compute the distinct non-consecutive segments of the cow-ref blockset N, which have to be compared to the human-ref blockset M by `MULTIZ`, in order to find the right combination of (g, m, n) and (w, x, y, z). Columns that have not been aligned per input block are reported. For further details on the algoritm, see the publication of Blanchette *et al.* [205].

The output produced by `MULTIZ` is in `MAF` format and is a required input file for our conservation analysis method, *cf.* Section 5.2.1.

EPO

The ENSEMBL alignments are generated by a pipeline called `EPO`. The three consecutively executed programs of `Enredo`, `Pecan` and `Ortheus`, first applied to the input genomes and then feeding into each other, create a whole-genome multiple alignment. Here we use slightly different vocabulary.

Segment. An unoriented contiguous sequence of DNA basepairs of a single input

genome.

Directed segment. Segments that are oriented in relation to another segment.

Segment-group. A group of homologous colinear aligned directed segments.

In the first step, `Enredo` splits the whole input genomes into groups of non-overlapping homologous colinear segments based on a segmentation graph, which identifies complex rearrangement events, like duplications and deletions. The construction of the used graph resembles the `Mercator` orthology constructor [221]. Edges are build to represent homology between genomic regions. A set of non-overlapping short segment-groups, computed by a local-alignment program, is used as "genome point anchors" to construct the initial segmentation graph. This is followed by non-trivial modifications of merging and removing edges to build the final graph. We refer to the original publication of Paten *et al.* [206] for a more detailed description of this program.

The resulting segment-groups are then given to `Pecan`, which aligns the colinear segments from each group through a probabilistic consistency-based alignment method. It combines the same underlying objective function as in `ProbCons` ([219], and briefly reviewed in [206]) with a framework of a constrained MSA [222] to work on a larger scale.

In the final stage, `Ortheus`, an evolutionary alignment modeller, generates a genome-wide ancestral sequence reconstruction. By using a phylogenetic tree as additional input to the MSA – produced by the previous steps – the program infers the evolutionary history of the MSA. The algorithm is based on a probabilistic progressive alignment variation of the Forward algorithm by Durbin *et al.* [214], which incorporates the generation of sequence graphs. This enables the method to distinguish insertion from deletion events. For a detailed description of the algorithm, we refer to the method paper of Paten *et al.* [207].

4.2. Maximum entropy models of RNA splice sites: MaxEntScan

The necessity to assess whether a found splice site ortholog is likely to be functional or not, prompted us to use `MaxEntScan`, a strong statistical tool introduced by Yeo and Burge [211]. In order to estimate the likelihood of a proper splicing signal, the program assigns a log-odd ratio score (`MaxEntScan` score) to the splice site sequences based on probabilistic models of acceptor and donor motifs. These models are developed using the "Maximum Entropy Principle" (MEP), which in contrast to position

weight matrices or (inhomogeneous) Markov models, accounts for both adjacent and non-adjacent dependencies between positions. The resulting gain in accuracy has been shown to reliably predict mis-splicing mutations [189]. The `MaxEntScan` score indicates the degree of similarity to a typical canonical major spliceosome splice site motifs. A higher score means higher probability of a true splice site and also indicates a strong splice junction.

In this section we will review the work of Yeo and Burge [211]. We briefly introduce the concept of MEP and explain how maximum entropy models (MEM) of splice junctions have been developed from constrained maximum entropy distributions (MED) of donor and acceptor motifs from known human transcript data.

We use the same variables as defined in [211] to describe the mathematical model:

$X = \{X_1, X_2, \ldots, X_\lambda\}$ denotes a random sequence of length λ, whose values are taken from the nucleotide alphabet $\{A, C, G, T\}$.

$x = \{x_1, x_2, \ldots, x_\lambda\}$ is a specific DNA sequence.

$p(X)$ represents the joint probability distribution $p(X_1 = x_1, X_2 = x_2, \ldots, X_\lambda = x_\lambda)$.

$P(X = x)$ is the probability of a state in the distribution.

4.2.1. Maximum entropy method

The MEP, first introduced by Jaynes [223, 224], states that, given a set of possible distributions, the best approximation of the true distribution is the distribution with the highest Shannon entropy

$$H(\hat{p}) = -\sum \hat{p}(x) \log_2 \hat{p}(x)$$

that is the sum over the probabilities of all possible sequences, x. The Shannon entropy can be seen as a measure of the uncertainty in X. By maximizing the entropy, we choose the least informative possible distribution, which assumes nothing about the world that is not known, given a set of constraints on sequences of length λ.

When a background distribution q is known, the Kullback-Leiber divergence can be used to estimate the logarithmic difference between probability $p(x)$ and $q(x)$. This is the minimum relative entropy (MRE) principle, where the distribution is chosen by the lowest relative entropy,

$$D_{KL}(\hat{p}) = \sum \hat{p}(x) \log \frac{\hat{p}(x)}{q(x)}$$

This is equivalent to a maximal Shannon entropy H, when q is a uniform distribution of all sequences.

4.2.2. Marginal constraints

The set of constraints used to determine the MED is derived from the marginal nulceotide frequencies of the emprical distributions. They are expected values or bounds on these values, which are consistent with features of the empirical distribution. Therefore, they represent statistics about the true distribution without assuming more than what can be reliably estimated from the available data. To form the desired specific MED, an initial uniform distribution, where all sequences are equally likely, is altered to satisfy the well-estimated constraints. It is distinguished between complete and specific constraints.

Complete constraints define position dependencies. Let S_x be a set of all lower-order marginal distribution of $p(X)$, which is a joint distribution over a proper subset of sequence X. The subsets $S_s^m \subseteq S_x$ are complete constraints specified through marginal-order m, and skips s of the distribution. The first-order constraint, S_0^1, always represents the empirical frequencies $p(X_i)$ of $\{A, C, G, T\}$ at all positions i in sequence X. The second-order constraint, S_s^2 additionally determines the dependencies in all possible dinucleotide combinations with neighboring distance s.

Specific constraints are requirements of a specific nucleotide frequency at certain sequence positions derived from the observed frequency of an element for the respective member of a complete constraint. For instance, $p(X_1)$ as a part of a complete first-order constraint, has four specific constraints: $\{A, C, G, T\}$. One for each possible nucleotide at position X_1. The number of possible specific constraints increases exponentially with the marginal-order m. In general a member of S_s^m will have 4^m specific constraints.

4.2.3. Maximum entropy model

It is now possible to distinguish between decoys and true signals by generating an MEM based on a distribution derived from probability distributions of true signals and decoys with a chosen set of constraint. The iterative scaling of an initial distribution with a ranked set of constraints specifies the MEDs. This step is described in more detail in [211].

When $P^+(X = x)$ and $P^-(X = x)$ are the probabilities of occurrence of the specific

sequence x for the distribution of signals (+) and decoys (−), respectively, and

$$L(X = x) = \frac{P^{+}(X = x)}{P^{-}(X = x)}$$

then sequence motifs for which $L(X = x) >= C$ are predicted to be true signals. C is the desired threshold at certain true-positive rate.

4.2.4. Models of the 5' and 3' splice site

Yeo and Burge [211] compute specific models for donor (5') and acceptor (3') sites based on a large data set of $1,821$ human transcripts with $12,715$ introns. They excluded non-canonical splice junctions from the computation.

For the model of the donor motif 9-mer sequences were extracted of which three nucleotides are from the exonic region and the remaining six from the intronic region. The best 5'model was achieved with a second-order marginal constraint with a maximum skip of $s = 5$. This accounts for all pairwise dependencies, which positions are closer than 6 nt to each other.

For the acceptor model a 23 nt long sequence was extracted, containing again three nucleotides from the exonic region and an intronic fraction of 20 nt in length. Due to the significantly longer consensus sequence, the sequence was segmented into nine overlapping fragments. The construction of an "overlapping" maximum entropy with a second-order marginal constraint achieved the best model for the acceptor site. This time with a maximal distance of two nucleotides.

These models are the least biased approximation for distributions of short sequence motifs, consistent with a set of estimated constraints. Therefore, we integrated parts of a freely available `perl` script[3] in our implementation of generating a splice site map. It employs the described models to compute the `MaxEntScan` score of donor and acceptor sequences, which will serve as an indication whether the ortholgs of the splice sites are likely to be conserved.

[3] `http://genes.mit.edu/burgelab/maxent/Xmaxentscan_scoreseq.html`

Contents

Chapter 5

Comparative splice site conservation map

The method developed during this work aims to trace the evolution of transcripts, especially those whose evolutionary history could not be detected by the conservation of their primary sequence alone. As explained in Chapter 2, splice sites, as an important element of gene structure, represent a mighty feature of evolution and therefore can be employed instead of pure primary sequence conservation to assess the conservation of a transcript. Here we will explain how we implemented this approach to work on a large scale.

5.1. Compilation of the splice site database

At the beginning of a project, we are usually interested in the conservation of a specific data set from a certain species. As a basic requirement for our approach, we need a compilation of all splice sites contained in these data. Therefore, depending on the format of the available data, this collection of splice sites is composed as the first step of our pipeline. The resulting data set contains the exact location of each splice sites determined by chromosome, strand and position. Additional information, like site type (donor/acceptor), the surrounding genomic sequence of 20 nt up- and down-

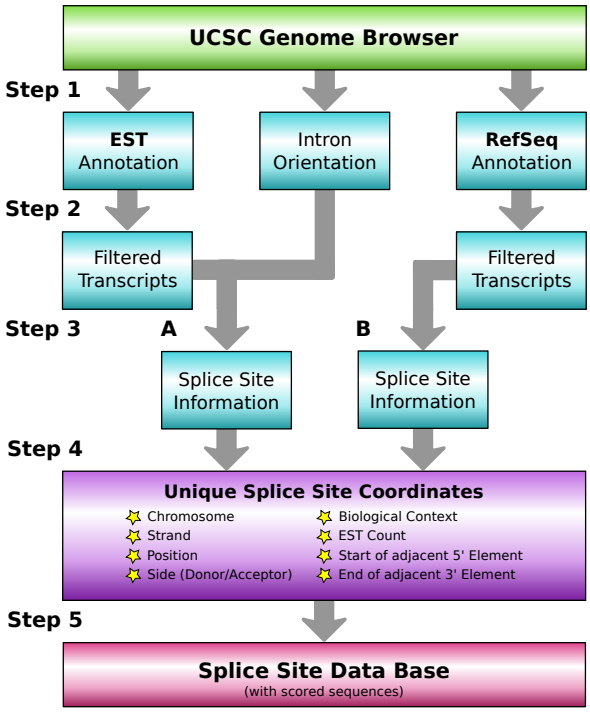

Figure 5.1.: Work flow scheme of splice site database generation. (1) Download data sets for desired species and additional information for intron orientation. (2) Choose relevant spliced transcripts. (3A) Calculate the correct reading direction for transcripts in the EST annotation. Then for both sets (3A and 3B) extract essential information of each splice site (marked with a star in the graphic). Omit splice sites, whose adjacent introns are shorter than 20 nt. (4) Sort both sets and fuse them uniquely. (5) Get the sequence information for each splice site with `twoBit2Fasta` tool and score them via `MaxEntScan`.

stream of the splice site and if possible EST count and biological context, are stored as well. Furthermore all possible start and end positions of adjoining introns and exons are noted, respectively. If the length of an associated intron is less than 20 nt, the splice site is omitted. This approach excludes too short introns, that are likely artefacts [225], and ensures the required length for the application of `MaxEntScan`.

5.1.1. RefSeq and EST

In a rather broad approach, we can compile a substantial list of splice sites from already existing annotations from diverse sources. We implemented a pipeline that retrieves RefSeq annotation as well as expressed sequence tag (EST) data from the UCSC genomes browser. These annotation tracks are automatically parsed into a list of splice site coordinates. The pipeline outlined in Figure 5.1 is repeated for a designated set of species, that can be specified as an input parameter. By this means we can quickly and conveniently generate a comprehensive database of splice sites for multiple species, which will subsequently serve as evaluation of conservation of homologous sites (Section 5.3).

In the following we will describe in detail how the splice sites are extracted from the EST and RefSeq annotation, which are specfied in the files `all_est.txt` and `refGene.txt`, respectively. Both files annotate one transcript per line by exon blocks with equivalent content information in the first five columns.

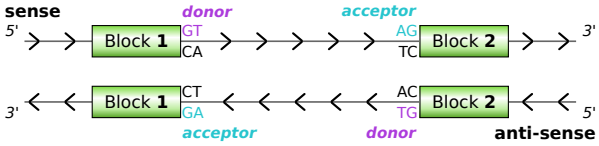

Figure 5.2.: Reading direction. The gap between two exons blocks marks the position of an intron. Start and end positions are generally given in an order corresponding to the location on the genome. For our method it is essential to correctly determine which side of the implied intron is donor and which side is acceptor. This can either be directly identified by the given strand information or indirectly inferred by searching for the canonical splice motifs.

1. **Name:** Transcript name or query sequence
2. **Chromosome:** Reference chromosome/scaffold
3. **Strand:** Direction of transcription (might differ for `all_est.txt`)
4. **Transcript start:** Start position of alignment in target (EST) or start of transcription
5. **Block starts:** List of start coordinates of each mapped EST block or exon

The two files differ in the content of column 6.

6. **Block ends** or **Block sizes**. While `refGene.txt` directly gives the end coordinate of the exon block, `all_est.txt` is giving the length of the block.

Example 5.1 (all_est.txt)**.**

```
EC091556        chr2    -       301409677       301409677,301411674,301414056,301417313,       358,132,142,75,
ES605411        chr8    -       205288907       205288907,205289546,205290546,205291058,       44,245,137,179,
DR041720        chr5    +       20929682        20929682,20930045,20931234,     42,212,71,
EC091498        chr4    +       362115005       362115005,362120755,362134327,362136672,362138864,      79,115,113,68,93,
EC091576        chr1    +       172178005       172178005,172178152,172178479,172180683,172180962,      8,94,110,176,20,
```

Example 5.2 (refGene.txt)**.**

```
NM_001032983    chr5    +       265536177       265535888,265536173,265541804,  265535924,265536336,265542055,  -1,0,0,
NM_001032967    chr6    +       110070412       110070340,110070666,    110070539,110070867,    0,1,
NM_001204849    chr1    -       648642372       648642372,648643642,648660954,  648642447,648643845,648662860,  0,1,0,
NM_001198553    chr8    -       221975982       221975982,221981131,221988375,  221976854,221981393,221988675,  1,0,0,
NM_001034076    chr6    -       36299984        36297247,36299948,36304318,     36297397,36300248,36304611,     -1,0,0,
```

This kind of exon annotation defines the exact coordinates of splice sites. Two consecutive exons mark the boundaries for a single intron. The end position of the 5' exon specfies the donor and the start position of the following 3' exon specifies the acceptor. To be precise, we define the coordinate of the first and the last nucleotide in the intron as the position of the donor and acceptor splice site, respectively. These are usually guanine nucleotides for canonically spliced introns. For our method, it is crucial to correctly determine whether a listed splice site is either donor or acceptor. Therefore, the reading direction of the transcript is an essential information, as blocks are always listed sorted according to the starting position in the genome (Figure 5.2).

While the correct reading direction for RefSeq transcripts is given in column 3 of

`refGene.txt`, getting the strand information for EST supported transcripts is not as trivial. In fact, column 3 of `all_est.txt` refers to the strand of the genomic sequence to which an EST aligns. This is not necessarily the direction of transcription, as ESTs can be sequenced both in 5'—3' and in 3'—5' direction. Therefore, further information is needed to determine the direction of transcription, which is given by an additional file – `estOrientInfo.txt` from UCSC Genome Browser. To provide this information UCSC does some calculations to determine the direction of transcription for each EST sequence, or give the most likely variant. This is implemented by deducting the number of `CT/AC` pairs from the number of `GT/AG` splice site pairs in the relevant EST sequence, which is then noted as the value *intronOrientation* in the aforementioned file. If *intronOrientation* is a negative value, the number of complementary `CT/AC` sites is higher than proper canonical sites. Thus, it is likely that the considered EST sequence is given as the reverse complement of the real transcript and the given strand information in `all_est.txt` has to be switched. A positive *intronOrientation* indicates that the given reading direction is correct. However, the closer the value to zero, the more uncertain is the strand information.

All splice site coordinates are listed uniquely. In other words, if two transcripts share the same splice site, this site is listed only once in the data set. To retain the possibility of identifying presumably alternative splice variants, the start and end positions of all available adjoining introns and exons are noted. Furthermore, for EST supported splice sites, it is counted how many ESTs mapped on this position. For RefSeq annotated splice sites it is also possible to note the context. This information is extracted from column 7 "exon frames" of `refGene.txt`, which contains information about the ORF for each exon. They are specified by {0,1,2}, or -1 for no ORF. Based on this information, it is possible to ascertain if a splice site is located on a non-coding transcript, on a protein-coding transcript or a non-coding region in a protein-coding transcript. If all frames of a transcript are -1, it is a non-coding transcript. Some transcripts have a mixed combination of exon frames starting or ending with -1 frames. These are protein coding transcripts, where the non-coding frames correspond to exons of 5'- or 3'-UTRs at the beginning or end of the transcript, respectively.

5.1.2. Other sources of annotation

Besides the official RefSeq and EST tracks from UCSC, there are plenty of established file formats that are common to use for annotating transcripts. We therefore developed adapted scripts that parse the most common formats, `bed12` and `gtf/gff` (e.g. GENCODE), into an equivalent list of splice sites. While the format of `bed12`

resembles the table schema from RefSeq/EST very strongly and only differs in the order of relevant columns and some 0-based instead of 1-based coordinates, the parsing of `gtf` is more complex. In this case, each exon block is given in an extra line. Thus, it is first parsed into a single line format and thereafter translated into the list of splice sites.

5.1.3. Data from split read mapping

Recent advances in high-throughput sequencing sparked a raise in the number of studies that generate experimentally obtained RNA-seq data. A variety of tools exists, that can map the resulting sequenced fragments of the transcriptome (reads) back onto the reference genome. In cases of spliced RNAs, the sequence of a read will map to distant positions on the genomes. It is called a split read. Since we are particularly interested in the position of splice junctions defined through those split reads, we use the `segemehl` mapping tool [226, 227], which contains an algorithm that is specialized on detecting splits. It reports beginning and end position of splits in `bed`-files. The following is a mixed example of different types of possible output lines.

Example 5.3 (splits.bed)**.**

```
1 chr10 226068  255829  splits:10:12:11:N:P                    0       +
2 chr10 298965  299114  splits:1:1:1:C:P                       0       +
3 chr10 138427  138427  distsplice:chr19:47506458:1:1:1:L:P     0       +
4 chr10 162424  162424  diststrandsplice:chr16:731812:1:1:2:L:P 0       +
```

Conventionally, the parts of the reads map linearly separated by a gap marking the location of an intron. Those "normal" splice junctions are tagged with the letter `N` in column 4 of the output `bed`-file from `segemehl` (Example 5.3, 1). The given start and end positions represent donor or acceptor site, depending on the reading direction. In cases where the RNA-seq protocol is not strand-specific, we can infer the direction of transcription by comparing the `MaxEntScan` scores of the possible splice site consensus sequences of the reported read with its reverse complement (compare with Figure 5.2).

Due to the high sensitivity of the split mapping algorithm in `segemehl`, it can also effectively detect infrequent atypical splice events, such as in circular or trans-spliced transcripts. In the `bed`-file circular splice junctions are tagged with the letter `C` in column 4 (Example 5.3, 2). In this special case the start position refers to the acceptor and the end position to the donor when the reading direction is sense and vice versa for antisense (Figure 5.3). With that knowledge we can determine the correct direction of transcription in the same way as for normal splice junctions,

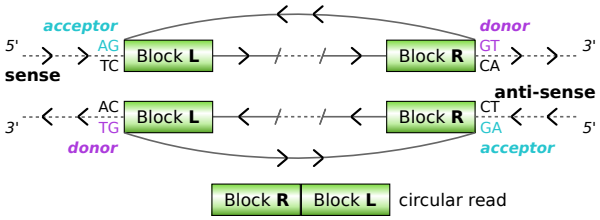

Figure 5.3.: Circular splice junction. In cases of circular transcripts, fragments of a read map in an order contrary to its reading direction. The reported split will therefore encompass exonic regions, putting donor and acceptor sites outside of the split interval. Compared to "normal" splice junctions (Figure 5.2), the sides of donor and acceptor are switched relative to their position on the reference genome.

given a canonical splice motif is present. Each trans-splice event (`dist`) is described by two entries in the output file. They connect positions on distinct chromosomes or scaffolds and are labeled as the "left" (`L`) or "right" (`R`) part of the split (relative to the genomic position). This facilitates the identification of the donor and acceptor side.

A splice sites database obtained by this type of data is considerably more comprehensive than those obtained from existing annotations. This is a great asset, especially when it comes to the investigation of splice events that are less frequent than the regular linear splicing.

5.2. Comparative map of splice sites

The created database of splice sites can now be used together with orthology information from multiple sequence alignments to compare splice sites across species. We establish a comparative map of splice sites by tracing each splice site of a reference genome set in the alignment and list all aligned positions. Again additionally to coordinates we store information on splice type, `MaxEntScan` score (s_{mes}) [211] and whether there is experimental evidence for the functionality of the splice site ortholog (either from existing annotation or available RNA-seq data). This section explains the multiple sequence alignment format, which we use and how we extract the coordinates of the orthologous positions. The pipeline is outlined in Figure 5.4.

5.2.1. Multiple sequence alignment

Our method is designed to search splice sites from a reference genome set in a `MAF` multiple sequence alignment. Such a file consists of alignment blocks separated by blank lines. The first line of a block starts with an '`a`' optionally followed by a

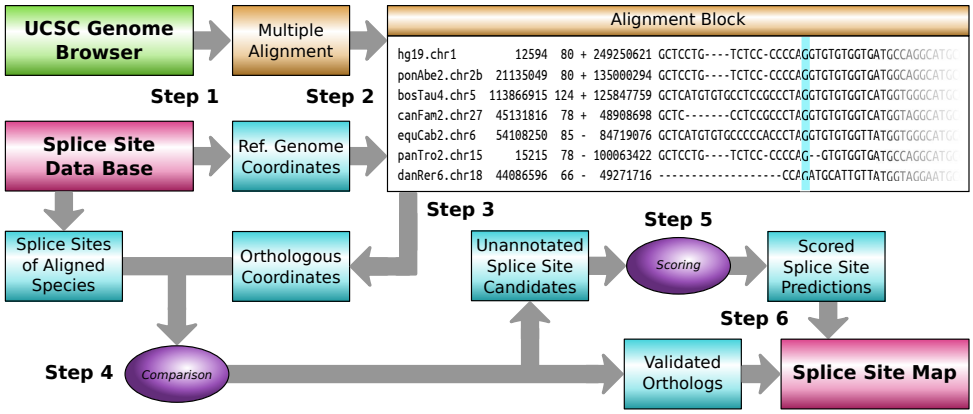

Figure 5.4.: Creating the map. (1) Choose reference genome set of splice sites, from the generated database. Download a multiple alignment, which is centered in the chosen reference genome. (2) Search the position of each splice site from the reference genome set in the multiple alignment. (3) Compute the corresponding positions in all aligned species. (4) Compare the resulting coordinates with the splice site database and discriminate between experimentally validated sites and new candidates. (5) Extract sequences and score candidates (*cf.* Figure 5.1 (5)) (6) Add validated as well as unsupported sites with all relevant information to the map.

'`name=value`' pair, which is usually `score=value`. Subsequently listed are lines with different types of data. The lines that are relevant for our method start with '`s`'. These lines contain the aligned sequences and thus represent the actual sequence alignment. They provide the following information[1] organized in six distinct fields:

1. **Source name.** The name of the source sequence, which is usually given as 'genome.chromosome'
2. **Start position.** The start position of the aligned region in the source sequence. If the aligned region is located on the minus strand, then this position is relative to the reverse-complemented source sequence.
3. **Length.** The length of the aligned region in the source sequence. This equals the number of nucleotides minus the number of gaps in this line.
4. **Strand.** The aligned sequence is from the plus or minus strand, but always displayed in reading direction (5' to 3').
5. **Source size.** The entire length of the whole source sequence (usually the chromosome size).
6. **Alignment text.** The aligned nucleotides and dashes as gaps.

[1]Description adopted from UCSC Genome Browser. For a more detailed description on `MAF` format see `http://genome.ucsc.edu/FAQ/FAQformat.html#format5`

Example 5.4 (Alignment Block)**.**

```
0 a score=42330.0
1 s hg19.chr2          18627  39  +  243199373  GGAAGGGCTCAGTCATCACAGATGCGAAAAG------GCAGTGTG
2 s canFam2.chr17    3030673  44  +   67347617  ACAAGGGCTTAGCCATCACCCATCCCCAAAGGCAGATGGAA-TCA
3 s equCab2.chr15      11627  41  -   91571448  GGGAGGACTCGGCTATCACAGACACAGGAAGGCAAATG----CCA
4 s rheMac2.chr13      16081  45  +  138028943  GGAAGGATTCAGTCATCACAGATGCCGAAAGGCAAACGCAGTGTG
5 s ponAbe2.chr13      33087  44  -  117095149  GAAAAGACTCAG-CATCACAGATGCCGAAAGGCAAACGCAGTGTG
```

The aligned sequence of *Canis familiaris* in line 2, starts from position 3,030,673 in *chr17* and ends at position 3,030,717 ($= startposition + length$). In line 3 the aligned region of *Equus caballus* is displayed in reading direction of the negative strand. The given start position thus refers to the reverse-complemented sequence of *chr15*. Hence the aligned sequence starts at the *absolute* position 91,559,822 ($= sourcesize - startposition + 1$) and ends at 91,559,781 ($= sourcesize - startposition - length + 1$).

The `MAF` file is required to be formatted in a certain layout to properly and effectively work with our search algorithm. (1) The first species of all alignment blocks has to be identical with the reference genome of the splice site data set for which we want to investigate conservation. (2) Furthermore the order of alignment blocks has to be sorted by the start position in the reference genome. The sorting ensures an effective $O(n)$ runtime of the search algorithm. All UCSC alignments are generated with `MULTIZ` and therefore fulfill these requirement (Section 4.1.2). MSAs from ENSEMBL are computed with `EPO` and thus are present in `EMF` (Ensembl Multi Format) and need to be converted into `MAF` with a parser[2] before using it as input file.

5.2.2. Calculation of orthologous sites

The search of splice sites in the multiple sequence alignments is straight forward. The algorithm loops through the list of splice sites of the reference genome data set, which is sorted by genomic position. A separate loop is initiated in parallel, iterating block by block over the multiple sequence alignment until the alignment interval of the current block matches with the current splice site position. It is possible that a genomic region is used completely or partially in more than one alignment block. Therefore, subsequent blocks of a matching block have to be controlled for a compatible alignment frame as well. When all eligible blocks have been detected, we calculate the orthologous positions of the current splice site in each of the aligned species.

First we count out the alignment column of the reference splice site by adding up

[2]Program `emf2maf.pl`, provided by ENSEMBL at `ftp://ftp.ensembl.org/pub/ensembl-compara/scripts/dumps/emf2maf.pl`

each nucleotide (omitting the gaps) to the start position until we reach the column of the splice site. To determine the orthologous position in the aligned species, count off the nucleotides until we reach the determined alignment column. Depending on the strand information of the aligned sequence, the number of nucleotides is either added up to or deducted from the starting position of the aligned sequence (see Example 5.3 for calculation of starting position). An aligned gap is considered as absence of an ortholog in this species.

In the example of Figure 5.4 the algorithm searches for orthologs of the human acceptor site, annotated in *chr1*, at position 12,612 on the plus strand. It found a relevant alignment block, where the splice site is located within the interval of the aligned human reference sequence in the first line ($12,594 \leq 12,612 < 12,594 + 80$). Now the alignment column of the acceptor site is determined by counting off the nucleotides. It is located in column 23, which accounts for 18 nt ($= 12,612 - 12,594$) and 5 gaps. Therefore the orthologous site in dog (canFam2) would be in *chr27* at position $45,131,816+23-7 = 45,131,832$ (7 gaps). The aligned sequence here is from the plus strand, hence the reading direction corresponds to the one displayed in the alignment. For zebrafish (danRer6) the aligned sequence corresponds to the minus strand of the genome. The displayed reading direction is reverse to the numbering of the sequence. The orthologous position is calculated by subtraction of the nucleotides and addition of occurring gaps. It is located on *chr18* at position $49,271,717 - 44,086,596 - 23 + 19 = 5,185,117$.

