TECHNISCHE UNIVERSITÄT MÜNCHEN

Lehrstuhl für Genomorientierte Bioinformatik

Applications of NGS : Regulatory actions of lincRNAs and SNP analysis of Restless legs syndrome

Nadine Isabel Albrecht

Vollständiger Abdruck der von der Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt der Technischen Universität München zur Erlangung des akademischen Grades eines

Doktors der Naturwissenschaften

genehmigten Dissertation.

Vorsitzender: Univ.-Prof. Dr. H.-R. Fries

Prüfer der Dissertation:

1. Univ.-Prof. Dr. H.-W. Mewes
2. Univ.-Prof. Dr. D. Frischmann

Die Dissertation wurde am 17.09.2012 bei der Technischen Universität München eingereicht und durch die Fakultät Wissenschaftszentrum Weihenstephan für Ernährung, Landnutzung und Umwelt am 17.01.2013 angenommen.

Contents

List of Figures

List of Tables

Abstract

The advent of Next Generation Sequencing (NGS) technology makes the generation of sequence data for whole genomes and exomes possible. New challenges in Bioinformatics arise with more and more NGS data. One of these challenges is the adaptation of methods and tools coping with this new data type. Another challenge is the application of NGS. NGS opens new possibilities answering biological questions, otherwise difficult to solve with existing approaches. One important advantage of the NGS approach itself lies in the detection of novel transcripts (coding and noncoding RNA) and mutations. Differential quantitative expression levels of e.g. novel ncRNA and its target genes can be calculated in a cell line. In this work we focus on two applications of NGS: (i) Regulatory actions of large intergenic noncoding RNAs (lincRNAs) and (ii) SNP analysis of Restless legs syndrome (RLS).

The first application is based on recently published RNA-seq data from three mouse cell lines containing novel lincRNA reconstructions. Potential interaction partners between novel lincRNA reconstructions and annotated coding and noncoding RNA were identified via sequence similarity searches. One main action of lincRNA interactions with sequence complementary pre-mRNA is the influence on their target genes' expression and alternative splice events. Alternate splice variants could be explained by lincRNA:mRNA duplexes for the first time in this dissertation. lincRNAs are predominately sequence complementary to introns of their target genes, especially 5' introns. Exons, close to target sites, show a significant decrease in expression levels. According to an enrichment of sites at 5' introns, strongest influence of duplexes on expression and splice events was observed at 5' exons. This novel finding is of importance since duplexes represent a new and unprecedented mechanism for splicing regulation.

Secondly, we analysed exome sequencing data of two pedigrees in cooperation with Prof. Juliane Winkelmann and Daniel Ellwanger. Individuals are known to be affected by the disease RLS. The primary (idiopathic) form of RLS suggests an autosomal dominant inheritance and is the focus of our analysis. In our part of the cooperation raw data was processed to detect and interpret RLS associated SNPs. Candidate genes with novel and non-synonymous (deleterious) SNPs could be identified.

Zusammenfassung

Die Einführung der Next Generation Sequencing (NGS) Methode macht die Generierung von Sequenzdaten für ganze Genome und Exome möglich. Neue Herausforderungen in der Bioinformatik ergeben sich mit mehr und mehr NGS Daten. Eine Herausforderung ist die Anpassung von Methoden und Programmen die dem neuen Datentyp gerecht werden. Eine andere Herausforderung ist die Anwendung von NGS. NGS eröffnet neue Möglichkeiten um biologische Fragestellungen zu beantworten, die andererseits mit existierenden Ansätzen schwierig zu lösen sind. Ein wichtiger Vorteil des NGS Ansatzes selbst liegt in der Entdeckung von neuen Transkripten (kodierende und nicht kodierende RNA) und Mutationen. Differentielle quantitative Expressionslevel von z.B. neuer ncRNA und deren Zielgene können in einer Zelllinie berechnet werde. In dieser Arbeit konzentrieren wir uns auf zwei Anwendungen von NGS: (i) Regulatorische Aktivitäten von large intergenic noncoding RNAs (lincRNAs) und (ii) SNP Analyse von Restless legs syndrome (RLS).

Die erste Anwendung basiert auf neu veröffentlichten RNA-seq Daten von drei Maus Zelllinien, die neue lincRNA Rekonstruktionen beinhalten. Potentielle Interaktionspartner zwischen neuen lincRNA Rekonstruktionen und annotierten kodierenden und nicht kodierenden RNA wurden mit Sequenzähnlichkeitssuchen identifiziert. Eine Hauptaktivität der lincRNA Interaktionen mit sequenzkomplementärer pre-mRNA ist der Einfluss auf die Expression und alternative Splice Ereignisse Ihrer Zielgene. Alternative Splice Varianten konnten zum ersten mal in dieser Dissertation durch lincRNA:mRNA Duplexe erklärt werden. lincRNAs sind überwiegend sequenzkomplementär zu Introns von Ihren Zielgenen, vor allem 5' Introns. Exons, in der Nähe von Zielstellen, zeigen eine signifikante Abnahme der Expressionslevel. In Übereinstimmung mit einer Anreicherung von Stellen an 5' Introns wurde der stärkste Einfluss der Duplexe auf die Expression und Splice Ereignisse an 5' Exons beobachtet. Diese neue Erkenntnis ist von Wichtigkeit, da Duplexe einen neuen und nie dagewesenen Mechanismus der Splicing Regulierung repräsentieren.

Zweitens, haben wir Exom Sequenzdaten von zwei Stammbäumen in Kooperation mit Prof. Juliane Winkelmann und Daniel Ellwanger untersucht. Einzelpersonen sind von der Krankheit RLS betroffen. Die primäre (idiopathische Form) von RLS geht von einer

autosomal dominanten Vererbung aus und ist der Schwerpunkt unserer Analyse. In unserem Teil der Kooperation wurden die Rohdaten prozessiert um RLS assoziierte SNPs zu entdecken und interpretieren. Kandidatengene, mit neuen nichtsynonymen schädlichen SNPs, konnten identifiziert werden.

Acknowledgements

In the first place I would like to thank my doctoral advisor and head of IBIS (Institute of Bioinformatics and Systems Biology) Prof. Hans-Werner Mewes. I appreciate that he gave me the opportunity to work at his department for my diploma and doctoral thesis. He is the one who raised the interesting topic of my work. It was a pleasure to me that he made it possible to join IRTG (International Research Training Group) RECESS (Regulation and Evolution of Cellular Systems) for my dissertation.

RECESS is funded by DFG and affords me and other PhD students great opportunities within this program. One of the opportunities is to visit Moscow (Russia) for conferences, wetlab courses and retreats in close interaction with colleagues of the Moscow State University. Additional thanks to the joint RECESS initiators Prof. Frishman, Prof. Zimmer of the according advisory board in Munich (Germany), Prof. Gelfand in Moscow (Russia) and my co-supervisor Prof. Mironov (Russia). I would like to further appreciate especially Prof. Frishman for supervision and taking part in my dissertation committee.

Further, appreciation goes to my former group leader Dr. Thorsten Schmidt (NGS group, Helmholtz Zentrum München) for the scientific supervision in bioinformatics. I am indebted to a colleague of mine Jonathan Hoser who passed through the dissertation with me and provided support in handling large amounts of NGS data. It has been an honour for me to supervise and work with the students Veit Hoehn, Kerstin Haase and Julia Krumhoff. Thanks go to Dr. Volker Stümpflen and Daniel Ellwanger (BIS group, Helmholtz Zentrum München) for the accomplishment of a joint cooperation with Prof. Juliane Winkelmann (Neurologische Klinik und Poliklinik, Klinikum rechts der Isar).

Finally, I owe gratitude to my family and friends who accompanied and supported me in every respect during my academic studies.

Dedication

To my parents, for their continuous encouragement and support over the developmental stages of my scientific career.

Thesis outline

The following dissertation is subdivided into two parts. The content of the first and main part delivers insight into the complex regulatory machinery of lincRNAs (large intergenic ncRNA). We focus on interactions of lincRNA transcriptome reconstructions with coding and noncoding transcripts and their associated functions. These functions are: (i) Regulation of their target alternative transcripts' expression and events (ii) Generation and/or interference of the activity of small ncRNAs. The second part is about a basic NGS (Next Generation Sequencing) analysis of RLS (Restless Legs Syndrom) affected patients and the identification of novel non-synonymous disease relevant SNPs.

It has to be mentioned that we analyzed several distinct data sets during the scientific development of this dissertation (first part). Accordingly the methodology of our large scale analysis has been improved during the developmental stages. In the following we provide a short insight into the stages:

In the very beginning of both, my doctor and diploma thesis in the year 2009 we started with data of a public database fRNAdb [2] providing data of distinct types of ncRNA. Since few lncRNAs (long ncRNAs) were reported in literature and the lack of a public database explicitly of lncRNAs at this time to our knowledge, we solely restricted the sequence lengths of ncRNAs to be > 200 nt according to the consideration of Kapranov et al. [3] to exclude small ones. In the year 2010 during my doctoral thesis to our knowledge the first RNA-seq data including mouse lincRNA (large intergenic ncRNA) transcriptome reconstructions came online by Guttman et al. [1] allowing for a deeper and fine tuned NGS analysis. For example quantitative expression levels could be determined for all coding and noncoding transcripts in a cell line for the first time [1]. With increasing amounts of lincRNAs more characteristics of lincRNAs came up in an equal measure [1, 4]. These upcoming characteristics are additionally of importance for fine tuning of our analysis of their associated regulatory actions and target effects [1, 4].

The content of the first part of our work highlights the most promising findings of our NGS group based on the RNA-seq data of Guttman et al. [1]. The field of lincRNAs is in an early stage of research, more and more RNA-seq data sets across an increasing amount of cell lines are coming up and will pave the way for ongoing bioinformatics anal-

ysis. We lay the focus on the biological component of bioinformatics in this dissertation - gaining knowledge about the regulatory manner of lincRNAs.

In general, the content of each part is structured in the conventional chapters: Introduction, Methods, Results and Discussion. The Motivation raises the biological issue and the aim of the part. The Introduction further provides a brief overview of the analyzed topic (part 1 mainly: RNA-sequencing, lincRNAs and alternative splicing regulation; part 2: RLS and exome sequencing). In the Methods chapter the NGS data and analysis steps of the according semi-automatic pipeline are explained step by step. The Results chapter illustrates the findings revealed in the NGS data using the implemented pipeline and biological interpretations. These results are discussed in the end of a part. For example yet unsolved and still remaining open biological questions in this thesis are addressed in the first part. The field of lincRNAs is in the very beginning and we are aware that we could unravel just a piece of the puzzle of the whole functionality of lincRNAs in this thesis. Another example for the second part is one weak point in the current NGS (exome sequencing) technique itself applied for the RLS patients. SNPs in lincRNA loci are neglected in this approach, but SNPs in these so called 'gene deserts' are currently gaining importance [5, 4]. The estimated frequency of disease causing SNPs in these loci is still increasing [5, 4] (one linkage to the first part denoting the impact of lincRNAs' SNPs and regulatory actions on disease such as cancer [6])

Part I

Regulatory actions of lincRNAs

1 Introduction

1.1 Motivation

Advances in transcriptomics and RNA-sequencing technologies paved the way for various research fields, including the reconstruction and identification of novel coding and noncoding transcripts [7]. Hence the amount of ncRNA of eukaryotic transcriptomes is still increasing by RNA-sequencing [1, 4]. ncRNAs are recently categorized according to their sequence length in small (<= 200 nt) and long (> 200 nt) by Kapranov et al. [3]. It turns out that lncRNAs (long ncRNAs) can originate from intronic or intergenic loci [6]. lincRNAs (large intergenic ncRNAs) account for a large fraction of noncoding RNA [8] and were primary analyzed in our work. These lncRNAs are capable to interact both with transcripts or splice variants of protein-coding genes and small ncRNAs [9]. This indicates that different coding and noncoding transcripts are not disconnected from each other, rather act in a joint regulatory manner [9]. Regulatory actions dependent on distinct interactions that were analyzed in our work are explained in the following subsections: (i) Regulation of their target alternative transcripts' expression and events 1.1.1 (ii) Generation and/or interference of the activity of small ncRNAs 1.1.2 and additionally exemplified in Figure 1.1.

1.1.1 Regulation of their target alternative transcripts' expression and events

Alternative splicing is a key regulator in eukaryotic gene function [10]. The majority of protein-coding genes (> 90 %) is subject to this mechanism [11]. The dominant hypothesis how alternate splicing is controlled is based on the action of splicing factors and their regulatory sites [12, 13, 14]. If splicing factors are common to the process, why then can they conditionally and specifically regulate a large number of alternate transcripts such as the many known tissue specific variants? Other explanations such as RNA Pol II elongation rate and chromatin modifications are rather unspecific and not suited to explain splicing regulation as a process or tissue specific regulatory mechanism [10].

Eukaryotic transcriptomes harbour a large fraction of non-coding transcripts (ncRNAs) of which a significant proportion exceeds lengths of more than 200 nucleotides obviously

not involved in miRNA guided translational control [3]. Such ncRNAs are commonly referred to as lncRNAs (long ncRNAs) [3]. Sound evidence shows that lncRNAs are functional and conserved in mammals [15, 16, 9]. For example, lncRNAs are involved in regulatory functions such as chromatin modification [17] and nuclear import [18]. In addition, several lncRNAs were reported to participate in the regulation of splicing [19, 20, 21]. For instance, the lncRNA Saf sequence is complementary to an intron of the human protein-coding gene Fas and induces alternative splicing of the Fas transcript [20]. Moreover, Kishore and colleagues showed that the snoRNA HBII forms a duplex RNA structure with the protein-coding serotine-receptor transcript and leads to an alternatively spliced isoform [22].

The formation of specific mRNA:ncRNA duplex structures to regulate alternative splicing is an attractive hypothesis for the specific regulation of alternative splicing. We demonstrate, that a substantial part of alternative protein isoforms can be explained by the formation of RNA:RNA duplex structures paired between mRNAs and lncRNAs. While analysing deep-sequencing data, we found evidence for the convincing hypothesis that the formation of RNA duplex structures controls alternate splicing in a specific, almost digital way, while other mechanisms may act rather global. In this work, we investigate the regulatory impact of lncRNAs on alternate splicing for the first time systematically at large scale. RNA-sequencing allows the reconstruction and quantification of the expression levels of (i) coding and non-coding transcripts [7] and (ii) of mixtures of transcript isoforms [23, 11]. Based on experimental RNA-seq data, we analyzed the influence of antisense lincRNAs (large intergenic ncRNAs) on splicing regulation of their targets. We used recently published strand-specific RNA-seq data obtained from three mouse cell lines: embryonic stem cells (ESC), neural progenitor cells (NPC) and mouse lung fibroblasts (MLF) published by Guttman and colleagues [1]. As targets of a lincRNA we identified all protein-coding transcripts that are candidates to form RNA duplex structures with complementary lincRNA. Functional characteristics of targets and target sites, including for example KEGG pathways and the GC content were analyzed. In the main part of our analysis for determining the influence of antisense lincRNAs on splicing regulation, we derived the quantitative change in expression and splice events of targeted protein-coding isoforms (part 1: functional role (i) in our work).

lincRNAs that are significantly sequence complementary to protein-coding transcripts are predominantly located at introns. The majority of complementary lincRNAs target multiple genes. We found that the expression of lincRNAs correlates with the expression of their target genes. Moreover, we demonstrate that lincRNAs regulate the expression of protein-coding alternative splice forms. lincRNAs significantly down-regulate the expression of coding exons which are close to target sites. A prevalence of lincRNAs to

target terminal introns, especially 5' introns, was observed. Down-regulation via complementary lincRNAs is strongest at 5' exons. Alternative transcript events: AFE and ALE (Alternative First and Last Exon) are significantly more frequently observed than other splice types. These results were confirmed in three different cell lines and implies a strong and widespread influence of lincRNAs on splicing regulation.

1.1.2 Generation and/or interference of the activity of small ncRNAs

An interaction between a long and small ncRNA is associated with distinct functions [9]. For example, a few lncRNAs are reported to serve as precursor for small ncRNAs and to influence the generation of small ncRNAs (like endo siRNAs) [9]. Additionally some lincRNAs interfere the activity of small ncRNAs (like miRNAs) [9]. The examination of this interaction (functional role (ii) in our work) is interesting of the following reasons: (i) Identification of lincRNA:small ncRNA interactions (ii) Unravelling whether one class of small ncRNAs is favoured more than other classes by lincRNAs and (iii) Clarifying whether these lincRNAs have an aberrant influence on alternative splicing regulation (functional role (i) in our work).

To answer these issues we primary determined the interplay of lincRNAs with distinct types of small ncRNAs using RNA-sequencing data including the reconstructions of novel lincRNAs across three mouse cell types [1], at large scale for the first time. To assess whether an interaction is existent, we run sequence similarity searches per cell line. As databases for small ncRNAs we included DeepBase[1] and fRNAdb[2]. The sequence similarity search analysis and calculation of quantitative expression levels is equivalent to the analysis of lincRNAs:protein-coding gene duplexes in context of the alternative splicing regulation (functional role i). Further we analyzed whether the regulatory influence of antisense lincRNAs on their coding transcripts' expression and events is changed in relation to interacting and non-interacting with an additional small ncRNA.

We found that a notable fraction of lincRNAs is sequence similar to at least one small ncRNA in all cell lines, especially in the ESC cell line. Interestingly this lincRNA:small ncRNA interaction is revealed for various distinct types of small ncRNAs, such as miRNAs and snoRNAs. We found that e.g. the frequency of small ncRNAs per lincRNA deviates across the small ncRNA types. This finding indicates a relation of the behaviour of a lincRNA to the class of a small ncRNA. The functionality of these interactions needs further experimental examination, but the interaction could be shown for RNA-seq data

[1] http://deepbase.sysu.edu.cn/
[2] http://www.ncrna.org/frnadb/

across three cell lines provided by Guttman et al. [1]. The influence of lincRNAs on alternative splicing regulation is retained irrespective of further interactions with small ncRNAs.

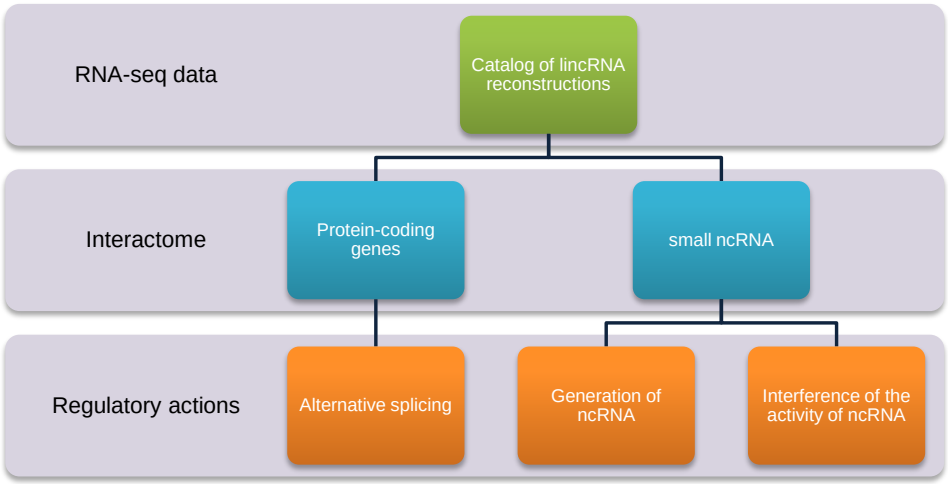


Figure 1.1: lincRNAs' regulatory actions at a glance. lincRNA exons can target either protein-coding genes or ncRNA. The lincRNA:gene interactions are associated with functional role (i): Regulation of their target alternative transcripts' expression and events; lincRNA:ncRNA interactions with (ii): Generation and/or interference of the activity of small ncRNAs.