5.2.3. Maximum entropy scoring of splice sites

The evolution of splice sites cannot be studied meaningfully based only on the annotated splice sites as the transcriptomes of many species are poorly covered in current databases, in particular in their non-coding regions. We therefore integrated `MaxEntScan` scores (see Section 4.2) in our map generation pipeline, in order to determine whether a splice site candidate is likely to be functional or not. This will serve as indicator for the estimation of conservation rate later.

By employing the perl wrappers provided for download by Burge Lab[3], the method computes s_{mes} for all aligned orthologs. It requires information as to whether the putative splice site is a donor or an acceptor, and a short surrounding sequence of a certain length as input.

Figure 5.5 shows a graphical representation of a scored comparative splice site map in the region of the GAS5 locus. The first column constitutes the reference genome

[3] `http://genes.mit.edu/burgelab/maxent/download`

Figure 5.5.: Splice site map of the GAS5 locus. Each line represents a splice site, each column a vertebrate genome arranged in increasing phylogenetic distance from human. The respective species name for the genome assembly abbreviations can be found on page 147*f*. `MaxEntScan` scores for splice site quality are color coded. Missing data, where no sequence is aligned, are indicated as gray background. Light green entries indicate validated splice sites present in RefSeq or sites that are experimentally confirmed by at least two ESTs.

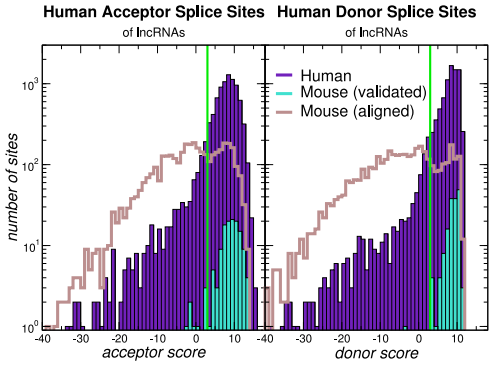

Figure 5.6.: Conservation of human lncRNA splice sites in mouse. Filled curves designate the distributions of `MaxEntScan` scores for human splice sites (purple) and orthologous positions that are known to be splice sites in mouse (cyan). The score distribution of all aligned positions in mouse (brown) is a superposition of conserved functional splice sites and positions that have been destroyed by substitutions. The cutoff value of 3.0 is indicated by a green line.

set of splice sites. The other columns in each line represent the orthologs to the reference splice site in the named species. The color of each pixel reflects s_{mes} of the respective site.

5.3. Assessment of splice site conservation

All splice site orthologs are compared to the previously established database of annotated splice sites. If a match is found, the ortholog is considered to be a *validated* functional splice site. In a case where no match is found, we invoke s_{mes} to assess the conservation. A splice site is *predicted* to have a functional ortholog if there is an orthologous site in the relevant genome with $s_{mes} > 3.0$. This cutoff is estimated from score distributions illustrated in Figure 5.6. It shows the distribution of donor and acceptor scores of all splice sites in a human lncRNA set as well as the scores of all aligned and all validated orthologs in the UCSC human-mouse alignment. While the majority of known splice sites features scores > 3 (cyan), we observe a clearly bimodal distribution for the non-validated sites (brown) composed of a large peak conforming to functional splice sites and a second broader distribution of scores ≤ 3 belonging to positions that most likely have lost their capability of acting as splice donors or splice acceptors.

In summary the conservation of splice sites can be classified into two compatible categories: *validated* and *predicted*. A *conserved* splice site therefore has a *predicted* or *validated* functional ortholog in the relevant genome or both.

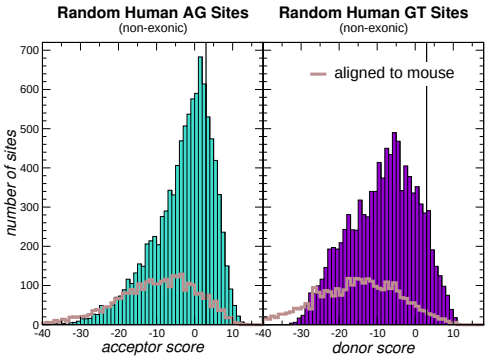

Figure 5.7.: Distribution of s_{mes} for random human non-exonic GT-AG sites. The distribution of `MaxEntScan` scores for the ortholog mouse sequence is displayed in brown.

To make a more intensive use of the established database of annotated splice sites, the implementation of the generation of the comparative splice site map holds the possibility to induces a less stringent search for *validated* orthologs in the multiple alignment by permitting misalignments around the splice site by a given nucleotide range.

5.3.1. False positive rate estimation

We sampled random non-exonic positions from the human genome with the additional requirement of a present canonical motif (`GT`/`AG`). About 31 % of these sites were alignable to mouse. In order to make an estimation on the false discovery rate on ortholog sites, we scored the aligned sequences with `MaxEntScan`. Figure 5.7 shows the distribution of the described scores. Only 1.2 % and 3.0 % of all of the random `GT` and `AG` sites, respectively, had a score > 3 in the aligned mouse sequence. It is expected that more distant species exhibit even lower false discovery rates. We emphasize that the score cutoff > 3 is restrictive and will tend to underestimate the number of conserved splice sites, since the `MaxEntScan` scores are gauged so that sites with positive scores are more likely to be functional than not [211]. This is also consistent with the results from Table 6.1, when comparing the *predicted* and *validated* splice site conservation of human coding regions in mouse.

5.4. Estimation of conservation on transcript level

Conservation rates on the transcript level are derived from its splice sites. We consider a transcript to be *conserved* if a particular fraction of its splice sites are *conserved* in the respective organism. Hereafter we will refer to that fraction c as the degree

of conservation. Since c is expressed as a percentage, it weighs each gene equally regardless of its length or number of splice sites it contains. Hence, this approach facilitates the comparability between genes of different sizes (number of exons).

Different values of c highlight different aspects of conservation and evolutionary change: At $c > 0\%$ we assay only presence or absence of a gene, and thus its evolutionary origin. The other extreme, $c = 100\%$, focuses on the exact conservation of the gene structure. By investigating the conservation of transcripts for different degrees of conservation c and comparing the results, we gain insights into the evolution on the structural level.

Part III.

Biological applications

Contents

Chapter 6

Conservation of human lncRNAs

Long mRNA-like transcripts that do not code for proteins are a big part of the human transcriptome and high-throughput sequencing reveals more and more of those molecules outpacing the exploration of their functionality. Studying the evolutionary history of these lncRNAs is essential to comprehend their role in the human cell. This, however, is a challenging task since their low level of sequence conservation precludes comprehensive homology-based surveys and makes them nearly impossible to align. With the method developed in the course of this work, we are able to trace the evolution of lncRNAs using the conservation of splice sites.

We show that more than 85 % of the human GENCODE lncRNAs were already present at the divergence of placental mammals and many hundreds of these RNAs date back even further. Nevertheless, we observe a fast turnover of intron/exon structures. We conclude that lncRNA genes are evolutionary ancient components of vertebrate genomes that show an unexpected and unprecedented evolutionary plasticity. We offer a public web service[1] that allows to retrieve sets of orthologous splice sites and to produce overview maps of evolutionarily conserved splice sites for visualization and further analysis. An electronic supplement containing the ncRNA data sets used in this study is available at `http://www.bioinf.uni-leipzig.de/publications/supplements/12-001`.

[1] `http://splicemap.bioinf.uni-leipzig.de`

6.1. Data

6.1.1. Transcriptome annotations

As a basis set of human transcripts we obtained a RefSeq track (10/2012, 40,373 transcripts) from UCSC as well as the GENCODE v14 collection of transcripts [228]. In addition we extracted all splice sites supported by at least one expressed sequence tag (EST) in the data collection of the UCSC genome browser (downloaded 08/2012).

To achieve a maximum coverage in assessing the experimental validation of splice site orthologs, we compiled splice site lists for all species of the UCSC alignment, for which equivalent EST and RefSeq annotation data were available. The generated splice site database comprised 20 species: *Homo sapiens*, *Pan troglodytes*, *Pongo abelii*, *Macaca mulatta*, *Callithrix jacchus*, *Mus musculus*, *Rattus norvegicus*, *Cavia porcellus*, *Oryctolagus cuniculus*, *Bos taurus*, *Equus caballus*, *Felis catus*, *Canis familiaris*, *Monodelphis domestica*, *Ornithorhynchus anatinus*, *Gallus gallus*, *Xenopus tropicalis*, *Fugu rubripes*, *Gasterosteus aculeatus* and *Danio rerio*.

6.1.2. Reference data sets of lncRNAs

Since many RefSeq non-coding transcripts are associated with coding loci, we focus our analysis on a restrictively filtered subset of the GENCODE data to ensure conservative estimates of lncRNA conservation. In order to ascertain a high-quality set of human lncRNAs we applied a series of filtering steps to an initial data set of 21,271 well-characterized "GENCODE v14 lncRNA" transcripts.

We discarded transcripts that overlapped within annotated protein-coding sequences or pseudogenes in sense or anti-sense direction annotated by at least one of GENCODE, ENSEMBL, UCSC, or RefSeq. For GENCODE, we could rely on the annotation with biotype classification for transcripts and genes. In the case of ENSEMBL, RefSeq and UCSC we employed the annotation of coding exons. Since some of the transcripts overlapping in sense-direction might just be non-coding isoforms of protein-coding transcripts, we opted to remove them. We also excluded transcripts located in anti-sense direction of these coding sequences since conservation of the coding sequence also constrains the sequence of the opposing transcripts, even though they are annotated as non-coding. We used `RNAcode` [229], a tool that efficiently detects conserved open reading frames in multiple sequence alignments, and `Tblastn` [230] to remove transcripts with putative coding regions. We only kept those transcripts that did not contain exons overlapping with significant `RNAcode` hits ($p < 0.05$) or, if an exon could not be scored by `RNAcode` due to low sequence conservation, `Tblastn` hits

(*E-value* < 0.05). We also removed all unspliced entries. At this stage we retained $5,703$ transcripts. The last filtering step included the application of `PhyloCSF` [231]. All remaining transcripts with a `PhyloCSF` score > 100 and a possible ORF of length ≥ 30 were sorted out. These cutoffs were chosen accordingly to [15]. This affected another 290 transcripts. The final data set comprises $5,413$ spliced transcripts with $17,163$ splice sites.

Alternative data sets

Besides the described main data set we additionally investigated the conservation of one similar data set from Cabili *et al.* [15] and three other more specific data sets of microRNA and snoRNA host genes, as well as mouse and zebrafish lncRNAs. By comparing the conservation results, the evalutation of method performance and consistency can be corroborated.

The main data set exhibits substantial overlap with the integrative compilation of $14,274$ spliced human non-coding transcripts from different sources covering 24 tissues and cell types by Cabili *et al.* [15]. $3,145$ of them are identically (99 % reciprocal strand-specific overlap) represented in our set; the agreement increases to $3,924$ loci when a sequence overlap of at least 70 % is required. We will refer to this collection of lncRNAs as the Cabili data set.

As an important subclass of spliced lncRNAs with well-understood function, we generated a set of microRNA and snoRNA host genes. We identified lncRNAs that overlapped known microRNAs and snoRNAs as annotated by ENSEMBL. This resulted in 128 transcripts hosting microRNAs (containing 602 unique splice sites) and 73 transcripts hosting snoRNAs (335 unique splice sites). Interestingly, snoRNA host genes and, to a lesser extent also microRNA host genes, on average have more introns than other lncRNAs (3.7 *vs.* 2.9 *vs.* 2.0 introns/transcript in all lncRNAs).

Guttman *et al.* [232] described a set of mouse lncRNAs involved in the circuitry controlling pluripotency and differentiation. It comprises $2,076$ spliced transcripts with $6,975$ splice sites, a major fraction of 77 % of them are also validated by EST or RefSeq data.

Pauli *et al.* [233] reported a conservative set of $1,133$ lncRNAs expressed in zebrafish embryos. A second, smaller set of 691 zebrafish lincRNAs expressed during brain development is described by Ulitsky *et al.* [16]; of which only 449 are spliced. Due to the small overlap of the two sets, we consider their union consisting of $1,508$ spliced transcripts with $5,415$ splice sites for conservation analysis.

6.1.3. Multiple sequence alignments

We will employ four distinct reference alignments for the search of orthologs. For the human data sets we use (1) the `MULTIZ`-based alignment [205] of 46 vertebrate genomes provided through the UCSC genome browser and (2) the `EPO` [206] multiple alignment of 12 eutherian mammals downloaded from ENSEMBL (Release 63). We reduce the latter alignment to those 8 species for which ENSEMBL and UCSC utilize the same genome versions: *Homo sapiens*, *Pan troglodytes*, *Pongo abelii*, *Mus musculus*, *Rattus norvegicus*, *Bos taurus*, *Equus caballus* and *Canis familiaris*. In the following we will refer to these two multiple sequence alignments as the UCSC and the ENSEMBL alignment, respectively.

While both species, mouse and zebrafish, are present in the 46way `MULTIZ` alignment, a conservation analysis with a projected alignment thereof would only contain sequences that are alignable to the human genome. Thus we use (3) the 8-way zebrafish `MULTIZ` alignment (containing five teleosts, frog, mouse and human) and (4) a mouse centered `MULTIZ` alignment (reduced to mouse, rat, human, dog, horse and cow) from UCSC to investigate the conservation of the respective data sets.

6.2. Results

In this study we aimed for a far-reaching traceability of old RNAs. Therefore we used the conservation degree of $c > 0\%$ throughout this result section, unless specified otherwise. In other words all results regarding the conservation of a transcript are under the premise, that a transcript is considered to be *conserved* if at least one of its splice sites corresponds to a *predicted* or *validated* ortholog.

6.2.1. Predicted conservation of protein-coding splice sites shows specificity of the method

The splice site conservation between human and mouse is summarized by Table 6.1 and Figure 6.1. We observed similar results for other mammalian species (see Supplemental Table A.1 and Figure A.1). The RefSeq data set overwhelmingly defines splice sites of coding exons. Of these are more than 95 % alignable, and nearly 92 % have experimentally validated orthologous splice sites in mouse. The high specificity of the cutoff being $s_{mes} > 3.0$ is highlighted by the fact, that the fraction of experimentally validated splice site orthologs tallies the fraction of those that are computationally predicted, with a tendency to even slightly underestimate that actual conservation rate.

Table 6.1.: Conservation of splice sites between human and mouse. We report the conservation of splice sites for different annotation sets. We give an overview on the total number (N) of splice sites present in human, the number of *aligned*, *predicted*, *validated* and *conserved* splice sites. The latter attribute is the union of *predicted* and *validated* sites.

	Human	Mouse			
Data set	N	Aligned	Predicted	Validated	Conserved
RefSeq coding	355,573	340,327	325,323	326,401	333,661
RefSeq 5'-UTR	16,035	11,737	8,200	6,908	8,339
RefSeq 3'-UTR	1,124	828	680	607	693
GENCODE lncRNAs	17,163	7,339	2,179	295	2,188
miRNA host	602	282	105	40	108
snoRNA host	335	141	83	46	85

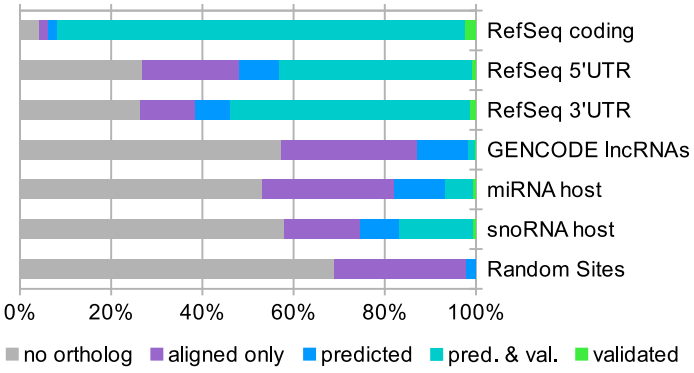

Figure 6.1.: Conservation of splice sites between human and mouse in different contexts. In non-gray colors the fraction of all alignable splice sites is shown. Colors from green to blue display the estimated conservation rate. The remaining fraction of alignable but likely non-conserved splice sites is shown in purple. The overlap of our predictions with validated splice sites is displayed in turquoise. In protein-coding RNAs 95 % of the splice sites are at least alignable to mouse, and of those almost all are conserved. While in lncRNAs the rate of alignable sites drops to around 40 %. The fraction of validated splice sites amongst predicted sites decreased from nearly 98 % to only 13 %, indicating that there is a high number of unannotated splice sites.

Only a small fraction of the RefSeq splice sites falls into UTRs, with more than 14-fold difference between 5'- and 3'-UTRs. Merely about three quarters of these regions are aligned between human and mouse in the UCSC alignments. Still, most of the predicted splice sites are backed up by experimental data. The strong depletion of introns in the 3'-UTRs has been described previously and can be explained as a consequence of nonsense-mediated decay (NMD) or a larger tolerance for intron retention, see e.g. [234].

6.2.2. Conservation of splice sites provides lower bounds on the number of conserved lncRNAs

Only a tiny fraction of about 3 % of the splice sites of human lncRNAs are orthologous to known splice sites of annotated transcripts in other non-primate Eutheria. This estimate is consistent with the observation that about 12 % of the lincRNAs compiled in [15] are syntenically paired with a corresponding transcript in another mammalian species as detectable by `TransMap` [235]. Furthermore non-coding transcripts are typically expressed at lower levels than their coding counterparts and are often restricted to specific cell lines or tissues [83].

Clearly, the poor sequence conservation of the lncRNAs [20] limits the number of human splice sites for which sequences from other eutherian families can be aligned. As a consequence, we can only determine a lower bound on the numbers of evolutionarily conserved splice sites in lncRNAs. The estimates therefore are limited by alignment coverage and quality. See Section 6.3 for a more detailed comparison of UCSC and ENSEMBL alignment.

This small fraction of conserved lncRNAs, however, is mainly the result of the incompleteness of the transcript catalogs in non-human species. We therefore use the conservation of splice sites as measured by `MaxEntScan` scores to obtain more accurate estimates. As detailed in Section 5.3, a cutoff of $s_{mes} > 3.0$ is sufficiently specific that we already tend to underestimate the number of conserved splice sites.

Intron-rich lncRNAs, such as GAS5 in Figure 5.5, tend to have an overrepresentation of poorly conserved splice sites with only marginal support and low `MaxEntScan` score. At least some of these are probably mapping artefacts that artificially reduce the estimates of splice site conservation from our data set. Since we consider a lncRNA as conserved if at least one splice site of the human transcript corresponds to a predicted or experimentally known splice site ($c > 0\%$), the high-scoring splice sites are sufficient to establish the ancient origin of lncRNAs. The biases introduced by spurious and low-scoring splice sites in the GENCODE data thus have little impact on the results at transcript level. Furthermore, we observe no strong dependence of

Table 6.2.: Multi-exonic lncRNAs. In dependence of the number of exons per transcript, we provide the number of lncRNAs, the underlying number of splice sites and their average human `MaxEntScan` score. For each splice site, we furthermore report the average as well as the maximum number of species in which we found it. The splice site scores only slightly increase with the number of exons per transcript. Furthermore, we observed some "ultra-conserved" splice sites which can be traced in nearly all vertebrate genomes.

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

Table 6.3: Conservation of GENCODE lncRNAs in the UCSC alignment. The number of *conserved* and *validated* splice sites and transcripts in selected species gives an overview of the conservation of human lncRNAs in vertebrates. A *validated* splice site is defined as a known splice site orthologous to the reference, whereas the category *conserved* additionally includes the *predicted* functional orthologs. *Union 5* refers to conservation in either mouse, rat, cow, or dog; *Union 15* refers to conservation in at least one of the following species: mouse, guinea pig, rabbit, cow, horse, dog, elephant, armadillo, opossum, chicken, frog, fugu, zebrafish and lamprey.

Species	Splice sites		Transcripts	
Human	17,163		5,413	
	Cons.	Val.	Cons.	Val.
Mouse	2,188	295	1,910	308
Rat	2,005	164	1,777	185
Cow	3,856	300	2,845	268
Dog	4,234	146	3,053	146
Union 5	6,541	515	3,862	462
Union 15	9,047	506	4,511	462

splice site conservation on the number of exons n_i per transcript, although the average splice site score slightly increases in transcripts with more exons (from 7.1 for $n_i = 2$ to 7.6 for $n_i \geq 4$), see Table 6.2.

6.2.3. More than half of the GENCODE lncRNAs are conserved across the Eutheria

We summarize the results for several mammalian species that have the best transcriptome annotation coverage in Table 6.3.

These data indicate that more than 38 % $(6,541 \,/\, 17,163)$ of the individual splice sites and 71 % $(3,862 \,/\, 5,413)$ of the transcripts are conserved across the major eutherian families. When we include 15 available non-primate vertebrate genomes, this

number increases further to 4,511 transcripts (83 %) and 53 % of the splice sites. This reveals a massive gap to an estimation of only 3 % (506 / 17,163) conservation of splice sites and 9 % (462 / 5,413) of transcripts, where only orthologs in annotated transcripts are considered as conserved.

In 2014 a subset of 1,898 GENCODE lncRNAs expressed in a certain collection of human tissues was investigated for conserved expression in five other mammalian species (chimp, rhesus, cow, mouse, and rat) [27]. Expression from orthologous loci was observed for 35 % (rat) to 80 % (chimp) of the human transcripts. In these RNAs, conservation of between 20 % to 60 % of the observed human splice junctions were directly confirmed as conserved by dedicated transcriptome sequencing data. This is in good agreement with the estimated conservation of mouse splice sites in Table 6.1. Our numbers, furthermore, are in agreement with the estimate that 60-70 % of the intergenic lncRNAs are conserved between human and mouse [29]. This estimate is based on the comparison of lncRNA expression from syntenically conserved loci, without regard to gene structure. Thus we do expect our estimate to be appreciably more conservative.

A surprisingly large number of lncRNAs can be traced even further: 784 transcripts (14.5 %) are conserved in at least one of the two marsupials (opossum, wallaby) and 446 can be found in the platypus genome.

6.2.4. Nearly 80% of the human lncRNAs may be older than the primates

By discarding all unaligned positions as missing data and considering only the conservation of splice sites of those sequences that are present in the multiple sequence alignments, we can estimate a crude upper bound on the conservation of lncRNAs. Table 6.7 summarizes the upper bound estimates in mouse, rat, cow and dog. As expected, these rates are substantially larger than the conservative estimates of Table 6.3, which interprets all missing data as non-conservation (for GENCODE transcripts conserved in mouse, 50.7 % compared to 35.3 %). Surprisingly, the discrepancy, however, is rather small for the number of transcripts that are conserved in at least one of the four species: 79.6 % *versus* 71.3 %, see Figure 6.2.

One could argue that a conservation degree of $c > 0\%$ might be too low a threshold. To check the effect of the choice of c we repeated our analysis for $c > 40\%$. A comparison of Figure 6.2B and 6.2C shows that the results change marginally when employing this much more stringent cutoff. Although the absolute number of conserved lncRNAs drops, the relative conservation (disregarding non-aligned sites) still covers more than one third for mammals.

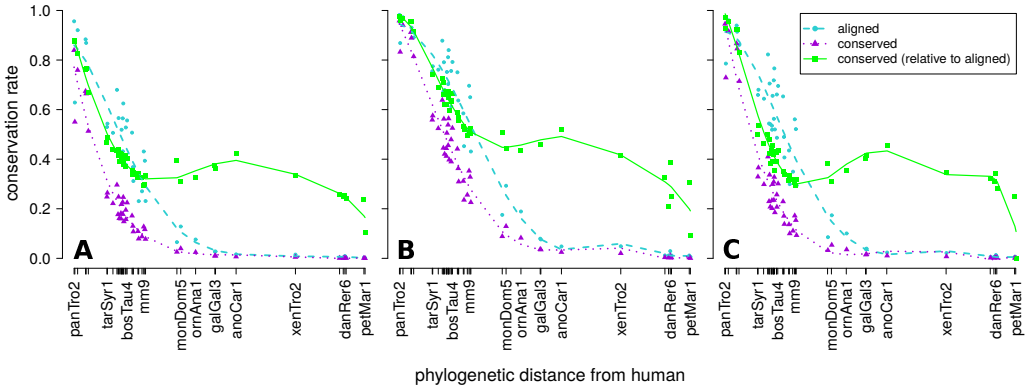

Figure 6.2.: Conservation of lncRNAs across 46 vertebrates. Indicated in blue is the fraction of aligned splice sites, in purple the fraction of splice sites that are validated and/or predicted to be a functional splice site in the respective species. The genome assembly abbreviations are listed on page 147*f*. The upper bounds on the fraction of conserved splice sites are shown in green. The numbers are estimated from the fraction of conserved splice sites within aligned sequence blocks only. Panel **(A)** shows the conservation rate of 17,163 single splice sites, while panel **(B)** illustrates the conservation on the level of transcripts for 5,413 lncRNAs. Panel **(C)** shows conservation of those transcripts if a conservation degree of $c > 40\%$ is required for a transcript to be considered as conserved.

6.2.5. Most human lncRNAs either date back to the origin of the Eutheria or are primate-specific

We inferred gains and losses of human GENCODE lncRNAs across the vertebrates by the parsimony criterion, summarized in Figure 6.3. Since the evolutionary distances within the primate clade are too small to distinguish between splice sites under stabilizing selection and chance conservation due to short divergence time, we left primate subtree unresolved in this analysis. More than 54 % (2,905 / 5,413) of the transcripts arose with the *Eutheria* and another 21 % (1,114 / 5,413) can be traced back to the origins of the *Theria*, while only 6.3 % (343 / 5,413) are primate specific.

6.2.6. Lineage-specific losses of lncRNAs are common

In contrast to 71 % of the transcripts that are conserved between human and at least one of four eutherian species (*Union 5* in Table 6.3), there are few transcripts that are ubiquitously present. In 2011 Rose *et al.* [32] introduced a method that detects novel evolutionarily conserved splice sites and provided a collection of predicted splice sites that are well-conserved across the Eutheria. 2,061 GENCODE lncRNAs have at least one splice site that is contained in this set of predictions. This fits well with 814 transcripts that are conserved between human and all four eutherian species listed in Table 6.3. This suggests that lineage specific losses are frequent.

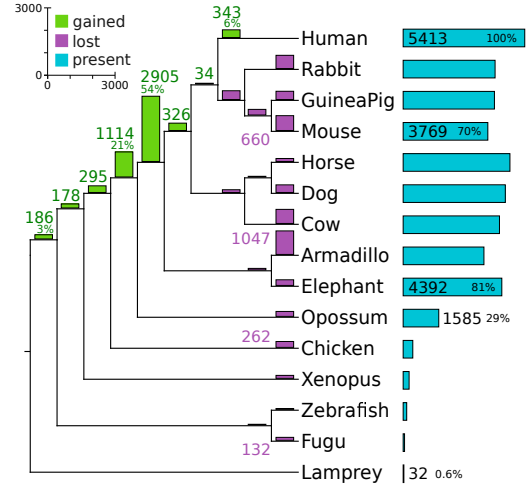

Figure 6.3.: Gains and losses of human GENCODE lncRNAs across the vertebrates. Event counts are based on the parsimony criterion: A loss of a gene is annotated at the edge before a maximal subtree without occurrences at any leaf; a gain event is annotated at the edge before the last common ancestor of all observed occurrences. The vertebrate phylogeny is the `phyloFit` tree provided by the UCSC browser. The primate subtree is omitted.

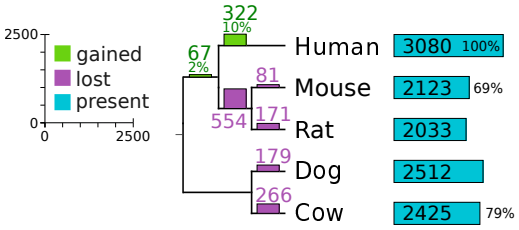

Figure 6.4.: Turnover of individual lncRNA splice sites. Illustration of the number of *gained* and *lost* splice sites from 814 lncRNAs that have at least one splice site conserved between human and all four depicted mammals.