1.2 A novel type of ncRNAs: Long ncRNAs

1.2.1 General characteristics

Current publications report that the human genome encodes just 20,000–25,000 protein-coding genes representing $< 2\%$ of total DNA, based on findings of the International Human Genome Sequencing Consortium (IHGSC) [24]. These statistics denote that protein-coding genes alone are insufficient to entirely explain the complexity of human genomes. It was reported that $> 90\%$ of the human genome is transcribed [25]. Gaining knowledge about ncRNA is one important component for a deeper understanding of the addressed complexity.

Kapranov et al. introduced a classification schema for RNAs in the year 2007 [3]. In this schema RNAs were assigned to following categories: long and small based on their transcript size [3]. This is to our knowledge one of the first publications introducing the term long RNAs (lRNAs). According to Kapranov, lncRNAs were considered to have an approximate lengths of > 200 nt [3]. Current estimates suggest that lncRNAs show lengths even up to 100 kb [6]. One of the first studied and reported lncRNAs was

discovered in the same year 2007 by Rinn et al. and termed HOTAIR [17] (according to a recent review [8]). HOTAIR is reported to play an important functional role in epigenetic gene silencing and is further associated with human cancers [26, 27]. LncRNAs are critically associated with diseases, such as cancer [6] (see subsection 1.2.4) underlining the importance of their target effects. According to the time scale of Gibb et al. the discovery of the first (cancer-associated) lncRNA H19 goes back to the year 1990 [6, 28]. The time scale can be found in the Appendix A Figure A.1, taken from the publication of Gibb and colleagues (Page 3 of 17) [6].

These lncRNAs were observed to be transcribed from distinct genomic regions [6]. Depending on these regions of origin and in relation to protein-coding genes, lncRNAs are often subclassified into e.g. intronic and intergenic [6]. The frequencies of distinct types of long and small ncRNAs are shown in diagram 1.2, taken from the review of Baker et al. [8]. This diagram illustrates the large amount of the lncRNA subclass, namely lincRNAs (large intergenic ncRNAs) in transcriptomes in relation to other types of ncRNA. 5,089 of 16,592 noncoding RNA genes are assigned to lincRNAs. We mainly focus on this subclass of lncRNAs consistently in the first part of our work, since we analyzed their functionality.

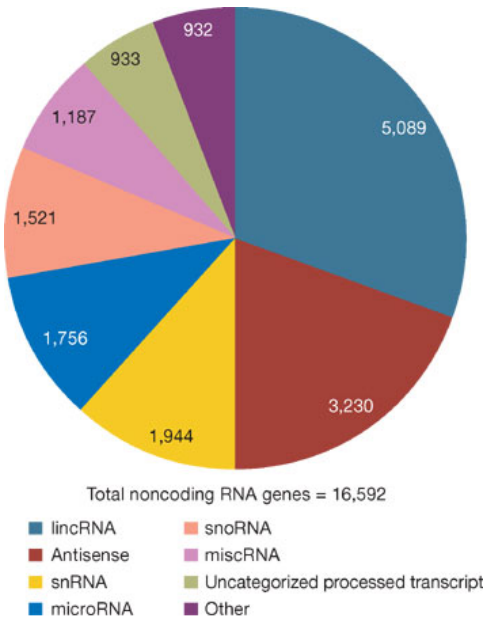


Figure 1.2: Overview about the frequency of distinct types of noncoding RNA genes. Each type is listed in the legend in a different colour. lincRNAs take a huge proportion of RNA genes into account. This Figure is taken from the review of Baker et al. [8].

Distinct techniques exist to date for the identification of lincRNAs in transcriptomes, such as chromatin state maps [29] and RNA-sequencing[1, 4]. The RNA-sequencing technique offers various opportunities [7]. One opportunity is the identification of novel transcripts as e.g. achieved in the work of Guttman and Cabili et al. [1, 4] in a high throughput manner. Ab initio transcriptome reconstruction applicable for RNA-seq data is a powerful method for the detection of large amounts of novel lincRNA loci among reconstructions [1, 4] (see subsection 1.2.2).

Not very long ago a debate was raised as addressed e.g. in the review of Mercer et al. [15]: LncRNAs were suggested to be non functional in the recent past mainly due to low conservation on the sequence level found in a few experiments. It turns out that this suggestion has to be reconsidered [15]. There is recent evidence that these lncRNAs rather yield as functional key players [15, 16, 9]. Their regulatory actions range from epigenetic gene silencing [26] to alternative splicing regulation [19, 20, 21].

A recent citation by Rinn addresses that lncRNA can be even considered as a new type of genes on the one hand. On the other hand the citation notes the underestimated and widespread functionality of this new type. The following citation of Rinn is taken from the review of Baker et al. [8]: "I don't know why people think that lncRNAs are all doing one thing", says Rinn. "They are just new types of genes, and their repertoire of functions I think will rival the proteome".

lincRNAs seem to have shared characteristics, with both protein-coding genes and ncRNAs. One shared characteristic with protein-coding genes is that many lincRNAs are reported to be spliced. lincRNA transcripts are multiexonic with on average about 3-4 exons [1, 4]. For comparison, protein-coding transcripts have on average 10-11 exons [1, 4]. Hence lincRNAs show a gene structure according to Guttman and Cabili et al. [1, 4]. In contrast to protein-coding genes, lincRNAs show as one example an increased tissue and cell line specificity [1, 4]. Likewise to other ncRNA types lincRNAs are reported to have decreased coding potentials and lack large ORFs [1, 4].

The functional annotation and classification of lncRNAs will be one of the important goals in this novel field. In subsection 1.2.3 in the Introduction a short overview about an extract of functional roles of lncRNAs is provided. We especially go into detail of functions relevant for the understanding of the first part of our work. It has to be noted, that on the one hand we have a huge and increasing amount of lncRNA data, but we have a lack of functionally classified lncRNAs on the other hand. A first database came online in the year 2011, termed lncRNAdb [30] trying to solve this issue and to provide a platform explicitly for the annotation of this type of ncRNAs. For example 112 lncRNAs

are available for Homo sapiens (human) and 88 for Mus musculus (mouse)[3] (download: April 25th, 2012).

1.2.2 NGS analysis

Next Generation Sequencing (NGS) generates millions of short reads from a sequence library. The most common second and third HT NGS platforms are listed in the following adopted from Table 1 of the review of Pareek et al. [31]. For a comparison of NGS platforms features such as read length, raw accuracy and sequencing run time are crucial. A detailed explanation and comparison of platforms can be found in reviews (e.g. [31, 32]). The read length is shown in rectangle in the following Enumeration.

NGS technologies

1. Roche GS FLX (400 bases)
2. Illumina (36 bases)
3. Life Technologies (35 bases)
4. Helicos Biosciences (Longer than 1000)
5. Pacific Biosciences (Longer than 1000)

RNA-seq and transcriptome reconstruction

RNA-sequencing is defined as the use of a NGS technology to sequence cDNA for transcriptome profiling. This method affords the reconstruction and quantification of whole transcriptomes (see reviews e.g. [7, 33, 34]). Determining the structure and expression level of transcripts from RNA-seq reads is an important issue for further scientific challenges. Transcriptome reconstruction approaches can be split into two strategies: (i) 'align-then-assemble' and (ii) 'assemble-then-align'. The strategies are illustrated in Figure 1.3, taken from Haas et al. [7]. A list of transcriptome reconstruction tools can be found in the Appendix A Table A.1, adopted from Garber et al. [33].

[3] http://lncrnadb.com/

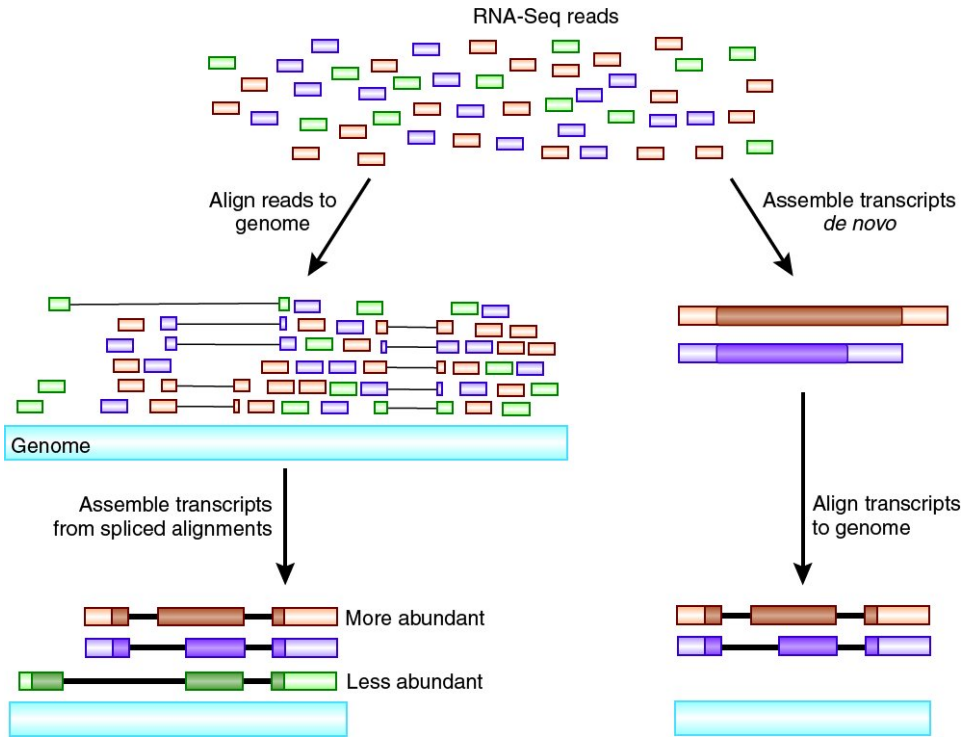


Figure 1.3: Illustration of transcriptome reconstruction strategies: align-then-assemble (left) and assemble-then-align (right). This Figure is taken from the review of Haas et al. [7].

Determination of novel lincRNA reconstructions

We focus on ab initio transcriptome reconstructions and the challenge: identification of novel lincRNAs since we analyzed the regulatory actions and target effects of lincRNAs. In the following we describe the idea behind this challenge (no explicit pipeline), split in four main computational tasks (see Figure 1.4). This idea is based on recent publications and described in detail by e.g. Guttman and Cabili et al. [1, 4].

To achieve these four addressed tasks (1-4), RNA-seq reads are mapped onto a reference genome with a spliced aligner in a first task (1). A common spliced aligner is e.g. Tophat [35]. Next all transcripts are assembled using an ab initio (assemble-then-align) transcriptome reconstruction tool (e.g. Scripture [1]) in a second task (2). lincRNAs are discovered among reconstructions using specific criteria (3) (i-vi) [1, 4], such as e.g.: (i) multiexonic (ii) intergenic (iii) low coding potential (vi) transcript length $>$ as applied threshold (e.g. 200 nt) (..). These lincRNA reconstructions are often classified in the categories [4] (4): (i) known or annotated and (ii) novel. Established annotation references are used (e.g.RefSeq NR*, GENCODE 4 and UCSC Non-coding) to retrieve known annotations [4]. An overview of RNA-seq analysis programs can be found in the

Appendix A Table A.1, taken from Garber et al. [33].

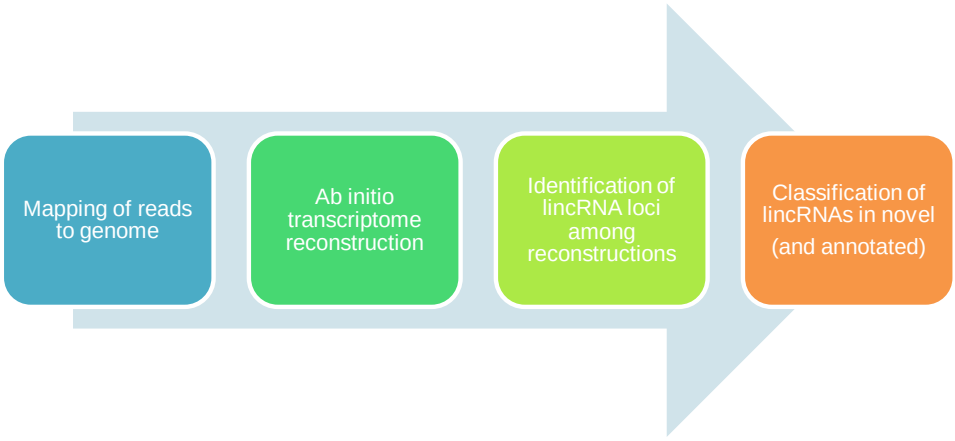


Figure 1.4: Overview of the discovery of lincRNA genes based on RNA-sequencing. The discovery is split in four steps (1-4). First reads are mapped onto genome (1). These pre-aligned reads are used to reconstruct the transcriptome (2). Next, lincRNA transcripts are identified (3) and categorized using annotation references (4).

One pipeline achieving this challenge was recently published by Cabili et al. built on this concept [4]. The lincRNA classification pipeline takes RNA-seq data and annotation sources as input. The pipeline is applicable for any RNA-seq data sets and is available online [4]. The steps can be summarized as follows (see Figure 1.5): First, RNA-seq data is assembled with the tools Scripture [1] and Cufflinks [36] and annotated lincRNAs are incorporated. Next, a unique set of reconstructed transcripts is generated with Cuffcompare. This unique set of isoforms is further filtered (i) known non-lincRNAs annotations (ii) transcripts with protein domain (iii) transcripts with positive coding potential and (iv) size selection.

Cabili and colleagues applied their implemented pipeline on human RNA-seq data across 24 tissues and cell lines [4]. A stringent set of 4662 human lincRNAs was generated. This set contains lincRNAs that were assembled in at least two different tissues or by two ab initio transcriptome reconstruction tools in the same tissue. These identified lincRNAs were further analyzed in the work of Cabili et al. [4]. Cabili and colleagues e.g. found that many lincRNAs show the presence of a K4-K36 domain [4]. A K4-K36 domain is a chromatin signature. Histone-3 Lys4 trimethylation modifications (H3K4me3) are reported to mark promoter regions and histone-3 Lys36 trimethylation modifications (H3K36me3) to mark the transcribed regions.

The pipeline of Cabili et al. shows several advanced possibilities [4] in contrast to other studies. For example Guttman and colleagues primary implemented and published the ab initio transcriptome reconstruction tool Scripture and provide a Scripture walkthrough [1]. They applied their implemented pipeline to identify novel lincRNAs. The tool Scripture and the data of reconstructed novel lincRNAs is available online. The technique for the identification of lincRNA loci is explained, but not provided as a tool by Guttman et al. [1].

Cabili et al. combine distinct ab initio transcriptome reconstruction tools (Figure 1.4, task 2) [4]. Further the discovery of both annotated and novel lincRNAs (Figure 1.4, task 3 and 4) is incorporated in one pipeline [4]. The pipeline and determined exon/intron structure of a reconstructed lincRNA of Cabili et al. [4] is shown in Figure 1.5 **A** and **C** for illustration purposes. This lincRNA is reconstructed by the ab initio reconstruction tools: Scripture [1] and Cufflinks [36]. Further the ncRNA fulfils several of addressed criteria (such as e.g. multiexonic) and is checked for novelty (in this case, already annotated in GENCODE and UCSC).

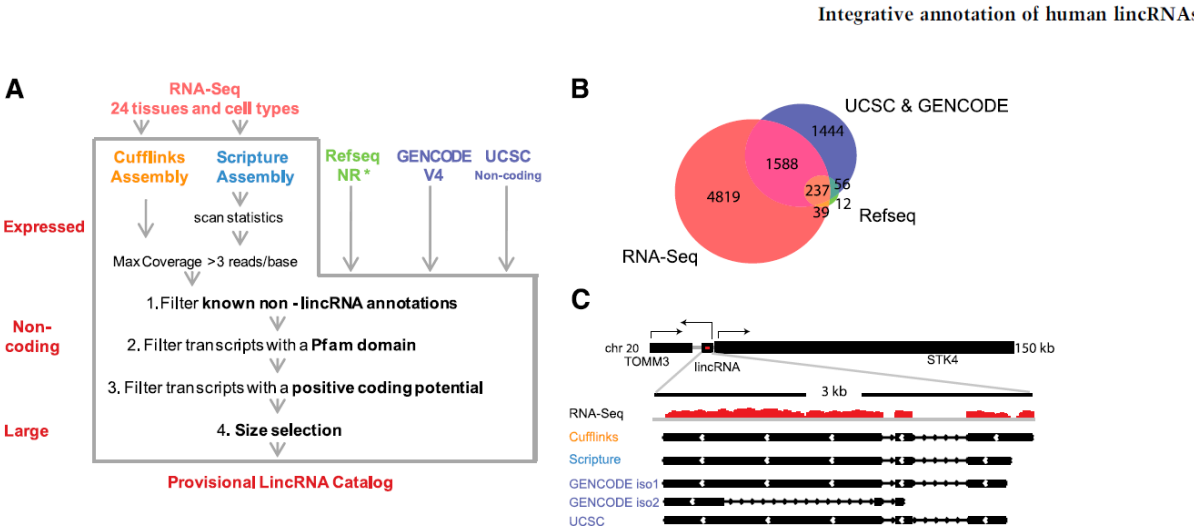


Figure 1.5: Study of Cabili et al. for the generation and analysis of lincRNA transcripts [4]. **A** lincRNA classification pipeline, as described above and by Cabili and colleagues. **B** The number of lincRNA reconstructions overlapping annotation sources (as Venn diagram). **C** Both ab initio transcriptome reconstruction tools identified a lincRNA loci, already annotated as noncoding RNA in GENCODE and UCSC. This Figure is taken from Cabili and colleagues [4].

1.2.3 Functionality

To provide a first brief overview about an extract of lncRNAs' regulatory actions we took the Figure of a review over by Wilusz et al. [9], see Figure 1.6. 8 actions are illustrated in this Figure including 2 regulatory actions that were analyzed in our work (slightly modified for the interaction with small ncRNAs), namely: 3. Modulate alternative splicing patterns and 8. Small RNA Precursor.

3. corresponds to the function (i) Regulation of their target alternative transcripts' expression and events and 8. to the function (ii) Generation and/or interference of the activity of small ncRNAs in our work. We explain these two functions based on recent findings in this subsection (2 findings per function). These findings are just shown in particular experiments and provide the building stones of ideas for the construction of our large scale analysis of lincRNA transcriptome reconstructions.