Indeed, we miss 12.2 % (660 / 5,413) of the ancestral lncRNAs in mouse and more than 19 % (1,047 / 5,413) in armadillo. These numbers have to be taken with caution, however. Although a conservation degree of $c > 0\%$ is sufficient to deem a transcript conserved, our conservative cutoff tends to over-emphasize losses and misplace origination events towards the tips of the tree, especially for intron-poor transcripts.

6.2.7. Gene structures of conserved lncRNAs evolve rapidly

Conserved lncRNAs exhibit a rapid evolution of their gene structure. To estimate the turnover of individual splice sites we consider 814 human transcripts conserved in all of mouse, rat, cow, and dog. They comprise 3,080 splice sites. Of these, 87 % were ancestrally present. Most novel splice sites were gained throughout primate evolution. Complementarily, a comparable number of donors and acceptors have been lost in Glires (Figure 6.4). In some examples the changes of transcript structure are quite dramatic. In the *ANRIL* isoforms, entire groups of exons are primate specific, while only a few splice sites, mostly located at the 5' and the 3' ends, are at least as old as the Eutheria, see Figure 6.7A. The visibly higher conservation until marmoset, is consistent with the finding that *ANRIL* is first fully developed in simians, after it went through a two-stage evolution [236]. Another famous example is *HOTAIR*, where the 5'-most exons appear to be lacking in mouse [112].

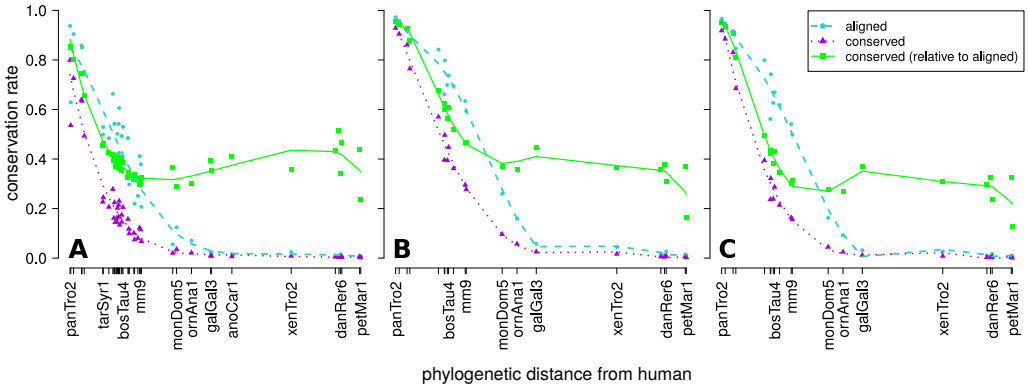

Figure 6.5.: Conservation of lncRNAs from Cabili *et al.* [15]. The estimated conservation on the level of **(A)** 32,515 single splice sites, and 14,274 transcripts with a required conservation rate of **(B)** at least one splice site per transcript and **(C)** more than 40 % of splice sites per transcript - is similar to the estimation resulting from our filtered lncRNA data set. In Panel (B) and (C) only the results of 22 species of the 46 vertebrates of the UCSC alignment are plotted in the graphs.

6.2.8. Alternative data sets lead to consistent results

Conservation of lncRNAs from Cabili et al. [15]

The Cabili data set [15] yields very similar results as the filtered GENCODE data, see Figure 6.5. The nearly constant conservation rate of about 30 % suggests that there is a population of highly conserved splice sites in ancient lncRNAs. On the other hand, it also indicates that sequence conservation in the remaining about 70 % of these highly conserved loci is unrelated to splicing and may not be conserved because of a function at the transcript level.

Conservation of microRNA and snoRNA host genes

MicroRNAs and snoRNAs are subgroups of small structural RNAs with well-defined functions. They are typically rather well conserved at least across the Eutheria. This is also true for their host genes, Table 6.4. There is little difference in the predicted conservation rate of snoRNA and microRNA host genes, even though microRNAs can be processed from both exonic and intronic parts of a primary transcript [237], while snoRNAs are obligatorily intronic at least in mammals [238]. Interestingly, a much larger fraction of snoRNA host genes has experimentally validated conserved splice sites compared to microRNAs. This is probably due to their different expression patterns: microRNAs are often tissue or cell-type specific, while snoRNAs are required ubiquitously.

Table 6.4.: Conservation of special subsets. We tabulate the number of conserved lncRNAs in selected species and in at least one of five Eutheria (human, mouse, rat, cow, dog), four Eutheria (mouse, human, cow, dog), Teleostei (tetraodon, stickleback) or Tetrapoda (human, mouse, frog). We decided to disregard rat for the mouse lncRNA subset calculations, as the two species are too closely related.

	Aligned	Predicted	Validated
128 human transcripts hosting microRNAs			
Mouse	102	63	19
Dog	118	92	3
5 Eutheria	122	110	26
73 human transcripts hosting snoRNAs			
Mouse	56	49	35
Dog	66	59	20
5 Eutheria	69	63	41
2,076 mouse lncRNAs [232]			
Human	1,770	1,113	446
Dog	1,628	944	185
4 Eutheria	1,776	1,237	472
1,508 zebrafish [16, 233]			
Teleostei	953	513	112
Tetrapoda	476	170	56

Conservation of mouse lncRNAs from Guttman et al. [232]

The fractions of alignable positions and predicted splice sites among murine pluripotency lncRNAs [232] is comparable to the GENCODE data. At the level of transcripts we again find substantial conservation across the Eutheria: more than half of the transcripts are predicted to be conserved in human, and 40 % of these have experimental evidence.

Conservation of zebrafish lncRNAs from Ulitsky et al. [16], Pauli et al. [233]

For zebrafish lncRNAs, a much lower conservation level of 34 % is observed among other teleosts. The divergence of zebrafish and Euteleostei is much older than the divergence of major eutherian groups (150 My *vs.* 95 My from paleontological data [239], or 230-333 My [240] *vs.* about 100-120 My [241] estimated from molecular data). This readily explains the smaller fraction and the lower conservation of alignable splice sites. Interestingly, more than 11 % of transcripts are conserved also in Tetrapoda.

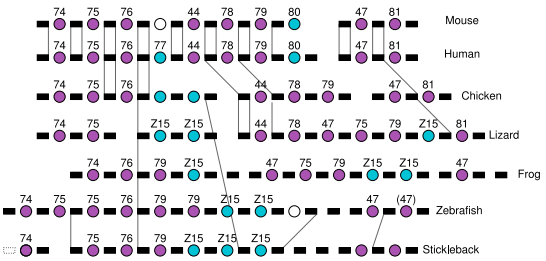

Figure 6.6.: Gene structure conservation of the GAS5 lncRNA. The GAS5 snoRNA host gene is among the most highly conserved lncRNAs. Its homologs are easily identifiable via the well-conserved snoRNAs (circles) located within its introns. Members of the SNORD80/Z15 family are shown in blue. Black boxes indicate the major exons supported by RefSeq and/or EST data. Gray lines indicate splice sites that can be traced manually in at least one of the genome-wide alignments available in the UCSC browser. Note that only a subset of these is represented in any individual alignment, *cf.* Figure 5.5. The transcript structure as well as its snoRNA payload has changed also by means of duplications and deletions.

6.2.9. Many lncRNAs are conserved throughout the vertebrates

Host genes of snoRNAs and microRNAs are found among the best conserved lncRNAs. We found 10 snoRNAs and 14 microRNAs among 271 non-coding transcripts, which are conserved in at least one of the Sauropsida. The deep conservation of host genes does not come as a surprise since their payload is conserved at least throughout the vertebrates in many cases [242–245].

The probably best-studied snoRNA host gene, GAS5, harbors about ten distinct snoRNAs in its introns [131]. It has recently attracted considerable attention since its in general poorly conserved exonic product acts as a riborepressor that binds to the DNA-binding domain of the glucocorticoid receptor [106, 132]. Its chicken homolog is described in detail in [246]. Large clusters of ESTs are easily identified as GAS5 homologs in frog (xenTro2, scaffold_1:6,870,168-6,878,818) and zebrafish (ENSDARG00000092337). The example of GAS5 clearly shows the limitations of genome-wide alignments. Although GAS5 is conserved and functional (at least) across the gnathostomes, Figure 6.6, the 46-way `MULTIZ` alignment does not contain the regions around the splice sites outside the Amniota; even in Sauropsida most parts are missing. Other well-studied examples of deeply conserved snoRNA host genes include UHG/SNHG1 (Figure 6.7E), and U87HG [196].

A well-studied microRNA precursor is Rmst, which harbors mir-1251. The human ortholog was described as differentially expressed in rhabdomyosarcoma subtypes [248]. The mouse ortholog appeared as Pax-2 related gene in early hind-brain development [249]. Its evolution was investigated in detail in [150], demonstrat-

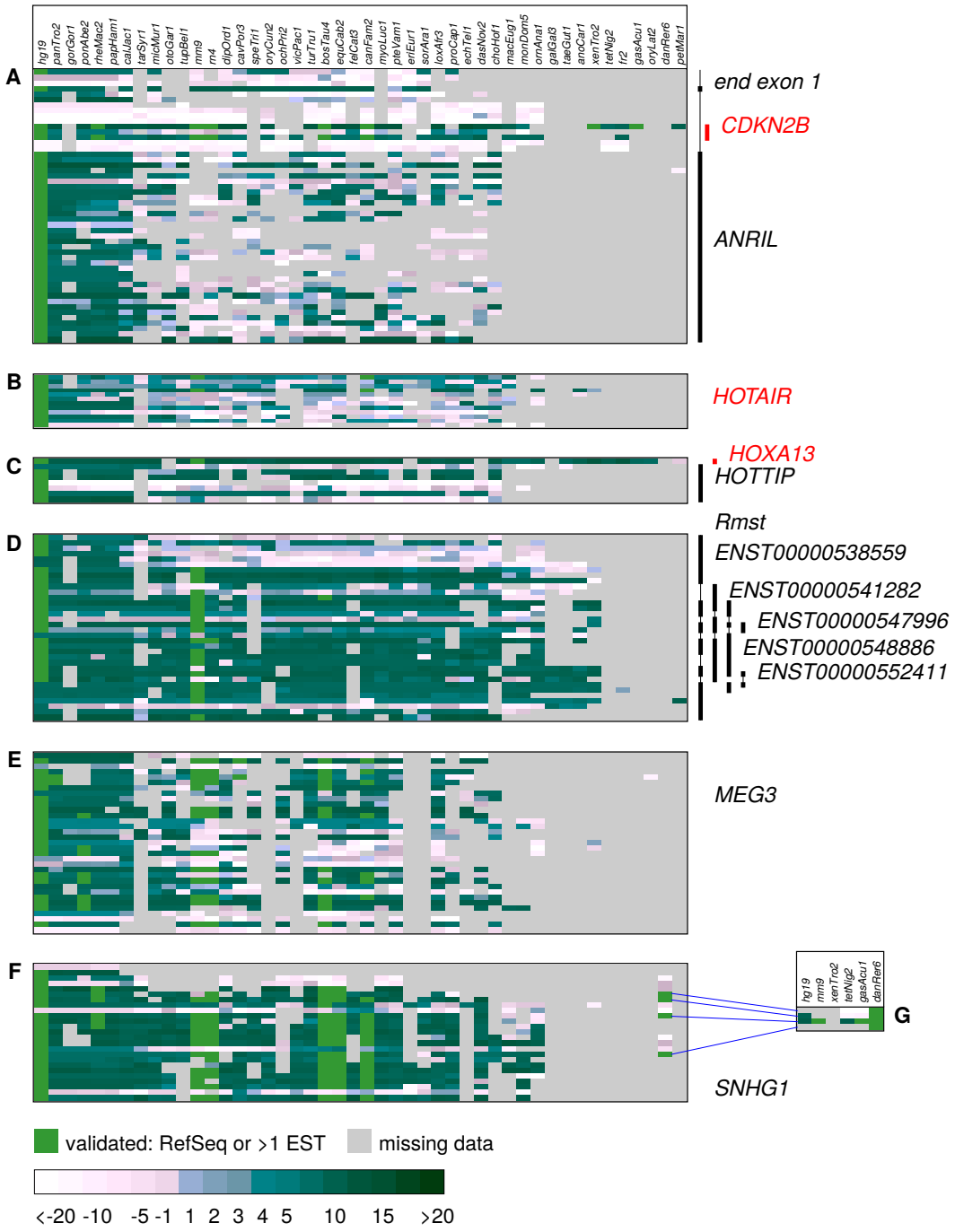

Figure 6.7.: Variation of splice site conservation. The patterns of splice site conservation vary substantially between different lncRNAs, even when their evolutionary age is comparable. The main panel refers to the UCSC 46-way alignment. In the case of *ANRIL*, only a few splice sites are conserved outside the primates (**A**). Although the mouse ortholog shares at least some functions with human *ANRIL* [247], there are only four shared conserved splice sites. Splice site conservation pattern of the HOTAIR transcript (**B**) shows that the 5' end of the lncRNA is much less well conserved than its 3' half. The first exon and intron (splice site at the bottom row of data) overlaps with the protein-coding transcript HOXC11. *HOTTIP*, with few exons that are partially conserved, is also a rather typical chromatin-related lincRNA (**C**). In contrast, the overwhelming majority of splice sites is conserved in Rmst (**D**). MEG3 shows an intermediate pattern, with more lineage-specific losses (**E**). The snoRNA host gene SNHG1 contains several splice sites that are deeply conserved among vertebrates (**F**). Some are even found in teleosts. Experimentally known splice sites from zebrafish SNHG1 were searched also in the 6-way zebrafish `MULTIZ` alignment (**G**). Additional homologous splice sites in two teleosts demonstrate once more the limitations arising from alignment quality. The color scheme is explained in Figure 5.5. Thick vertical bars on the right mark splice sites that belong to a specific transcript (black: plus strand, red: minus strand). Thin lines between these bars indicate conserved splice sites, that are not part of the annotated transcripts.

ing conservation of both the transcript and its expression patterns in opossum and chicken brains. The comparative splice site map shows that Rmst is conserved also in Xenopus, Figure 6.7D. The imprinted MEG3 lncRNA exhibits a large number of differentially expressed isoforms [117]. It is an eutherian innovation apparently associated with the emergence of imprinting at the Dlk1 locus [250]. Indeed, only a single splice site close to the 3' end of the transcripts is shared with a putative evolutionary precursor in the marsupials, Figure 6.7E. It hosts the snoRNA SNORD112 as well as the microRNA mir-770.

The majority of the lncRNAs implicated in chromatin-based regulation can be traced throughout the Eutheria, although it is very likely that many of them are evolutionarily even older. A good example is *HOTTIP* [113], Figure 6.7C, where we lose the sequence conservation in most parts of the locus outside of the placental mammals. Although there are a few deeply conserved elements, these do not include one of the splice site sequences. Nevertheless, the transcript functions also in chick limb-buds [113], suggesting that the gene is considerably older than the Eutheria.

Two zebrafish lncRNAs that are conserved across vertebrates were investigated in detail [16]. *cyrano* (oip5 antisense transcript) is required for proper embryonic development. Our splice site map identifies conservation of splice sites across mammals. However, the sequence is not conserved enough to support an alignment between teleosts and tetrapods. *megamind* (located antisense in an intron of birc6) regulates brain morphogenesis and eye development. The last acceptor site is conserved across gnathostomes in the 8-way zebrafish centered alignment, Figure 6.7G.

In the GENCODE data set three splice sites from three lncRNAs show conservation in every species through to lamprey, namely AC011995.1-001, RP11-423H2.3-003, and RP11-123M21.1-001. These are neither microRNA nor snoRNA host genes. We find 87 conserved transcripts (including one snoRNA host genes) in at least one of the teleosts. 26 % of them even are experimentally validated.

6.3. Alignment coverage and quality limit conservation estimates

The multiple sequence alignment underlying the splice site map has a major influence on the estimates of splice site conservation. We computed separate splice site maps from the UCSC and the ENSEMBL alignments to investigate the impact of alignment coverage and quality. The observed splice site conservation differs significantly between the two genome-wide alignments. Even though the total coverage of the two alignments is quite similar: About 31 % of the whole human genome is aligned to a

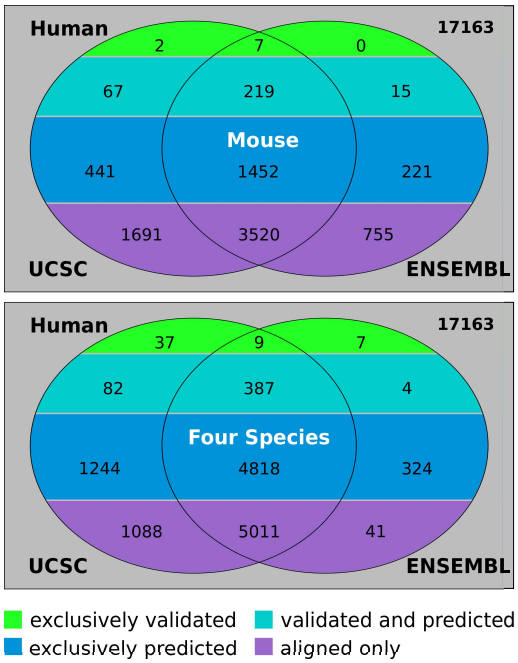

Figure 6.8.: Comparison of UCSC and ENSEMBL alignment regarding influence on the estimates of splice site conservation. All splice sites of 17,163 human lncRNAs, aligned to the considered species are shown distinguished in four groups within the Venn diagram. Top: number of human splice sites found in mouse. Bottom: number of human splice sites present in at least one out of mouse, rat, dog, and cow.

mouse sequence in the UCSC alignments, while the fraction is 27 % in the ENSEMBL alignments. This small difference cannot explain the discrepancy of about one fifth in the coverage of splice sites.

6.3.1. Differences in lncRNA sets

For the majority of the human GENCODE lncRNA splice sites, no aligned mouse sequence is reported in either alignment. Figure 6.8 shows the overlaps between the two alignments. Surprisingly, the alignable sequence fragments differ quite a bit between the two different alignments. Although the coverage of the UCSC alignment is larger ($\sim$ 4 %), there are still nearly one thousand human splice sites for which the ENSEMBL alignment proposes homologous sequence while no sequence at all is aligned in the UCSC alignment. Integrating over the four eutherian species, however, increases the overlap by 16 % to more than 78 %.

As we expected, the larger coverage of the UCSC alignment results also in a greater number of alignable lncRNA splice sites. The estimated upper bounds on the conservation rates are comparable for both alignments. Interestingly, most (89 %) loci that are alignable in the ENSEMBL alignment *only*, correspond to conserved splice sites in at least one the four non-primate mammals, Figure 6.8. When the results of the two alignments are combined, we obtain a lower bound estimate of 40 % for the fraction of splice sites in lncRNAs that originated early in the evolution of placental mammals. Although both alignments show a substantial overlap, the fact that we

Table 6.5.: Conservation of miRNA and snoRNA host genes based upon ENSEMBL alignment. We tabulate the number of conserved lncRNAs in selected species and in at least one of five Eutheria (human, mouse, rat, cow, dog).

	Aligned	Predicted	Validated
128 human transcripts hosting microRNAs			
Mouse	82	53	12
Dog	106	81	1
5 Eutheria	109	99	17
73 human transcripts hosting snoRNAs			
Mouse	47	42	26
Dog	62	54	19
5 Eutheria	63	57	34

Table 6.6.: Conservation of RefSeq splice sites between human and mouse based upon ENSEMBL alignment. For a comparison with the results based on the UCSC alignment see Table 6.1.

	Human	Mouse			
Data set	N	Aligned	Predicted	Validated	Conserved
RefSeq coding	355,573	260,507	249,588	251,385	256,045
RefSeq 5'-UTR	16,035	8,622	6,022	5,024	6,120
RefSeq 3'-UTR	1,124	608	501	445	511

can find hundreds of splice sites whose conservation is visible only in the more stringent ENSEMBL alignment strongly suggests that the actual numbers might still be higher.

For the alternative data set of microRNA and snoRNA host genes, the data for UCSC and ENSEMBL alignments are also quite similar, *cf.* Table 6.5. Here again the coverage is a bit smaller for the ENSEMBL alignments.

6.3.2. Differences in RefSeq annotated sets

Table 6.6 outlines major differences in the observed conservation rates of splice sites compared to the data in Table 6.1, which are computed on the base of the UCSC alignment. For splice sites in coding regions it makes a difference of nearly 12 %, for UTRs even up to 15 % change of estimated conservation rate.

Table 6.7.: Upper bounds on the percentage of conserved splice sites and transcripts in lncRNAs. The numbers are an estimation based on the fraction of conserved splice sites amongst alignable sequence only.

Alignment	Mouse	Rat	Cow	Dog	*Union*
Splice sites					
UCSC	29.6	29.7	40.4	39.5	51.6
ENSEMBL	30.9	30.7	41.4	40.5	52.3
Transcripts					
UCSC	50.7	50.5	66.3	67.1	79.6
ENSEMBL	54.5	53.7	68.5	68.6	79.5

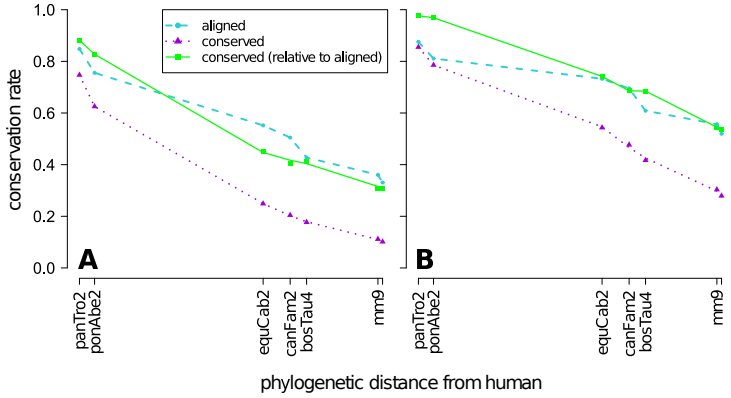

Figure 6.9.: Conservation of lncRNAs across eight mammals according to ENSEMBL alignment. The estimated conservation on the level of **(A)** 17,163 single splice sites and **(B)** 5,413 transcripts ($c > 0\%$) is similar to the estimation resulting from the UCSC alignment.

6.3.3. Differences in upper bound estimation

By disregarding the non-aligned sites, the resulting upper bounds on conservation rate are almost the same for both alignments. Interestingly, the upper bounds in ENSEMBL are slightly higher (up to 0.3 %) than in UCSC alignments. Hence the ENSEMBL alignments contain relatively more conserved splice sites than the UCSC data. This difference is even enhanced when data are aggregated to the level of transcripts, *cf.* Table 6.7. While in coding sequences the gap between the estimated upper bounds on the level of single splice sites is only 0.3 %, the difference increases by 10-fold on the transcript level of lncRNAs.

6.4. SpliceMap web service

The precomputed splice site maps derived from the mentioned multiple sequence alignments are the base for the SpliceMap web service. This service provides a tabular view of conserved splice site coordinates from a given region and produces corresponding visualizations such as those in Figures 5.5 and 6.7. The results can be exported as a text file as well as a custom track for visual inspection in the UCSC genome browser. A list of either splice site coordinates or genomic intervals serve as input.

The underlying algorithm that creates the visualizations of the splice site maps, extracts the relevant lines from the chosen map determined by the input intervals or coordinates, and translates the `MaxEntScan` scores of the orthologous sites in all available species into a specific color code. The species that are considered depend on the available species in the multiple alignment, that has been used to calculate the chosen map. The color scheme translates negative scores to colors from light pink ($s_{mes} < 0$) to white ($s_{mes} < -20$). These splice sites are highly likely lost in this species. Scores in the range of $[0, 3]$ can not be unequivocally deemed conserved nor lost. They are displayed in an intermediate coloring of light blue. Orthologous sites that exceed the defined cutoff for conservation $s_{mes} > 3$, are shown in turquoise to dark green. If a predicted orthologous splice site is validated by RefSeq data or more than one EST, the site is displayed in bright green regardless of the `MaxEntScan` score. A species with no aligned sequence for this site gets a neutral gray color to represent the missing data.

The web site and the computation results are served by a set of Python scripts and rendered into static HTML using the Mako template engine. The jobs are scheduled in a queued fashion. Upon completion, the results are transferred to the web server and available under a personalized link for two weeks. The service can be accessed at `http://splicemap.bioinf.uni-leipzig.de`.

6.5. Discussion

The majority of the human long non-coding RNAs dates back at least to the radiation of the Eutheria, and thousands of these transcripts arose even earlier. The conservation of parts of their transcript structure constitutes compelling evidence for stabilizing selection, despite the often negligible constraints on the sequence itself. Utilizing the conservation of splice sites rather than measures of sequence similarity, furthermore, disentangles for a given locus the selective pressures on DNA elements

from those that refer to the transcript. Our analysis, which suggests that some 70 % of human lncRNAs date back to the eutherian ancestor is in agreement with an independent estimate of the conservation of lincRNAs conservation between man and mouse [29] and with a direct comparison of lncRNA expression in six diverse mammals [27].

Despite the conservation at transcript level we observed a surprising amount of turnover at the level of individual splice sites, again in agreement with [27]. We observe that many of the lncRNA loci exhibit a large number of splicing isoforms. As a consequence of the lack of detailed transcriptomics data for most species, it is currently impossible to trace the evolution of individual isoforms. The discrepancies among individual splice sites, however, leads us to hypothesize that differential selection of isoforms caused the observed rapid divergence of transcript structures. Together with a prolific innovation of new splice sites this process can quickly obscure the evolutionary relationships. Our analysis may still drastically underestimate the evolutionary age of lncRNAs.

We suspect that, as in the case of HOTAIR or ANRIL, major changes of transcript structure go hand in hand with functional changes. This view is supported by major differences between isoforms e.g. in the association of their expression levels with disease phenotypes [103, 251, 252] or the change of function of HOTAIR in mouse that correlates with the loss of several exons [112]. If our hypothesis is true, lncRNAs are likely to be the root cause for rapid phenotypic evolution, as their often chromatin-associated mode of action is subject to large functional changes by easy-to-achieve changes in gene structure. The selective inclusion or exclusion of protein binding sites would affect the composition of complexes of enhancers and chromatin modifiers, see e.g. [253], and thus rapidly alter the rules of transcriptional regulation without affecting the proteins machinery. A similar scenario can be drawn for the post-transcriptional regulation of the pool of microRNA composition by sponges such as HULC [254]. We conclude that lncRNAs are an ancient component of vertebrate genomes with an unexpected and unprecedented evolutionary plasticity.

Contents

Chapter 7

Conservation of atypical latimerian RNAs

Coelacanths (*Latimeria*), as one of the two surviving species of the lobe-finned vertebrate lineage (sarcopterygian), have the potential to unveil many evolutionary aspects of the transition of their ancient relatives from aquatic to terrestrial animals. As anticipated, due to their morphological stasis, the protein-coding sequence of the African species, *L. chalumnae*, was shown to have a retarded evolutionary rate compared to tetrapods, while most other genomic features evolve at comparable speed. Many prominent changes relative to genomes of bony fish can be attributed to land adaptation during vertebrates evolution [255]. Here, we will be concerned with global patterns of the coelacanth's transcriptome.

High throughput transcriptome sequencing provides a view on the RNA content of a sample in unprecedented depth and detail. Although the technology as such promises largely unbiased data, it requires elaborate processing of raw data. It is at this step that preconceptions about what we expect to see in a transcriptome can guide quality control and noise filtering procedures. As a result, these are more often than not neglected as parts of the data set that do not fit to the established paradigms. In this contribution we therefore focused on this blind spot and aimed to identify those reads that do not map locally and colinearly to their reference genome.

In fact, several classes of "atypical" transcripts – circular and apparently trans-spliced RNAs – have been observed in previous studies as abundant types of transcripts in mammalian transcriptome data.