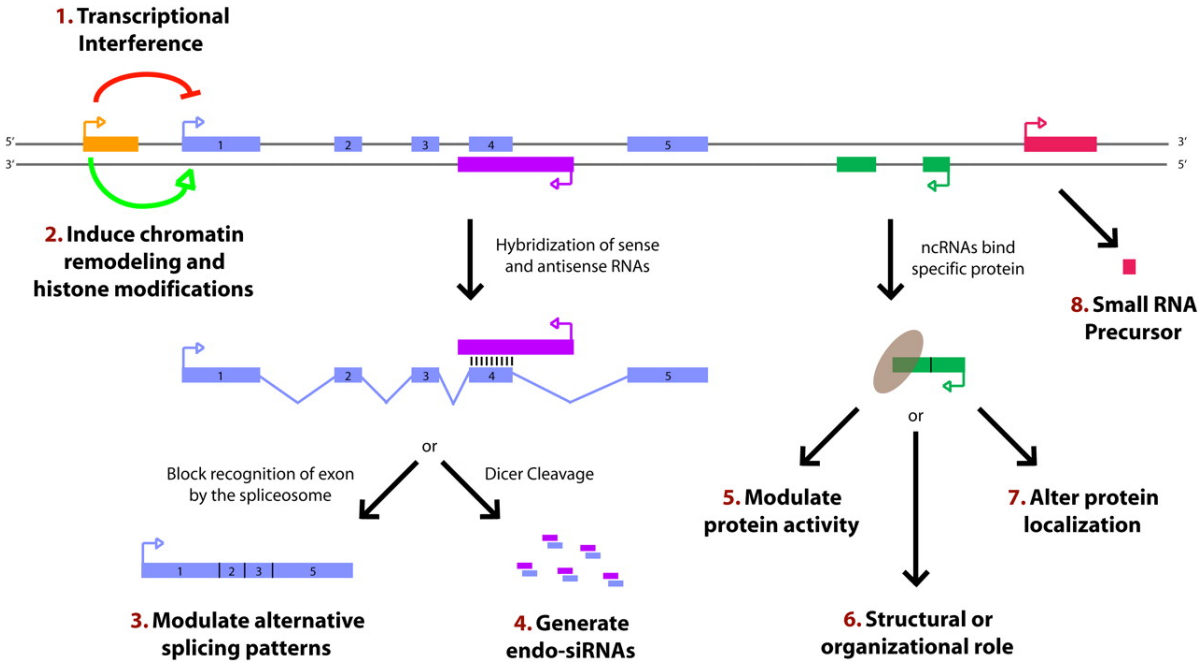


Figure 1.6: Illustration of the regulatory actions of lncRNAs. Eight actions are exemplified. This Figure is taken from the review of Wilusz et al. [9]. 3. and 8. of depicted actions were analyzed in our work. 3. LncRNAs are reported to be capable to influence alternative splicing regulation. 8. In addition lncRNAs serve as precursors for small ncRNAs.

LncRNAs (antisense) are capable to bind complementary to distinct regions of their targets (sense): entirely to an intron or to an exon/intron boundary (splice junction). These regions of complementarity indicate to be crucial for the mode of splicing regulation of target genes: activatory or inhibitory. Activation can be associated with alternative transcript events such as e.g. skipped exon or mutually exclusive exon. Contrary, inhibition might show associations with other events e.g. intron retention. Splicing and splice types are summarized in the introduction, see section 1.3.

One example for splicing regulation (function 3.) is the ncRNA Saf analyzed in the work of Yan et al. [20]. Saf is antisense (opposite transcript orientation) to intron 1 of its sense counterpart protein-coding gene Fas. Further Saf is in close proximity to the next splice site of exon 1 of Fas. The Fas (Apo-1/ CD95) gene is a receptor of the tumor necrosis factor (TNF) and nerve growth factor family. Apoptosis is known to be induced by the association of receptor Fas with e.g. its natural ligand FasL. Saf shows an influence on the expression of alternative isoforms of the Fas gene. A significant upregulation was found for two of four isoforms or splice forms of Fas in Jurkat cells. This mechanism of the Saf:Fas RNA interaction as described might go along with e.g. isoforms lacking the death domain and thereby causing the observed inhibition of apoptosis. A

protection effect was observed under the influence of Saf:Fas. FasL-induced apoptosis was inhibited. 'Intronic' antisense:sense transcript pairs, such as Saf:Fas are obviously crucial for alternative splicing regulation.

Another example for the formation of a non-'intronic' RNA duplex or antisense:sense transcript pair is explained in the following. A ncRNA is reported by Beltran et al. [21] to show complementarity to the 5' splice site of 5' UTR intron (or intron 1) of the zinc finger homeobox mRNA of Zeb2 (Sip1). Zeb2 (Sip1) is a transcriptional repressor of E-cadherin (Calcium dependent adhesion molecules). E-cadherin is a transmembrane protein and important for cell-adhesion. This first intron, containing the internal ribosome entry site (IRE) is retained. This 'intron retention' leads to the translation and expression of Zeb2 (Sip1). An increase of Zeb2 protein levels causes a down-regulation of E-cadherin mRNA and protein. This demonstrates that non-'intronic' antisense:sense transcript pairs influence alternative splicing as well as 'intronic' ones. In this explicit case study the alternative transcript event is verified: Inhibition of splicing of exon/intron 1 of Zeb2 (Sip1). The first intron is retained (see Figure 1.7, taken from Mercer et al. [15]).

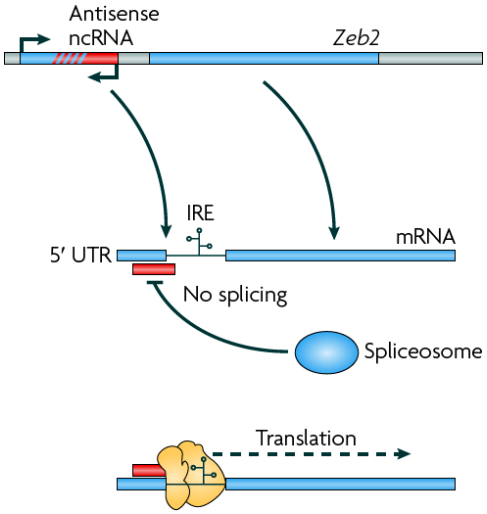


Figure 1.7: Example of an antisense lncRNA preventing splicing of an intron. The ncRNA is located at the 5' splice site of 5' UTR intron. This mechanism of base pairing between lnc,- and mRNA leads to the retention of the intron. This Figure is taken over from Mercer et al. [15].

Besides the capability of lncRNAs to target protein-coding genes, other targets such as small ncRNAs are reported. Small ncRNAs are suggested to be derived from lncRNAs (sense). Further small ncRNAs can be base paired by lncRNAs (antisense). Distinct small ncRNA types are reported as target candidates including miRNAs. The regula-

tory behaviour of small ncRNAs can be modulated via the effect of lncRNAs.

For example the first nc-exon of lncRNA H19 (involved in genomic imprinting) serves as precursor for miRNA miR-675 in human and mouse [37]. Further small ncRNAs' actions (such as targeting genes) can be inhibited through complementary lncRNAs (termed as 'microRNA sponges'), as illustrated in the work of e.g. Ebert et al. [38].

1.2.4 Connection to disease

There is increasing evidence for a role of ncRNA in diseases, including cancer. The rise of according publications per year is illustrated in Figure A.2 in Appendix A. This Figure is taken from the review of Gibb et al. [6] and is based on a pubmed search for the terms "ncRNA" or "non-coding RNA" or "noncoding RNA" or "non-protein-coding RNA" with cancer and annual (January 1 to December 31). LncRNAs influence cancer progression via distinct modes of action. An extract of reported mechanisms are highlighted in Figure 1.8 and summarized below including case studies of lncRNAs (reviewed in e.g. the publications of Gibb et al. [6] and Tsai et al. [27]) A list of cancer-associated lncRNAs including references could be found in Table A.2 in Appendix A, additionally taken from the review of Gibb et al. [6].

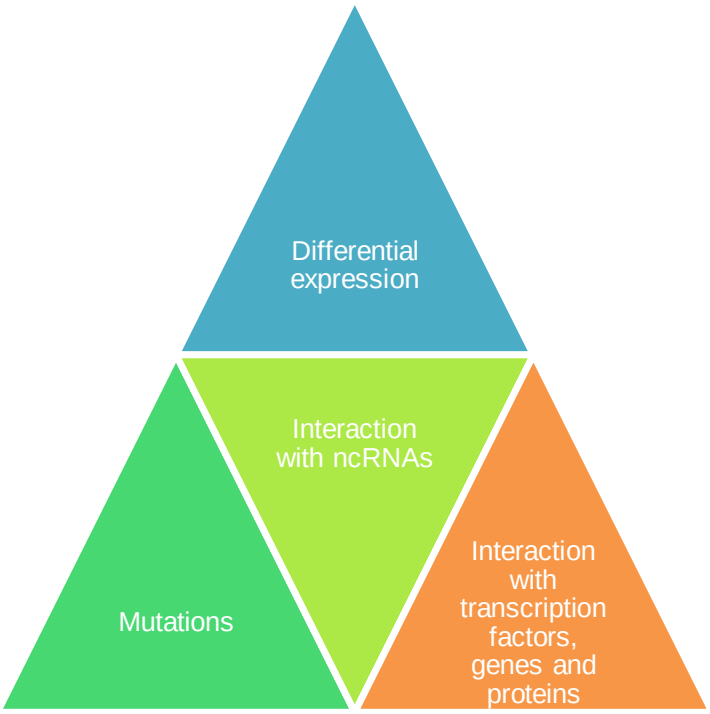


Figure 1.8: Overview about the mechanisms of lncRNAs in cancer progression. Differential expression: The expression of lncRNAs are reported to be up/down-regulated, Mutations: SNPs were found within e.g. intergenic loci on the genome in cancer tissues, Interaction with coding and noncoding transcripts: LncRNAs show the capability to target small ncRNAs and transcription factors that are associated with cancer.

lincRNAs are reported to show differential expression in cancer cells in comparison to normal cells. For example the lincRNA HOTAIR (HOX antisense intergenic RNA) is 2.2 kb in length, originates from Chromosme 12 and is up-regulated in one human cancer type (Breast, listed in Table A.2 in Appendix A). The mechanism of HOTAIR is shown in Figure 1.9, taken from Gibb et al. [6]. Upregulation of HOTAIR is correlated with cancer metastasis.

HOTAIR recruits the protein-complexes PRC2 and LSD1. Sequence motifs are crucial for these interactions, e.g. the 5' region of HOTAIR binds PRC2. This recruitment results in targeting of metastasis suppressor genes at Chromosome 2. Target genes are further silenced through H3K27 methylation and H3K4 demethylation leading to metastasis. The association of HOTAIR with cancer is reported [17, 39], listed in Table A.2 in Appendix A.

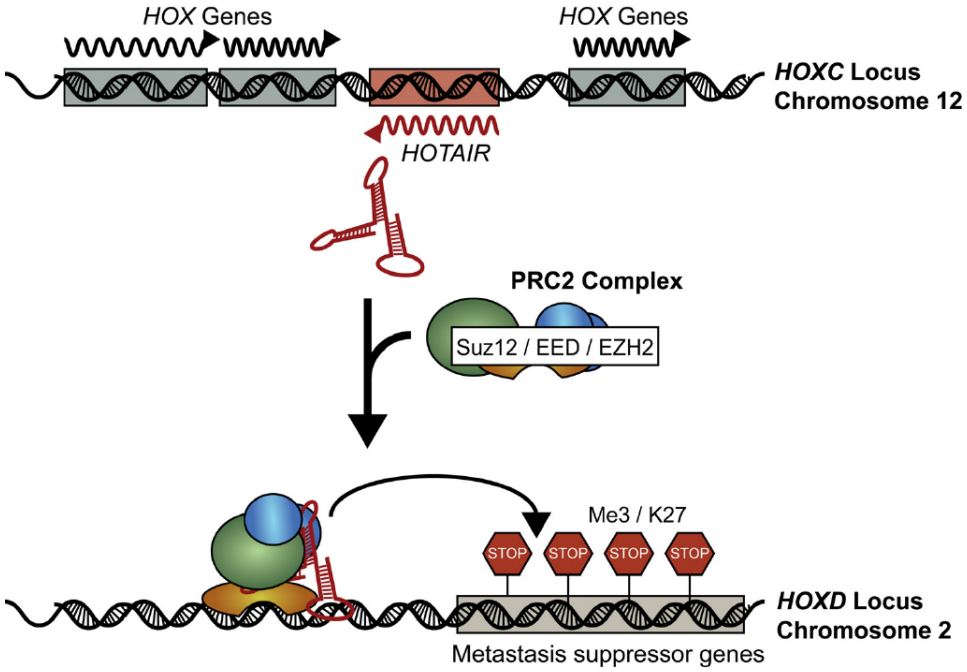


Figure 1.9: Example of a lincRNA involved in cancer progression. The lincRNA HOTAIR interacts with protein-complexes for silencing of metastasis suppressor genes at HOXD locus. This Figure is taken from the review of Gibb et al. [6]. According to Gibb and colleagues just the complex PRC2 is shown for simplification.

LncRNAs were found to harbour mutations such as SNPs. One example is PRCNR1 (prostate cancer non-coding RNA 1), 13 kb in length and transcribed from a 'gene desert' region on chromosome 8. SNPs between rs1456315 and rs7463708 are most significantly associated with PC (prostate cancer) susceptibility. PRCNR1 is up-regulated in PC cells. The findings of PRCNR1 are described in the work of Chung et al. [5], listed in Table A.2 in Appendix A. Likewise to HOTAIR, other cancer types have not been reported yet.

Besides differential expression and SNPs, lncRNAs interact with coding and noncoding transcripts. H19 is 2.3 kb in length and is in contrast to e.g. HOTAIR and PRCNR1 not just associated with one cancer type, but with over ten. This lncRNA H19 is reported to have both oncogenic and tumor suppressive potential. Interactions with genes (e.g. p53), transcription factors (e.g. c-Myc) and small ncRNAs (e.g. miR-675) are reported (see Table A.2 in Appendix A for references). As already mentioned in subsection 1.2.3 the first exon of H19 yields for example as precursor for miRNA miR-675. This generation of miR-675 is crucial since this miRNA targets the tumor suppressor retinoblastoma (RB) [40]. An inverse relation between the expression of RB and miR-675 was shown in e.g. human colorectal cancer (CRC) [40]. This is one example that the first lncRNA exon might be an important region of the lncRNA transcript for the generation of a small ncRNA. Further this example notes the severe regulatory interplay of long and

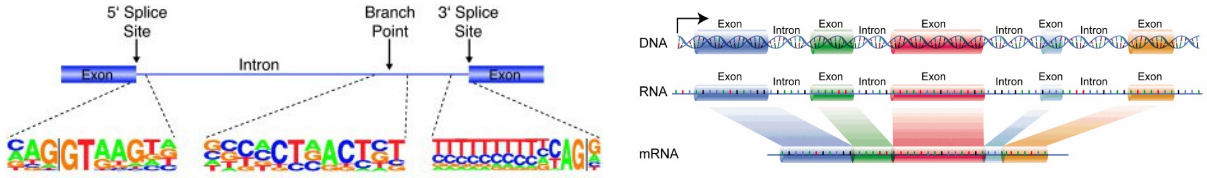
small ncRNA and association to disease.

1.3 RNA processing

1.3.1 Splicing

The process of gene expression is achieved by two main steps: transcription and translation. In a first step the genomic DNA is transcribed into pre-mRNA (mRNA precursor) in the nucleus: transcription. The basic segments of the pre-mRNA of a gene are the exons and introns. These segments are bounded by signals, namely splice sites. In detail, the 5' (or donor) splice site of an intron is marked by the dinucleotide *GU*, the 3' (or acceptor) site by *AG*. Another additional and important signal is the branch point *A*, residing upstream of the 3' splice site. The branchpoint is followed by a polypyrimidine tract. The sequence motifs of mentioned signals, are established under the term *cis*-acting regulatory elements and of importance for spliceosome assembly.

RNA processing is an important part of eukaryotic gene expression. The introns of the pre-mRNA are excised or spliced during this process. The exons of a gene are reconnected to the mRNA (mature RNA). The mRNA is finally transported into the cytoplasm and translated into protein in the second step of gene expression: translation. Explained fundamentals of gene expression and splicing are reviewed in e.g. [41, 12, 42]. For clarification, a detailed intron schema and an illustration of splicing are depicted in Figure 1.10. The references are listed in the Figure legend.



(a) (b)

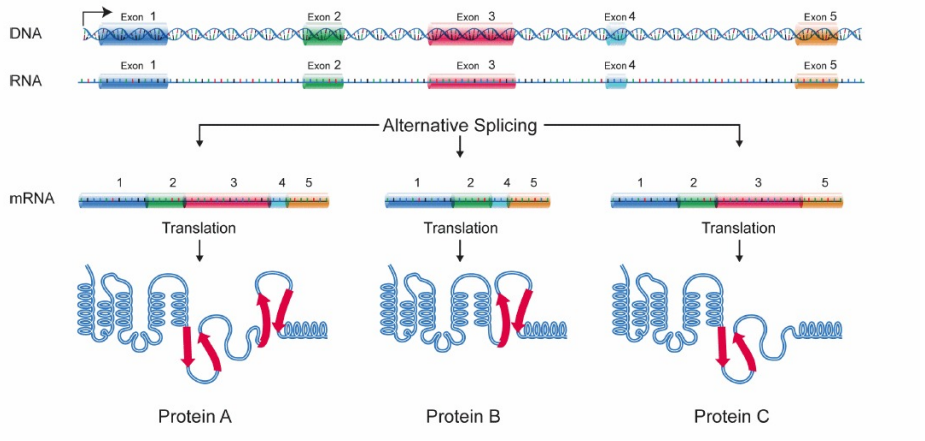
Figure 1.10: Schemata of the pre-mRNA and the splicing process. (a) An illustration of the pre-mRNA with its exons and introns. *cis*-acting regulatory elements are marked with an arrow. The Figure is taken from the review of McManus [41]. (b) Diagram of the splicing process (Figure is taken from: www.genome.gov). Briefly, the introns are spliced out and exons are ligated in this process.

1.3.2 Alternative splicing

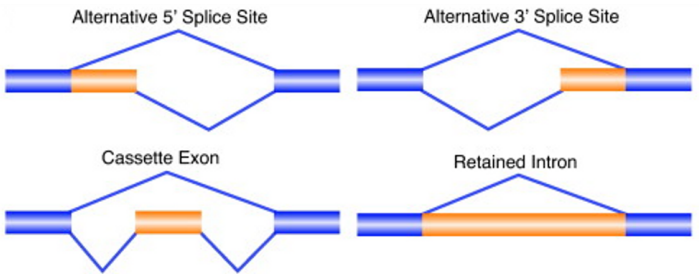
Splicing can be split into the categories constitutive and alternative. In the following, we further go into detail for the category alternative. The exons are reconnected in different ways during AS resulting is accordant mRNA transcripts, as reviewed by McManus [41]. These mRNA transcripts are also known as e.g. isoforms or splice variants. One single gene can thereby encode for multiple proteins [41]. This is in contrast to constitutive splicing (mentioned and illustrated above, see Figure 1.10) and goes along with an increase in proteomic diversity and complexity [41]. A few points are highlighted in the following in order to emphasize the importance of AS.

40-60% of human genes were estimated to be subject to AS (Alternative Splicing) in the year 2002 [43]. Just six years later in 2008, indication of even 92-94% was reported [11]. The human genome project found just twice as many genes as fruit fly (Drosophila) [44]. This high percentage of eukaryotic genes undergoing AS and the increased complexity of the proteome is one dominant explanation for the findings of this project. For example, humans show an increased rate of alternative splicing compared to fruit flies [45] to affirm this suggestion. In addition, specific introns can be retained through AS (AS types, see Figure 1.11). As previously mentioned in the introduction, ncRNA is associated with introns. On top both, AS and ncRNA are critically connected to disease as explained in the introduction.

Alternative splicing can be assigned to distinct events (as reviewed in e.g [41, 46]). These events are reported to be tissue specific [11, 41]. The four basic events are (i) exon skipping (cassette exon) (ii) alternative 5'ss (iii) 3'ss and (ii) intron retention [41]. A complex event is built by a combination of these basic events, such as Mutually exclusive exon (out of two cassette exons) [41]. According to recent reviews the most common event is cassette exon [46]. Cassette/alternative and constitutive exons differ in their biological patterns [46]. For example cassette exons tend to have (i) high conservation levels (ii) short exons and (iii) weak splice sites (e.g. for 5'ss the binding affinity to U1 snRNA) (reviewed by Kim et al. [46]). A schematic illustration of AS and its types is shown in Figure 1.11. References can be found in the Figure Legend.



(a)



(b)

Figure 1.11: Schemata of alternative splicing and splice types. (a) One gene (with Exon 1-5) can encode for multiple proteins (A,B,C) via alternative splicing (Figure is taken from: www.genome.gov) and the event: exon skipping (cassette exon). (b) Alternative splicing events are exemplified, namely alternative 5'ss, 3'ss, cassette exon and retained intron. This Figure is taken from the review of McManus et al. [41].

1.3.3 Spliceosome assembly

Splicing can be described as a two-step enzymatic reaction [42]. These two transesterification steps are carried out by the spliceosome [42]. The reaction is shown in Figure 1.12 and briefly summarized as follows, according to the review of Black et al. [42]. Two intermediates are produced in a first step. The 5' splice site of an intron is cleaved producing a detached 5' exon intermediate. A lasso like intermediate (or lariat) is built by concatenation of the 5' splice site with the branch point. This first step is an attack of the 2' hydroxyl group at the branchpoint on the phosphate at the 5' splice site. Secondly, the 3' site is cut, the two exons are joint and the intron released as lariat. This second step is an attack of the 3' hydroxyl group of the exon intermediate on the phosphate at the 3' splice site.

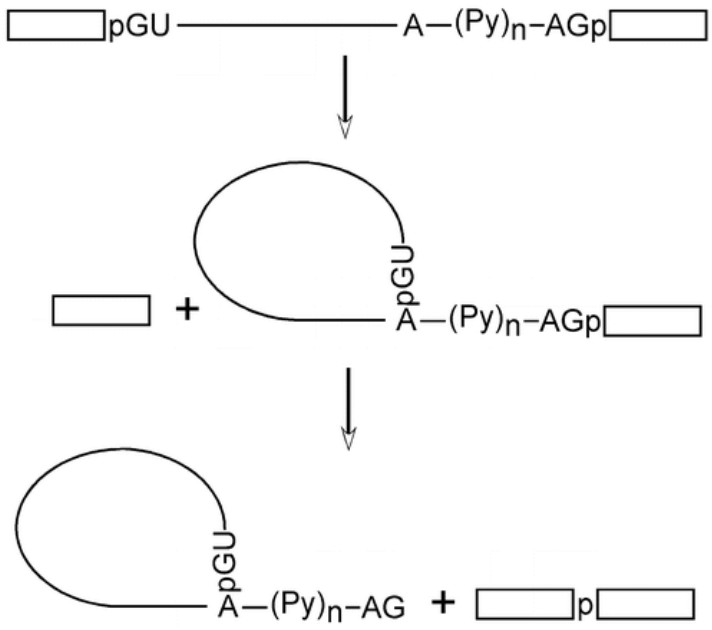


Figure 1.12: Exemplification of splicing as a two-step enzymatic reaction. First, an intron lariat is created by cleavage of the 5' splice site. Secondly, exons are ligated after cutting the 3' site. This Figure is taken from the review of Black et al. [42].

The ribonucleoprotein complex spliceosome is as mentioned responsible for the catalysis of splicing. This complex is mainly composed of five snRNPs (small nuclear ribonucleoproteins). The most common snRNPs are known as U1, U2, U4, U5 and U6. One snRNA in combination with proteins forms a particle. Regulation of splicing is traditionally explained by the combination of *cis* and *trans* components. *cis* components are sequence motifs of the pre-mRNA. *trans* components are factors, such as RNA or protein. Spliceosome assembly is directed by snRNPs recognizing according *cis*-acting sequence elements. For example U1 snRNP starts with binding to the 5' splice site and U2 follows with binding to the branchpoint. Complexes of components are built to fulfill spliceosome assembly (known as E, A, B and the catalytic C complex [42]). For example the Early (E) complex consists of U1 and the 5' splice site.

The previously mentioned *cis*-acting sequences as e.g. branchpoint are essential for the guidance of the splicing reaction. Further signals are located in exons and introns and of importance for alternative splicing regulation. The regulatory action of these additional signals can be either enhancing or silencing. Exonic signals are e.g. known as exon enhancers (ESEs) and silencers (ESSs), for intronic (ISEs) and silencers (ISSs) respectively. RNA binding proteins, such as SR proteins (serine/arginine-rich proteins) and hnRNPs (heterogeneous nuclear ribonucleoproteins), recognize these motifs. The basics of spliceosome assembly and signals are reviewed in e.g. [41, 42].

1.3.4 Complexity of splicing and connection to disease

The basic principles as addressed in this section of our work are simplified for illustration purposes. It has to be noted that splicing is a complex mechanism in many respects. Many full particulars of splicing regulation are still unknown. For example, regulation of splicing cannot be just explained by previously mentioned *cis* and *trans* components. There is recent evidence that chromatin structure and histone modifications play an important role in alternative splicing regulation [10].

In fact diseases, such as cancer, Pakinson‘s and Alzheimer‘s disease, are assumed to be caused by aberrant splicing. Cells and their genome in aged patients show alterations. Basically, mutations in *cis* and *trans* components are responsible for inaccuracies in splicing. According to the components, the classification in *cis* and *trans* effects is established. A *cis* effect is a mutation in a sequence motif of the pre-mRNA. This mutation might affect correct splicing in the sense that for example the binding affinity of a splicing factor is altered. For example one nulceotide of the 5’ (or donor) splice site of an intron, is point mutated. Consequently the sequence might not be recognized and bound any more by *trans* components. This misregulation to splice out an intron results in the AS type: Intron retention. The expression of one gene is affected. Contrary a *trans* acting effect is taking place if a *trans* component is mutated. As one possible result the splicing factor can loose its binding capability, for example to a splice site. This leads to accessibility and usage of this exposed splice site for splicing. The expression of multiple genes can be affected. Alternative splicing in association with disease is reviewed in e.g. [47, 48].

1.3.5 Regulatory mechanisms of alternative splicing

Recent reviews suggest additional mechanisms such as elongation rate and chromatin modifications to play a role in alternative splicing regulation (see [49, 10]). A link between elongation rate and splicing is supported by the ’promoter effect’ in the first place. The promoter effect can be explained as a consequence of different rates of RNA Pol II elongation (kinetic coupling). Alternative splicing outcome (e.g. cassette exon 33 inclusion levels) can be affected by different RNA Pol II promoters (e.g. of human fibronectin (FN)). Chromatin structure is reported to show an effect on splicing factor recruitment and splicing (e. g. histone methyltransferase CARM1 interacts with snRNP proteins). Further, nucleosomes and RNA Pol II have a differential distribution along genes. For example an enrichment of nucleosomes at splice sites was observed. The enrichment is important for exon definition. Additional higher enrichment was found for included in comparison to excluded alternatively spliced exons. This finding is supportive for a link to splicing.

As already mentioned, RNA and proteins can serve as *trans* components. ncRNA, such as lncRNAs and snoRNAs, are additionally and most recently suggested to regulate alternative splicing [49]. Previously explained mechanisms are supported in few experiments. It is still unclear how and to which extent splicing is explicitly regulated. In addition to this unclarity, ncRNAs can act in a more specific way: in target choice and regulation. The pre-mRNA of target genes can be accurately bound by sequence complementary ncRNAs. Examples of lncRNAs participating in alternative splicing regulation via formation of RNA:RNA interactions are provided detailed in the introduction (see 1.2.3). Alternative splicing regulation via lincRNA:mRNA interactions is analyzed in our work. Exons, close to binding sites, are specifically regulated and resulting splice variants can be explained by RNA:RNA interactions. In the following we additionally highlight for each type of ncRNA one example in context of splicing misregulation and disease (illustrated in Figure 1.13).

Another lncRNA, namely MALAT-1 (metastasis associated lung adenocarcinoma transcript 1), regulates alternative splicing via another mechanism [50, 19]. MALAT-1 is processed into a 61-nt tRNA-like small RNA (mascRNA, MALAT1-associated small cytoplasmic RNA) and the mature long MALAT1 transcript. The long MALAT1 transcript interacts with the *trans* components, SR proteins [50, 19] in the nucleus. Downregulation of the expression of MALAT-1 results in a rise of unphosphorylated SR proteins and exon inclusion events [19]. The expression of MALAT-1 is reported to be three-fold increased in NSCLC (non-small-cell lung cancer) metastasizing tumors [51].

Splicing regulation is for example shown for the human snoRNA HBII-52 [52, 22]. This snoRNA is processed to smaller ones, termed psnoRNAs [52, 22]. These psnoRNAs show sequence complementarity to an ESE (exon silencing element) in exon Vb [52, 22]. These RNA:RNA duplexes interfere with a splicing factor leading to exon inclusion [52, 22]. This observation is described with an additional Figure in a recent review [49] (see Figure 1.13). Interestingly, loss of the expression of HBII-52 in the Prader-Willi Syndrome is found to cause misregulation of splicing [52, 22].

Gaining more knowledge about regulatory mechanisms is an important step for the involvement splicing and diseases. The listed findings underline the concern to unravel splicing regulation of ncRNA as *trans* components and their associated effects.

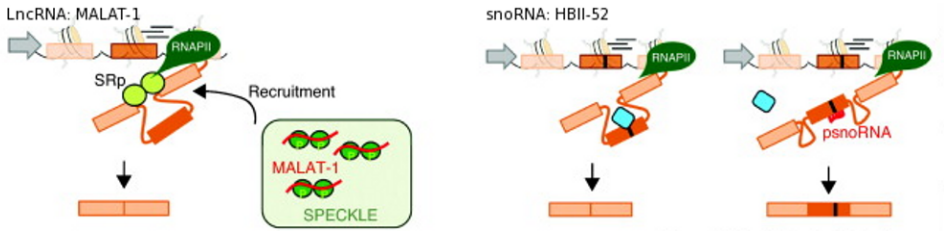


Figure 1.13: Long and small ncRNAs regulating alternative splicing. Downregulation of MALAT1 leads to an increase of SR proteins and exon inclusion (left). RNA:RNA duplex with psnoRNA and ESE (exon silencing element) is suggested to prevent binding of a splicing repressor leading to exon inclusion (right). This Figure is taken from Luco et al. [49].

2 Materials and Methods

2.1 Data sources

2.1.1 Novel lincRNA reconstructions and annotation resources

Strand-specific next generation sequencing data of mouse cell lines were taken as published by Guttman et al. [1]. This data consists of paired end RNA reads with length of 76 nucleotides [nt] of embryonic stem cells (ESC), neural progenitor cells (NPC) and mouse lung fibroblasts (MLF) sequenced on an Illumina GAII platform. The advantage using this RNA-seq data set is the coverage of distinct cell lines. In addition, Guttman and colleagues pre-aligned the reads and reconstructed the transcriptome. Novel lincRNA reconstructions were identified and are available from the original data. lincRNAs tend to be spliced and show an exon/intron structure. The exons of novel lincRNA reconstructions with strand orientation were taken as reference lincRNA set per cell line in our work.

Since Guttman et al. [1] provide just a catalogue of novel lincRNA reconstructions, annotation resources for known ncRNAs can be incorporated. We used the RefSeq gene annotation track and the mouse assembly (NCBI37/mm9) from UCSC genome browser[1] (download: October 6, 2010). We selected transcripts satisfying the following criteria as suggested by Cabili et al. [4]: transcript ids starting with the prefix NR* (RNA[2]), multiexonic and transcript size > 200 nt.

2.1.2 Coding transcripts

As genome reference sequence we used the mouse assembly (NCBI37/mm9). Information about protein-coding transcripts and their exon/intron structures were taken from the UCSC genome browser with track: RefSeq Genes and assembly: NCBI37/mm9 (download October 6, 2010; mRNA[3]).

[1] http://genome.ucsc.edu/cgi-bin/hgTables?command=start
[2] http://www.ncbi.nlm.nih.gov/RefSeq/key.html
[3] http://www.ncbi.nlm.nih.gov/RefSeq/key.html

2.1.3 small ncRNAs

Data of different types of small ncRNAs were obtained from the public databases: DeepBase[4] (download: October 14, 2011; track: miRDeep miRNA, snoSeeker snoRNA, nasRNA, pasRNA, sense rasRNA and sense easRNA) and fRNAdb[5] (download: October 14, 2011; track: snoRNA, snRNA, miRNAs of miRBase (pre-miRNA and mature-miRNA), piRNA). A list of the abbreviations and full names of small ncRNAs is given in Appendix A (see Table A.3). A comprehensive overview about our transcript types and according frequencies is shown in Figure 2.1.

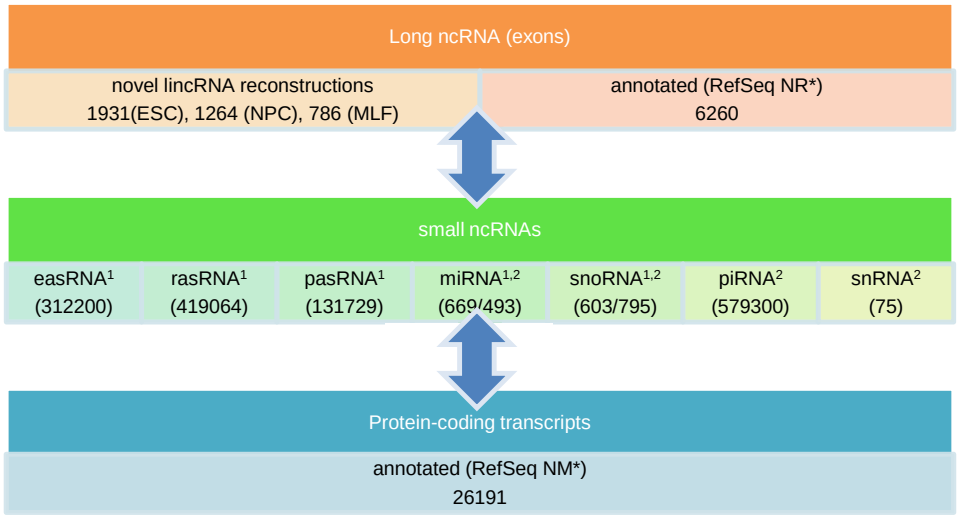


Figure 2.1: Overview of the distinct data sources for coding and non-coding transcripts. The frequencies of distinct transcripts are listed in this Figure. We show the statistics for lncRNAs, split in novel and annotated. Novel lincRNA reconstructions are provided per cell line by Guttman et al. [1]. We extended this data with RefSeq gene annotations to obtain a complete catalogue of lncRNAs. For coding:coding and coding:noncoding interactions we incorporate additionally data of small ncRNAs and protein-coding genes. Data for distinct types of small ncRNAs were taken from the public databases: DeepBase[1] and fRNAdb[2].

[4]http://deepbase.sysu.edu.cn/
[5]http://www.ncrna.org/frnadb/

2.2 Regulatory actions of lincRNAs

In this work lincRNAs were analyzed for an interaction with coding and noncoding transcripts. According to these interactions lincRNAs can be associated with distinct regulatory actions on the transcriptional level: (i) Regulation of their target alternative transcripts' expression and events and (ii) Generation and/or interference of activity of small ncRNAs. lincRNAs sequence complementary to protein-coding genes might be associated with the functional role: Influence on alternative splicing regulation (i). lincRNAs that are sequence similar to a small ncRNA might serve as precursor of small ncRNAs or activators/inhibitors for small ncRNAs' actions via base pairing (ii). An overview of functional roles analyzed in this work is shown in Figure 2.2. In the following we explain the methodology for the large scale analysis of lincRNA reconstructions and their regulatory actions.

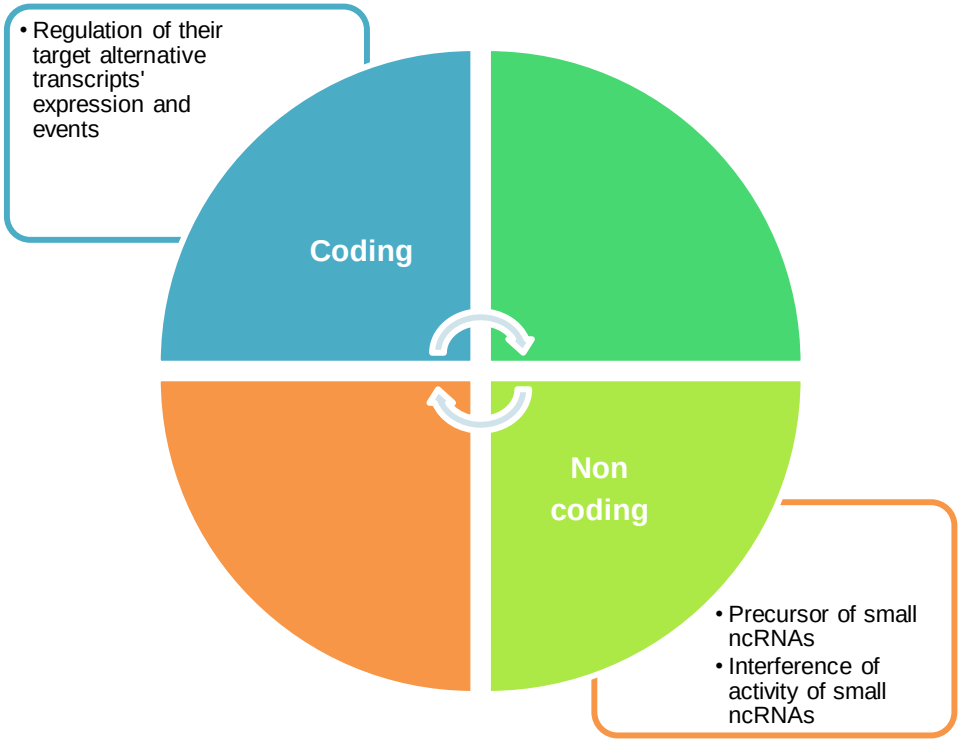


Figure 2.2: Overview of the functional ontology for lincRNAs interacting and regulating on the transcriptional level. lincRNAs are capable to interact with coding and noncoding transcripts. This leads to a variety of regulatory actions. We analyzed two functions in this work listed in rectangles. Each function is assigned to the according interaction with coding (blue frame) or noncoding (orange frame) RNA.