By re-analyzing RNA-seq data sets of different tissues from *L. chalumnae* [255] and *L. menadoensis* [256] with increased sensitivity of the employed mapping procedures, we reveal that both types of non-colinear RNAs are also abundant in the African and the Indonesian coelacanth. Section 7.1 will give an overview on the workflow of high-sensitivity split read mapping and postprocessing, that was used to refine and the considerably expand the existing coelacanth annotation. We observed more than 8,000 lincRNAs with normal gene structure and several thousands of circularized and trans-spliced products, showing that such atypical RNAs form a substantial contribution to the transcriptome. Surprisingly, the majority of the circularizing and trans-connecting splice junctions are unique to atypical forms, i.e., are not used in normal isoforms.

In order to investigate in detail the functional and evolutionary significance of these extraordinary transcripts, we perform a computational splice site conservation analysis – as described in the Methodology (Chapter 5) – on the newly established comprehensive coelacanth transcriptome, with special focus on the atypical transcripts. The results that highlight a potentially functional importance and emphasize the evolutionary relevance of these molecules, are presented in Section 7.2.

7.1. Identification of atypical transcripts via split read mapping

The majority of tasks regarding the work flow of read mapping, including the handling of the RNA-Seq libraries were performed by joint first author Gero Doose[1].

7.1.1. RNA-seq data sets

In this work four transcriptome data sets have been analyzed. Coelacanth RNA-seq samples were obtained based on liver (SRR576100) and testis tissue (SRR576101) from a single individual of *L. menadoensis* [256] and muscle tissue of a specimen of *L. chalumnae* (SRR401852) [255]. As reference data sets we downloaded the publicly available muscle RNA-seq data sets from human (SRR545711) and zebrafish (ERR145647) from the sequence read archive. All data were paired-end reads se-

[1]Interdisciplinary Center for Bioinformatics, University of Leipzig, Leipzig, Germany.
E-mail: gero@bioinf.uni-leipzig.de

quenced with a comparable, non-strand-specific sequencing protocol. The raw reads with length of 101 nt were quality trimmed with `FASTX-Toolkit` version 0.0.13 [257] and adapter clipped with `Cutadapt` version 1.2.1 [258]. For splice site discovery we mapped all available reads. In order to allow for a direct comparison of the relative abundance of circular and trans-spliced reads we down-sampled the data sets to approximately the same size. In this way we avoid artifacts that are caused by the use of coverage thresholds for the detection of splice junctions. Otherwise, the number of detected junctions would increase in a poorly controlled manner with the size of the mapped library.

L. chalumnae genome annotation, described in Amemiya *et al.* [255], was downloaded from ENSEMBL version 70.

7.1.2. Mapping and splice site detection

We used `segemehl` version 0.1.4 [226, 227] to map the reads onto the *Latimeria chalumnae* genome allowing explicitly for split reads. Throughout this chapter we strictly distinguish between splice sites, defined as the genomic positions of a splice donor or splice acceptor, and a splice junction, defined as a pair of donor and acceptor positions spanned by an observed transcript. The splice sites reported by `segemehl` were then filtered by `haarz`, a component of the `segemehl` suite, in order to accumulate high confidence splice sites. To further reduce the chance of mapping artifacts, only splice junctions supported by at least three split reads were kept. Splice sites not included in one of these junctions were also removed from further analysis.

We determined three types of splice junctions: (1) "normal" junctions with read fragments mapped colinearly with the genomic DNA to the same strand of the same scaffold and an insert size between 15 nt and 50 kb; (2) "circular" junctions on the same strand of the same scaffold with a distance less than 50 kb and with fragment order inverted relative to the genomic DNA; (3) "trans-splicing" junctions, where the two splice sites are located on different scaffolds. The relative orientation is of course irrelevant in this case. Spliced reads that connect two scaffolds can arise from normal, colinear splice events if the scaffolds are short or the splice sites are close to the ends of the scaffolds. In order to avoid contamination from such effects arising from the incompleteness of the genome assembly, we classified reads as trans-spliced only if those reads connect loci at least 50 kb from both ends of at least one of the two involved scaffolds. The number of unique junctions (after previously described filter steps), which could be assigned to each of these groups are summarized in Table 7.1.

Since a strand-unspecific RNA-seq protocol was used here, the reading direction of spliced reads could only be inferred indirectly. For reads splitting at canonical

Table 7.1.: Number of unique junction locations, that meet the mapping criteria and are supported by a minimum number of three reads.

Species	**Unique junction locations**		
	Normal	Circular	Trans
L. menadoensis (liver)	80,781	1,061	4,531
L. menadoensis (testis)	102,639	1,216	6,563
L. chalumnae (muscle)	53,895	1,309	3,296
H. sapiens (muscle)	112,183	9,217	8,172
D. rario (muscle)	78,613	1,285	2,715

splice junctions we used `MaxEntScan` [211] scores to compare the two putative reading directions. For both directions we computed the sum of the donor and acceptor score. If one direction had a positive sum, which was greater than the sum of the opposite direction plus 3, we defined this as the correct reading direction.

7.1.3. Transcriptome reconstruction and identification of novel lincRNAs

We used `cufflinks` version 2.0.2 [259] to reconstruct possible transcripts together with their isoforms. The mapping output of `segemehl` was modified to fit the input requirements of `cufflinks` using a custom script. Separate transcript assemblies for both the complete *Latimeria chalumnae* data set and the combined *Latimeria menadoensis* data sets were merged together with `cuffmerge` as proposed by Trapnell *et al.* [260]. Overlaps between transcript and annotation data were computed with the help of `BEDTools` [261]. In order to predict the coding potential of transcripts that were located at unannotated regions we applied `RNAcode` [229] to the coelacanth-centric multiple alignment described in Amemiya *et al.* [255]. Transcripts were classified as potentially coding if at least half of their exons showed a minimum overlap of 50 % with potentially coding regions. Transcripts that did not overlap with potentially coding regions were classified as potentially new lincRNAs. To confirm these lincRNA candidates they were compared against the non-redundant protein database version (07.03.2012) with `tblastx` [262]. Candidates that showed significant alignment hits were added to the potentially coding class. We operationally combined transcripts with the same reading direction separated by less than 5 kb into a single locus to account for the fact that many lincRNAs have rather low expression levels and thus may not be fully covered.

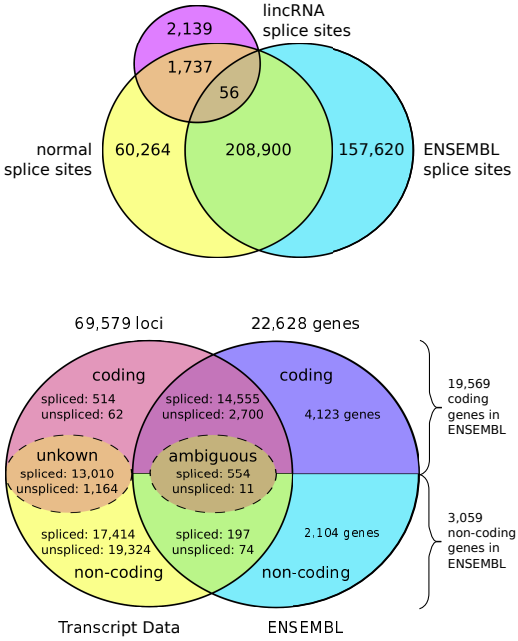

Figure 7.1.: Overview of splice sites and "loci" in comparison to the existing annotation. **(Top)** Venn diagram of unique single splice site positions, detected in our colinear mapped split reads ("normal splice sites"), annotated by ENSEMBL and reported in lincRNAs identified in the main paper [255]. **(Bottom)** Venn diagram comparing ENSEMBL gene annotation with expressed loci from our mapping data. Transcripts with a distance less then 5 kb to each other were merged to one loci, resulting in 69,579 loci. The intersection shows the number of loci, which overlap gene boundaries annotated by ENSEMBL. The distinction of these loci into coding and non-coding is determined by the biotype of the respective overlapping genes.

7.1.4. Splice junctions and transcripts

We made use of the enhanced sensitivity of `segemehl` in mapping split reads to extend the ENSEMBL 70.1 gene build for the latCha1 assembly. The extreme similarity between the two coelacanth species, comparable to human and chimp, justified to combine the RNA-seq data for the purpose of constructing transcript models.

For the *Latimeria chalumunae* (muscle complete) RNA-seq data 26,176,970 reads were mapped with local, colinear splits. For the *Latimeria menadoensis* (testis and liver) 14,201,048 reads were mapped. For the union of these sets 12,817,375 normal split reads that satisfied our filtering criteria were retained. Although the RNA-seq data had been produced with a non-strand-specific protocol, the reading direction could be determined with the help of `MaxEntScan` [211] for 98.8 % of these reads based on the canonical splice site motifs. This resulted in 270,957 unique splice sites, of which 208,956 exactly matched the splice sites of the ENSEMBL 70.1 gene build for the latCha1 assembly (Figure 7.1, Top). About 43 % of the ENSEMBL splice junctions were not visible in our transcriptome map because the corresponding genes were not expressed at sufficient levels to pass our filtering criteria in the three tissues considered here. Additionally, 1,793 sites matched to splice junctions from the lincRNA set reported in the Supplemental Material (Supplemental Data 1) of the coelacanth genome paper [255]. Another 17,801 mapped to novel splice junctions within the boundaries of genes annotated in ENSEMBL 70.1 in the correct reading direction. Since they did not match exactly to positions of annotated splice sites

Table 7.2.: Relation of splice junctions to annotation.

Species	Within	Combining	One site within	Outside
Unique splice junctions				
L. menadoensis (combined)	106,905 (83.4 %)	46 (0.1 %)	2,692 (2.1 %)	18,596 (14.5 %)
L. chalumnae (complete)	72,067 (87.7 %)	34 (0.0 %)	1,331 (1.6 %)	8,759 (10.7 %)
Union	119,802 (82.2 %)	72 (0.1 %)	3,360 (2.3 %)	22,424 (15.4 %)
Intersection	59,170 (91.4 %)	8 (0.0 %)	663 (1.0 %)	4,931 (7.6 %)
Read support				
L. menadoensis (combined)	7,555,360 (86.1 %)	513 (0.0 %)	115,255 (1.3 %)	1,106,975(12.6 %)
L. chalumnae (complete)	3,463,424 (89.1 %)	499 (0.0 %)	50,772 (1.3 %)	371,906 (9.6 %)
Union	11,018,784 (87.0 %)	1,012(0.0 %)	166,027 (1.3 %)	1,478,881 (11.7 %)
Intersection	7,071,592 (92.5 %)	430 (0.0 %)	63,500 (0.8 %)	506,130 (6.6 %)

of ENSEMBL 70.1, they are grouped outside of the ENSEMBL overlap and are shown included within the yellow section in Figure 7.1 (Top). This left 42,463 novel splice sites located outside annotated genes, corresponding to 22,424 distinct splice junctions that are located entirely outside of annotation. Furthermore, we identified 3,360 distinct junctions with only one side outside the published annotation. A detailed comparison of observed splice junctions is compiled in Tables 7.2 and 7.3, a graphical summary of the splice sites accounting for the exact matches only is given in Figure 7.1 (Top).

Assembled into transcripts with `cufflinks` and `cuffmerge`, these combined transcriptome data of *L. chalumunae* and *L. menadoensis* encompassed 126,235 distinct transcripts belonging to 109,761 genes. This amounts to an average of 2.54 exons. Of these, 86,203 (68.3 %) transcripts were intronless. 61.9 % of the transcripts (69,434) did not contain exons located within gene boundaries annotated by ENSEMBL. The majority of these, namely 58,058 transcripts, were intronless.

About 87 % (60,444) of these new transcripts can be considered as lincRNAs since they have no overlaps with `RNAcode` hits or `blastx` hits in the CCDS database with an E-value $e < 10^{-10}$. About 18 % of the rest, i.e., 1,586 new transcripts can be classified as potentially coding genes, since at least half of their exons overlap by at least 50 % of their sequence with `blastx` alignments or with regions found by `RNAcode`. If strand information was available, the overlap had to be strand-specific.

We found 22,424 novel splice junctions outside the published annotation corresponding to 41,139 unique splice sites. Of these, 32,467 matched exactly with splice sites in the collated transcript models produced by `cuffmerge`. 4,163 additional

splice sites were located within these transcripts, apparently corresponding to local variations in the exact splicing position.

It should be noted that a substantial fraction of splice sites from the raw data were not incorporated into transcript models by `cufflinks`. This explains e.g. why part of the splice sites in the lincRNAs annotated in Amemiya *et al.* [255] are not recovered in our analysis.

An overview of the transcriptome analysis relative to the previously available annotation is given in Figure 7.1 (Bottom), where transcripts were merged into loci according to a 5 kb window. Overall, we report here 50,644 novel expressed loci that were overlooked in previous analyses of the same data sets. Of these, 30,268 contain spliced transcripts. The vast majority of newly identified transcripts is non-coding. Nevertheless, we were able to identify more than 500 additional loci with coding capacity.

7.2. Splice site conservation analysis and results

In order to obtain evidence for the conservation of gene structure we used the 9-way coelacanth-centered multiple sequence alignment [255] of *Homo sapiens*, *Mus musculus*, *Canis familiaris*, *Monodelphis domestica*, *Anole carolenesis*, *Gasterosteus aculeatus*, *Xenopus tropicalis* and *Gallus gallus*to search for the homologous sequence positions of the set of latimerian splice sites, that we established by split read mapping data. We followed the regular approach regarding the assessment of conservation, see Methodology in Chapter 5.

7.2.1. Colinear splice sites

Of the 270,957 canonical splice sites in the combined data set, which includes 208,956 sites matching to ENSEMBL annotation (Table 7.3), about 77.8 % were alignable in at least one of eight other vertebrate genomes. More than 96 % of these were conserved according to splice site scores, and for 92.7 % there was experimental evidence for a functional splice site in at least one of these eight species Table 7.4. The overwhelming majority of these splice sites were located within protein-coding genes.

7.2.2. LincRNA transcript structure

We observed 23,065 splice sites in 8,066 spliced lincRNAs in the union of our lincRNAs and the lincRNAs reported in the coelacanth genome paper [255]. About 14 %

Table 7.3.: Comparison of observed splice junctions in *L. chalumnae* and *L. menadoensis*, resulting from reads mapped with local, colinear splits. The addition "filtered" describes the filtering by `haarz` mapping criteria and a minimum junction support of three reads.

	L. chalumnae (muscle complete)	*L. menadoensis* (combined)	Union
Normal split reads (total)	26,176,970	14,201,048	40,378,018
Normal split reads (filtered)	3,931,662	8,885,713	12,817,375
With determined reading direction	3,886,601	8,778,103	12,664,704
Unique junction locations	82,191	128,239	145,658
Unique splice sites	156,763	243,515	270,957
Match with ENSEMBL	130,924	192,405	208,956
Match with lincRNA [255]	1,416	1,054	1,793
Within unannotated regions	16,236	36,193	42,463

Table 7.4.: Conservation of normal latimerian splice sites. The first column shows, how many coelacanth splice sites could be "aligned" to the relevant species. The second column describes, the number of splice sites, which are annotated as splice sites in this species. The abbreviation "pred." refers to "predicted" splice sites, with a `MaxEntScan` score > 3 in the aligned sequence. The last column summarizes the "conserved" splice sites, as the union of the "annotated" and "predicted" ones. "H or M" = human or mouse, "8 Species" refers to presence of that splice site in at least one of the eight other vertebrates in the latimeria-centered 9-way multiple sequence alignment.

Species	**No. of unique splice sites**			
Coelacanth	270,957			
	align.	**annot.**	**pred.**	**cons.**
Human	174,191	168,623	165,826	169,941
Mouse	169,828	163,894	162,067	165,803
H or M	183,838	178,055	176,661	179,201
Frog	168,385	136,047	159,808	162,243
Stickleback	146,684	69,617	136,065	138,591
8 Species	210,794	195,388	203,219	203,944

Table 7.5.: Conservation of splice sites of coelacanth lincRNAs.

Species	**Splice sites**		**Transcripts**	
Coelacanth	23,065		8,066	
	align.	**annot.**	**align.**	**annot.**
Human	447	254	310	146
Mouse	350	195	253	117
H or M	540	292	374	166
Frog	823	334	514	190
Stickleback	315	79	229	52
8 Species	1,839	733	1,135	394

of the splice sites (1,839 sites in 1,135 transcripts) in this combined lincRNA set were alignable to sequence in at least one of the other eight vertebrate genomes included in the latimeria-centered MSA (Table 7.5). Of these, 40 % exactly correspond to an annotated splice site in at least one of these species, providing direct evidence for the partial conservation of 301 lincRNA loci (merged from 391 transcripts).

The rather poor conservation of lincRNAs as measured by splice sites does not come as a surprise, since only a small fraction of the observed splice junctions were included in the multiple sequence alignments in the first place. Their level of sequence conservation was very low compared to other functional transcripts [20, 22], although there is good evidence that, at least as a group, mRNA-like non-coding RNAs are under stabilizing selection [19–21, 197].

7.2.3. Circularized transcripts

For *L. menadoensis* and *L. chalumnae* we observed 5,760 circularizing junctions and 17,066 trans-splicing junctions. For a fraction of 10.6 % and 28.7 %, respectively, we were able to determine a reading direction, based on canonical splice motifs. Thus 610 circular junctions remain, consisting of 1,120 canonical splice sites. Almost half of these splice sites (501) are also utilized in regular, colinear splice junctions. They are surprisingly well conserved: more than 60 % are located in a region that is alignable in at least one other distant vertebrate and more than a third of these positions constitute a functional splice site according to the available experimental evidence, see Table 7.6. A comparison of circularizing splice junctions with recent reports of circular microRNA sponges in the human transcriptome [64, 65] did not provide evidence for the conservation of these particular RNAs between mammals and coelacanth, however.

Table 7.6.: Conservation of circular latimerian splice sites. Since 501 of the 1,120 circular splice sites are also involved in normal splice events, we only used the remaining 619 for the conservation statistic. For a column description see Table 7.4.

Species	**No. of unique splice sites**			
Coelacanth	619			
	align.	**annot.**	**pred.**	**cons.**
Human	282	117	103	132
Mouse	273	102	96	116
H or M	296	126	115	147
Frog	291	81	102	109
Stickleback	263	39	81	89
8 Species	375	147	173	202

Table 7.7.: Conservation of latimerian trans-splice sites. Since 1,116 of the 7,486 trans-splice sites are also involved in normal splice events, we only used the remaining 6,370 for the conservation statistic. For a column description see Table 7.4.

Species	**No. of unique splice sites**			
Coelacanth	6,370			
	align.	**annot.**	**pred.**	**cons.**
Human	1,887	1,616	1,607	1,653
Mouse	1,815	1,540	1,534	1,583
H or M	2,023	1,738	1,746	1,781
Frog	1,814	1,249	1,525	1,550
Stickleback	1,545	640	1,274	1,306
8 Species	2,483	1,964	2,128	2,150

7.2.4. Trans-spliced transcripts

In the combined Latimeria RNA-seq data we found 17,066 trans-splice junctions connecting different scaffolds. Among these are 338 that are backed by more than 100 split reads. The majority of these splice sites were unique to trans-splicing events.

Table 7.7 summarizes the conservation of the trans-splicing sites. Only a third of them could be aligned to homologous sequences in other vertebrates. In most of these cases we observed a functional splice site in the other species. However, in general, the specificity for non-local junctions does not appear to be as conserved across species as other splice sites.

7.3. Discussion

Atypical transcripts, characterized by mapping non-locally or non-linearly to the reference genome, become more and more recognized as a prevalent part of the RNA world. In this contribution we analyzed in detail the available RNA-seq data of two coelacanth species, *L. chalumnae* and *L. menadoensis*. The improved mapping algorithm implemented in `segemehl` [227], which deals efficiently with both typical and atypical transcripts, allowed us to paint a comprehensive picture of the diverse coelacanth transcriptome. In particular we report 51,488 additional expressed loci from which normal transcripts arise (576 protein-coding and 37,099 lincRNAs), together with 362 splice sites of circular RNAs and 4,698 of long-range (trans-spliced) connections. The very high fraction of junctions that use canonical splice sites is a strong indicator that the overwhelming majority of these transcripts cannot be dismissed as artificial products of RT-based technology but instead must be interpreted as biological reality.

The use of comparative splice site maps provides the most remarkable finding of this study, which is the unexpectedly high level of evolutionary conservation of splice sites involved in circularization. Especially, as the majority of these sites is exclusive to circularized transcripts. Their conservation indicates that they play an important key role in cell functionality. Recent reports of abundant, stable and often conserved circular RNAs in mammals have identified them as a crucial class of regulatory molecules [59, 64, 65]. Our results show that such "atypical" transcripts are evolutionarily old, dating back at least to an osteichthyan ancestor. Non-locally spliced transcripts are even less well understood. The statistical similarities in splice site usage and conservation between trans-spliced and circularized products, suggests that at least a subset is also functional. This observation is further strengthened by evolutionary conservation of a fraction of the non-local trans-splice sites, albeit a smaller one than with the circularizing sites. Future exploration of the functional significance of "atypical" transcripts, such as these, promises to yield many new insights.

Contents

Chapter 8

Evolution of Alzheimer associated genes

Alzheimer's disease (AD) is an age-related chronic neurodegenerative disorder of unknown cause with complex genetic and environmental traits. It is pathologically characterized by neurofibrillar aggregates of Aβ-peptides and the microtubule-associated protein tau. Transgenic mice models of AD have been successfully established for therapeutic research. However, the observations that have been made with these mice models could not be translated into effective therapies for AD patients by now. While AD is extremely prevalent in human elderly, both Aβ and tau pathology are less common in non-primate mammals, and even non-human primates develop only an incomplete form of the disease [263]. This human-specificity suggests a phylogenetic aspect of AD. Still, the evolutionary dimension of the AD pathomechanism remains difficult to prove and has not been established unequivocally so far. Defining those clear-cut phylogenetic traits of the AD pathomechanism, however, will have far reaching consequences with respect to our approaches of disease prevention and therapy including defining appropriate model systems.

To prove the contribution of brain evolution towards the AD pathomechanism, we applied the systematic analysis on the conservation of splice sites, as described in Chapter 5, to a data set of AD-associated protein-coding and non-coding genes. The

AD-associated genome-wide RNA profile, comprising both the protein-coding (cRNA) and non-protein-coding (ncRNA) transcripts, was established through microarray analysis in preceding work, which will be outlined in the following Section 8.1.

Genome-wide studies that systematically analyze the evolutionary age of protein-coding and non-protein-coding AD-associated genes have not been performed previously. While major evolutionary changes might have occurred at the transcriptomic level, they appear to be particularly pronounced for lncRNAs [28, 35]. As shown by analyses of sequenced genomes of a large variety of species, the relative amount of non-coding sequence increases consistently with complexity [12]. Thus, lncRNAs, most likely constitute a critical layer of gene regulation in complex organisms that have expanded during evolution [264]. However, the evolutionary histories of lncRNAs are hard to study due to their usually low level of sequence conservation (discussed in Chapter 3). This not only hampers comprehensive homology-based annotation efforts but also makes it nearly impossible to obtain the high fidelity sequence alignments that are required for in depth studies into their evolution.

As elucidated throughout this thesis, we can utilize the conservation of gene structure, or more precisely the conservation of splice sites to establish homology of lncRNAs. We have already shown in previous research [35] that lncRNAs, although clearly ancient components of vertebrate genomes, exhibit a rapid turnover of their intron/exon structures, which may be indicative of functional adaptation.

While the disease-relevance of lncRNAs is increasingly recognized, previous systematic gene expression profiling studies nevertheless focused predominantly on protein-coding genes. Consequently, so far, only a few individual AD-associated ncRNAs have been identified and functionally characterized [265].

8.1. Previous work

To detect the conservation of splice sites, specifically of AD-associated genes, we used a set of transcripts, that were identified to be differentially expressed in AD. This set was obtained via the employment of a microarray and subsequently the use of a variety of bioinformatic methods. This stages of the workflow were performed previously in the course of a collaboration of the *Paul-Flechsig-Institute for Brain Research* (PFI, Leipzig, Germany), together with the RNomics group at the *Fraunhofer Institute for Cell Therapy and Immunology* (IZI, Leipzig, Germany), and the Bioinformatics Group in the Department of Computer Science at *University of Leipzig* (Germany).

In the following the work stages and the methods that were used to obtain the data set, which was used to generate a comparative splice site map are explained.

Besides Appendix C.1, where the steps are described in more detail, it is refered to these individuals, who designed and performed the array experiments and performed the bioinformatic analysis of the expression data:

- Prof. Dr. Thomas Arendt, `thomas.arendt@medizin.uni-leipzig.de`
- Dr. Christian Arnold, `christian.arnold@embl.de`
- Dr. Kristin Reiche, `kristin.reiche@izi.fraunhofer.de`
- Dr. Jörg Hackermüller, `joerg.hackermueller@ufz.de`

An extensive description and evaluation of the process of identifying differentially expressed loci can be found in the Ph.D. thesis of Christian Arnold [266].

8.1.1. Microarray workflow

At first a custom array was designed comprising 931,898 probes derived from Agilent's Whole Human Genome Oligo array, lncRNA probes extracted from public databases, computationally predicted loci of structured RNAs, and lncRNA probes experimentally identified by transcriptome-wide expression variation studies based on the Affymetrix Human Tiling 1.0 array comparing AD patients with control samples. Applying this custom array to 19 AD patients and 22 age-matched control samples, we identified a differential expression of 154 multi-exonic cRNAs with a total of 4,162 splice sites and 141 multi-exonic lncRNAs with a total of 1,297 splice sites. We will refer to these loci as differentially expressed regions (Figure 8.1, blue barrel), which we used to obtain our final signal data sets in the next step.

8.2. Data sets

We compiled lists of splice sites from the GENCODE v14 annotation for human protein-coding genes and long non-coding RNAs. Genome wide multiple sequence alignments across 18 vertebrate genomes were used to construct two maps of conserved splice sites as background sets as described in Chapter 5. Independently for cRNAs and lncRNAs, the signal sets were obtained as intersection of the background maps with the differentially expressed regions. Thus the signal sets are designed as strict subsets of the background – a trait that ensures comparability. The splice site conservation rates of the background were then compared to the conservation rates of the signal. To evaluate the statistical significance of the differences in the conservation rates, we computed an empirical p-value by drawing 1,000 random samples of matching size from the GENCODE-derived backgrounds. We used $p < 0.05$ as

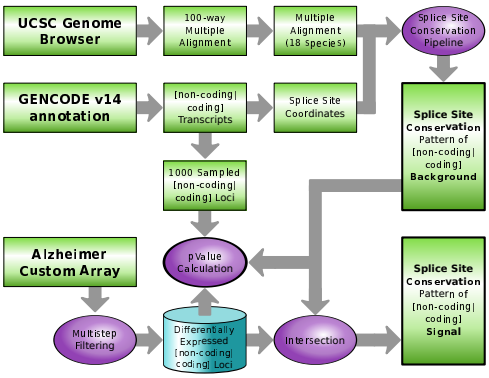

Figure 8.1.: Method workflow. We constructed background sets from the GENCODE v14 annotation and generated their splice site conservation maps with genome-wide multiple sequence alignment. The splice site conservation map for the AD-related genes is obtained as the intersection of the differentially expressed custom array loci with the background set. Empirical p-values are computed from sets of random loci of matching size drawn from the GENCODE-derived background to evaluate the statistical significance.

significance threshold. Compare with Figure 8.1 for a schematic work flow.

Additionally, a control set was obtained as intersection with regions that are expressed in human brain, to preclude a possible bias in the results of AD-associated regions towards generally brain-expressed transcripts. We used brain expression data from the study of Necsulea *et al.* [28] here.

8.3. Results

In order to compare the conservation of genes at a structural level, we classify the data by the "degree of conservation" c, which is the fraction of conserved splice junctions per gene. We ask – for a fixed value of c – whether loci that are differentially expressed in AD patients show signs of accelerated evolution compared to the set of genes, which are included in the GENCODE v14 annotation of the human genome.

8.3.1. Protein-coding AD-associated genes are not younger than background

Nearly all AD-associated protein-coding genes are evolutionarily old (Figure 8.2D). There were no differences in conservation rate at $c > 0\%$ between AD-associated and all protein-coding genes, i.e., AD-associated protein-coding genes did not originate later in evolution than other protein-coding genes. In line with previous reports [264], lncRNAs are much less well conserved and many have emerged in the course

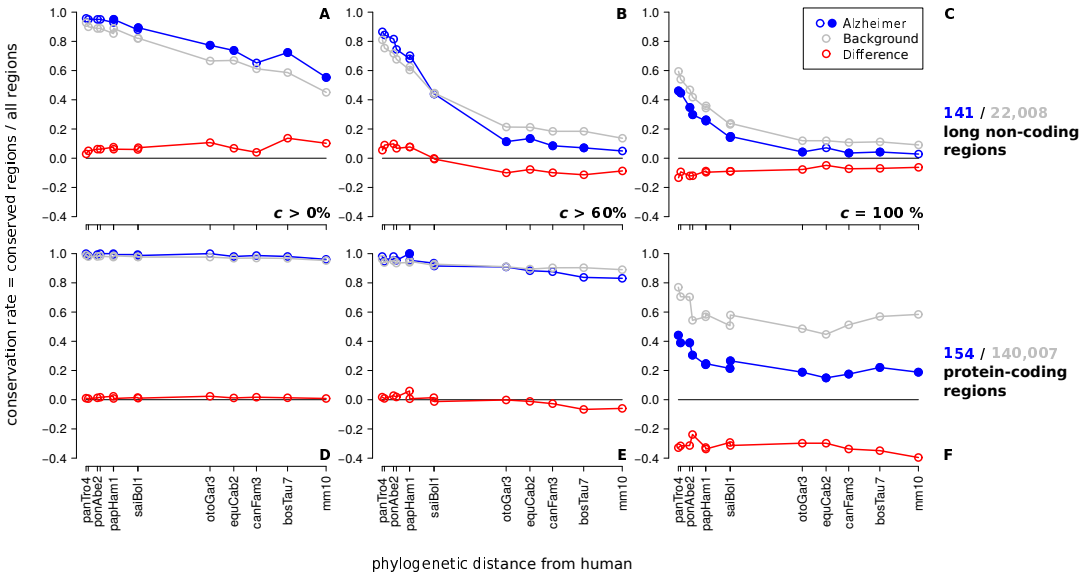

Figure 8.2.: Conservation rates of human AD-associated non-protein-coding (A-C) and protein-coding (D-F) regions for different conservation degrees ($c > 0\%$, $c > 60\%$, $c = 100\%$). On the horizontal axis mammalian species are indicated (denoted by the UCSC abbreviations, which can be found on page 147*f.*) at their phylogenetic distance from human. Distinct data points are connected by lines to guide the eye. Variations in assembly and alignment quality cause some non-monotonicity in the curves, the overall decrease of conservation with phylogenetic distance is nevertheless clearly visible. Statistical significance of differences is computed independently for each species. Filled circles indicate $p < 0.05$. The fraction of detectable conserved AD-associated non-coding genes is marginally higher than the conservation of the background set non-coding transcripts if only presence/absence of a transcript is considered (**A**). In contrast, if conservation of the entire gene structure is considered, AD-associated genes are significantly less conserved than the control. This is true for both lncRNAs (**C**) and protein-coding genes (**F**). Additional controls against possible confounding effects e.g. of alignment quality in Figure 8.3 and Figure 8.4 corroborate that the trends shown here are robust.

of mammalian evolution. The fraction of conserved lncRNAs thus decreases rapidly with evolutionary time (Figure 8.2A-C). As for protein-coding sequences we do not observe a significantly younger origin of AD-associated genes.

8.3.2. AD-associated genes are subject to accelerated change of gene structure

While there is no recognizable difference in the evolutionary age of origin between AD-associated genes compared to the transcriptome as a whole (Figure 8.2D), we observe significant, albeit more subtle differences in the evolution of AD-associated and general lncRNAs, concerning the changes in gene structure. With an increasing degree of conservation c, the initially higher conservation rate of AD-associated non-coding genes decreases and eventually falls distinguishably below the background level

(Figure 8.2A-C). The difference between AD-associated and general lncRNA genes becomes significant for $c > 60\%$ ($p < 0.05$) in the comparison with distantly related mammals. When complete conservation of gene structure is considered, $c = 100\%$, the lower conservation rate of AD-associated ncRNAs becomes significant even in primates. In other words, the fraction of transcripts that have the entirety of their splice sites conserved is smaller amongst AD-associated ncRNAs than amongst non-coding genes at large. AD-associated ncRNAs hence show an accelerated evolution of their gene structure. This is indicative of a more rapid functional adaptation of AD-associated non-coding genes.

Despite the very high conservation rates of protein-coding genes in general, we observe the same increase of splice site turnover in AD at $c = 100\%$. In fact, the relative effect is even stronger compared to non-protein-coding loci ($\approx$ 30-40% *versus* $\approx$ 5-15% difference, shown as red lines in Figure 8.2C and 8.2F, respectively). However, even a moderate level of splice site turnover is much less common for protein-coding genes than for non-coding genes. This is reflected by the negligible differences between the conservation rates of signal and background for $c > 60\%$. Since the same fraction of transcripts is already detectable at low conservation degrees, while the conservation rate decreases with higher c, we conclude that splice sites are systematically less conserved in human AD-associated regions compared to the typical behavior of the transcriptome. While protein-coding loci exhibit an enhanced rate of small changes in their gene structure, we observe large changes in lncRNAs, again with a significantly enhanced rate in the AD-associated ncRNAs. This suggests that in particular AD-associated non-coding genes play an important, as yet largely unexplored, role in the AD pathomechanisms.

8.3.3. Upper bounds of conservation rates are consistent with findings

Quality and completeness of the underlying alignment may influence the conservation results. Naturally, alignments do have gaps and not every splice site has an ortholog position aligned in each species. This is a concern in particular for non-model organisms. As a control, we therefore also calculated the percentage of positions that can be aligned independently of any splice site conservation (Figure 8.3) and the fraction of conserved genes among alignable genes (Figure 8.4), which represents the upper bounds of transcript conservation rate. The overall trends and the conclusion of the analysis remain unchanged, when taking into account the fraction of unaligned and annotated data.

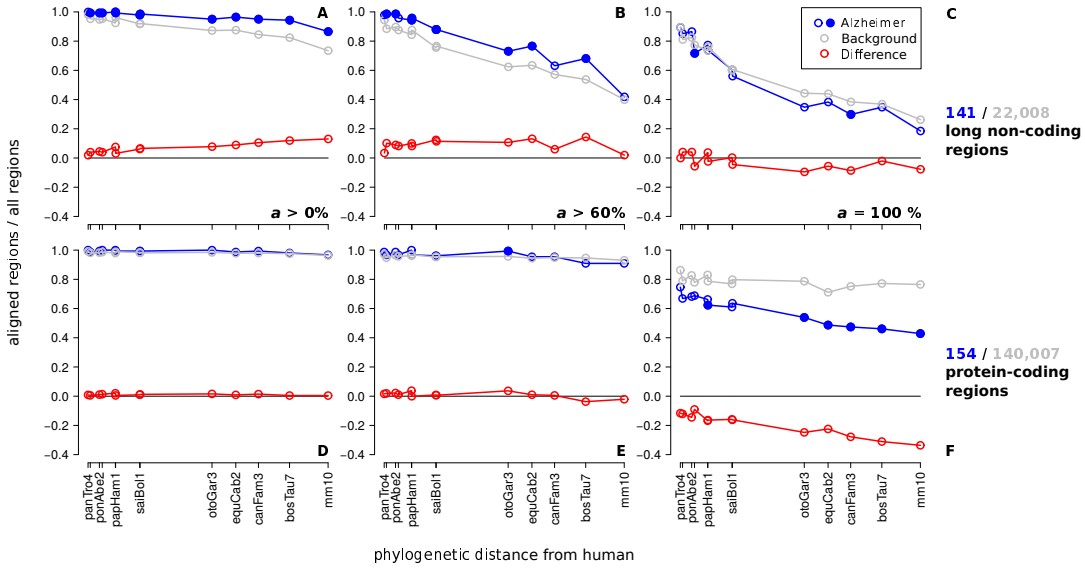

Figure 8.3.: Fraction of alignable human AD-associated non-protein-coding (A-C) and protein-coding (D-F) regions for different degrees of alignability ($a > 0\%$, $a > 60\%$,$a = 100\%$), that is the fraction of splice sites per region which are alignable to another species. On the horizontal axis mammalian species are indicated (denoted by the UCSC abbreviations) at their phylogenetic distance from human. Distinct data points are connected by lines to guide the eye. Statistical significance of differences is computed independently for each species. Filled circles indicate $p < 0.05$.

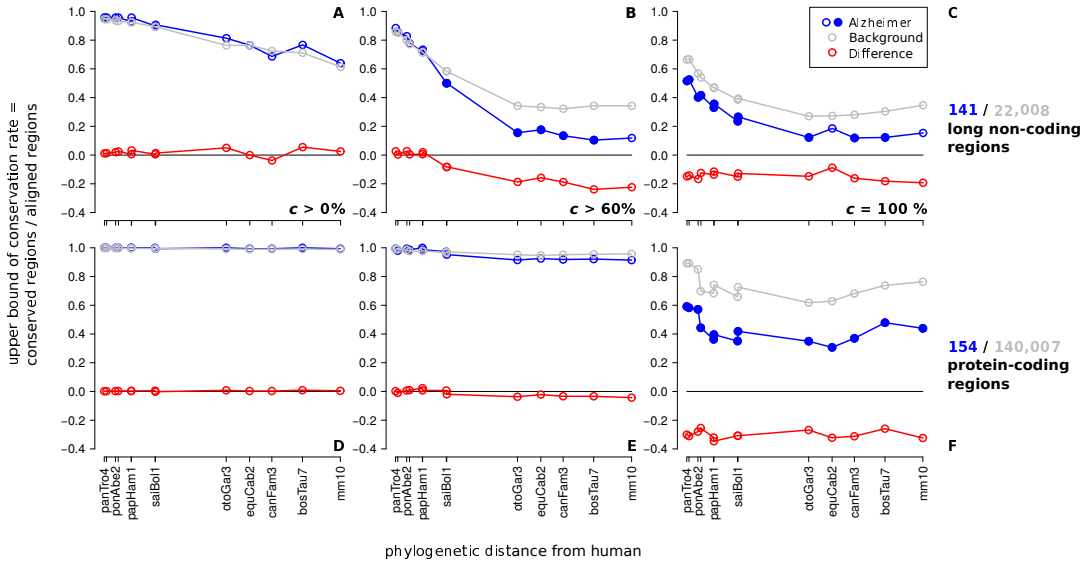

Figure 8.4.: Upper bound of conservation rates of human AD-associated non-protein-coding (A-C) and protein-coding (D-F) regions for different conservation degrees ($c > 0\%$, $c > 60\%$, $c = 100\%$). The amount of aligned regions is delimited by the degree of alignability $a = c$. The legend of this figure is analogous to Figure 8.3.

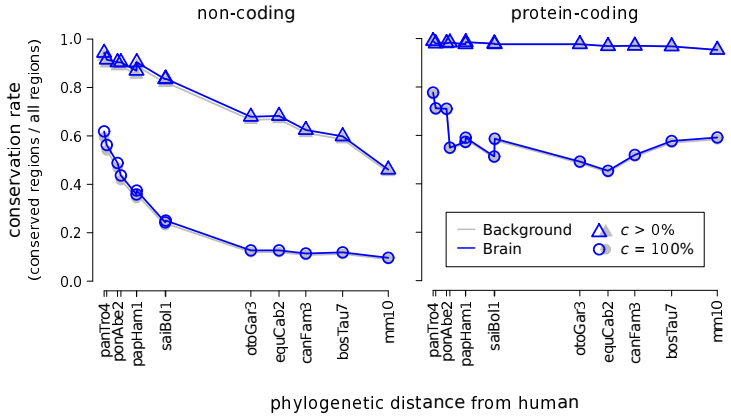

Figure 8.5.: Conservation rates of human brain-expressed non-coding and protein-coding genes in comparison with the respective background of GENCODE v14 annotated genes for different degrees of conservation ($c > 0\%$, $c = 100\%$).

8.3.4. No brain related bias in data

When we employ the control set of brain-expressed human transcripts as background, all found results remain valid, since the difference in conservation rates compared to GENCODE v14 annotated genes at large and the subset known to be expression in brain is marginal for all degrees of conservation c for both protein-coding and non-coding genes, Figure 8.5.

8.4. Discussion

We have shown here that gene structures of both lncRNAs and proteins associated with AD evolve faster than the genome at large, while there is no evidence that AD-associated genes originated particularly late in evolution.

The enhanced rate of gene structure evolution in AD-related genes hints a relation of AD to recent adaptive evolution, presumably in relation to the rapid evolution of the human brain, which may have caused changes of cerebral structure and function that have rendered the human brain sensitive to AD [267]. Importantly, replacing the background set by only genes expressed in brain does not affect the conclusions. Major phenotypic brain changes that have occurred in the course of recent human evolution, in particular between human and chimpanzee, appear to be mostly the result of an increase in gene expression and are, thus, reflected at the transcriptomic level. [268–270]. Genes whose expression has increased in human brain are mainly related to growth and differentiation [271] and frequently are involved in transcriptional regulation and RNA processing [268, 269]. The most significant differences

in gene expression between the human and non-human primate brain have been observed in the association cortex [268, 272], i.e., brain areas that have expanded during hominid evolution [273] and are affected in AD most early and most constantly [274]. Evolutionary expansion of the neocortex, and in particular phylogenetic shaping of association areas, is associated with a developmental deceleration and an extended period of high neuronal plasticity into adulthood [271]. The presence of these neurons which remain structurally immature throughout their lifespans might provide the prerequisite both for the human adaption to the "cognitive niche" and for a high vulnerability towards factors that lead to the development of AD [275–277].

Our data support the concept that neuronal vulnerability in AD is a result of the evolutionary legacies that have occurred during the course of evolution of the human brain, making AD an example of antagonistic pleiotropy. This evidence for a phylogenetic trait of AD highlights the necessity to reconsider our approaches to define the molecular pathology of AD and the appropriateness of current animal model systems [278] to develop disease-modifying strategies.

Chapter 9

Conclusion

This thesis introduces a method to reliably predict the conservation of RNA, based on the conservation of their splice sites. This enables a prediction independent from coding capacity in form of preserved open reading frames, what makes this method suitable particularly for investigating the conservation of non-coding RNAs. By focusing on the conservation of gene structure, represented through splice sites, it is possible to capture the evolution of lncRNA transcripts separately from other selective constraints such as regulatory DNA elements that may affect sequence conservation.

This approach employs comparative splice site maps, which are generated from transcriptomics data together with multiple sequence alignments. The splice site motif scoring method of `MaxEntScan` is used to assess the conservation of orthologous sites. The accuracy of the method will profit from future advances in large-scale in-depth sequencing technologies, that can provide more comprehensive transcriptome and genome assemblies to tone down the limiting factors of alignment imperfections and incomplete transcriptome annotations. However, when the method is applied to RefSeq annotated protein-coding RNAs, the prediction of conserved splice sites in the extensively explored model-organism mouse, almost perfectly matches with experimentally validated sites. This fact in combination with the low estimated false positive rate, corroborates the high precision and robustness of the method.

The specificity of the chosen reference transcriptomics data and the range of species in the employed multiple sequence alignment can be adapted to find answers for a broad scope of evolutionary questions.

In Chapter 6 we used comparative maps among non-coding splice sites to predict the conservation of human lncRNAs across 46 vertebrate genomes. The number of evolutionarily conserved single splice sites (e.g. $\sim 13\,\%$ of all splice sites between human and mouse), provide a lower bound on the estimated number of conserved lncRNA and are in good agreement with previously suggested conservation rates in the studies of Washietl *et al.* [27], Necsulea *et al.* [28] and Managadze *et al.* [29]. Considering a lncRNA transcript as evolutionarily conserved, if at least one of its splice sites is present in the other species ($c > 0\%$) we were able to trace more than $85\,\%$ of human lncRNAs back to the divergence of placental mammals. This number presumably constitutes an upper bound. More examples of lncRNAs with relatively high levels of sequence conservation that also exhibit completely or partially conserved gene structure between mouse and human can be found in the `slncky` browser [128], confirming also many of the computational findings in Chapter 6.

RNA trans-splicing and circularization increase the potential of genetic information to form various products which enrich the diversity of the proteome or the regulatory machinery. Although trans-splice events are reported to occur more frequently in lower species than in higher vertebrates, we identified more than $17,000$ trans-connecting splice junctions (donor–acceptor pairs) and nearly $5,800$ circularizing junctions in a joined data set of RNA-seq data from two latimerian species by employing the specialized mapping tool `segemehl`. The abundance of these atypical transcripts suggests that they are in fact a previously hidden component of vertebrate genomes. A decent fraction of those junctions uses canonical splice sites, indicating a spliceosome-mediated splicing process which eliminates the possibility that these transcripts are merely RTfacts.

The conservation analysis of the subsets of trans-splice and circular splice sites, that were canonical and exclusively involved in atypical splicing, yielded an unexpectedly high level of evolutionary conservation for both sets as further evidence against the hypothesis of "splicing noise" from aberrant transcription. Our results reveal that these transcripts are evolutionarily old and must have been present at least at the divergence of tetrapods and teleostei, as we find orthologs between coelacanth and human. This indicates that chimeric and circular RNAs are of importance for physiological cell functions which also suggests a pathological role.

Indeed, circular RNAs have been found to be associated with the occurrence of diverse human diseases [279]. In recent studies they have been described as abundant

stable transcripts of eukaryotic cells, with spatio-temporal specific expression patterns, that are especially enriched in the human brain [56, 59, 63]. Since circRNAs have been discovered to act as regulators of gene expression in the role of miRNA sponges [64, 65], there is emerging evidence that a dysregulation of circRNAs may impact the pathology of various human diseases, e.g. Alzheimer's disease [280].

By establishing an AD-associated genome-wide RNA-profile of both protein-coding and non-protein-coding transcripts, we were able to investigate the evolution of AD. Since AD is a young disease from the evolutionary perspective, we expected to find little conservation in distant species. However, we could show that AD-associated genes did not originate later than non-AD-associated genes. In fact, we detected the same up to a significantly ($p < 0.05$) higher fraction of these genes in the respective species ($c > 0\%$). Conversely, when comparing the conservation rates of transcripts with a completely conserved gene structure ($c = 100\%$), we saw significantly less conserved AD-associated than non-AD-associated genes for both protein-coding and non-protein-coding regions. This is striking evidence for an accelerated evolution of AD-related genes. Importantly, genes expressed in brain do not exhibit this peculiar conservation pattern and instead are nearly congruent with the background.

Changes in gene structure can be expected to have in general larger functional effects than point mutations. The enhanced evolution rate of gene structure in AD-associated genes supports the view of AD as a consequence of recent rapid adaptation of genes involved in functionality and cerebral structure of the human brain [267]. This phylogenetic trait highlights the necessity for a paradigmatic change of AD concepts and the need to reconsider the appropriateness of current animal-models to develop disease-modifying strategies. Non-coding genes in particular play an important, as yet largely unexplored role in AD.

At present we can only assert that lncRNAs in general show a high level of variability in their gene structure. In the absence of data that would allow us to locate specific functions or molecular interactions to individual exons, we can only speculate about the functional meaning of the observed rapid turnover. The most plausible view is that many lncRNAs act as "coat hangers" [281, 282], i.e., interaction partners, for several proteins and RNA partners. Turnover in gene structure thus would translate into different composition and thus likely modified molecular functions of ribonucleo-particles. Small genetic changes therefore may be amplified to substantial effects at the functional level, making lncRNAs and their isoforms a prime candidate to understand rapid, lineage-specific adaptations and exaptations.

Appendices

Contents

Appendix A

Conservation of human RNAs

A.1. Supplementary results

We used RefSeq annotated transcripts in order to investigate the conservation of protein-coding sequences. Table A.1 shows the absolute conservation of coding splice sites for four chosen mammals, namely mouse, rat, dog, and cow. The chart in Figure A.1 illustrates, that the level of predicted conservation (blue and cyan colored) is similar in all of these species. The high annotation rate of predicted conserved splice sites in mouse (cyan) suggests that to a large extent the predicted splice sites in the remaining species are unannotated protein-coding splice sites.

Table A.1.: Conservation of RefSeq splice sites. RefSeq annotated transcripts were used for estimation of coding transcripts only, since the majority of the non-coding RefSeq transcripts are still associated with coding loci.

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

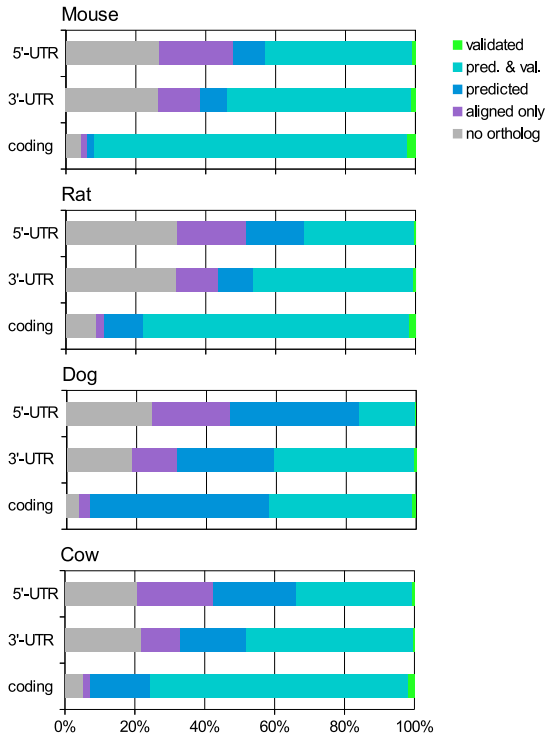

Figure A.1.: Conservation of RefSeq splice sites in mouse, rat, dog and cow. Graphical illustration of numbers displayed in Table A.1.

Contents

Appendix B

Identification of atypical transcripts in coelacanths

B.1. Supplementary methods

B.1.1. Variation calling

Variation within and between the two coelacanth species was quantified by determining SNPs of mapped transcriptome at all sites with a coverage of at least 8 reads. We used `GATK` version 2.3 [283] for SNP calling.

B.1.2. Circular motif search

In order to find a putative motif that is predominantly associated with circular junctions, we extracted 6 nt of DNA sequence at each splice site (3 nt in exon and intron) and combined it to form a 12 nt sequence pattern for each splice junction. This results in 5,561 unique patterns for 5,760 circular splice junctions and 27,311 unique patterns for 213,417 normal splice junctions. We employed `MEME` [284] for the motif search. It was run with the "zero or one match per sequence" option. As expected, the canonical splice junction motif was readily recovered. After removing about 700 12-mers that conform to a canonical or minor spliceosome motif, we again started the

`MEME` motif search to see if any additional characteristic patterns could be detected. This was not the case.

B.1.3. SHSs and RTfacts

We analyzed to which extent the splice junctions of the data sets can be explained by the short homologous sequences model proposed in Li *et al.* [45] or by RT PCR artifacts that show similar sequence homology (cf. [48]). We thus computed the maximal length of the homologous subsequences between the exonic regions of the donor and acceptor splice sites. An exact overlap of at least 4 nt was counted as "short homologous sequence" (SHS), which may indicate an RTfact.

B.1.4. Coverage estimation for splice junctions

In order to investigate the relationships between RNA expression and abundance of spliced RNA reads, we defined "coverage loci" as follows: We considered genomic regions with a minimum coverage of 8 reads and merged sites separated by less than 100 nt. Sites smaller than 50 nt were removed from further analysis. To account for inaccuracies in determining the boundaries of loci, we counted all spliced reads with a splice junction within 50 nt of a "coverage locus".

B.1.5. Validation experiments

The first-strand cDNA and genomic DNA from muscle of *L. menadoensis* was amplified by thermal cycling using the Takara ExTaq PCR kit (Takara, Japan). Primer pairs were designed to generate a PCR product that spanned the fusion site for these transcripts. Additional control primers were designed to amplify sequences present only in the local genomic contexts. Amplification was performed for 30 cycles at 94 °C for 15 s, 55 °C for 30 s, and 72 °C for 1 min, with a final elongation for 8 min at 72 °C. The amplified PCR products were sub-cloned into the pCRII-TOPO dual promoter vector (Invitrogen) and sequenced.

B.2. Supplementary results

Mapping and variation

Between 75 % and 80 % of the reads in the individual RNA-seq data sets could be mapped to the reference genome. Between 1/6 and 1/5 of these mapped with splits.

In addition, sub-sampled libraries were mapped to obtain comparable sample sizes for quantifying circular and trans-spliced reads.

The RNA-seq libraries of *L. chalumnae* muscle tissue and *L. menadoensis* liver and testis were of comparable size and quality, covering slightly more than 1 % of the genome assembly. Using these transcriptome data as a reference, the two Latimeria species were very similar. The *L. menadoensis* transcripts showed only about 0.3 % divergence from the *L. chalumnae* reference genome, while the number of heterozygous SNPs, i.e., the intra-specific variation in *L. menadoensis*, was about twice as large. The number of homozygous differences between transcriptome and reference genome barely exceeded 0.1 % and was consistent with about 0.4 % heterozygous SNPs in *L. chalumnae* RNA-seq data. The small divergence relative to the intra-specific diversity justified a joint analysis of all coelacanth transcriptome data in the following.

Comparison of normal and atypical transcripts

For a better comparison of the properties of circular and trans-spliced transcripts in the individual data sets we used sub-samples of equal size. In this way we obtained a comparable sequencing depth, which should at least alleviate the biases arising from very rare junctions in the largest data sets. While this simple normalization cannot account for differences in the expression profiles of the different tissues it should at least make the data sets qualitatively comparable.

Results for the two coelacanth species are very similar, hence we use their union. We compared atypical reads with normal (local and colinear) splice events for coelacanth, human, and zebrafish RNA libraries. As expected, the overwhelming majority of normal splice events utilizes canonical splice patterns. In contrast, circular and trans (long-range) splice events often use alternative sequence patterns, although a substantial fraction still conforms to the canonical motifs. We observed that in the coelacanth data, more circularizing splice junctions are off by 1 or 2 nt compared to both the human and the zebrafish data set. Adding these to the canonical subset, yielded nearly the same fraction of about $70-80$ % canonical splice motifs as zebrafish and human. We note that this fraction is substantially larger than the numbers reported in Li *et al.* [45]. Surprisingly, most of the circularizing and trans-joining splice junctions are disjoint from normal splice junctions. This effect is even more pronounced in coelacanth and zebrafish than in human. This pattern, which we observe for both the circularizing and the trans-junctions strongly suggests that the resulting unconventional transcripts are not merely a by-product of conventional, local splicing events.

Since a substantial fraction of the circular and trans-splice junctions did not fit the

canonical splice site motif, we searched for additional over-represented patterns in the remaining junctions. No significant pattern could be identified, however. We then searched for the "short homologous sequences", i.e., short sequences with four or more nucleotides, shared by the sequences surrounding the "splice junction". According to Houseley and Tollervey [48], however, these might be RTfacts. We found that such patterns are rare in our data, ranging from 0.7 % to 2.6 % of the circularized or trans-spliced transcripts. At the same time, the majority of atypical junctions are associated with canonical splice site motifs. We thus conclude that contamination levels in our data are low and the majority of both circular and trans-spliced RNAs cannot be explained as technical artifacts.

Contents

Appendix C

Spadework of the Alzheimer Project

The Alzheimer Project [37] was a collaboration between the PFI, IZI and the Bioinf Group of University Leipzig (see Section 8.1). Therefore multiple people contributed to the project at certain stages, like lab work and microarray processing. This chapter is based on the Supplement of the resulting publication [37] and provides a detailed description of all worksteps that were implemented by the co-authors previously to the computational conservation analysis, which was performed on the resulting data set of differentially expressed loci. Section C.2 specifies once more how the final data set was obtained.