2.2.1 Identification of RNA:RNA interactions

Potential duplex structures between lincRNAs and pre-mRNA sequences were revealed by sequence similarity searches using the tool Blat [53] version 34 with default settings (-minScore=30, -minIdentity=90) . Each lincRNA was mapped onto the complete set of all full pre-mRNA protein-coding sequences to identify target sites. The annotated RefSeq coding transcripts (NCBI version 37, mm9) were used as gene model reference. lincRNAs that show significant sequence complementarity are used for further analysis. The regulatory effect of lincRNA:target pairs on alternate splice products was then compared to the relative expression of alternate splice variants.

An analogue analysis was run for lincRNAs and small ncRNA sequences to identify lincRNA:small ncRNA interactions (both strand orientations: sense and antisense interactions). This differentiation of the strand orientation is necessary to distinguish the noted scenarios linked to one lincRNA:small ncRNA interaction as addressed in the work of Wilusz et al. [9] and shown in Figure 2.2.

2.2.2 Quantitative expression of sequence regions

All sequence reads were mapped onto the mouse reference genome NCBI37, mm9 per cell line using the tool TopHat version v1.0.13 [35] in accordance to the protocols of Guttman and colleagues [1]. We applied default parameters modified by option 'g 1' for the best or unique hit. To quantify transcripts, we applied the RPKM (Reads Per Kilobase of exon model per Million mapped reads) estimations that were introduced for the quantitative comparison of expression levels [54]. Accordingly, we calculated the expression level from the number of reads per nucleotide utilizing SAM tools version 0.1.7-18 [55]. To calculate RPKM values of any sequence segment on the pre-mRNA, the accumulated number of reads within the segment was taken and normalized by the length and total number of mapped reads in the experiment. We considered the strand orientation of mapped reads for this calculation since Guttman and colleagues used a strand-specific library for RNA-seq. For the large scale analysis, we calculated the RPKM for the following sequence regions: origin, target sites of each lincRNA (sites at targets, either protein coding gene or small ncRNA) and coding exons/introns of RefSeq transcripts. We eliminated all target transcripts where the corresponding maximal expressed exon has a lower mean read coverage than 10 to filter insufficiently expressed regions as previously introduced for transcriptional units by Zhang et al. [56].

Comparison of expression levels of RNA-seq data to Microarrays

We verified the expression levels of our RNA-seq data in comparing the mean coverage of mouse pre-mRNAs against microarray probe intensities from comparable experiments. Since we analyzed RNA-seq data of mouse embryonic stem cells (V6.5 cells) [1] we selected corresponding microarray experiments from GEO (Gene Expression Omnibus) [57]. We chose the GEO accession GSE3231 [58] and three experiments GSM72802, GSM72804 and GSM72806 as microarray expression data. The mean over these experiments was taken as the expression level for a probe. The probes were assigned to transcripts. This comparison was conducted by a colleague of mine, Kerstin Haase [59].

2.2.3 Regulation of their target alternative transcripts' expression and events

Analysis workflow

The influence of lincRNAs on transcript variants was analyzed for each cell line separately in six steps (see Figure 2.3). Our analysis is based on lincRNA reconstructions provided by Guttman et al. [1]. The first part (steps 1-3) of this Figure describes the 'Scripture Walkthrough' as introduced by Guttman et al. [1] and the subsequent identification of ncRNAs. Briefly Guttman and colleagues mapped RNA-seq reads onto the mouse genome in a first step (step 1). Secondly, the transcriptome was reconstructed to identify the exon/intron structure of all coding and non-coding transcripts (step 2). Next, novel lincRNA transcripts were identified based on multiple filtering criteria (step 3). In our work (steps 4-6), lincRNAs were mapped on protein-coding RefSeq pre-mRNA sequences to identify lincRNA:coding transcript duplexes (step 4). The antisense part is termed as antisense lincRNA and the sense counterpart of the protein-coding transcript is termed target. In addition, we calculated features such as GO enrichment and GC content for targets and unique target sites (step 5). Finally, we determined the influence of lincRNAs on splicing regulation by calculating the 'fraction of remaining expression' for exons, residing upstream and downstream of lincRNA target sites and predicting alternate events (step 6). These adjacent exons might be influenced in their expression or splice events by the presence of complementary lincRNAs.

To achieve this analysis workflow we set up a semi-automatic pipeline (mainly in the programming language Java). The illustrated workflow can be adapted to any RNA-seq data and ncRNA type. Pre-aligned RNA-seq reads, reference data sets for both ncRNA and protein-coding genes are in general required. As indicated by an arrow in Figure 2.3 annotated lncRNAs can be incorporated as ncRNA type, as well as our small ncRNA data sets.

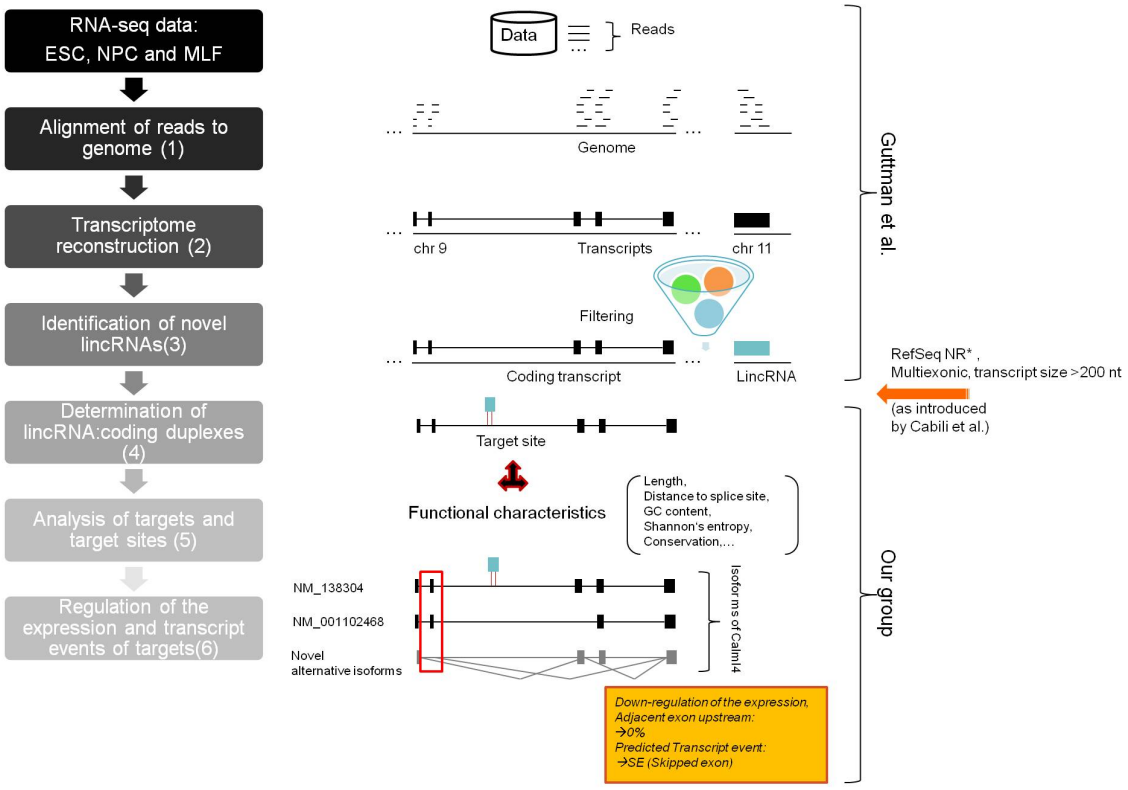


Figure 2.3: Analysis workflow for the reconstruction and functional analysis of expressed lincRNAs. Exons are shown as black boxes and introns as lines. The lincRNA target site is illustrated as light blue box. The workflow is split in the following six steps. (1)-(3) lincRNAs are reconstructed and (4) checked for sequence complementarity to protein-coding transcripts (antisense lincRNAs). (5) Functional characteristics of targets and target sites are examined. (6) Duplexes are analyzed in context of their regulatory action on alternative transcripts' expression and events. The arrow illustrates that not just novel lincRNAs can be analyzed with our workflow, but other noncoding references as well.

Characteristics of lincRNA targets and target sites

Distinct characteristics of 'intronic' lincRNA target sites and targets were investigated. The resulting distributions of target sites were compared to sequence regions up,- and downstream of target sites of 100 nt (according to the median length of target sites). To test the NULL-hypothesis we additionally selected 1000 random sampled introns. The list of these characteristics can be optionally modified and extended. The features are listed in the following, targets:

1. GO annotation and KEGG pathway enrichment: We used the web-based tool IGEPROS (Integrated GEne and PROtein annotation Server) tool[6] for gene sets

[6]http://www.biosino.org/iGepros/Gene/index.jsp

to identify enrichment. IGEPROS was applied to the data set of target genes with default parameters (model organism: mouse and query term: gene name). The GO enrichment analysis of IGEPROS identifies e.g. common GO terms among target genes. Hypergeometric distribution with 0.05 as P-value threshold is used for enrichment of GO terms. The pathway enrichment analysis is similar to GO enrichment.

2. Half-Lives: The half-lives of 19,977 non-redundant genes of mouse embryonic stem cells were obtained from the work of Sharova et al. [60]. This feature was analyzed just for ESC since half-lives are not available for the other cell lines [60]. Each target gene was assigned to half life in hours. Sharova and colleagues reported that mRNA species with short half-lives are enriched in genes with regulatory functions. Shorter half-lives among target genes in comparison to non-targeted genes (background) might support the presence of lincRNAs:gene pairs. Hence, the resulting distribution of targeted genes was compared to non-targeted genes (background) and the Wilcoxon rank sum test was applied.

, target sites:

1. Length of target sites in nucleotides [nt].

2. Expression of target sites quantified as [RPKM].

3. Distance from target site to adjacent splice sites in nucleotides [nt].

4. GC content: The GC content was calculated as the percentage of G and C nucleotides for each sequence. A sequence consisting exclusively out of G and C nucleotides has a GC content of 100%.

5. Shannon's entropy: We used the Shannon's entropy for the measurement of uncertainty: $H = \sum_{i=0}^{n} p_i \log_2 p_i$. Considering one DNA sequence the probability p_i corresponds to the proportion of one nucleotide with $i = A\,T\,G\,C$ and n = 4. The entropy ranges from 0 to $\log_2 n$. The maximal entropy would be observed for equally distributed nucleotides ($p_i = 0.25$) as expected for random sampled sequences.

6. Conservation: For the analysis of the conservation, we selected the conservation track from the UCSC table browser[7] to calculate the conservation per nucleotide of a certain sequence region. The conservation ranges from 0 (no conservation) to 1 (complete conservation among the 30 organisms).

7. Clustering of sequences: We used CD-HIT (Cluster Database at High Identity with Tolerance) [61] version 4.0 beta to cluster target sites according to their sequence

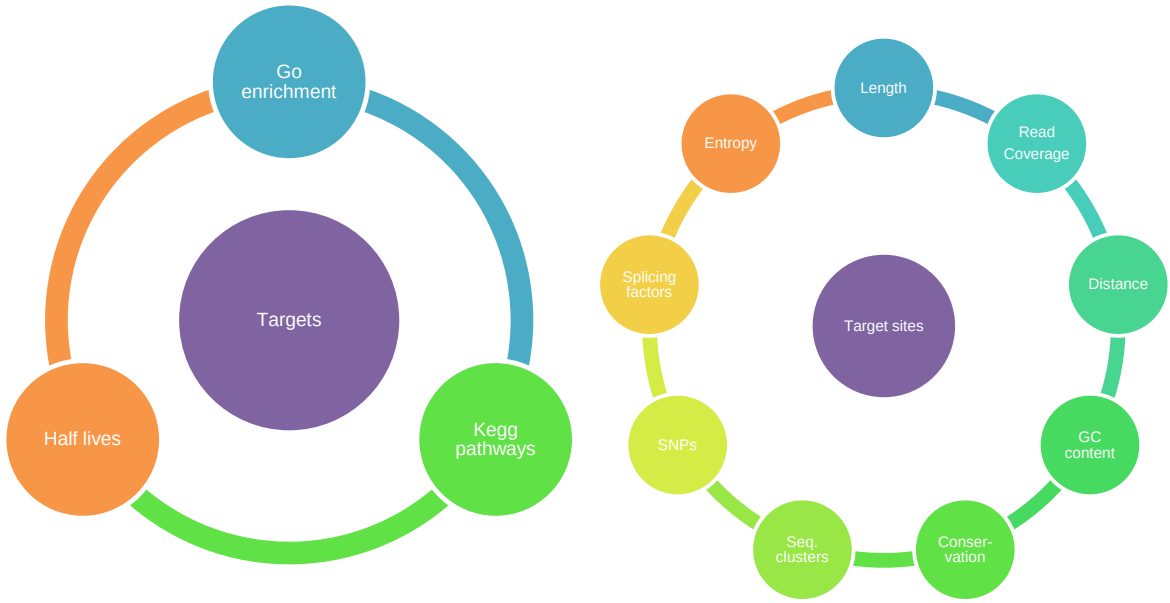
[7] http://genome.ucsc.edu/cgi-bin/hgTables?command=start

similarity (default parameters for DNA sequences). Target sites of one cluster might share common antisense elements or conserved binding motifs. Therefore we calculated the frequency of target sites per cluster to determine the divergence of sequences.

8. SNPs: We called SNPs using samtools [55] version 0.1.7-18 with the previously mapped RNA-seq reads of Guttman et al. [1]. The frequencies of these potential SNPs were calculated for each sequence and normalized by sequence length.

9. SF (Splicing Factor)-binding motif prediction: We used the tool SFMap [62, 63] to detect distinct SF binding sites. SFs, such as SR (Serine/arginine Rich proteins) are involved in alternative splicing regulation. The binding sites of these factors may be blocked by the presence of lincRNA:mRNA duplexes. SF-binding motifs incorporated in the prediction can be found in the publication of Akerman et al. [63] and in Appendix A (see Table A.4). The normalized frequencies of these motifs were calculated for each sequence region.

An overview of all characteristics or features is shown in Figure 2.4 for targets and target sites.



(a) (b)

Figure 2.4: Overview of all characteristics of targets and target sites analyzed in this work. Each characteristic or feature is shown in an outer circle, separated in (a) targets. (b) target sites (inner circle).

Fraction of remaining expression of adjacent coding exons

Determining the expression of exons individually allows for the detection of mixtures of expressed splice variants within a single sample. Here, we measure the ratio of splice variants for each exon. If one or more splice variants of a protein-coding gene are present in the same sample, the expression of a skipped exon will be significantly lower than the expression of the pre-mRNA and its other remaining exons. We calculated the quantitative expression as RPKM (Reads Per Kilobase of exon model per Million mapped reads) of an adjacent exon in relation to the maximal expressed exon as 'fraction of remaining expression'. Adjacent exons are exons residing upstream and downstream of 'intronic' lincRNA target sites. This measure indicates the regulatory change in expression levels via the functional influence of lincRNA target sites.

The significance of regulation was determined as follows: First, we randomly sampled the same amount of non-targeted introns as 'intronic' target sites. Secondly, the 'fraction of remaining expression' of adjacent exons, close to target sites (lincRNAs) was compared to adjacent exons, next to randomly sampled introns (background).

We observed a differential distribution of lincRNA target sites along the pre-mRNA. A preference of lincRNAs to bind at the 5' introns was found. This intron position might be crucial for accurate alternative splicing regulation. Thus we determined splicing regulation as the 'fraction of remaining expression' value in dependence on the intron position. Therefore, we generated a two dimensional matrix. The two dimensions are listed in the following:

1. Intron position: As intron positions of coding genes we chose the terminal introns 5' and 3' and internal ones (2-9, 5' to 3' direction).

2. Fraction of remaining expression: We used non-overlapping window sizes of ten percent ([0..10[,. . . ,[90..100[) for 'fraction of remaining expression' and the measure of splicing regulation.

In each entry of the matrix (intron position e.g. 1 and window of 'fraction of remaining expression' e.g. [0..10[) the frequency of both lincRNA target sites and randomly sampled introns was calculated. Next, the difference thereof was determined. Random sampling was repeated 100 times and results were averaged. An increased frequency of lincRNA target sites compared to randomly sampled introns in the first entry (1 and [0..10[) illustrates that the expression of adjacent coding exons is significantly down-regulated to a remaining expression level between 0% and $< 10\%$, reflecting the maximal affect of lincRNAs on down-regulation of expression levels. This down-regulation might be associated with splicing regulation and potential events, such as exon skipping or alternative

variants in the first/last exon.

Prediction of alternative transcript events

We used the tool MISO (Mixture-of-Isoforms) (version 0.1) for the prediction of 'alternative transcript events' based on RNA-seq data ('exon-centric' analysis) [23, 11] with pre-aligned reads per cell line. Alternative event annotations are incorporated in the tool MISO additionally expanding the RefSeq gene model [23, 11]. Target genes' events were derived for each lincRNA giving a detailed picture of alternative splicing regulation via complementary lincRNAs [23, 11]. Genes undergoing at least one event are indicative for the mode of regulation (category: 'Event' for simplification).

The percentage of genes assigned to the category 'Event' of targeted genes were compared to non-targeted genes (background). An increased percentage of targeted genes in comparison to non-targeted genes is supportive for an effect of lincRNAs on alternative transcript events. MISO is capable to predict eight distinct alternative transcript events as described in the work of Wang et al. [11]. The analysis was achieved for each of the eight events. The abbreviations for the events are listed in the following [11]:

1. SE: Skipped exon
2. RI: Retained intron
3. A5SS: Alternative 5' splice site
4. A3SS: Alternative 3' splice site
5. MXE: Mutually exclusive exon
6. AFE: Alternative first exon
7. ALE: Alternative last exon
8. TandemUTR: Tandem 3' UTR

2.2.4 Generation and/or interference of the activity of small ncRNAs

We further analyzed the interplay of lincRNAs with distinct types of small ncRNAs. To assess whether an interaction exists, we run sequence similarity searches per cell line and identified expressed sites. The sequence similarity search is explained in 2.2.1.

3 Results

3.1 Antisense lincRNAs are predominately targeting intronic regions

In order to determine the influence of lincRNAs on splicing regulation we used recently published RNA-seq data of Guttman et al. [1] from three cell lines: embryonic stem cells (ESC), neural progenitor cells (NPC), mouse lung fibroblasts (MLF) transcripts provided by Guttman and colleagues [1]. This data includes the transcriptome reconstructions of novel lincRNAs, available from the data of Guttman et al. [1].

Out of reconstructed transcripts, 1931 in ESC, 1266 (NPC) and 786 (MLF) are annotated as novel lincRNAs respectively (see Table 3.1). In the following, we refer to the sense protein-coding genes that may form a duplex with antisense lincRNAs as targets. In this context, we address the sites of the putative duplex structures as target sites. Antisense lincRNAs have a large number of targets and target sites, resulting in 149,171 potential duplexes in ESC, 9654 (NPC) and 7154 (MLF) in our data, listed in Table 3.1. The vast majority of target sites are indeed located within introns: 91% (135,295 of 149,171) in ESC, 93% (8990 of 9654) in NPC and 95% (6764 of 7154) in MLF (shown in Table 3.1). The distribution of target sites located within intronic and exonic regions of their targets is additionally shown in the Appendix A, Figure A.3. These 'intronic' antisense lincRNAs and target sites are selected for further analysis. 18% (344 of 1931) in ESC, 8% (98 of 1266) in NPC and 7% (52 of 786) in MLF of lincRNAs that are antisense to intronic regions target at least one transcript.