C.1. Supplementary methods

C.1.1. Patient and control samples

We used brain tissue from 41 deceased subjects, 22 of whom developed AD before they died; the other 19 were considered healthy controls with no history of neurological or psychiatric illness. The diagnosis of AD was made on the basis of both clinical and neuropathological evidence according to the criteria of the International

Working Group (IWG) for New Research Criteria for the diagnosis of AD [285, 286] in the revision of 2014 (IWG-2) [287], the NIA-AA diagnostic criteria in the revision of 2011 [288–291], and the NIA-AA guidelines for the neuropathological assessment of AD [292, 293]. Only cases with typical AD according to the IWG-2 criteria were included. All cases had undergone neuropsychological assessment during the final six months of their lives. Clinical Dementia Rating (CDR) scale scoring was based on neuropsychological testing (CERAD) [294], MMSE [295], and rating scales [296]. All cases were neuropathologically assessed for NFT stage according to Braak and Braak [274, 297] and Braak et al. [298], for Aβ/amyloid plaque score according to Thal et al. [299], and for neuritic plaque score according to CERAD [300]. NFTs and Aβ/amyloid plaques were detected by immunocytochemical labeling of phosphotau (anti-human PHF-tau monoclonal antibody AT8; Thermo Scientific) and Aβ (beta amyloid monoclonal antibody, 6E10; BioLegend), respectively. Severity of AD pathology was staged following the consensus guidelines for the neuropathologic evaluation of AD according to Hyman et al. [292] and Montine et al. [293].

Case recruitment, autopsy, and data handling were performed in accordance with the ethical standards as laid down in the 1964 Declaration of Helsinki and its later amendments as well as with the convention of the Council of Europe on Human Rights and Biomedicine, and were approved by the responsible Ethics Committee of Leipzig University.

C.1.2. RNA isolation

RNA was isolated by TRIzol^{TM} method (Invitrogen, Karlsruhe, Germany). 100 mg deeply frozen human brain tissue (temporal cortex) was homogenized in the presence of 1 ml Trizol in a glass-Teflon^{TM} homogenizer. The homogenate was transferred to a microtube and after adding chloroform, samples were centrifuged at 15,000 g (4 °C) for 15 min and the supernatant transferred to a fresh tube. Samples were mixed with equal amounts of isopropanol and centrifuged at 12,000 g (4 °C) for 15 min to precipitate the RNA. After washing, the pellet was air-dried and dissolved in water. RNA quality was assessed by denaturing formaldehyde agarose gel electrophoresis, by spectrophotometry (scanning at $220-320$ nm) and by analysis using Agilent 2100 bioanalyzer. Only samples with RIN > 5 were further processed. The RNA concentration was estimated spectrophotometrically by absorbance at 260 nm, concentration was adjusted to 1 mg/ml and RNA was stored at -80 °C until use.

C.1.3. Whole genome tiling arrays

Equal masses of total RNA derived from three patient and three control samples, respectively, were pooled. The Affymetrix Human Whole Genome Tiling Array 1.0 Set consisting of 14 arrays was used according to the manufacturer's instructions, except that separate labeling reactions were used for each array starting from 10 μg pooled total RNA.

We used the TileShuffle algorithm described in [301] to determine expressed and differentially expressed genomic intervals. Affymetrix Human Whole Genome Tiling Array 1.0 Set raw signal intensities were mapped to human genome version NCBI36 using Affymetrix BPMAP files[1]. Expressed segments were detected with the TileShuffle parameter settings: window size = 200, the window score was defined as the arithmetic mean trimmed by the maximal and minimal values over signal intensities of all probes in a window, number of permutations = 10,000 and number of GC classes = 4. All windows with an adjusted $p < 0.05$ according to Benjamini and Hochberg [302] were defined to be significantly expressed. DE-TARs are differentially expressed TileShuffle intervals with adjusted $p < 0.05$ (window size = 200, the window score was defined as the log-fold-change discarding all probes with converse behavior as observed for the relevant significantly expressed windows, number of permutations = 100,000 and number of GC classes = 1). Finally, the genome coordinates of all significantly expressed and all significantly differentially expressed segments were lifted over[2] to GRCh37 (hg19).

C.1.4. Design of the `Alzheimer Custom Microarray`

Genomic intervals that were found expressed in the tiling array approach in AD or control were combined with regions we found differentially expressed in tiling array experiments on p53 induction, STAT3-signaling, cell cycle phases, and macroRNAs called STAiRs, described in [303], a list of manually curated AD-associated genes from literature, and other sources of annotated or predicted ncRNAs for probe design: Known lncRNAs retrieved from public databases — *NONCODE* [304], *lncRNAdb* [305], *fRNAdb* [306], *RNAdb* [307], *H-InvDB* [308], GENCODE v4 [309], *RefSeq* [310], from literature — lncRNAs originating from actively transcribed genes [311], chromatin-associated RNAs [312], snoRNAs from the `snoBoard` database [313], intronic RNAs identified in [314], and genomic intervals with RNA secondary structure under stabilizing selection (`RNAz` [315], `Evofold` [157]). Since natural antisense tran-

[1] `http://www.affymetrix.com/analysis/downloads/lf/tiling/Hs35b_MR_v02-2_NCBIv36_v2.bpmap.zip`

[2] `http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/liftOver`

scripts appear to regulate transcription and translation of neighboring genes (e.g. [316]), we designed probes antisense to protein-coding genes (GENCODE v4). MRNAs were represented by Agilent's 026652 catalog probe set, which is based on human RefSeq mRNA sequence. Also, we designed probes for all protein-coding genes found additionally in GENCODE v4.

Custom microarray probe design is a non-trivial task for pervasively transcribed genomes. The `CEM-designer` pipeline [317] was therefore used to facilitate (i) the collection and generation of a set of unified target sequences and (ii) the selection of a set of sensitive and specific probes that represent the target sequences best while meeting space constraints of the array. Target sequences shorter than 60 bp and duplicate target sequences (i.e., identical start and end positions) were discarded. Parts of non-coding annotations that overlapped coding sequences were removed to enable a clear separation between probes interrogating non-coding and coding transcripts.

Probe design was performed using Agilent's `eArray` platform, using standard parameters for expression arrays, in particular 60 bp probes and the base composition methodology, which aims at equally distributing probes across the target sequence. Probe uniqueness was checked against human genome assembly version hg19, rigorously discarding non-uniquely mapping probes using BLAT with options that maximize sensitivity (`-stepSize=5 -repMatch=1000000 -fine -minIdentity=90`). This design strategy ensured that probes were unique both on the DNA and RNA level (according to human genome version GRCh37/hg19 and all known RefSeq transcripts, respectively).

The number of probes per target sequence was set in dependence to target length. Target sequences were represented by exactly one probe if the length was $60 \leq l < 300$, three probes if the length was $300 \leq l < 600$, and five probes if the length was $600 \leq l < 1000$. Target sequences longer than 1000 bp were split into intervals of 60 bp overlapping ≤ 1000 bp chunks to ensure that probes may also be designed in the vicinity of the split positions. Each subsequence was then treated as an individual region subject to the design strategy as described above. For target sequences with an unknown reading strand (e.g., sequences originating from the various tiling array experiments, ncRNA predictions, and chromatin-associated ncRNAs), we designed probes for both strands.

Overall, the `Alzheimer Custom Microarray` contains 931,898 probes of which 905,197 are custom probes. A summary of the genomic distribution of probes is shown in Table C.1.

Table C.1.: Genomic distribution of probes for the `Alzheimer Custom Microarray`, based on GENCODE version 4, which was used for probe design. A probe corresponds to a category if it overlaps strand-specifically to at least 95 % (57 nucleotides) with at least one annotation (i.e., feature or sequence) of the category. For introns and intergenic regions, the strand information has been ignored. 5'UTRs and 3'UTRs correspond to 5' and 3' untranslated regions of mRNAs. CDS corresponds to the coding exons of mRNAs. The relative fraction is defined according to overall number of probes on the `Alzheimer Custom Array`. The total numbers in the last column may not add up to 100 % due to the mandatory control probes and probes that overlap with no category with at least 95 %.

Annotation category	Number of probes	Relative fraction (in %)
5'UTRs (sense)	39,233	4.21
5'UTRs (antisense)	38,021	4.08
CDS (sense)	70,451	7.56
CDS (antisense)	43,799	4.70
3'UTRs (sense)	101,297	10.87
3'UTRs (antisense)	73,340	7.87
Introns	388,881	41.73
Intergenic regions	162,803	17.47
Pseudogenes	8,201	0.88
Repeats	17,706	1.90

C.1.5. Processing of the `Alzheimer Custom Microarray`

Total RNA quality was checked using Agilent's 2100 Bioanalyzer and only samples with a RIN ≥ 5.0 were retained for microarray analysis. For 19 patient and 22 control samples 1 μg of total RNA was labeled using the Quick Amp Labeling Kit (Agilent, Waldbronn, Germany), according to the manufacturer's instructions with the adaptation of using 120 pmol of a random $N_6 - T7$ primer (Metabion, Planegg, Germany) instead of a polyT-T7 primer. cRNA quantity was checked using a NanoDrop ND-1000 UV-VIS Spectrophotometer, as enlisted in the manufacturer's instructions. 1.65 μg of labeled cRNA was used for hybridization following manufacturer's instructions. After hybridization the arrays were washed according to the manual and scanned using the Agilent G2565CA Microarray Scanner System with Agilent Scan Control Software (Version A851) following settings for scanning: Profile: AgilentG3_GX_1Color; Channels Green; Scan Region: Agilent HD (61×21.6 mm); Resolution 3 μm double pass; Tiff: 20 bit; Green PMT Gain: 100 %. Result tables were extracted after grid placement using Agilent Feature Extraction Software (Version 11.5.1.1)

C.1.6. Identification of differentially expressed probes

Differential expression analysis was performed using `R` and the `Bioconductor` package `Limma` [318]. Quality control of arrays were performed by checking distribution of "bright corner", "dark corner" probes, and relative spike-in concentration versus normalized signal. The controls confirmed high quality of the results and consequently all microarray data were included in the downstream analysis. Initially, independent filtering was performed, removing probes (i) with signal intensity above the background in less than one third of all arrays and (ii) exhibiting an interquartile range of log2 signal intensity across all samples of less than 1. Background expression was defined by the mean intensity plus three times the standard deviation of negative control spots (Agilent's 3xSLv spots). 113,047 out of 931,898 probes were retained after filtering. Signal intensities were quantile normalized [319] but not background corrected, due to the low background intensities of Agilent arrays.

Differential expression between AD and control samples was determined using a linear model that includes age because on average, individuals from the AD group were older than controls (∼81 and 65 years, respectively):

$$E[X_i] = \alpha \times \mathrm{AD} + \beta \times \mathrm{Age} + \epsilon \tag{C.1}$$

where $E[X_i]$ is the expected expression of probe i, ϵ an error term, α the coefficient modeling the impact of AD on the expression variance of probe i, and β the coefficient modeling the influence of the patient's age. The linear model was fitted using the `R` package `Limma` and reliable variance estimates were obtained by Empirical Bayes moderated t-statistics. False discovery rate was controlled by a modified Benjamini-Hochberg procedure that incorporates an estimated proportion of the null p-values [320] to compute q-values using the `fdrtool R` package [321, 322].

With $q < 0.2$, a comparably relaxed cutoff for controlling the false discovery rate of individual probes was chosen, because individual probes were subsequently aggregated for each annotated item, as described below. Probes meeting this cutoff and uniquely mapping to the genome defined the set s_{diff} that we used for all subsequent analyses.

C.1.7. Identification of differentially expressed loci

Subsequently, we identified differentially expressed loci aggregating differentially expressed probes. The rationale behind this step was to identify the set of genes that show particularly trustworthy signs of differential expression. We argue that the differential expression of an individual probe may not be a sufficient criterion for the

corresponding gene to be deemed differentially expressed. For example, consider the following case for a particular gene g for which one differentially expressed probe P_{diff} mapping to g has been identified. Among all probes that map to g, P_{diff} may be a false positive, and all other probes do not show signs of differential expression. Thus, further incorporating g may not be useful because other genes show much stronger and homogeneous signals with respect to differential expressions of probes.

Therefore, we deemed loci as differentially expressed if a significant fraction of probes overlapping the locus in sense direction exhibited differential expression in "the same direction", i.e., with same sign of log fold change, according to a binomial test ($p < 0.05$). Sources for annotations were equal to those used for probe design except for GENCODE, where v14 was used. GENCODE was used as primary annotation and all non-overlapping annotations from the other sources were used in addition. Probes were considered, if at least 95 % of its sequence overlapped with a particular annotation. Annotations were considered per gene, i.e., we considered overlaps with all exons of a gene and did not test for individual transcripts. If a probe mapped to multiple distinct annotated genes, we tested each gene individually but recorded the ambiguity to avoid losing potentially relevant signals. We considered only sense and discarded antisense overlap because the transcript structure is not known for transcripts that are antisense to annotated transcripts unless they map to known antisense transcripts, which were already included in the various annotation sources as listed above. However, for annotation items with an unknown reading direction (e.g., loci from the tiling array experiments, ncRNA predictions, caRNAs [312]), we ignored the strand information and considered all overlaps.

For each probe $P \in s_{\text{diff}}$, we determined whether p was located in a locus with known transcripts (protein-coding, non-coding, or pseudogenes, as described before). A probe was mapped to a particular gene if it was located in (i) an exon of at least one annotated splice variant (only for protein-coding transcripts because non-coding and pseudogene transcripts may exist in an unspliced and/or spliced version), (ii) the UTR of that gene, or (iii) in a putative previously unrecognized exon (no overlap with annotated exons but located in an exon of at least two spliced ESTs). Probes located exclusively intronic of a protein-coding gene (i.e., no overlap with annotated exons and less than two overlaps with exonic ESTs) were classified as putative intronic transcripts and therefore added to the non-coding list. If multiple introns overlapped, we used the cluster of overlapping introns as loci.

For each differentially expressed probe $P_i \in s_{\text{diff}}$ that overlapped with a particular differential expression candidate i (i.e., a locus with known or unknown transcript structure) in sense direction, we then identified the set of probes P_{all_i} that also overlapped with i with the criteria as described above (with respect to their genomic

location such as exonic or intronic) and recorded the fraction of probes for which the expression level change was in the same direction as P_i (i.e., up- or downregulated as compared to the control group). We then used a one-tailed binomial test to identify differentially expressed loci with a significance threshold of $p < 0.05$. As this threshold can only be met with a minimal sample size of five probes, we separately recorded cases with less than five overlapping probes but more than 50 % of the overlapping probes had a expression level change in the same direction.

Additionally, we recorded transcripts that achieved borderline significance (4 out of 5, 5 out of 6, and 6 out of 7 probes changing in the same direction). These loci should be treated with caution, however, because they may contain an increased amount of false positives. For probes located in loci with unknown transcript structures, we checked if the probe overlapped with spliced ESTs. If more than one spliced EST overlapped with the probe, we used the full overlapping EST cluster as locus rather than the original locus for the subsequent significance test. Lastly, for each of the four classes (three types of known transcripts and unknown transcripts), we filtered the list and only retained loci for which either the binomial test was significant or for which at least one probe had a differential expression $q < 0.05$. Although this procedure eliminates potentially relevant signals, it reduces the number of false positives due to the relatively high initial q-value.

C.2. Supplementary results

We followed a multi-step approach to identify AD-associated changes in gene expression: Initially, a whole genome tiling array was used to identify expressed regions in pooled AD and control samples, respectively. An `Alzheimer Custom Microarray` was designed, which interrogated the intervals identified as expressed in the tiling array approach, additional intervals found differentially expressed in response to several pathways and cell cycle described in [303] and additional ncRNA annotations from literature and databases. Subsequently, this custom array was applied to a set of AD and control samples for identifying AD-associated coding and non-coding genes.

C.2.1. Tiling arrays identify expressed regions in AD and control samples

We used whole genome tiling arrays to identify non-annotated transcripts in in three pooled AD and three pooled control samples, respectively. Using the `TileShuffle` algorithm we identified 64,488 and 48,412 transcribed fragments (transfrags) in AD and control samples, respectively, expressed significantly higher than background ($FDR < 0.05$). Again using `TileShuffle`, we found 1,459 transfrags that were sig-

nificantly expressed in at least AD or control samples and significantly differentially expressed between both conditions ($FDR < 0.05$)

C.2.2. Differentially expressed loci in Alzheimer's disease

Applying a custom expression microarray specifically designed for this study (*cf.* Section C.1) to 19 AD and 22 control samples, we identified 4,184 probes differentially expressed between AD and control ($q < 0.2$). Of these, 4,095 mapped uniquely to the genome. Using a multi-step approach, we identified a set of 764 differentially expressed genomic loci, 31 of which were associated with at least three distinct differentially expressed probes. Dependent on the genomic location of the differentially expressed probe(s), we then associated each genomic locus with one of the four following classes: protein-coding, non-coding, pseudogenes, and uncharacterized. The first three classes corresponded to known transcripts, whereas the latter represented loci with uncharacterized transcript structure and strand. In summary, we identified 162 differentially expressed protein-coding genes, 460 differentially expressed non-coding genes or non-coding loci, 29 differentially expressed pseudogenes, and 113 differentially expressed loci with unknown/uncharacterized transcript structure and type that did not overlap with any known genes or transcripts. The intersection of the identified differentially expressed loci with the constructed splice site conservation map of the GENCODE-derived background, resulted in a set of 4,162 splice sites falling in 154 multi-exonic protein-coding transcripts and a set of 1,297 splice sites falling in 141 multi-exonic non-protein-coding transcripts. Those data sets were used to compute the splice site conservation.

List of Abbreviations

AD	Alzheimer's Disease
ATP	Adenosine triphosphate
BPS	Branch point site
CDS	Coding DNA sequence
chr	Chromosome
ENSEMBL	Ensembl project database [209]
ESE	Exonic splicing enhancer
EST	Expressed sequence tag
FDR	False discovery rate
GTP	Guanosine triphosphate
MED	Maximum entropy distribution
MEM	Maximum entropy model
MGA	Multiple genome alignment
MLL	Mixed lineage leukemia/Trx complex
MSA	Multiple sequence alignment
ORF	Open reading frame
PCR	Polymerase chain reaction
PRC	Polycomb repressive complex
RBP	RNA-binding protein
RNAP	RNA polymerase
RNP	Ribonucleoprotein
RT	Reverse transcription
SHS	Short homologous sequence
SNP	Single nucleotide polymorphism
SR protein	Serine/arginine-rich protein
TF	Transcription factor
UCR	Ultraconserved region
UCSC	University of California Santa Cruz (Genome Browser) [208]
UTR	Untranslated region

Genome assemblies

anoCar	*Anolis carolinensis* (Lizard)
bosTau	*Bos taurus* (Cow)
calJac	*Callithrix jacchus* (Marmoset)
canFam	*Canis lupus familiaris* (Dog)

cavPor ..... *Cavia porcellus* (Guinea pig)
choHof ..... *Choloepus hoffmanni* (Sloth)
danRer ..... *Danio rerio* (Zebrafish)
dasNov ..... *Dasypus novemcinctus* (Armadillo)
dipOrd ..... *Dipodomys ordii* (Kangaroo rat)
echTel ..... *Echinops telfairi* (Tenrec)
equCab ..... *Equus caballus* (Horse)
eriEur ..... *Erinaceus europaeus* (Hedgehog)
felCat ..... *Felis catus* (Cat)
fr ..... *Takifugu rubripes* (Fugu)
galGal ..... *Gallus gallus* (Chicken)
gasAcu ..... *Gasterosteus aculeatus* (Stickleback)
gorGor ..... *Gorilla gorilla* (Western Gorilla)
hg ..... *Homo sapiens* (Human)
latCha ..... *Latimeria chalumnae* (Coelacanth)
loxAfr ..... *Loxodonta africana* (Elephant)
macEug ..... *Macropus eugenii* (Wallaby)
micMur ..... *Microcebus murinus* (Mouse lemur)
mm ..... *Mus musculus* (Mouse)
monDom ..... *Monodelphis domestica* (Opossum)
myoLuc ..... *Myotis lucifugus* (Microbat)
ochPri ..... *Ochotona princeps* (Pika)
ornAna ..... *Ornithorhynchus anatinus* (Platypus)
oryCun ..... *Oryctolagus cuniculus* (Rabbit)
oryLat ..... *Oryzias latipes* (Medaka)
otoGar ..... *Otolemur garnettii* (Bushbaby)
panTro ..... *Pan troglodytes* (Chimp)
papHam ..... *Papio hamadryas* (Baboon)
petMar ..... *Petromyzon marinus* (Lamprey)
ponAbe ..... *Pongo pygmaeus abelii* (Orangutan)
proCap ..... *Procavia capensis* (Rock hyrax)
pteVam ..... *Pteropus vampyrus* (Megabat)
rheMac ..... *Macaca mulatta* (Rhesus)
rn ..... *Rattus norvegicus* (Rat)
sorAra ..... *Sorex araneus* (Shrew)
speTri ..... *Spermophilus tridecemlineatus* (Squirrel)
taeGut ..... *Taeniopygia guttata* (Zebra finch)
tarSyr ..... *Tarsius syrichta* (Tarsier)
tetNig ..... *Tetraodon nigroviridis* (Tetraodon)
tupBel ..... *Tupaia belangeri* (Tree shrew)
turTru ..... *Tursiops truncatus* (Dolphin)
vicPac ..... *Vicugna pacos* (Alpaca)
xenTro ..... *Xenopus tropicalis* (Western clawed frog)

RNA and DNA

caRNA ..... Chromatin-associated RNA
cDNA ..... Complementary DNA
circRNA ..... Circular RNA
cRNA ..... Protein-coding RNA

DNA	Deoxyribonucleic Acid
l(i)ncRNA	Long (intergenic) non-coding RNA
miRNA	Micro RNA
mRNA	Messenger RNA
ncRNA	Non-coding RNA
pre-mRNA	Precursor messenger RNA
RNA	Ribonucleic Acid
rRNA	Ribosomal RNA
SL RNA	Spliced leader RNA
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
snRNP	Small nuclear ribonucleoprotein
tRNA	Transfer RNA

Units

bp	Base pairs
kb	Kilo bases
Mb	Mega bases
My	Million years
nt	Nucleotides

Variables

c	Degree of conservation (*cf.* p. 72)
FDR	False discovery rate
p	p-value
s_{mes}	`MaxEntScan` score

List of Figures

List of Tables

Bibliography

[1] Watson J, and Crick F (1953). A structure for deoxyribose nucleic acid. *Nature* 171:737–738.

[2] Chow LT, Gelinas RE, Broker TR, and Roberts RJ (1977). An amazing sequence arrangement at the 5' ends of adenovirus 2 messenger RNA. *Cell* 12:1–8.

[3] Berget SM, Moore C, and Sharp PA (1977). Spliced segments at the 5' terminus of adenovirus 2 late mRNA. *Proceedings of the National Academy of Sciences* 74:3171–3175.

[4] Stein LD, *et al.* (2003). The genome sequence of caenorhabditis briggsae: a platform for comparative genomics. *PLoS Biol* 1:e45.

[5] Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, and Lander ES (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences* 104:19428–19433.

[6] Sequencing TC, Consortium A, *et al.* (2005). Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

[7] Prüfer K, *et al.* (2012). The bonobo genome compared with the chimpanzee and human genomes. *Nature* 486:527–531.

[8] Consortium EP (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74.

[9] FANTOM Consortium and the RIKEN PMI and CLST (DGT) (2014). A promoter-level mammalian expression atlas. *Nature* 507:462–470. doi: 10.1038/nature13182.

[10] Djebali S, *et al.* (2012). Evidence for transcript networks composed of chimeric RNAs in human cells. *PLoS One* 7:e28213.

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

[12] Taft RJ, Pheasant M, and Mattick JS (2007). The relationship between non-protein-coding DNA and eukaryotic complexity. *Bioessays* 29:288–299.

[13] Iyer MK, *et al.* (2015). The landscape of long noncoding RNAs in the human transcriptome. *Nature genetics* 47:199–208.

[14] Clark M, Amaral P, Schlesinger F, Dinger M, Taft R, *et al.* (2011). The reality of pervasive transcription. *PLoS Biology* 9:e1000625.

[15] Cabili MN, Trapnell C, Goff L, Koziol M, Tazon-Vega B, and Regev JL Avivand Rinn (2011). Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. *Genes Dev* 25:1915–1927.

[16] Ulitsky I, Shkumatava A, Jan CH, Sive H, and Bartel DP (2011). Conserved function of lincRNAs in vertebrate embryonic development despite rapid sequence evolution. *Cell* 147:1537–1550.

[17] Lorenzen JM, and Thum T (2016). Long noncoding RNAs in kidney and cardiovascular diseases. *Nature Reviews Nephrology* 12:360–373.

[18] Esteller M (2011). Non-coding RNAs in human disease. *Nature Reviews Genetics* 12:861–874.

[19] Ponjavic J, Ponting CP, and Lunter G (2007). Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs. *Genome Res* 17:556–565.

[20] Marques AC, and Ponting CP (2009). Catalogues of mammalian long noncoding RNAs: modest conservation and incompleteness. *Genome Biol* 10:R124.

[21] Guttman M, *et al.* (2009). Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature* 458:223–227.

[22] Pang KC, Frith MC, and Mattick JS (2006). Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genetics* 22:1–5.

[23] Suzuki M, and Hayashizaki Y (2004). Mouse-centric comparative transcriptomics of protein coding and non-coding RNAs. *Bioessays* 26:833–843.

[24] Baldo L, Santos ME, and Salzburger W (2011). Comparative transcriptomics of Eastern African cichlid fishes shows signs of positive selection and a large contribution of untranslated regions to genetic diversity. *Genome Biol Evol* 3:443–455.

[25] Bräutigam A, *et al.* (2011). An mRNA blueprint for C4 photosynthesis derived from comparative transcriptomics of closely related C3 and C4 species. *Plant Physiol* 155:142–156.

[26] Hashimshony T, and Yanai I (2010). Revealing developmental networks by comparative transcriptomics. *Transcr* 1:154–158.

[27] Washietl S, Kellis M, and Garber M (2014). Evolutionary dynamics and tissue specificity of human long noncoding RNAs in six mammals. *Genome Res* 24:616–628.

[28] Necsulea A, Soumillon M, Warnefors M, Liechti A, Daish T, Zeller U, Baker JC, Gr??tzner F, and Kaessmann H (2014). The evolution of lncRNA repertoires and expression patterns in tetrapods. *Nature* 505:635–640. doi: 10.1038/nature12943. URL `http://dx.doi.org/10.1038/nature12943`.

[29] Managadze D, Lobkovsky AE, Wolf YI, Shabalina SA, Rogozin IB, and Koonin EV (2013). The vast, conserved mammalian lincRNome. *PLoS Comput Biol* 9:e1002917. doi: 10.1371/journal.pcbi.1002917. URL `http://dx.doi.org/10.1371/journal.pcbi.1002917`.

[30] Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, and Ulitsky I (2015). Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. *Cell reports* 11:1110–1122.

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

[32] Rose D, Hiller M, Schutt K, Hackermüller J, Backofen R, and Stadler PF (2011). Computational discovery of human coding and noncoding transcripts with conserved splice sites. *Bioinformatics* 27:1894–1900.

[33] Nitsche A (2012). *Comparative Transcriptomics of Long Non-Coding RNAs*. Master's thesis, Universität Leipzig, Fakultät für Mathematik und Informatik, Institut für Informatik. (Diplomarbeit).

[34] Nitsche A, and Stadler PF (2017). Evolutionary clues in lncRNAs. *Wiley Interdisciplinary Reviews: RNA* 8. doi: 10.1002/wrna.1376.

[35] Nitsche A, Rose D, Fasold M, Reiche K, and Stadler PF (2015). Comparison of splice sites reveals that long non-coding RNAs are evolutionarily well conserved. *RNA* 21:801–812. doi: 10.1261/rna.046342.114.

[36] Nitsche A, Doose G, Tafer H, Robinson M, Saha NR, Gerdol M, Canapa A, Hoffmann S, Amemiya CT, and Stadler PF (2014). Atypical RNAs in the coelacanth transcriptome. *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 322:342–351.

[37] Nitsche A, Reiche K, Ueberham U, Arnold C, Hackermüller J, Horn F, Stadler PF, and Arendt T (2017). Alzheimer related genes show accelerated evolution. bioRxiv: 10.1101/114108. submitted.

[38] Elliott D, and Ladomery M (2017). *Molecular biology of RNA*. Oxford University Press.

[39] Hertel KJ (2014). *Spliceosomal Pre-mRNA Splicing: Methods and Protocols*. Humana Press New York.

[40] Yoshimi A, and Abdel-Wahab O (2017). Molecular pathways: Understanding and targeting mutant spliceosomal proteins. *Clinical Cancer Research* 23:336–341.

[41] Pan Q, Shai O, Lee LJ, Frey BJ, and Blencowe BJ (2008). Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nat Genet* 40:1413–1415. doi: 10.1038/ng.259. URL `http://dx.doi.org/10.1038/ng.259`.

[42] Zhou HL, Luo G, Wise JA, and Lou H (2014). Regulation of alternative splicing by local histone modifications: potential roles for RNA-guided mechanisms. *Nucleic acids research* 42:701–713.

[43] Gingeras TR (2011). Implications of chimaeric non-co-linear transcripts. *Nature* 461:206–211.

[44] Zaphiropoulos PG (2011). Trans-splicing in higher eukaryotes: Implications for cancer development? *Front Genet* 2:92.

[45] Li X, Zhao L, Jiang H, and Wang W (2009). Short homologous sequences are strongly associated with the generation of chimeric RNAs in eukaryotes. *J Mol Evol* 68:56–65.

[46] Cocquet J, Chong A, Zhang G, and Veitia RA (2006). Reverse transcriptase template switching and false alternative transcripts. *Genomics* 88:127–131.

[47] Roy SW, and Irimia M (2008). When good transcripts go bad: artifactual RT-PCR 'splicing' and genome analysis. *Bioessays* 30:601–605.

[48] Houseley J, and Tollervey D (2010). Apparent non-canonical trans-splicing is generated by reverse transcriptase in vitro. *PLoS ONE* 5:e12271.

[49] McManus CJ, Duff MO, Eipper-Mains J, and Graveley BR (2010). Global analysis of trans-splicing in *Drosophila*. *Proc Natl Acad Sci USA* 107:12975–12979.

[50] Frenkel-Morgenstern M, *et al.* (2012). Chimeras taking shape: potential functions of proteins encoded by chimeric rna transcripts. *Genome Res* 22:1231–1242.

[51] Frenkel-Morgenstern M, and Valencia A (2012). Novel domain combinations in proteins encoded by chimeric transcripts. *Bioinformatics* 28:i67–i74.

[52] Ma L, Yang S, Zhao W, Tang Z, Zhang T, and Li K (2012). Identification and analysis of pig chimeric mRNAs using RNA sequencing data. *BMC Genomics* 13:429.

[53] Dorn R, Reuter G, and Loewendorf A (2001). Transgene analysis proves mRNA trans-splicing at the complex mod(mdg4) locus in drosophila. *Proc Natl Acad Sci U S A* 98:9724–9729.

[54] Gao Y, Wang J, Zheng Y, Zhang J, Chen S, and Zhao F (2016). Comprehensive identification of internal structure and alternative splicing events in circular RNAs. *Nature communications* 7.

[55] Zhang XO, Wang HB, Zhang Y, Lu X, Chen LL, and Yang L (2014). Complementary sequence-mediated exon circularization. *Cell* 159:134–147.

[56] Rybak-Wolf A, *et al.* (2015). Circular RNAs in the mammalian brain are highly abundant, conserved, and dynamically expressed. *Molecular cell* 58:870–885.

[57] Ivanov A, *et al.* (2015). Analysis of intron sequences reveals hallmarks of circular RNA biogenesis in animals. *Cell reports* 10:170–177.

[58] Salzman J, Gawad C, Wang PL, Lacayo N, and Brown PO (2012). Circular RNAs are the predominant transcript isoform from hundreds of human genes in diverse cell types. *PLoS ONE* 7:e30733.

[59] Jeck WR, Sorrentino JA, Wang K, Slevin MK, Burd CE, Liu J, Marzluff WF, and Sharpless NE (2013). Circular RNAs are abundant, conserved, and associated with ALU repeats. *RNA* 19:141–157.

[60] Zaphiropoulos PG (1996). Circular RNAs from transcripts of the rat cytochrome P450 2C24 gene: correlation with exon skipping. *Proc Natl Acad Sci USA* 93:6536–6541.

[61] Caldas C, So CW, MacGregor A, Ford AM, McDonald B, Chan LC, and Wiedemann LM (1998). Exon scrambling of MLL transcripts occur commonly and mimic partial genomic duplication of the gene. *Gene* 208:167–176.

[62] Surono A, Takeshima Y, Wibawa T, Ikezawa M, Nonaka I, and Matsuo M (1999). Circular dystrophin RNAs consisting of exons that were skipped by alternative splicing. *Hum Mol Genet* 8:493–500.

[63] Salzman J, Chen RE, Olsen MN, Wang PL, and Brown PO (2013). Cell-type specific features of circular RNA expression. *PLoS Genet* 9:e1003777.

[64] Memczak S, *et al.* (2013). Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495:333–338. doi: doi:10.1038/nature11928.

[65] Hansen TB, Jensen TI, Clausen BH, Bramsen JB, Finsen B, Damgaard CK, and Kjems J (2013). Natural RNA circles function as efficient microRNA sponges. *Nature* 495:384–388. doi: doi:10.1038/nature11993.

[66] Guo JU, Agarwal V, Guo H, and Bartel DP (2014). Expanded identification and characterization of mammalian circular RNAs. *Genome biology* 15:409.

[67] Brinster RL, Allen JM, Behringer RR, Gelinas RE, and Palmiter RD (1988). Introns increase transcriptional efficiency in transgenic mice. *Proceedings of the National Academy of Sciences* 85:836–840.

[68] McKenzie RW, and Brennan MD (1996). The two small introns of the drosophila affinidisjuncta adh gene are required for normal transcription. *Nucleic acids research* 24:3635–3642.

[69] Swinburne IA, and Silver PA (2008). Intron delays and transcriptional timing during development. *Developmental cell* 14:324–330.

[70] LIu K, Sandgren EP, Palmiter RD, and Stein A (1995). Rat growth hormone gene introns stimulate nucleosome alignment in vitro and in transgenic mice. *Proceedings of the National Academy of Sciences* 92:7724–7728.

[71] Proudfoot NJ, Furger A, and Dye MJ (2002). Integrating mRNA processing with transcription. *Cell* 108:501–512.

[72] Moore M, and Proudfoot N (2009). Pre-mRNA processing reaches back to transcription and ahead to translation. *Cell* 136:688–700.

[73] Collins LJ, Macke TJ, and Penny D (2004). Searching for ncRNAs in eukaryotic genomes: maximizing biological input with `RNAmotif`. *J Integ Bioinf* 1:2004–08–04. URL `http://journal.imbio.de/index.php?paper_id56`.

[74] Kapranov P, Willingham AT, and Gingeras TR (2007). Genome-wide transcription and the implications for genomic organization. *Nat Rev Genetics* 8:413–423.

[75] The FANTOM Consortium, R I K E N Genome Exploration Research Group, Genome Science Group (Genome Network Project Core Group), Carninci P, *et al.* (2005). The transcriptional landscape of the mammalian genome. *Science* 309:1559–1563.

[76] Pertea M, and Salzberg SL (2010). Between a chicken and a grape: estimating the number of human genes. *Genome Biol* 11:206. doi: 10.1186/gb-2010-11-5-206.

[77] Ezkurdia I, Juan D, Rodriguez JM, Frankish A, Diekhans M, Harrow J, Vazquez J, Valencia A, and Tress ML (2014). Multiple evidence strands suggest that there may be as few as 19 000 human protein-coding genes. *Human molecular genetics* 23:5866–5878.

[78] Hangauer MJ, Vaughn IW, and McManus MT (2013). Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet* 9:e1003569.

[79] St Laurent G, Shtokalo D, Tackett MR, Yang Z, Eremina T, Wahlestedt C, Urcuqui-Inchima S, Seilheimer B, McCaffrey TA, and Kapranov P (2012). Intronic RNAs constitute the major fraction of the non-coding RNA in mammalian cells. *BMC Genomics* 13:504.

[80] Derrien T, *et al.* (2012). The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. *Genome Res* 22:1775–1789. doi: 10.1101/gr.132159.111. URL `http://dx.doi.org/10.1101/gr.132159.111`.

[81] Carpenter S (2016). Long noncoding RNA: Novel links between gene expression and innate immunity. *Virus research* 212:137–145.

[82] Kashi K, Henderson L, Bonetti A, and Carninci P (2016). Discovery and functional analysis of lncRNAs: Methodologies to investigate an uncharacterized transcriptome. *Biochimica et Biophysica Acta (BBA)-Gene Regulatory Mechanisms* 1859:3–15.

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

[84] Maeda N, *et al.* (2006). Transcript annotation in FANTOM3: Mouse Gene Catalog based on physical cDNAs. *PLoS Genetics* 2:e62.

[85] Rinn JL, and Chang HY (2012). Genome regulation by long noncoding RNAs. *Annual review of biochemistry* 81:145–166.

[86] Kampa D, *et al.* (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome research* 14:331–342.

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

[88] Tian B, Hu J, Zhang H, and Lutz C (2005). A large-scale analysis of mRNA polyadenylation of human and mouse genes. *Nucleic acids research* 33:201–212.

[89] Ponting C (2008). The functional repertoires of metazoan genomes. *Nature Reviews Genetics* 9:689–698.

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

[91] Kapranov P, Ozsolak F, and Milos PM (2012). Profiling of short RNAs using Helicos single-molecule sequencing. *Methods Mol Biol* 822:219–232.

[92] St Laurent G, Wahlestedt C, and Kapranov P (2015). The landscape of long noncoding RNA classification. *Trends Genet* 31:239–251. doi: 10.1016/j.tig.2015.03.007.

[93] Brosius J (2005). Waste not, want not–transcript excess in multicellular eukaryotes. *TRENDS in Genetics* 21:287–288.

[94] Rands CM, Meader S, Ponting CP, and Lunter G (2014). 8.2% of the human genome is constrained: Variation in rates of turnover across functional element classes in the human lineage. *PLoS Genet* 10:e1004525. doi: 10.1371/journal.pgen.1004525.

[95] Pheasant M, and Mattick JS (2007). Raising the estimate of functional human sequences. *Genome research* 17:1245–1253.

[96] van Bakel H, Nislow C, Blencowe BJ, and Hughes TR (2010). Most "dark matter" transcripts are associated with known genes. *PLoS Biol* 8:e1000371. doi: 10.1371/journal.pbio.1000371.

[97] Palazzo AF, and Lee ES (2015). Non-coding RNA: what is functional and what is junk? *Front Genet* 6:2.

[98] Cech TR, and Steitz JA (2014). The noncoding RNA revolution—trashing old rules to forge new ones. *Cell* 157:77–94.

[99] Ulitsky I, and Bartel DP (2013). LincRNAs: Genomics, evolution, and mechanisms. *Cell* 154:26–46.

[100] Yang L, Froberg JE, and Lee JT (2014). Long noncoding RNAs: fresh perspectives into the RNA world. *Trends in biochemical sciences* 39:35–43.

[101] Flynn RA, and Chang HY (2014). Long noncoding RNAs in cell-fate programming and reprogramming. *Cell stem cell* 14:752–761.

[102] Congrains A, Kamide K, Ohishi M, and Rakugi H (2013). ANRIL: molecular mechanisms and implications in human health. *Int J Mol Sci* 14:1278–1292. doi: 10.3390/ijms14011278. URL `http://dx.doi.org/10.3390/ijms14011278`.

[103] Holdt LM, *et al.* (2013). Alu elements in ANRIL non-coding RNA at chromosome 9p21 modulate atherogenic cell functions through trans-regulation of gene networks. *PLoS Genet* 9:e1003588.

[104] Grote P, *et al.* (2013). The tissue-specific lncRNA fendrr is an essential regulator of heart and body wall development in the mouse. *Developmental cell* 24:206–214.

[105] Lai KMV, *et al.* (2015). Diverse phenotypes and specific transcription patterns in twenty mouse lines with ablated LincRNAs. *PloS one* 10:e0125522.

[106] Kino T, Hurt DE, Ichijo T, Nader N, and Chrousos GP (2010). Noncoding RNA GAS5 is a growth arrest- and starvation-associated repressor of the glucocorticoidreceptor. *Sci Signal* 3:ra8.

[107] Zhao J, Ohsumi TK, Kung JT, Ogawa Y, Grau DJ, Sarma K, Song JJ, Kingston RE, Borowsky M, and Lee JT (2010). Genome-wide identification of polycomb-associated RNAs by RIP-seq. *Molecular cell* 40:939–953.

[108] Kallen AN, *et al.* (2013). The imprinted h19 lncRNA antagonizes let-7 microRNAs. *Molecular cell* 52:101–112.

[109] Monnier P, Martinet C, Pontis J, Stancheva I, Ait-Si-Ali S, and Dandolo L (2013). H19 lncRNA controls gene expression of the imprinted gene network by recruiting MBD1. *Proceedings of the National Academy of Sciences* 110:20693–20698.

[110] Rinn JL, *et al.* (2007). Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129:1311–1323.

[111] Li L, *et al.* (2013). Targeted disruption of HOTAIR leads to homeotic transformation and gene derepression. *Cell Rep* 5:3–12. doi: 10.1016/j.celrep.2013.09.003. URL `http://dx.doi.org/10.1016/j.celrep.2013.09.003`.

[112] Schorderet P, and Duboule D (2011). Structural and functional differences in the long noncoding RNA HOTAIR in mouse and human. *PLoS Genet* 7:e1002071.

[113] Wang KC, *et al.* (2011). A long noncoding RNA maintains active chromatin to coordinate homeotic gene expression. *Nature* 472:120–124.

[114] Thakur N, Tiwari VK, Thomassin H, Pandey RR, Kanduri M, Göndör A, Grange T, Ohlsson R, and Kanduri C (2004). An antisense RNA regulates the bidirectional silencing property of the kcnq1 imprinting control region. *Molecular and cellular biology* 24:7855–7862.

[115] Redrup L, Branco MR, Perdeaux ER, Krueger C, Lewis A, Santos F, Nagano T, Cobb BS, Fraser P, and Reik W (2009). The long noncoding RNA kcnq1ot1 organises a lineage-specific nuclear domain for epigenetic gene silencing. *Development* 136:525–530.

[116] Gutschner T, Hämmerle M, and Diederichs S (2013). MALAT1 – a paradigm for long noncoding RNA function in cancer. *J Mol Med (Berl)* 91:791–801.

[117] Zhang X, Rice K, Wang Y, Chen W, Zhong Y, Nakayama Y, Zhou Y, and Klibanski A (2010). Maternally expressed gene 3 (MEG3) noncoding ribonucleic acid: Isoform structure, expression, and functions. *Endocrinology* 151:939–947.

[118] Han L, Zhang E, Yin D, Kong R, Xu T, Chen W, Xia R, Shu Y, and De W (2015). Low expression of long noncoding RNA PANDAR predicts a poor prognosis of non-small cell lung cancer and affects cell apoptosis by regulating bcl-2. *Cell death & disease* 6:e1665.

[119] Hung T, *et al.* (2011). Extensive and coordinated transcription of noncoding RNAs within cell-cycle promoters. *Nature genetics* 43:621–629.

[120] Lin N, *et al.* (2014). An evolutionarily conserved long noncoding RNA TUNA controls pluripotency and neural lineage commitment. *Mol Cell* 53:1005–1019. doi: 10.1016/j.molcel.2014.01.021. URL `http://dx.doi.org/10.1016/j.molcel.2014.01.021`.

[121] Johnsson P, Lipovich L, Grandér D, and Morris KV (2014). Evolutionary conservation of long non-coding RNAs; sequence, structure, function. *Biochim Biophys Acta* 1840:1063–1071. doi: 10.1016/j.bbagen.2013.10.035. URL `http://dx.doi.org/10.1016/j.bbagen.2013.10.035`.

[122] Stadler PF (2010). Evolution of the long non-coding RNAs MALAT1 and MENβ/ϵ. In CE Ferreira, S Miyano, PF Stadler, editors, *Advances in Bioinformatics and Computational Biology, 5th Brazilian Symposium on Bioinformatics*, volume 6268 of *Lecture Notes in Computer Science*, pages 1–12. Springer Verlag, Heidelberg, pages 1–12.

[123] Hutchinson J, Ensminger AW, Clemson CM, Lynch CR, Lawrence JB, and Chess A (2007). A screen for nuclear transcripts identifies two linked noncoding RNAs associated with SC35 splicing domains. *BMC Genomics* 8:39.

[124] Wilusz JE, and Spector DL (2010). An unexpected ending: noncanonical 3' end processing mechanisms. *RNA* 16:259–266.

[125] Yue M, Charles Richard JL, and Ogawa Y (2015). Dynamic interplay and function of multiple noncoding genes governing X chromosome inactivation. *Biochim Biophys Acta* doi: doi:10.1016/j.bbagrm.2015.07.015.

[126] Duret L, Chureau C, Samain S, Weissenbach J, and Avner P (2006). The *Xist* RNA gene evolved in eutherians by pseudogenization of a protein-coding gene. *Science* 312:1653–1655.

[127] Elisaphenko EA, Kolesnikov NN, Shevchenko AI, Rogozin IB, Nesterova TB, Brockdorff N, and Zakian SM (2008). A dual origin of the *Xist* gene from a protein-coding gene and a set of transposable elements. *PLoS One* 3:e2521.

[128] Chen J, Shishkin AA, Zhu X, Kadri S, Maza I, Guttman M, Hanna JH, Regev A, and Garber M (2016). Evolutionary analysis across mammals reveals distinct classes of long non-coding RNAs. *Genome Biol* 17:19.

[129] Bejerano G, Pheasant M, Makunin I, Stephen S, Kent WJ, Mattick JS, and Haussler D (2004). Ultraconserved elements in the human genome. *Science* 304:1321–1325.

[130] Calin GA, *et al.* (2007). Ultraconserved regions encoding ncRNAs are altered in human leukemias and carcinomas. *Cancer cell* 12:215–229.

[131] Smith CM, and Steitz JA (1998). Classification of GAS5 as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminaloligopyrimidine gene family reveals common features of snoRNA host genes. *Mol Cell Biol* 18:6897–6909.

[132] Williams GT, Mourtada-Maarabouni M, and Farzaneh F (2011). A critical role for noncoding RNA GAS5 in growth arrest and rapamycin inhibition in human T-lymphocytes. *Biochem Soc Trans* 39:482–486.

[133] Pollard KS, *et al.* (2006). Forces shaping the fastest evolving regions in the human genome. *PLoS Genet* 2:e168. doi: 10.1371/journal.pgen.0020168. URL `http://dx.doi.org/10.1371/journal.pgen.0020168`.

[134] Pollard KS, *et al.* (2006). An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* 443:167–172. doi: 10.1038/nature05113. URL `http://dx.doi.org/10.1038/nature05113`.

[135] Beniaminov A, Westhof E, and Krol A (2008). Distinctive structures between chimpanzee and humanin a brain noncoding RNA. *RNA* 14:1270–1275.

[136] Hubisz MJ, and Pollard KS (2014). Exploring the genesis and functions of human accelerated regions sheds light on their role in human evolution. *Current opinion in genetics & development* 29:15–21.

[137] Ponting CP, Oliver PL, and Reik W (2009). Evolution and functions of long noncoding RNAs. *Cell* 136:629–641.

[138] Slater GSC, and Birney E (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* 6:31. doi: 10.1186/1471-2105-6-31.

[139] Mainguy G, Koster J, Woltering J, Jansen H, and Durston A (2007). Extensive polycistronism and antisense transcription in the mammalian HOX clusters. *PLoS One* 2:e356.

[140] Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, Lan F, Shi Y, Segal E, and Chang HY (2010). Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329:689–693.

[141] Tanzer A, Amemiya CT, Kim CB, and Stadler PF (2005). Evolution of microRNAs located within *HOX* gene clusters. *J Exp Zool: Mol Dev Evol* 304B:75–85.

[142] Lee AP, Koh EG, Tay A, Brenner S, and Venkatesh B (2006). Highly conserved syntenic blocks at the vertebrate HOX loci and conserved regulatory elements within and outside HOX gene clusters. *Proc Natl Acad Sci USA* 103:6994–6999.

[143] Punnamoottil B, Herrmann C, Pascual-Anaya J, D'Aniello S, Garcia-Fernàndez J, Akalin A, Becker TS, and Rinkwitz S (2010). Cis-regulatory characterization of sequence conservation surrounding the HOX4 genes. *Dev Biol* 340:269–282.

[144] Natale A, Sims C, Chiusano ML, Amoroso A, D'Aniello E, Fucci L, Krumlauf R, Branno M, and Locascio A (2011). Evolution of anterior HOX regulatory elements among chordates. *BMC Evol Biol* 11:330.

[145] He S, Liu S, and Zhu H (2011). The sequence, structure and evolutionary features of HOTAIR in mammals. *BMC Evol Biol* 11:102.

[146] Yu H, Lindsay J, Feng ZP, Frankenberg S, Hu Y, Carone G Dawn andShaw, Pask AJ, O???Neill A Rachel andPapenfuss, and Renfree MB (2012). Evolution of coding and non-coding genes in HOX clusters of a marsupial. *BMC Genomics* 13:251.

[147] Haerty W, and Ponting CP (2015). Unexpected selection to retain high GC content and splicing enhancers within exons of multiexonic lncRNA loci. *RNA* 21:333–346.

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

[149] Ng SY, Bogu GK, Soh BS, and Stanton LW (2013). The long noncoding RNA RMSt interacts with SOX2 to regulate neurogenesis. *Molecular Cell* 51:349–359. ISSN 1097-2765. doi: 10.1016/j.molcel.2013.07.017. URL `http://dx.doi.org/10.1016/j.molcel.2013.07.017`.

[150] Chodroff RA, Goodstadt L, Sirey TM, Oliver PL, Davies KE, Green ED, Molnár Z, and Ponting CP (2010). Long noncoding RNA genes: conservation of sequence and brain expression among diverse amniotes. *Genome Biol* 11:R72.

[151] Pesole G, Mignone F, Gissi C, Grillo G, Licciulli F, and Liuni S (2001). Structural and functional features of eukaryotic mRNA untranslated regions. *Gene* 276:73–81.

[152] Wilusz JE, Freier SM, and Spector DL (2008). 3énd processing of a long nuclear-retained noncoding RNA yields a trna-like cytoplasmic rna. *Cell* 135:919–932.

[153] Fontana W, Konings DAM, Stadler PF, and Schuster P (1993). Statistics of RNA secondary structures. *Biopolymers* 33:1389–1404.

[154] Schultes EA, Spasic A, Mohanty U, and Bartel1 DP (2005). Compact and ordered collapse of randomly generated RNA sequences. *Nature Struct Mol Biol* 12:1130–1136. doi: 10.1038/nsmb1014.

[155] Rivas E, and Eddy SR (2001). Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2:8.

[156] Washietl S, and Hofacker IL (2004). Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J Mol Biol* 342:19–30.

[157] Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, Kent J, Miller W, and Haussler D (2006). Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Computational Biology* 2:e33.

[158] Washietl S, Hofacker IL, and Stadler PF (2005). Fast and reliable prediction of noncoding RNAs. *Proc Natl Acad Sci USA* 102:2454–2459.

[159] Gesell T, and Washietl S (2008). Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* 9:248. doi: 10.1186/1471-2105-9-248.

[160] Yao Z, Weinberg Z, and Ruzzo WL (2006). CMfinder – a covariance model based RNA motif finding algorithm. *Bioinformatics* 22:445–452.

[161] Gorodkin J, and Hofacker IL (2011). From structure prediction to genomic screens for novel non-coding RNAs. *PLoS Comp Biol* 7:e1002100. doi: 10.1371/journal.pcbi.1002100.

[162] Torarinsson E, Sawera M, Havgaard JH, Fredholm M, and Gorodkin J (2006). Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res* 16:885–889.

[163] Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, Tommerup N, Ruzzo WL, and Gorodkin J (2008). Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. *Genome Res* 18:242–251.

[164] Stadler PF, *et al.* (2009). Evolution of vault RNAs. *Mol Biol Evol* 26:1975–1991.

[165] Bernhart SH, Hofacker IL, Will S, Gruber AR, and Stadler PF (2008). `RNAalifold`: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics* 9:474.

[166] Washietl S, *et al.* (2007). Structured RNAs in the ENCODE selected regions of the human genome. *Gen Res* 17:852–864.

[167] Smith MA, Gesell T, Stadler PF, and Mattick JS (2013). Widespread purifying selection on RNA structure in mammals. *Nucleic Acids Res* 41:8220–8236.

[168] Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, and Stadler PF (2005). Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nature biotechnology* 23:1383–1390.

[169] Rose D, Hackermüller J, Washietl S, Reiche K, Hertel J, Findeiß S, Stadler PF, and Prohaska SJ (2007). Computational RNomics of drosophilids. *BMC genomics* 8:406.

[170] Hofacker IL, Fekete M, and Stadler PF (2002). Secondary structure prediction for aligned RNA sequences. *Journal of molecular biology* 319:1059–1066.

[171] Knudsen B, and Hein J (1999). RNA secondary structure prediction using stochastic context-free grammars and evolutionary history. *Bioinformatics* 15:446–454.

[172] Havgaard JH, Lyngsø RB, Stormo GD, and Gorodkin J (2005). Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics* 21:1815–1824.

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

[174] Will S, Joshi T, Hofacker IL, Stadler PF, and Backofen R (2012). LocARNA-P: Accurate boundary prediction and improved detection of structured RNAs for genome-wide screens. *RNA* 18:900–914.

[175] Will S, Yu M, and Berger B (2013). Structure-based whole-genome realignment reveals many novel noncoding RNAs. *Genome Res* 23:1018–1027. doi: doi:10.1101/gr.137091.111.

[176] Managadze D, Rogozin IB, Chernikova D, Shabalina SA, and Koonin EV (2011). Negative correlation between expression level and evolutionary rate of long intergenic noncoding RNAs. *Genome Biol Evol* 3:1390–1404. doi: 10.1093/gbe/evr116. URL `http://dx.doi.org/10.1093/gbe/evr116`.

[177] Schüler A, Ghanbarian AT, and Hurst LD (2014). Purifying selection on splice-related motifs, not expression level nor RNA folding, explains nearly all constraint on human lincRNAs. *Mol Biol Evol* 31:3164–3183.

[178] Mortimer SA, Kidwell MA, and Doudna JA (2014). Insights into RNA structure and function from genome-wide studies. *Nature Reviews Genetics* 15:469–479.

[179] Kwok CK, Tang Y, Assmann SM, and Bevilacqua PC (2015). The RNA structurome: transcriptome-wide structure probing with next-generation sequencing. *Trends Biochem Sci* 40:221–232. doi: 10.1016/j.tibs.2015.02.005.

[180] Lanz RB, Razani B, Goldberg AD, and O'Malley BW (2002). Distinct RNA motifs are important for coactivation of steroid hormone receptors by steroid receptor RNA activator (SRA). *Proc Natl Acad Sci U S A* 99:16081–16086.

[181] Leygue E (2007). Steroid receptor RNA activator (SRA1): unusual bifaceted gene products with suspected relevance to breast cancer. *Nuclear Receptor Signaling* 5:e006.

[182] Cooper C, Vincett D, Yan Y, Hamedani MK, Myal Y, and Leygue E (2011). Steroid receptor RNA activator bi-faceted genetic system: Heads or tails? *Biochimie* 93:1973–1980. doi: 10.1016/j.biochi.2011.07.002.

[183] Novikova IV, Hennelly SP, Tung CS, and Sanbonmatsu KY (2013). Rise of the RNA machines: exploring the structure of long noncoding RNAs. *J Mol Biol* 425:3731–3746. doi: 10.1016/j.jmb.2013.02.030.

[184] Wongtrakoongate P, Riddick G, Fucharoen S, and Felsenfeld G (2015). Association of the long non-coding RNA steroid receptor RNA activator (SRA) with TrxG and PRC2 complexes. *PLoS Genet* 11:e1005615. doi: 10.1371/journal.pgen.1005615.

[185] McKay DB, Xi L, Barthel KK, and Cech TR (2014). Structure and function of steroid receptor RNA activator protein, the proposed partner of SRA noncoding RNA. *J Mol Biol* 426:1766–1785. doi: 10.1016/j.jmb.2014.01.006.