The increased numbers of transcripts and duplexes in ESC fit to some illustrated results in the publication of Guttman et al. [1]. Guttman and colleagues reported that for example 497 novel exons in ESC, 274 in NPC and 76 in MLF were found. Reasons remain unclear. The results of our work are mainly discussed and illustrated for novel lincRNAs expressed in the ESC cell line for simplification since we obtained similar results for the other cell lines, NPC and MLF.

Cell line	Frequency of lincRNA:coding duplexes in introns of total sites	Frequency of lincRNAs antisense to introns
ESC	91% (135295 of 149171)	18% (344 of 1931)
NPC	93% (8865 of 9522)	8% (97 of 1264)
MLF	95% (6764 of 7154)	7% (52 of 786)

Table 3.1: Data of reconstructed lincRNAs with significant expression that are annotated as novel lincRNAs according to Guttman et al. [1] for each cell line: ESC, NPC and MLF. The majority of RNA:RNA duplexes are located within introns (listed in the second column).

3.2 Complexity of multiple lincRNA:protein-coding RNA interactions

One of the first observations in our data of lincRNA:coding duplexes was that lincRNAs are capable to target multiple protein-coding genes resulting in various distinct duplexes per lincRNA in all cell lines. In addition to this capability of lincRNAs, protein-coding genes are obviously targeted by multiple distinct lincRNAs, as well.

A schematic illustration of this scenario is shown in Figure 3.1 in the first two parts. In the top part we illustrate that one lincRNA targets multiple genes. Distinct domains of a lincRNA might be crucial for the interactions. In the middle part the analogue illustration could be seen for protein-coding genes, targeted by multiple antisense lincRNAs. In the bottom part we indicate the probable target specificity of lincRNAs.

In our work we analyse the impact of lincRNA:coding duplexes on alternative splicing regulation of their target genes. Since lincRNAs are assumed to be involved in a variety of functions [15, 16, 9] just a subset of our data of lincRNAs may be involved in this specific regulatory action. The choice of the protein-coding RNA partner might be one of the crucial conditions for further regulatory actions. This subset e.g. might especially act in nuclear compartments since alternative splicing (constitutive and alternative) is reported to take place in the nucleus [64]. The distribution of the frequency of lincRNA:protein-coding RNA interactions in relation to both scenarios is explained in the following.

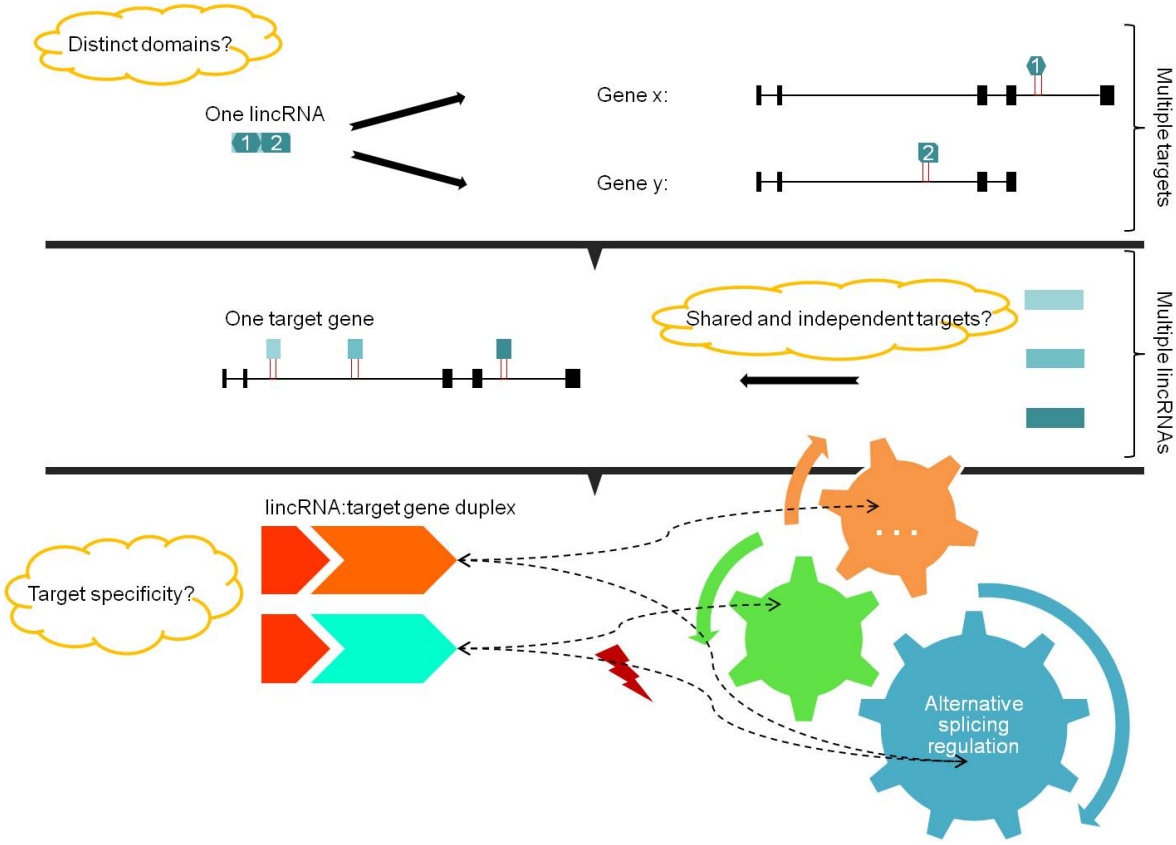
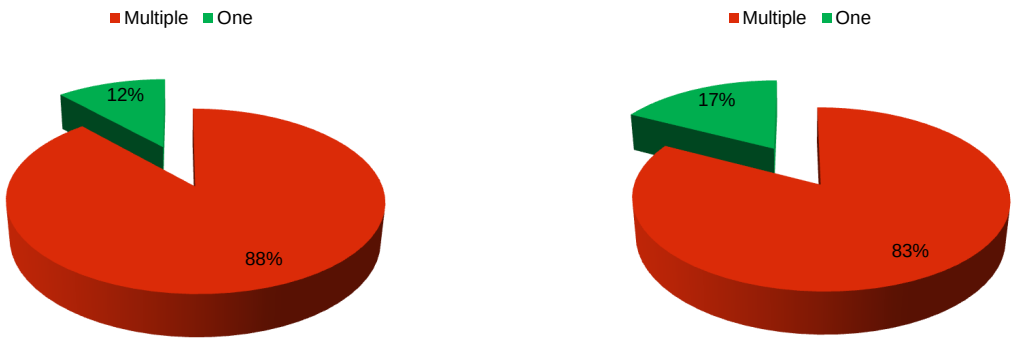


Figure 3.1: Scenario of the diversity of interactions of lincRNAs and coding transcripts. One lincRNA is indicated by a box (blue). One lincRNA might harbour distinct sequence motifs (marked by 1 ond 2, corresponding to 2 domains). Different shades of the colours correspond to distinct lincRNAs. The exons of a protein-coding gene are illustrated as boxes (black) and introns as lines. In the top part we illustrate that one lincRNA can target multiple protein-coding genes, resulting in distinct target sites (red lines). Domains of lincRNAs might be crucial for the proper target (written in the cloud). In the middle part we show that distinct lincRNAs are capable to target common and different protein-coding genes. The cloud raises the question whether lincRNAs tend to share a common set of targets. The bottom part summarizes the complexity of lincRNA:coding duplexes. A lincRNA is shown as red arrow and different coding genes as potential counterpart arrows in distinct colours. Since lincRNAs are reported to be target specific in the context of chromatin modifications the arrangement of the adequate pairs might be of importance for further regulatory actions. One action is shown in one wheel, bottom part (right side). In dependence of the target choice of a lincRNA distinct regulatory actions might be introduced. In this thesis we primary focus on the regulatory action - Regulatory influence of lincRNAs on target alternative transcripts' expression and events.

More than 90% of predicted base-pairing duplexes between linc- and mRNAs are located at introns in all cell lines. Duplexes were identified using the tool Blat [53] for sequence similarity searches as explained in the methods. These 'intronic' lincRNAs are found to be capable to target multiple protein-coding genes. Furthermore, the predicted duplexes involve several lincRNAs targeting the same intron, indicating alternate splicing being controlled by multiple input signals.

88% (229 of 259) of these 'intronic' antisense lincRNAs have multiple distinct target transcripts in the ESC cell line (see Figure 3.2). We obtained similar results for the other cell lines respectively: 79% (NPC) and 87% (MLF). One lincRNA targets on average about 201 distinct coding transcripts and 152 genes in ESC. In total 30% (8355 of 27636) of all RefSeq coding transcripts are targeted by a lincRNA. This distribution is shown in Figure 3.2 for illustration purposes. 83% (6909 of 8355) of these expressed coding transcripts in ESC are targeted by more than one lincRNA (on average about 6 lincRNAs per coding transcript). The percentage of targeted transcripts is decreased in the cell lines NPC with 36% and MLF 20% in comparison to ESC.

Since the current lack of knowledge about sequence domains or binding motifs of lincRNAs we run sequence similarity searches using Blat [53] for the whole original sequence. Thereby we identify target sites of high sequence similarity for the whole stretch of a lincRNA. In section 3.7 we explain the results of some characteristics of target sites analyzed in our work, including sequence motifs. For example, the target choice might be dependant on binding motifs.



(a) (b)

Figure 3.2: Distribution of the frequencies of coding and noncoding transcripts and their interactions (shown for the ESC cell line). lincRNAs are capable to target multiple protein-coding genes. Likewise to ncRNAs, genes can be targeted by multiple distinct lincRNAs. In this Figure we show the frequencies for one(=1)/multiple(>1) transcripts for these two possibilities. (a) Frequency of lincRNAs targeting genes. The majority of lincRNAs target multiple protein-coding transcripts and genes. (b) Frequency of targets. Transcripts are most often targeted by multiple lincRNAs

3.3 RNA-seq expression of protein-coding genes is comparable to microarrays

Different technologies exist to date for differential gene expression analysis [33, 65]. Two well established technologies are for example RNA-sequencing and microrrays [33, 65]. Several studies provide detailed comparisons of the advantages and disadvantages [65], but this is not the aim of our analysis. The choice of the technology depends on the scientific issue someone is interested in. One of the main advantages of RNA-seq over microrrays is the detection of novel coding and noncoding transcripts, as achieved e.g. in the work of Guttman et al. [1]. Guttman and colleagues detected novel lincRNAs across three mouse cell lines using the Illumina RNA-sequencing technology and an ab initio transcriptome reconstruction approach (Scripture) [1]. Since we are interested in alternative splicing regulation of these novel lincRNA reconstructions [1], RNA-seq brings a further advantage: the detection and incorporation of novel splice variants (of lincRNAs' target genes in our work). Disregarding the differences in the technologies the expression values are reported to be correlated in other studies [1]. Guttman and colleagues found a significant correlation in their Illumina RNA-seq data (with Affymetrix expression arrays) for the expression ranks of protein-coding genes [1].

Since we modified the alignment procedure (of reads - provided by Guttman et al. [1] - onto the mouse genome) we repeated the comparison of the expression values of anno-

tated protein-coding transcripts of RNA-seq with microarrays. In agreement to reported findings of Guttman and colleagues [1], a correlation is existent in the ESC cell line, shown in Figure 3.3 (Pearson correlation coefficient: 0.7). This supports our methodology of the quantification of expression levels of our data (as described in the methods section) as a reliable measure. This comparison was achieved in teamwork with a colleague of mine, Kerstin Haase [59].

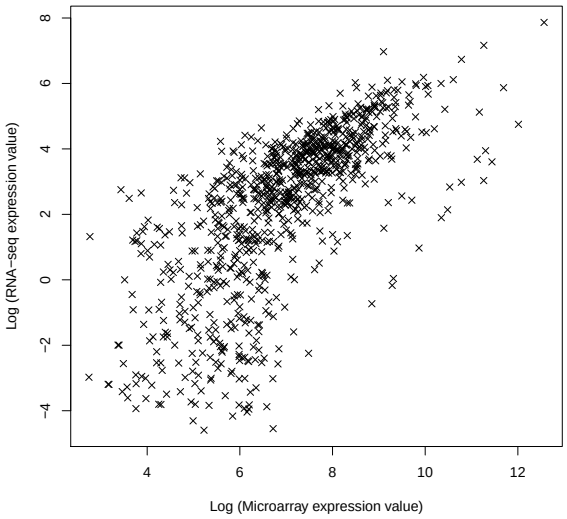


Figure 3.3: Comparison of the expression of 1000 randomly sampled RefSeq transcripts of RNA-seq with microarrays (shown for the ESC cell line). The expression of transcripts of RNA-seq is correlated with microarrays. Each data point describes the expression of one RefSeq transcript with the probe intensities of microarrays (as MAS5) on the x-axis and read coverage of RNA-seq (as mean read coverage) on the y-axis. The Pearson correlation coefficient is 0.7.

3.4 Expression of ncRNAs is correlated with their targets' expression

lincRNAs have multiple distinct target genes and sites (as previously explained in section 3.2) with on average 153 genes in ESC, 63 in NPC and 124 in MLF per lincRNA. The expression of a lincRNA exon correlates beyond doubt with the expression of its target genes in all cell lines. The expression level was calculated as RPKM. This correlation is shown for the ESC cell line in Figure 3.4 (Pearson correlation coefficient: 0.9). This significant correlation holds also for the other cell lines investigated.
We additionally proved whether the expression of randomly sampled introns is correlated

with the pre-mRNA of non-targeted protein-coding genes to avoid any bias in our result. We observed a slight correlation in our random model, see Appendix A Figure A.4 (Pearson correlation coefficient: 0.5) but the correlation of lincRNA target sites with a coefficient of 0.9 is significantly increased compared random sites with a coefficient of 0.5.

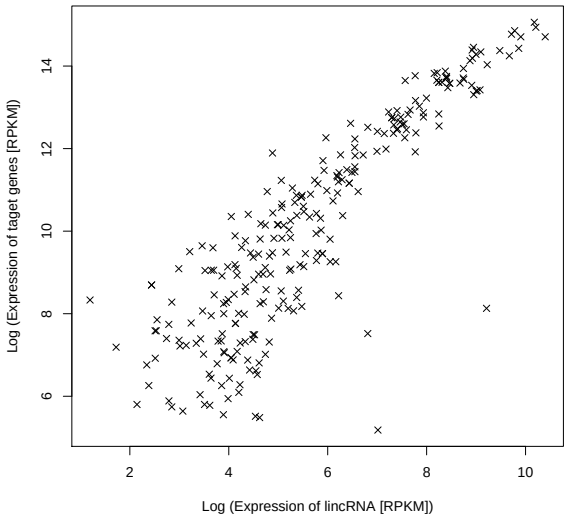
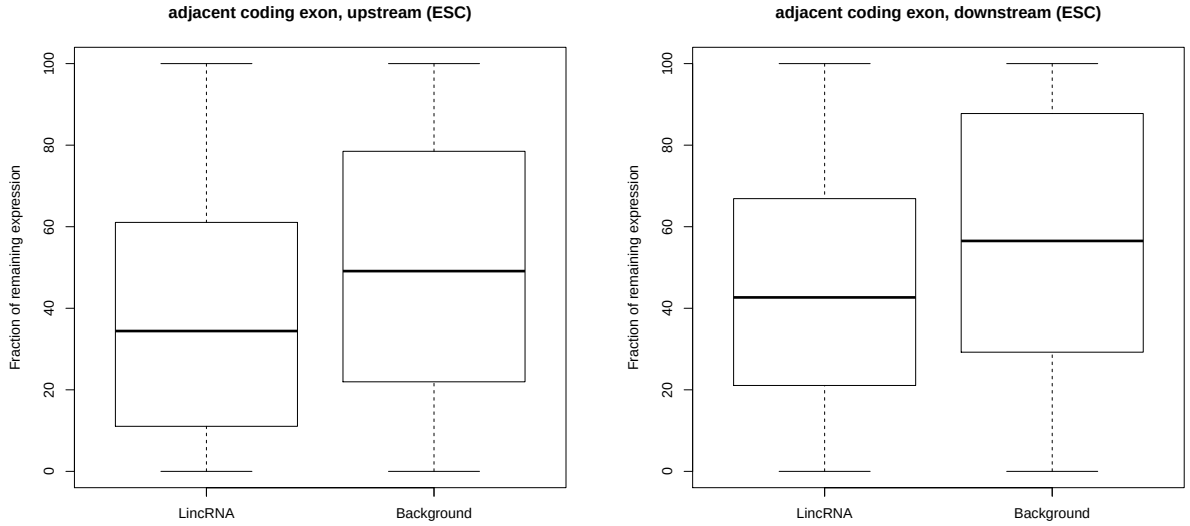


Figure 3.4: Comparison of the expression of lincRNAs with their target genes' expression. The expression as RPKM of lincRNAs shows a positive correlation with the expression of their target genes (shown for the ESC cell line, expression [RPKM] > 0). Each data point describes one lincRNA with the expression of a lincRNA on the x-axis and the expression of its target genes on the y-axis. The expression of a lincRNA is composed of origin and target sites. The expression of one target corresponds to the maximal expressed exon of the regarded transcript (pre-mRNA). The Pearson correlation coefficient is 0.9.

3.5 Exons, next to target sites, show a significant decrease in expression

To determine the influence of lincRNAs on splicing regulation, we investigated whether the expression of adjacent exons is changed. Therefore we calculated the expression as RPKM and determined the 'fraction of remaining expression' per adjacent coding exon (as explained in the methods section). We compared the resulting distribution of coding exons, close to 'intronic' target sites to exons, close to randomly sampled introns (background).

The expression of adjacent coding exons, especially of the exons upstream, is significantly down-regulated by nearby target sites. The 'fraction of remaining expressions' for ESC are shown in Figure 3.5 as boxplots for upstream and downstream exons (lincRNA and Background). We applied the Wilcoxon rank sum test for lincRNAs and background, upstream and downstream (p-value $<$ 2.2e-16) to determine the significance of the down-regulation of the expression levels via lincRNAs. Findings (e.g. enrichment of lincRNAs at 5' introns) have to be carefully considered in the calculation of the background distribution. For example the intron position has to be taken into account in random sampling of introns. Significance holds for an adjusted background distribution. The same distribution of the fractions of remaining expressions is shown for e.g. stratified intron positions (for lincRNAs and background) in ESC (see Figure A.5 in the Appendix A).



(a) (b)

Figure 3.5: Comparison of the fraction of remaining expression of adjacent exons of target sites with random introns (shown for the ESC cell line). The expression of adjacent coding exons of target sites (lincRNA) is significantly decreased. The comparison is shown for (a) upstream. (b) downstream exons. The p-value of the Wilcoxon rank sum test for upstream and downstream in ESC is $<$ 2.2e-16. The median is shown as black line in each boxplot.

3.6 Targets show functional characteristics, such as GO enrichment

We confirmed the hypothesis that common lincRNA targets should act in a common functional context. Thus we performed a GO enrichment analysis using the tool IGEPROS[1] to identify enriched GO annotations among target genes. This analysis results in 75 Biological Process (BP), 187 Molecular Function (MF) and 129 Cellular Component (CC) GO categories (ESC cell line and 0.05 as P-value threshold). For simplification we just show the 5 most significant GO categories for each of the three classes, listed in Table 3.2. Significance is assigned by a ranking of p-values. Each row in Table 3.2 describes one enriched GO annotation. Target genes are enriched in many GO annotations indicating a widespread functionality of targets. For example enrichment of target genes was observed in the cellular component (CC): nucleus (GO ID: GO:0005634 and Term: nucleus).

[1] http://www.biosino.org/iGepros/Gene/index.jsp

GO ID	GO Term	P-value
Biological Process (BP)		
GO:0043170	macromolecule metabolic process	$1.61e-42$
GO:0008104	protein localization	$1.60e-41$
GO:0044238	primary metabolic process	$8.11e-41$
GO:0044257	cellular protein catabolic process	$1.39e-32$
GO:0070727	cellular macromolecule localization	$5.28e-29$
Molecular Function (MF)		
GO:0000166	nucleotide binding	$5.68e-71$
GO:0001882	nucleoside binding	$1.89e-54$
GO:0005524	ATP binding	$7.00e-54$
GO:0032555	purine ribonucleotide binding	$1.17e-53$
GO:0030554	adenyl nucleotide binding	$1.41e-52$
Cellular Component (CC)		
GO:0005737	cytoplasm	$1.48e-146$
GO:0044424	intracellular part	$1.85e-65$
GO:0005634	nucleus	$7.57e-51$
GO:0005623	cell	$1.15e-46$
GO:0005622	intracellular	$2.11e-30$

Table 3.2: GO enrichment in target genes. lincRNA target genes are enriched in three GO classes: 375 (BP), 187 (MF) and 129 (CC). For illustration purposes just the Top 5 enriched GO annotations are listed for each class (ESC cell line). Top 5 correspond to the most significant enrichments (assigned by p-value). Each row describes one GO annotation. Columns represent (1-3): the id, name and p-value of one GO annotation. Target genes are enriched in splicing relevant GO annotations, such as cellular component: nucleus.

In addition we identified enriched KEGG pathways among target genes using IGEPROS[2] (ESC cell line and 0.05 as P-value threshold). Target genes are enriched in 82 distinct pathways (5 most significant KEGG pathway enrichments listed in Table 3.3). Enrichments include KEGG pathways associated with diseases, such as cancer (e.g. 05200: Pathways in cancer, $6.93e-06$). These results hold for the other cell lines NPC and MLF.

It was reported that there is a relation of half-life to regulatory function. Sharova et al. provide the information about half-lives of 19,977 genes across pluripotent and differentiating mouse embryonic stem cells [60]. mRNA species with short half-lives show an

[2]http://www.biosino.org/iGepros/Gene/index.jsp

KEGG ID	KEGG Term	P-value
04120	Ubiquitin mediated proteolysis	$3.79e-19$
01100	Metabolic pathways	$1.73e-136$
00310	Lysine degradation	$5.73e-10$
04110	Cell cycle	$1.13e-09$
00280	Valine, leucine and isoleucine degradation	$7.85e-08$

Table 3.3: KEGG enrichment in target genes. lincRNA target genes are enriched in 82 KEGG pathways. Top 5 enriched pathways are listed. Each row describes one KEGG pathway. Columns represent (1-3): the id, name and p-value of one KEGG pathway.

enrichment for genes with regulatory functions, such as transcription factors. Half-lives of targeted genes do not significantly differ from non-targeted genes (background), as shown as boxplots in Figure 3.6. As background we randomly sampled the same amount of introns as 'intronic' target sites.

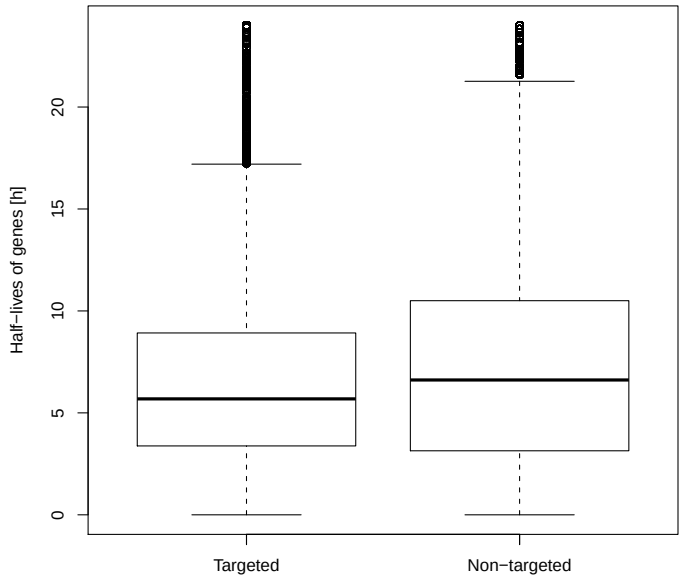


Figure 3.6: Comparison of half-lives of targeted genes with non-targeted genes (shown for the ESC cell line). The p-value of the Wilcoxon rank sum test in ESC is $< 2.2e\text{-}16$ ($2.201203e\text{-}132$). The median is shown as black line in each boxplot. (Median of half lives in hours: targeted 5.7 and non-targeted 6.6)

3.7 lincRNA target sites show a significant increase in GC content

We analyzed distinct characteristics of target sites as described in the methods. The results of this analysis are summarized in Table 3.4 for selected features and sequence regions. In general lincRNA target sites have a median lengths of 103 [nt] and expression of 0 [RPKM]. The expression is comparable to the observed expression in intronic regions of coding genes (background). These target sites are in close proximity to the adjacent splice sites (median distance to next splice site: 2271 [nt] in comparison to median intron length, targeted: 10760 nt and all RefSeq: 1355 nt). The increased targeted intron length is a result of e.g. the enrichment of target sites at first introns (see section 3.8). It is reported that first introns show increased intron sizes [66]. This is of interest since the distance to splice sites might be one criterion whether a lincRNA is involved in splicing regulation or not. We compared each feature (listed in Table 3.4) such as GC content of each target site to its surrounding up,- and downstream sequence regions on pre-mRNA. This comparison is additionally shown in the Appendix A for this feature (see Figure A.6). We applied the Wilcoxon rank sum test to assign the significance as p-value per feature. Features that differed significantly from the surrounding sequence are marked with a star in the first column in Table 3.4.

We observed significance only for the GC content. lincRNA target sites show an increased GC content compared to surrounding regions (p-value $<$ 2.2e-16). An increased GC content might reflect an increased stability of duplexes supporting the functionality of target sites. Target sites and surrounding sequences are not conserved. The median is equal to zero. The low conservation might be the result of the location of target sites at introns since the distribution fits to our background model (random sampled introns).

The C/D box snoRNAs are known to harbour conserved motifs, known as the C (UGAUGA) and D (CUGA) boxes residing next to the 5' and 3' ends. These motifs are essential for the regulatory action on alternative splicing [22]. lincRNAs might show conserved motifs for proper binding and regulation, as well. Hence the program CD-hit was applied to identify clusters of target sites with sequence similarity [61]. Target sites in a cluster might share common binding motifs and can be used for Motif Finder tools. CD-hit revealed 18742 clusters of 46900 unique target sites in ESC. The median frequency of sites is just 1.00. Since target sites in our data are distributed in many small clusters and are not conserved the detection of significant motifs remains difficult.

We are mainly interested in regulation of transcript variants. There might be a cor-

relation between target sites and SF binding sites. We used the tool SFMap [62, 63] for the prediction of motifs. The normalized frequency of motifs within target sites is not significantly increased compared to surrounding sequence regions and background. Significance was additionally not observed for SNPs and entropy.

Feature	Target sites	Up,-Downstream	Background
GC content [%]*	47.15	43.00	44.57
Shannon's entropy	1.96	1.95	1.97
Conservation	0.00	0.00	0.00
Sequence similarity	1.00	1.00	1.00
SNPs	0.00	0.00	0.00
Splicing factors	0.11	0.11	0.22

Table 3.4: List of the median of analyzed functional characteristics of target sites (for the ESC cell line). The median is shown for distinct features. The features are listed in the first column and distinct sequence regions are listed in the first row. As sequence regions we used the lincRNA target sites, surrounding up,- downstream regions and randomly sampled introns (Background). To determine whether a feature is significant we compared the distribution of lincRNA target sites with their surrounding sequence regions on pre-mRNA and applied the Wilcoxon rank sum test to determine the p-value. A feature with a significant p-value is marked with a star. A significant increase within target sites could be observed for the GC content. The GC content is more increased than expected by chance. The p-value of the GC content of the target sites with its surrounding sequence region is p-value $<$ 2.2e-16.

3.8 The 5' introns are most frequently targeted by ncRNAs and the according exons show the strongest down-regulation

The majority of distinct lincRNAs targeting protein-coding genes is antisense to the 5' introns of their counterpart transcripts (see Figure 3.7). We determined the total number of potential target sites and distinct coding/non-coding transcripts in relation to their intron position. The declining distribution of frequencies is shown in Figure 3.7 for distinct positions and the ESC cell line. As intron positions of coding genes we chose the terminal introns 5' and 3' and internal ones (2-9, 5'− >3'). As shown in Figure 3.7, 208 lincRNAs and 15236 target sites are associated with the 5' intron. These results show that nearby 5' exons of protein-coding transcripts might be influenced in the first place.

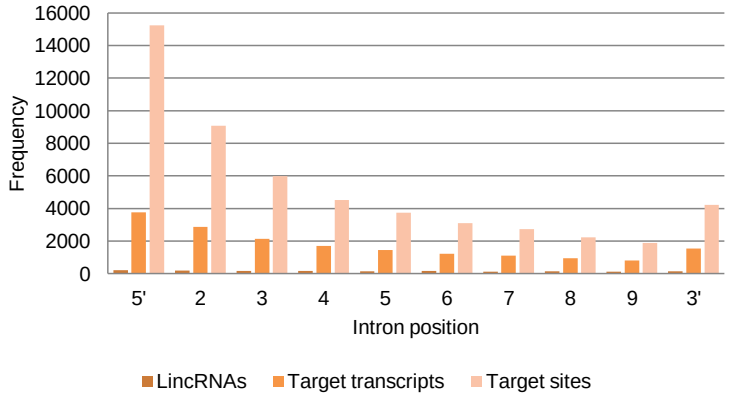


Figure 3.7: Distribution of lincRNAs along distinct intron positions of their targets (shown for the ESC cell line). lincRNAs, target transcripts and sites are most significantly enriched at 5' introns. The declining distribution of frequencies is shown for distinct intron positions. As intron position of coding genes we chose the terminal introns 5', 3' and internal ones (2-9, 5' − > 3'). lincRNAs are especially enriched at the 5' intron, indicated by the maximal frequency. The frequency of duplexes declines with increasing exon/intron position of protein-coding genes.

We further determined the influence of duplexes on splicing regulation in relation to their location on the pre-mRNA. First, we determined the frequency of lincRNAs and background (as explained in methods) for each intron position. Next, the frequencies of each intron position are split in non-overlapping windows of the 'fractions of remaining expressions' for adjacent coding exons. We additionally normalized the values by the total frequency of the regarded intron position in each window. The differences of lincRNAs and random sampled introns are shown per window and intron position in Figure 3.8. The expression of the 5' exons of protein-coding genes is most significantly down-regulated (Figure 3.8). For example in the ESC sample, 15236 lincRNA target sites are located within the 5' intron. On average 14047 randomly sampled 5' introns were found. The normalized frequencies of 5' lincRNA target sites and random introns are shown for the windows of the 'fractions of remaining expressions' in the supplement (5' exon, see Figure Appendix A A.7). The maximal difference could be observed in the window of [0..10[with (0.28 = 0.5425 - 0.2641). This window corresponds to a down-regulation of the expression of the 5' exon to a remaining expression of <10% if the corresponding lincRNAs are active.

Our results hold for the other cell lines and demonstrate that i) lincRNAs target the first exons more frequently as expected from a random distribution and ii) that the down-regulation in the 5' exons is strongest. Further this indicates that alternative transcript events such as alternative first exons or 5' splice sites as described in the work of Katz et al. [23] is caused by lincRNAs.

	Intron position									
Remaining expression [%]	5'	2	3	4	5	6	7	8	9	3'
Adjacent exon, upstream										
0..10	0.28	0.11	0.04	0.05	0.01	0.02	0.03	0.02	0.02	-0.03
10..20	0	0.07	0.04	0.02	0.03	0	0	0	0.01	-0.03
20..30	-0.01	0.04	0.03	0.02	0.01	0.04	0.02	0.02	0	-0.01
30..40	-0.04	0	0.04	0.03	0.05	0.04	0.02	0.01	0.01	0
40..50	-0.03	0	0.03	0	0.01	0	0.01	0.02	-0.01	0.02
50..60	-0.03	-0.02	0	0.01	0	0.02	0.01	0.03	0.02	0.02
60..70	-0.02	-0.01	-0.01	-0.02	-0.03	-0.03	-0.02	-0.01	0	0
70..80	-0.02	-0.04	-0.02	-0.02	0	-0.02	0	-0.02	-0.01	0.03
80..90	-0.01	-0.02	-0.03	-0.02	-0.04	-0.01	-0.04	-0.03	-0.03	0.01
90..100	-0.01	-0.02	-0.02	0	-0.01	-0.01	0	0	0.01	0.01
downstream										
0..10	0.07	0.03	0.01	0	-0.01	0	0.01	0	-0.01	-0.02
10..20	0.09	0.07	0.02	0.02	0.01	0.03	-0.01	-0.02	0.01	0.06
20..30	0.06	0.04	0.05	0.03	0.05	0.02	0.05	-0.01	0.01	0.02
30..40	0.01	0.05	0.05	0.02	0.03	0.02	0.06	0.04	0.01	0.02
40..50	0.03	0.04	0.03	0.01	0.02	0.02	0	0.02	0.03	0
50..60	0.02	0	0	0.01	-0.01	-0.01	0	0.05	0.02	-0.01
60..70	-0.01	-0.01	-0.01	0	0.02	-0.02	0	-0.04	-0.02	0
70..80	-0.02	-0.03	-0.01	0	-0.01	0	-0.03	0.01	-0.01	0.02
80..90	-0.01	-0.03	-0.02	-0.01	-0.01	-0.01	-0.02	0	-0.01	0.01
90..100	-0.02	0	0	0.03	0	0.03	-0.02	0.02	-0.01	0.01

Figure 3.8: Matrix of frequencies of target sites in dependence of the fractions of remaining expressions and intron position. lincRNAs, antisense to the 5' intron positions of their target genes have the most impact on splicing regulation of their adjacent exons, up,- and downstream in all samples. In this table we show the results for the cell line ESC. In the first column the non-overlapping window sizes of ten percent ([0..10[,...[90..100[) and in the first row the intron positions are listed. As intron positions of coding genes we chose the terminal introns 5', 3' and internal ones (2-9, 5' $->$ 3'). We calculated the normalized frequency of 'intronic' target sites and random sampled introns as described in the methods per intron position and window. The difference of lincRNAs and background is shown in each unit. The distribution is shown for upstream and downstream adjacent coding exons. Positive values, shown in red, correspond to an excess of lincRNA target sites compared to random introns. Negative values correspond to an excess of background. The maximal decrease in the expression can be observed for the adjacent exon, upstream with the target sites located within the 5' intron of target genes. The maximal difference is 0.28 with the maximal excess of lincRNAs in the bin [0..10[. Hence the 5' exon is significantly down-regulated to a remaining expression of <10 %.

3.9 Alternative transcript events AFE/ALE are most significantly fostered by the regulation of complementary lincRNAs

First 'alternative transcript events' were predicted for each of the eight included events using the tool MISO [23, 11], see methods. This prediction was run for each cell line. Based on resulting predictions, the frequency of genes undergoing an event was calculated for each event in the next step (category: 'Event' for simplification). The percentage of genes with an event was calculated for distinct sets of genes: namely target genes, non-targeted genes and all RefSeq genes. An increased percentage in the set of target genes undergoing an event compared to the other sets would be indicative. This analysis was performed for each of the eight 'alternative transcript events' separately. Non-targeted genes were selected from our previous curated background model (see above).

One important observation was that lincRNAs are enriched primary at 5' introns and additionally but not in the first place 3' (last) introns. We restricted our analysis on genes with target sites located at 5' (first) or 3' (last) introns to focus the analysis on nearby exons. The following events were merged: AFE with ALE and A5SS with A3SS. Merging of events was applied to reduce marginal noise e.g. from distinct annotation sources. The same restriction was applied to random sampling (background model).

Events such as for example AFE or ALE (AFE/ALE) are preferred. The distribution is shown in Figure 3.9 for the ESC cell line and the significant merged alternative transcript event: AFE/ALE. In the ESC cell line 40% (1430 of 3618) of target genes undergo AFE/ALE. This percentage is two-fold increased compared to non-targeted genes: 22% (754 of 3375). The results hold for the other cell lines: NPC (targeted: 43%, non-targeted: 26%) and MLF with (targeted: 43%, non-targeted: 23%). The percentage of target genes undergoing AFE/ALE is significantly increased in comparison to non-targeted genes. Thus there is evidence that lincRNAs especially antisense to 5' (first) introns are primary involved in alternative promoter usage. The other merged event: A5SS/A3SS is not significant with targeted: 23%, non-targeted: 27%. Just a minority of lincRNAs is distributed along internal introns. These lincRNAs seem not to be involved in alternative splicing regulation since other possible appropriate events, such as Skipped Exon and Mutually Exclusive Exon are not significant in all cell lines.

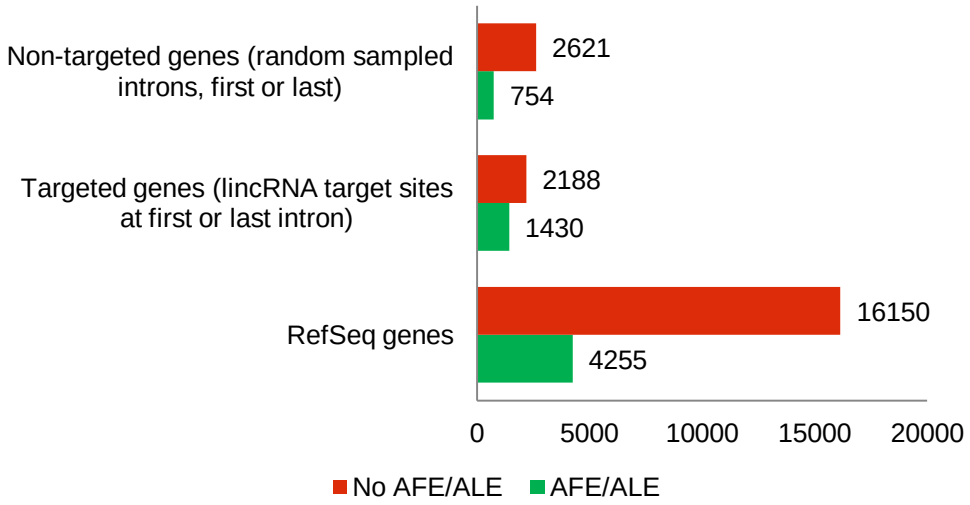


Figure 3.9: Distribution of the frequencies of genes undergoing an alternative transcript event. In this Figure the results are shown for AFE/ALE and the ESC cell line. We calculated the distribution for three distinct sets of genes: targeted, non-targeted genes and all RefSeq genes. We selected targets with target sites assembled at first and last introns of all sites for this analysis. The same restriction was set for random sampling. The frequencies of genes are split in undergoing an event and no event. This event is significantly fostered by lincRNAs. 40 % of target genes undergo this merged event. This percentage is 2-fold increased compared to random sampling.