[186] Meyer IM, and Miklós I (2005). Statistical evidence for conserved, local secondary structure in the coding regions of eukaryotic mRNAs and pre-mRNAs. *Nucleic Acids Res* 33:6338–6348.

[187] Findeiß S, Engelhardt J, Prohaska SP, and Stadler PF (2011). Protein-coding structured RNAs: A computational survey of conserved RNA secondary structures overlapping coding regions in drosophilids. *Biochimie* 93:2019–2023.

[188] Ulveling D, Francastel C, and Hubé F (2011). When one is better than two: RNA with dual functions. *Biochimie* 93:633–644.

[189] Eng L, Coutinho G, Nahas S, Yeo G, Tanouye R, Babaei M, Dörk T, Burge C, and Gatti RA (2004). Nonclassical splicing mutations in the coding and noncoding regions of the ATM gene: maximum entropy estimates of splice junction strengths. *Hum Mutat* 23:67–76.

[190] Zhang XHF, Leslie CS, and Chasin LA (2005). Computational searches for splicing signals. *Methods* 37:292–305.

[191] Hertel J, and Stadler PF (2015). The expansion of animal microRNA families revisited. *Life* 5:905–920.

[192] Kehr S, Bartschat S, Tafer H, Stadler PF, and Hertel J (2014). Matching of soulmates: coevolution of snoRNAs and their targets. *Mol Biol Evol* 31:455–467.

[193] Hoeppner MP, White S, Jeffares DC, and Poole AM (2009). Evolutionarily stable association of intronic snoRNAs and microRNAs with their host genes. *Genome Biol Evol* 1:420–428. doi: 10.1093/gbe/evp045.

[194] Zemann A, op de Bekke A, Kiefmann M, Brosius J, and Schmitz J (2006). Evolution of small nucleolar RNAs in nematodes. *Nucleic acids research* 34:2676–2685.

[195] Schmitz J, Zemann A, Churakov G, Kuhl H, Grützner F, Reinhardt R, and Brosius J (2008). Retroposed SNOfall—a mammalian-wide comparison of platypus snoRNAs. *Genome research* 18:1005–1010.

[196] Makarova JA, and Kramerov DA (2009). Analysis of C/D box snoRNA genes in vertebrates: The number of copies decreases in placental mammals. *Genomics* 94:11–19.

[197] Young RS, Marques AC, Tibbit C, Haerty W, Bassett AR, Liu JL, and Ponting CP (2012). Identification and properties of 1,119 candidate lincRNA loci in the *Drosophila melanogaster* genome. *Genome Biol Evol* 4:427–442.

[198] Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ, and Pachter L (2010). Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol* 28:511–515.

[199] Diederichs S (2014). The four dimensions of noncoding RNA conservation. *Trends Genet* 30:121–123. doi: 10.1016/j.tig.2014.01.004. URL `http://dx.doi.org/10.1016/j.tig.2014.01.004`.

[200] Ritz J, Martin JS, and Laederach A (2012). Evaluating our ability to predict the structural disruption of RNA by SNPs. *BMC Genomics* 13:S6. doi: 10.1186/1471-2164-13-S4-S6.

[201] Sabarinathan R, Tafer H, Seemann SE, Hofacker IL, Stadler PF, and Gorodkin J (2013). `RNAsnp`: Efficient detection of local RNA secondary structure changes induced by SNPs. *Hum Mut* 34:546–556.

[202] Salari R, Kimchi-Sarfaty C, Gottesman MM, and Przytycka TM (2012). Sensitive measurement of single-nucleotide polymorphism-induced changes of rna conformation: application to disease studies. *Nucleic Acids Res* 41:44–53. doi: 10.1093/nar/gks1009.

[203] Chatzou M, Magis C, Chang JM, Kemena C, Bussotti G, Erb I, and Notredame C (2015). Multiple sequence alignment modeling: methods and applications. *Briefings in bioinformatics* :bbv099.

[204] Böckenhauer HJ, and Bongartz D (2003). *Algorithmische Grundlagen der Bioinformatik: Modelle, Methoden und Komplexität (Leitfäden der Informatik) (German Edition)*. Vieweg+Teubner Verlag. ISBN 3519003988.

[205] Blanchette M, *et al.* (2004). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome Res* 14:708–15.

[206] Paten B, Herrero J, Beal K, Fitzgerald S, and Birney E (2008). Enredo and Pecan: genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res* 18:1814–28. doi: 10.1101/gr.076554.108.

[207] Paten B, Herrero J, Fitzgerald S, Beal K, Flicek P, Holmes I, and Birney E (2008). Genome-wide nucleotide-level mammalian ancestor reconstruction. *Genome Research* 18:1829–1843.

[208] Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, and Haussler D (2002). The human genome browser at UCSC. *Genome research* 12:996–1006.

[209] Hubbard T, *et al.* (2002). The Ensembl genome database project. *Nucleic acids research* 30:38–41.

[210] Hubbard TJ, *et al.* (2009). Ensembl 2009. *Nucleic acids research* 37:D690–D697.

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

[212] Needleman SB, and Wunsch CD (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology* 48:443–453.

[213] Smith TF, and Waterman MS (1981). Identification of common molecular subsequences. *Journal of molecular biology* 147:195–197.

[214] Durbin R, Eddy SR, Krogh A, and Mitchison G (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press. ISBN 0521629713.

[215] Hogeweg P, and Hesper B (1984). The alignment of sets of sequences and the construction of phylogenetic trees: an integrated method. *Journal of molecular evolution* 20:175–186.

[216] Feng DF, and Doolittle RF (1987). Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *Journal of molecular evolution* 25:351–360.

[217] Thompson JD, Higgins DG, and Gibson TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research* 22:4673–4680.

[218] Notredame C, Higgins DG, and Heringa J (2000). T-Coffee: A novel method for fast and accurate multiple sequence alignment. *Journal of molecular biology* 302:205–217.

[219] Do C, Mahabhashyam M, Brudno M, and Batzoglou S (2005). ProbCons: Probabilistic consistency-based multiple sequence alignment. *Genome Research* 15:330–340.

[220] Schwartz S, Kent WJ, Smit A, Zhang Z, Baertsch R, Hardison RC, Haussler D, and Miller W (2003). Human–mouse alignments with BLASTZ. *Genome research* 13:103–107.

[221] Dewey C, *et al.* (2007). Aligning multiple whole genomes with Mercator and MAVID. *METHODS IN MOLECULAR BIOLOGY-CLIFTON THEN TOTOWA*- 395:221.

[222] Myers G, Selznick S, Zhang Z, and Miller W (1997). Progressive multiple alignment with constraints. In *Proceedings of the first annual international conference on Computational molecular biology - RECOMB '97*, pages 220–225. Association for Computing Machinery (ACM), pages 220–225.

[223] Jaynes ET (1957). Information theory and statistical mechanics. *Physical review* 106:620.

[224] Jaynes ET (1957). Information theory and statistical mechanics. II. *Physical review* 108:171.

[225] Hong X, Scofield DG, and Lynch M (2006). Intron size, abundance, and distribution within untranslated regions of genes. *Mol Biol Evol* 23:2392–2404.

[226] Hoffmann S, Otto C, Kurtz S, Sharma C, Khaitovich P, Vogel J, Stadler PF, and Hackermüller J (2009). Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comp Biol* 5:e1000502.

[227] Hoffmann S, *et al.* (2014). A multi-split mapping algorithm for circular RNA, splicing, trans-splicing and fusion detection. *Genome biology* 15:R34.

[228] Harrow J, *et al.* (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol* 7 Suppl 1:S4.1–9. doi: 10.1186/gb-2006-7-s1-s4.

[229] Washietl S, Findeiß S, Müller SA, Kalkhof S, von Bergen M, Hofacker IL, Stadler PF, and Goldman N (2011). RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* 17:578–594.

[230] Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (1990). Basic local alignment search tool. *Journal of molecular biology* 215:403–410. URL `http://www.sciencedirect.com/science/article/pii/S0022283605803602`.

[231] Lin MF, Jungreis I, and Kellis M (2011). PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27:i275–i282. doi: 10.1093/bioinformatics/btr209. URL `http://dx.doi.org/10.1093/bioinformatics/btr209`.

[232] Guttman M, *et al.* (2011). lincRNAs act in the circuitry controlling pluripotency and differentiation. *Nature* 477:295–300.

[233] Pauli A, Valen E, Lin MF, Garber M, Vastenhouw NL, Levin JZ, Fan L, Sandelin A, Rinn A John Land Regev, and Schier AF (2012). Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis. *Genome Res* 25:1915–1927. doi: 10.1101/gr.133009.111.

[234] Scofield DG, Hong X, and Lynch M (2007). Position of the final intron in full-length transcripts: Determined by NMD? *Mol Biol Evol* 24:896–899.

[235] Zhu J, Sanborn JZ, Diekhans M, Lowe CB, Pringle TH, and Haussler D (2007). Comparative genomics search for losses of long-established genes on the human lineage. *PLoS Comput Biol* 3:e247.

[236] He S, Gu W, Li Y, and Zhu H (2013). ANRIL/CDKN2B-AS shows two-stage clade-specific evolution and becomes conserved after transposon insertions in simians. *BMC Evolutionary Biology* 13:247. ISSN 1471-2148. doi: 10.1186/1471-2148-13-247. URL `http://dx.doi.org/10.1186/1471-2148-13-247`.

[237] Kim VN (2005). MicroRNA biogenesis: coordinated cropping and dicing. *Nat Rev Mol Cell Biol* 6:376–338.

[238] Maxwell ES, and Fournier MJ (1995). The small nucleolar RNAs. *Annu Rev Biochem* 64:897–934.

[239] Benton MJ, and Donoghue PCJ (2007). Paleontological evidence to date the tree of life. *Mol Biol Evol* 24:26–53.

[240] Yamanoue Y, Miya M, Inoue JG, Matsuura K, and Nishida M (2006). The mitochondrial genome of spotted green pufferfish *Tetraodon nigroviridis* (Teleostei: Tetraodontiformes) and divergence time estimation among model organisms in fishes. *Genes & Genetic Systems* 81:29–39.

[241] Hasegawa M, Thorne JL, and Kishino H (2003). Time scale of eutherian evolution estimated without assuming a constant rate of molecular evolution. *Genes & Genetic Systems* 78:267–283.

[242] Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, and The Students of Bioinformatics Computer Labs 2004 and 2005 (2006). The expansion of the metazoan microRNA repertoire. *BMC Genomics* 7:25.

[243] Sempere LF, Cole CN, McPeek MA, and Peterson KJ (2006). The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. *J Exp Zoolog B Mol Dev Evol* 306:575–588.

[244] Lestrade L, and Weber MJ (2006). snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. *Nucleic Acids Res* 34:D158–D162.

[245] Marz M, *et al.* (2011). Animal snoRNAs and scaRNAs with exceptional structures. *RNA Biology* 8:938–946.

[246] Shao P, Yang JH, Zhou H, Guan DG, and Qu LH (2009). Genome-wide analysis of chicken snoRNAs provides unique implications for the evolution of vertebrate snoRNAs. *BMC Genomics* 10:86.

[247] Pasmant E, Laurendeau I, Sabbagh A, Parfait B, Vidaud M, Vidaud D, and Bièche I (2010). The amazing story of ANRIL, a long non-coding RNA. *Médecine sciences: M/S* 26:564.

[248] Chan AS, Thorner PS, Squire JA, and Zielenska M (2002). Identification of a novel gene NCRMS on chromosome 12q21 with differential expression between rhabdomyosarcoma subtypes. *Oncogene* 21:3029–3037.

[249] Bouchard M, Grote D, Craven SE, Sun Q, Steinlein P, and Busslinger M (2005). Identification of PAX2-regulated genes by expression profiling of the mid-hindbrain organizer region. *Development* 132:2633–2643.

[250] Weidman JR, Maloney KA, and Jirtle RL (2006). Comparative phylogenetic analysis reveals multiple non-imprinted isoforms of opossum Dlk1. *Mamm Genome* 17:157–167.

[251] Burd CE, Jeck WR, Liu Y, Sanoff HK, Wang Z, and Sharpless NE (2010). Expression of linear and novel circular forms of an INK4/ARF-associated non-coding RNA correlates with atherosclerosis risk. *PLoS Genet* 6:e1001233.

[252] Holdt LM, Beutner F, Scholz M, Gielen S, Gäbel G, Bergert H, Schuler G, Thiery J, and Teupser D (2010). ANRIL expression is associated with atherosclerosis risk at chromosome 9p21. *Arterioscler Thromb Vasc Biol* 30:620–627.

[253] Mercer TR, Dinger ME, and Mattick JS (2009). Long noncoding RNAs: insights into function. *Nat Rev Genet* 10:155–159.

[254] Wang J, Liu X, Wu H, Ni P, Gu Z, Qiao Y, Chen N, Sun F, and Fan Q (2010). CREB up-regulates long non-coding RNA, HULC expression through interaction with microRNA-372 in liver cancer. *Nucleic Acids Res* 38:5366–5383.

[255] Amemiya CT, *et al.* (2013). The african coelacanth genome provides insights into tetrapod evolution. *Nature* 496:311–316.

[256] Pallavicini A, *et al.* (2013). Analysis of the transcriptome of the Indonesian coelacanth *Latimeria menadoensis*. *BMC Genomics* 14:538.

[257] Lab H (2011). FASTX Toolkit. URL `http://hannonlab.cshl.edu/fastx_toolkit/`.

[258] Martin M (2011). `Cutadapt` removes adapter sequences from high-throughput sequencing reads. *EMBnetjournal* 17.

[259] Roberts A, Pimentel H, Trapnell C, and Pachter L (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. *Bioinformatics (Oxford, England)* 27:2325–2329.

[260] Trapnell C, Roberts A, Goff L, Pertea G, and Kim D (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* .

[261] Quinlan AR, and Hall IM (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)* 26:841–842.

[262] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, and Lipman D (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.

[263] Perez SE, Raghanti MA, Hof PR, Kramer L, Ikonomovic MD, Lacor PN, Erwin JM, Sherwood CC, and Mufson EJ (2013). Alzheimer's disease pathology in the neocortex and hippocampus of the western lowland gorilla (Gorilla gorilla gorilla). *Journal of Comparative Neurology* 521:4318–4338.

[264] Mattick JS (2005). The functional genomics of noncoding RNA. *Science* 309:1527–1528.

[265] Zhang Z (2016). Long non-coding RNAs in alzheimer's disease. *Curr Top Med Chem* 16:511–519.

[266] Arnold C (2014). *The Eukaryotic Chromatin Computer*. Ph.D. thesis, Universität Leipzig, Fakultät für Mathematik und Informatik, Institut für Informatik.

[267] Rapoport S (1989). Hypothesis: Alzheimer's disease is a phylogenetic disease. *Medical Hypotheses* 29:147–150.

[268] Cáceres M, Lachuer J, Zapala MA, Redmond JC, Kudo L, Geschwind DH, Lockhart DJ, Preuss TM, and Barlow C (2003). Elevated gene expression levels distinguish human from non-human primate brains. *Proceedings of the National Academy of Sciences* 100:13030–13035.

[269] Uddin M, Wildman DE, Liu G, Xu W, Johnson RM, Hof PR, Kapatos G, Grossman LI, and Goodman M (2004). Sister grouping of chimpanzees and humans as revealed by genome-wide phylogenetic analysis of brain gene expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 101:2957–2962.

[270] Fukuda K, *et al.* (2013). Regional DNA methylation differences between humans and chimpanzees are associated with genetic changes, transcriptional divergence and disease genes. *Journal of Human Genetics* 58:446–454.

[271] Somel M, *et al.* (2009). Transcriptional neoteny in the human brain. *Proceedings of the National Academy of Sciences* 106:5743–5748.

[272] Cáceres M, Suwyn C, Maddox M, Thomas JW, and Preuss TM (2007). Increased cortical expression of two synaptogenic thrombospondins in human brain evolution. *Cerebral Cortex* 17:2312–2321.

[273] Stephan H (1983). Evolutionary trends in limbic structures. *Neuroscience & Biobehavioral Reviews* 7:367–374.

[274] Braak H, and Braak E (1991). Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica* 82:239–259.

[275] Arendt T (2000). Alzheimer's disease as a loss of differentiation control in a subset of neurons that retain immature features in the adult brain. *Neurobiology of aging* 21:783–796.

[276] Bianchi S, *et al.* (2013). Synaptogenesis and development of pyramidal neuron dendritic morphology in the chimpanzee neocortex resembles humans. *Proceedings of the National Academy of Sciences* 110:10395–10401.

[277] Bufill E, Blesa R, and Agustí J (2013). Alzheimer's disease: an evolutionary approach. *Journal of Anthropological Sciences* 91:1–23.

[278] Zahs KR, and Ashe KH (2010). 'too much good news'–are Alzheimer mouse models trying to tell us how to prevent, not cure, Alzheimer's disease? *Trends in Neurosciences* 33:381–389.

[279] Ghosal S, Das S, Sen R, Basak P, and Chakrabarti J (2013). Circ2Traits: a comprehensive database for circular RNA potentially associated with disease and traits. *Frontiers in genetics* 4:283.

[280] Lukiw WJ (2013). Circular RNA (circRNA) in Alzheimer's disease (AD). *Front Genet* 4:307.

[281] Hogg JR, and Collins K (2008). Structured noncoding RNAs and the RNP renaissance. *Curr Opin Chem Biol* 12:684–689. doi: 10.1016/j.cbpa.2008.09.027.

[282] Marz M, and Stadler PF (2011). RNA interactions. In LJ Collins, editor, *RNA Infrastructure and Networks*, volume 722 of *Advances in Experimental Medicine and Biology*, pages 20–38. Landes Biosciences, Springer-Verlag, Berlin, pages 20–38.

[283] DePristo MA, *et al.* (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature genetics* 43:491–498.

[284] Bailey TL, and Elkan C (1995). Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Machine Learning* 21:51–80.

[285] Dubois B, *et al.* (2010). Revising the definition of alzheimer's disease: a new lexicon. *The Lancet Neurology* 9:1118–1127.

[286] Dubois B, *et al.* (2007). Research criteria for the diagnosis of alzheimer's disease: revising the NINCDS–ADRDA criteria. *The Lancet Neurology* 6:734–746.

[287] Dubois B, *et al.* (2014). Advancing research diagnostic criteria for alzheimer's disease: the IWG-2 criteria. *The Lancet Neurology* 13:614–629.

[288] Albert MS, *et al.* (2011). The diagnosis of mild cognitive impairment due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia* 7:270–279.

[289] Jack CR, Albert MS, Knopman DS, McKhann GM, Sperling RA, Carrillo MC, Thies B, and Phelps CH (2011). Introduction to the recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & Dementia* 7:257–262.

[290] McKhann GM, *et al.* (2011). The diagnosis of dementia due to alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia* 7:263–269.

[291] Sperling RA, *et al.* (2011). Toward defining the preclinical stages of alzheimer's disease: Recommendations from the national institute on aging-alzheimer's association workgroups on diagnostic guidelines for alzheimer's disease. *Alzheimer's & dementia* 7:280–292.

[292] Hyman BT, *et al.* (2012). National institute on aging–alzheimer's association guidelines for the neuropathologic assessment of alzheimer's disease. *Alzheimer's & Dementia* 8:1–13.

[293] Montine TJ, *et al.* (2012). National institute on aging–alzheimer's association guidelines for the neuropathologic assessment of alzheimer's disease: a practical approach. *Acta neuropathologica* 123:1–11.

[294] Morris J, Heyman A, Mohs R, Hughes J, Van Belle G, Fillenbaum G, Mellits E, and Clark C (1989). Investigators, c. (1989). the consortium to establish a registry for alzheimer's disease (cerad). part i. clinical and neuropsychological assessment of alzheimer's disease. *Neurology* 39:1159–1165.

[295] Folstein MF, Folstein SE, and McHugh PR (1975). "mini-mental state": a practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research* 12:189–198.

[296] Reisberg B, Ferris SH, de Leon MJ, and Crook T (1982). The global deterioration scale for assessment of primary degenerative dementia. *The American journal of psychiatry* .

[297] Braak H, and Braak E (1996). Development of Alzheimer-related neurofibrillary changes in the neocortex inversely recapitulates cortical myelogenesis. *Acta Neuropathologica* 92:197–201.

[298] Braak H, Alafuzoff I, Arzberger T, Kretzschmar H, and Del Tredici K (2006). Staging of alzheimer disease-associated neurofibrillary pathology using paraffin sections and immunocytochemistry. *Acta neuropathologica* 112:389–404.

[299] Thal DR, Rüb U, Orantes M, and Braak H (2002). Phases of aβ-deposition in the human brain and its relevance for the development of AD. *Neurology* 58:1791–1800.

[300] Mirra SS, *et al.* (1991). The consortium to establish a registry for alzheimer's disease (cerad) part II. standardization of the neuropathologic assessment of alzheimer's disease. *Neurology* 41:479–479.

[301] Otto C, Reiche K, and Hackermüller J (2012). Detection of differentially expressed segments in tiling array data. *Bioinformatics* 28:1471–1479.

[302] Benjamini Y, and Hochberg Y (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B (Methodological)* 57:289–300.

[303] Hackermüller J, *et al.* (2014). Cell cycle, oncogenic and tumor suppressor pathways regulate numerous long and macro non-protein coding RNAs. *Genome Biol* 15:R48.

[304] Bu D, *et al.* (2012). NONCODE v3.0: integrative annotation of long noncoding RNAs. *Nucleic Acids Research* 40:D210–D215.

[305] Amaral PP, Clark MB, Gascoigne DK, Dinger ME, and Mattick JS (2011). lncRNAdb: a reference database for long noncoding RNAs. *Nucleic Acids Research* 39:D146–D151.

[306] Kin T, Yamada K, Terai G, Okida H, Yoshinari Y, Ono Y, Kojima A, Kimura Y, Komori T, and Asai K (2007). fRNAdb: a platform for mining/annotating functional RNA candidates from non-coding RNA sequences. *Nucleic Acids Research* 35:D145–D148.

[307] Pang KC, Stephen S, Dinger ME, Engström PG, Lenhard B, and Mattick JS (2007). RNAdb 2.0–an expanded database of mammalian non-coding RNAs. *Nucleic Acids Research* 35:D178–D182.

[308] Yamasaki C, *et al.* (2008). The H-Invitational Database (H-InvDB), a comprehensive annotation resource for human genes and transcripts. *Nucleic Acids Research* 36:D793.

[309] Harrow J, *et al.* (2012). GENCODE: the reference human genome annotation for the ENCODE project. *Genome Research* 22:1760–1774.

[310] Pruitt KD, Tatusova T, Brown GR, and Maglott DR (2012). NCBI reference sequences (RefSeq): current status, new features and genome annotation policy. *Nucleic Acids Research* 40:D130–D135.

[311] Khalil A, *et al.* (2009). Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. *Proceedings of the National Academy of Sciences of the United States of America* 106:11667–11672.

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

[313] Bartschat S, Kehr S, Tafer H, Stadler PF, and Hertel J (2013). `snoStrip`: A snoRNA annotation pipeline. *Bioinformatics* :btt604.

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

[315] Hofacker I, and Stadler P (2010). RNAz 2.0: Improved noncoding RNA detection. In *Pacific Symposium on Biocomputing*, volume 15, pages 69–79. pages 69–79.

[316] Guil S, and Esteller M (2012). *Cis*-acting noncoding RNAs: friends and foes. *Nature Structural & Molecular Biology* 19:1068–1075.

[317] Arnold C, Externbrink F, Hackermüller J, and Reiche K (2014). CEM-designer: design of custom expression microarrays in the post-ENCODE era. *J Biotechnol* 189:154–156. doi: 10.1016/j.jbiotec.2014.09.012. URL `http://dx.doi.org/10.1016/j.jbiotec.2014.09.012`.

[318] Smyth GK, *et al.* (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology* 3:3.

[319] Bolstad BM, Irizarry RA, Åstrand M, and Speed TP (2003). A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19:185–193.

[320] Storey JD (2002). A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 64:479–498.

[321] Strimmer K (2008). A unified approach to false discovery rate estimation. *BMC Bioinformatics* 9:303.

[322] Strimmer K (2008). fdrtool: a versatile R package for estimating local and tail area-based false discovery rates. *Bioinformatics* 24:1461–1462.

Curriculum Scientiae

Personal Information

Name	Anne Nitsche
Birth	June 17, 1984
Birthplace	Hoyerswerda

Education

PhD Student	**University of Leipzig** *(since 2012)* Bioinformatics Group of Prof. Peter F. Stadler Thesis: "Tracing the evolution of long non-coding RNAs – Principles of comparative transcriptomics for splice site conservation and biological applications"
Diploma Student	**University of Leipzig** *(2003 – 2012)* Studies in Computer Science with focus on Bioinformatics Thesis: "Comparative Transcriptomics of Long Non-Coding RNAs" (Dipl.-Inf.)

Working Experience

Visiting Student	**UC Berkeley** *(01/2017 – 05/2017)* Department of Integrative Biology
Scientific Staff	**University of Strasbourg** *(02/2016 – 10/2016)* Institute of Molecular and Cellular Biology
Scientific Staff	**University of Leipzig** *(10/2012 – 01/2016)* Faculty of Medicine

Student Assistant	**University of Leipzig** *(11/2010 – 02/2011, 04/2011 – 07/2011, 11/2011 – 02/2012, 04/2012 – 08/2012)* Faculty of Mathematics and Computer Science, Department of Computer Science
Intern	**Max Planck Institute for Human Cognitive and Brain Sciences** *(10/2009 – 05/2010)* Department of Neurophysics

IT-Knowledge

Operating systems	UNIX, Linux, Windows
Programming	Perl, Java, Python, R, Bash, Shell
Other	Latex, HTML, MySQL

Languages

German	native speaker
English	fluent

Publications

First Author

Nitsche A, Doose G, Tafer H, Robinson M, Saha NR, Gerdol M, Canapa A, Hoffmann S, Amemiya CT, and Stadler PF (2014). **Atypical RNAs in the coelacanth transcriptome.** *Journal of Experimental Zoology Part B: Molecular and Developmental Evolution* 322:342–351

Nitsche A, Rose D, Fasold M, Reiche K, and Stadler PF (2015). **Comparison of splice sites reveals that long non-coding RNAs are evolutionarily well conserved.** *RNA* 21:801–812. doi: 10.1261/rna.046342.114

Nitsche A, and Stadler PF (2017). **Evolutionary clues in lncRNAs.** Wiley *Interdisciplinary Reviews: RNA* 8. doi: 10.1002/wrna.1376

Nitsche A, Reiche K, Ueberham U, Arnold C, Hackermüller J, Horn F, Stadler PF, and Arendt T (2017). **Alzheimer related genes show accelerated evolution.** *bioRxiv*: 10.1101/114108. submitted

Collaborations

Amemiya CT, et al. (2013). **The african coelacanth genome provides insights into tetrapod evolution.** *Nature* 496:311–316

Hackermüller J, et al. (2014). **Cell cycle, oncogenic and tumor suppressor pathways regulate numerous long and macro non-protein coding RNAs.** *Genome Biol* 15:R48

Schmid M, et al. (2015). **Third report on chicken genes and chromosomes 2015.** *Cytogenetic and genome research* 145:78–179

Conferences

Head of organizing commitee	**13. Herbstseminar der Bioinformatik 2015** Doubice, Czech Republic *(09 – 10/2015)*
Head of organizing commitee	**12. Herbstseminar der Bioinformatik 2014** Doubice, Czech Republic *(10/2014)*
Presentation	**11. Herbstseminar der Bioinformatik 2013** *Atypical Splicing – circular RNA* Doubice, Czech Republic *(10/2013)*
Presentation	**28th TBI winter seminar 2013** *Splice site maps reveal evolution of non-coding RNA* Bled, Slovenia *(02/2013)*
Presentation	**9. Herbstseminar der Bioinformatik 2011** *Evolution of splice sites in lncRNAs* Vysoka Lipa, Czech Republic *(10/2011)*
Presentation	**8. Herbstseminar der Bioinformatik 2010** *Splice Site Evolution* Vysoka Lipa, Czech Republic *(09/2010)*

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den

(Anne Nitsche)