We included our list of features of targets and target sites in this analysis. There is no significant trend obvious for a relation between a feature, such as e.g. GO category and the merged event AFE/ALE or 'fraction of remaining expression' value. Restricting our analysis to genes with the assignment of the GO enriched cellular component e.g.: nucleus for instance, increases the percentages of genes undergoing an event on the one hand (from 40 to 43% in ESC). On the other hand this increase is not significant since it is observed for affected (43%) as well as non-affected genes (27%) (ESC).

3.10 Case study of a lincRNA involved in alternative splicing regulation

An example of one expressed lincRNA in ESC is shown in Figure 3.10. This lincRNA originates from chromosome 11 and its target site is located within the second intron of the protein-coding gene Calml4 on chromosome 9. The target site in this example is 61 nt in length, 2368 nt apart from the exon upstream of isoform NM_138304 and has a significantly increased GC content of 49% (compared to surrounding sequence regions). further characteristics of the target site are shown in Table 3.5. Other features were investigated, but significance of target sites not observed. The expression of this exon upstream is down-regulated compared to pre-mRNA ('fraction of remaining expression'=0%). This exon is not spliced out in the annotated transcript NM_001102468, but the observation indicates the existence of novel isoforms where this down-regulated exon is skipped. Therefore we run an additional analysis using the tool MISO [23, 11] to predict alternative transcript events. This exon is predicted to undergo exon skipping.

Feature	Upstream	Target site	Downstream
GC content [%]	31.00	49.18	47.00
Entropy	1.86	1.99	1.99
Conservation	0.00	0.00	0.00
SNPs	0.00	0.00	0.00
Splicing factors	0.11	0.10	0.16

Table 3.5: Characteristics of one lincRNA target site regulating the alternative transcript event SE (Skipped Exon) predicted by MISO [23, 11]. The features are listed in the first column and distinct sequence regions are listed in the first row. As sequence regions we used the lincRNA target site and surrounding up,- downstream regions.

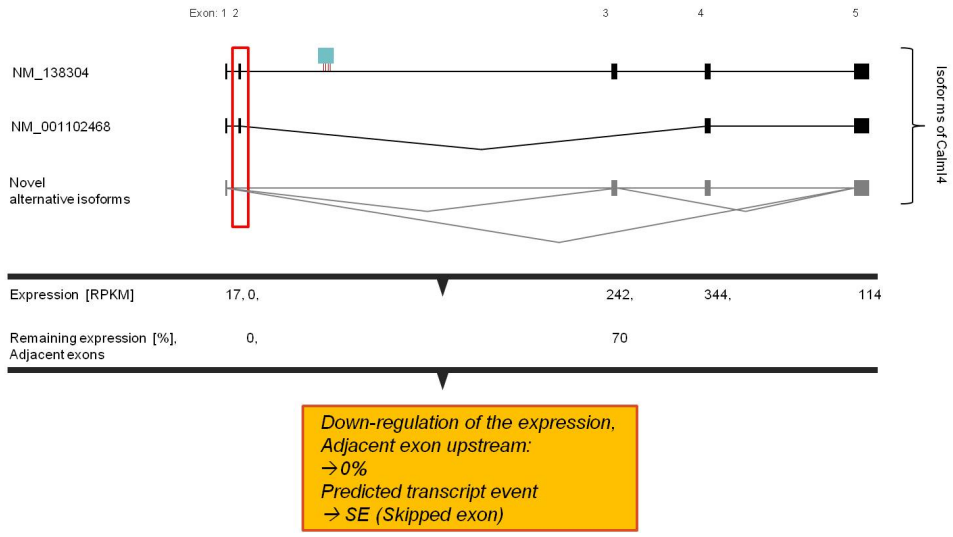


Figure 3.10: Schematic illustration of lincRNAs influencing target alternative transcripts' expression and events. The lincRNA target site is shown as light blue box. The exons of protein-coding gene Calml4 are shown as black boxes and introns as lines. Exon 1 corresponds to the 5' coding exon of a gene and exons 2,-3 to the adjacent exons, up,-downstream of the target site. The expression is shown per coding exon in the bottom part as [RPKM] and fraction of remaining expression [%]. The lincRNA is transcribed from a different genomic location than its target site, namely from chr11. As found by sequence complementarity searches, the lincRNA could form a putative duplex structure with intron of Calml4. From the experimental expression levels, it can be seen that the known splice variants NM_138304 and NM_001102468 could be found in the sample. The second exon corresponds to the adjacent exon, upstream. This adjacent exon is down-regulated via the influence of the lincRNA target site to a remaining expression level of zero and 0%. This suggests a novel splice variant where the exon is skipped and most important that lincRNA fosters the alternative splicing of the adjacent exon and alternative transcript events such as exon skipping. This event and the novel variant are confirmed by the tool MISO [23, 11]. This tool is as already mentioned capable to detect distinct alternative transcript events including (Skipped Exon) SE.

3.11 Illustration of the complexity of lincRNA:small ncRNA interactions

There is indication that a lincRNA could be sequence related to distinct targets, protein-coding transcripts and small ncRNAs. An interaction between lincRNAs and small ncRNAs could be associated with several functions. A few lincRNAs are reported to yield as precursor for small ncRNAs [9]. Besides these observations, some lincRNAs are suggested to interfere the activity of small ncRNAs (like miRNAs) by base pairing [9]. The complexity of distinct scenarios of the interaction of lincRNAs with small ncRNAs is shown in Figure 3.11.

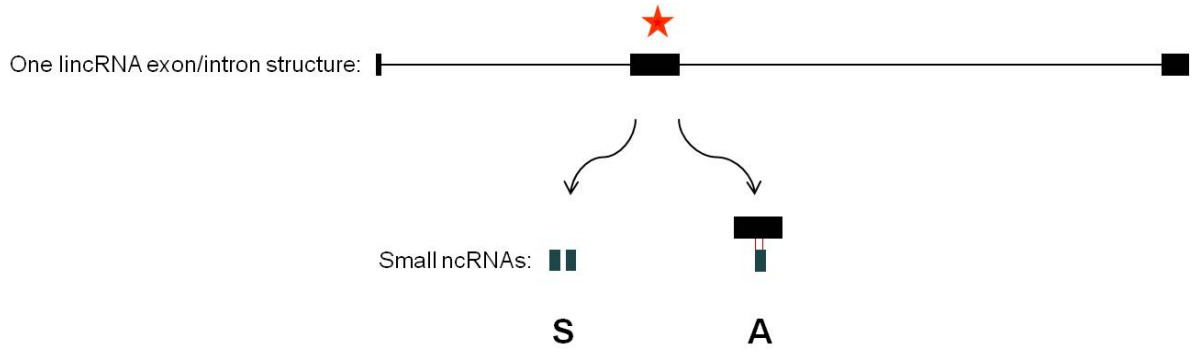


Figure 3.11: Illustration of the complexity of lincRNAs' regulatory actions. In the top part we illustrate one lincRNA exon/intron like structure. The exons of a lincRNA are illustrated as boxes and introns as lines (black). We go into detail of one exon of a lincRNA (marked with a red filled star) since we analyzed each part of a lincRNA in context of regulatory actions in this work. Sense (S) : lincRNAs are assumed to yield as precursor of a small ncRNA. Antisense (A): Antisense lincRNAs could inhibit a small ncRNA's regulation via base pairing. These two scenarios are indicated by a directed arrow.

3.12 A small fraction of lincRNAs potentially interplay with small ncRNAs

We run sequence similarity searches to determine the interaction of long with small ncRNAs as explained in the methods. Different database tracks corresponding to distinct types of small ncRNAs from two databases DeepBase[3] and fRNAdb[4] were included for this analysis. 9% (179 of 1931) of lincRNA loci in ESC, 5% (62 of 1264) in NPC and 4% (31 of 786) in MLF show sequence similarity to at least one small ncRNA included in DeepBase or fRNAdb. The percentage of interacting lincRNAs is shown for each cell line in Figure 3.12).

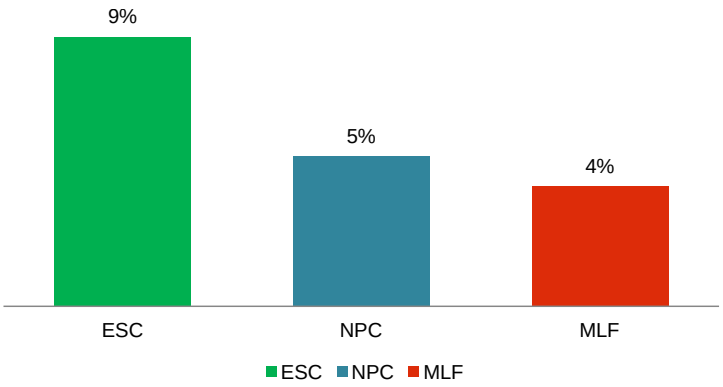


Figure 3.12: Distribution of the frequencies of lincRNAs interacting with at least one small ncRNA included in DeepBase or fRNAdb. The frequencies are shown as percentages for each cell line separately: ESC, NPC and MLF (bars in distinct colours: green, blue and red). The percentage is maximal in the ESC cell line with 9%.

[3]http://deepbase.sysu.edu.cn/
[4]http://www.ncrna.org/frnadb/

A detailed list of distinct statistics for interacting and expressed lincRNAs, subdivided in sense (S) and antisense (A) interaction, is given in Table 3.6. The table for lincRNAs irrespective of expression is additionally shown in the Appendix A (see Table A.5) for completeness. Sense/antisense was assigned according to the strand orientation reported by the sequence similarity search using Blat [53]. Each row describes the statistics for one type of small ncRNA, with small ncRNAs of DeepBase[1] and fRNAdb[2]. lincRNAs seem to interplay sense with a distinct set of small ncRNAs compared to antisense.

For example 103 sense lincRNAs tend to have sequence similarity with rasRNAs[1] and 48 with snoRNAs[2] in the first place. This is in contrast to lincRNAs sequence complementary to small ncRNAs (antisense). Just 48 lincRNAs show complementary to rasRNAs[1] and 44 to snoRNAs[2]. In the case of pasRNAs[1], easRNAs[1] and rasRNAs[1] the sets obviously have the strongest discrepancies.

None of lincRNAs interact with snRNAs of fRNAdb, nor with nasRNAs of DeepBase (excluded in Table 3.6). It also has to be noted that e.g. the frequency per lincRNA varies across small ncRNA types. The average frequency of small ncRNAs ranges from 1 to 35. It has to be noted that the statistics of the ncRNA types: easRNAs and rasRNAs have to be carefully considered.

For example the relatively high number of lincRNAs interacting with easRNAs might be just a result of the terminology. We used in our data the exons of lincRNAs and easRNAs (exon-associated small RNAs) overlap according to the methodology applied for the detection of these small ncRNAs by Yang et al. with exons [67]. The sequence similarity might be the result of the underlying sequence compositions of the exonic regions of lincRNA and easRNAs. Hence these lincRNA:easRNA interactions are of minor importance. The interaction of lincRNAs with miRNAs and snoRNAs is more interesting. For example in total 180 lincRNAs are sequence similar to at least one snoRNA of fRNAdb. Thereof 48 and 44 are assigned to sense and antisense transcript orientations. A lincRNA has on average one, rather than multiple, small ncRNA interaction partner.

small ncRNA type	Frequency of lincRNAs		Average frequency of small ncRNAs	
	S	A	S	A
snoRNA[1]	1	1	1	1
miRNA[1]	26	25	2	2
pasRNA[1]	1	22	1	3
easRNA[1]	82	48	7	5
rasRNA[1]	103	48	28	35
miRNA[2]	7	-	1	-
piRNA[2]	17	20	2	2
snoRNA[2]	48	44	1	1

Table 3.6: Statistics of the relation between our lincRNA data set with distinct types of small ncRNAs. The number of expressed lincRNAs with sequence similarity to a small ncRNA type are listed in this table. One lincRNA can interact with multiple distinct types of small ncRNAs. Each row describes the statistics of the interaction of lincRNAs (ESC) with one type of small ncRNA. Column 1 lists the distinct types of small ncRNAs, with small ncRNAs of DeepBase[1] and fRNAdb[2]. Columns 2-4 show the statistics of lincRNA:one type of small ncRNAs interactions: (2) the frequency of lincRNAs with sequence similarity to at least one small ncRNA (expressed interaction sites), (3) average frequency of small ncRNAs a lincRNA is interacting with. The frequencies are separated in sense: S and antisense: A interactions (strand orientation + and -).

3.13 Adjacent coding exons of interacting lincRNAs (with small ncRNAs) retain their decreased expression levels

We are further interested whether the lincRNA or the generated small ncRNA is the regulator of alternative splicing. A lincRNA serving as precursor (i) could additionally regulate alternative splicing regulation of a target gene (ii) as one possible scenario. Contrary such a lincRNA might not be directly involved in this second specific function (ii), but rather indirectly in the sense that generated small ncRNA takes the function over.

To gain further evidence about these complex scenarios we proceeded with our work in context of lincRNAs and their influence on alternative transcripts' expression and events. We already reported that 344 lincRNAs are antisense to introns of protein-

coding genes in ESC (see section 3.1). 259 (of 344) lincRNAs are expressed and 53 (of 259) are sequence similar sense/antisense to a snoRNA of fRNAdb for instance. We observed as one of the most promising findings that adjacent exons, close to lincRNA target sites, show less expression than expected by chance. To verify the previously explained scenarios we compared the 'fractions of remaining expressions' (measure for the change in expression) of interacting with non-interacting lincRNAs for these snoRNAs. There is no difference apparent in all cell lines. The median for interacting lincRNAs is 34.15 and non-interacting: 34.61 in the ESC cell line. The same holds for snoRNAs of DeepBase.

These observations indicate that one lincRNA could interact with both, small ncRNAs and protein-coding genes (i) the influence on alternative target genes' expression remains unchanged indicating that the lincRNA itself (not the small ncRNA) might maintain its role as alternative splicing regulator (ii). The regulatory effect of snoRNAs on alternative splicing regulation was examined by Kerstin Haase in her diploma thesis [59]. snoRNAs show similar key results to lincRNAs: (i) Enrichment of target sites at introns, especially 5' introns (ii) Downregulation of the expression of nearby exons (iii) and effect on alternative transcript events AFE/ALE.

3.14 Case study of one lincRNA as potential precursor for a miRNA

A lincRNA yielding as precursor might overlap sense with an annotated small ncRNA. Further characteristics might be an increased expression or conservation at the site of the generated small ncRNA (interaction site) among others. Hence we compared these two features of an interaction site with surrounding sequence regions, up,- and downstream. At large scale we could not observe any significant increase within interaction sites.

For example the lincRNA loci chr7:149762733-149764040 shows sequence similarity (sense) to the miRNA mir675 included in fRNAdb (miRNA track: pre-miRNA and mature miRNA). The pre-miRNA is e.g. 84 nt in length and has a 100% match on the lincRNA loci. According to this identical match there is indication that this lincRNA yields as precursor. It was particularly shown that this microRNA miR-675 is processed from the first nc-exon of lncRNA H19. Obviously this specific miRNA is additionally generated by other lncRNA transcripts besides H19. The first nc-exon criteria for the processing of miRNAs, but this is just for 1 of 8 interacting lncRNAs observed in our data. Interestingly the lincRNA loci is not sequence complementary to any protein-coding transcript in our data. Hence this lincRNA is according to our analysis workflow not capable to

be associated with the regulatory action - Regulatory influence of lincRNAs on target alternative transcripts' expression and events.

4 Discussion

4.1 Regulatory influence of lincRNAs on target alternative transcripts' expression and events

lincRNAs are assigned to a variety of distinct functional roles in recent studies [15, 16, 9]. In this work, we present evidence for the role of lincRNAs in alternate splicing. ncRNAs are candidates for the specific selection of protein-isoforms. In our analysis, RNA-seq data of Guttman et al. [1] including the transcriptome reconstruction of mouse lincRNAs in three cell lines, embryonic stem cells (ESC), neural progenitor cells (NPC) and mouse lung fibroblasts (MLF) for our large scale analysis were used.

The results show that a subset of lincRNAs is antisense to protein-coding genes, potentially forming RNA:RNA duplexes. These antisense lincRNAs target a large number of protein-coding genes. The majority of duplexes is located almost exclusively within intronic regions, with 91% in ESC-, 93% in NPC- and 95% in MLF- cells respectively. We found a positive correlation between the expression of a lincRNA and its protein-coding targets' expression levels. We could show that protein-coding exons next to potential antisense and intronic lincRNAs sites are significantly less expressed compared to exons in close proximity to random sampled introns. This finding holds for all cell lines investigated. Most of these duplexes are notably assembled on the 5' part of protein-coding genes and 5' exons show the strongest decrease in expression.

Not just a clear correlation of linc- and mRNAs duplexes and down-regulation of adjacent exons was detected but splice events, especially the event: AFE/ALE. We are aware of some remaining limitations in predicting alternative transcript events. A fraction of genes is targeted by lincRNAs, but not predicted to undergo an event. Some genes are not targeted by a lincRNA, but are assigned to an event in our background model. These interacting pairs can be used for further investigation. It is unclear to date whether all duplexes must result in an alternative splicing event and what other types of small ncRNA participate.

Our findings are confirmed by experimental data published recently. Recently a sin-

gle case of a RNA:RNA duplex influencing the regulation of alternative splicing has been described [20]. The example is the human long ncRNA Saf, sequence complementary to the 5' intron of its target and 521 nt apart from the 5' exon [20]. Alternative splice variants of Fas could be explained by the regulation of Saf [20]. Furthermore Saf is expressed in several tissues such as heart and cancer cell lines including Jurkat (acute T-cell leukemia) [20]. Yan and colleagues explicitly show that Saf over-expression affects the expression of Fas isoforms [20].

Until now, the impact of lincRNAs on alternative splicing was just shown in specific studies and for single ncRNAs, but not on a large scale [19, 20]. Getting a more comprehensive insight into splicing regulation in context of RNA:RNA duplexes is of importance since the majority of eukaryotic genes undergo alternative splicing events and mis-regulation of splicing is associated with diseases, such as cancer [10, 49]. As an example the lincRNA HOTAIR is frequently altered in human cancers [27]. Although the causative effect of lincRNAs on the malignant properties of cancer cells is unclear, gaining knowledge about the regulatory effect on alternate isoforms might be of medical importance for cancer treatment [27].

Altogether, our analysis indicates a wide and general role of transcript variant regulation by lincRNAs and opens new perspectives for the systematic investigation of the conditional expression of isoforms and their cellular functions. The expression level of coding exons appears to be down-regulated by the influence of near binding lincRNA target sites. Alternative splice variants of protein-coding genes can be explained by such RNA:RNA duplexes. The systematic analysis of published deep-sequencing data as well as the design of hypothesis-driven experiments is needed to get a more detailed insight into the specific role of lnRNAs to control protein isoforms in various tissues and various conditions of differentiation.

The conditional control of protein isoforms needs to be specific and dynamic. The traditional view of interacting proteins such as enhancers, silencers or other factors involved in the splicing machinery is not sufficiently convincing to rationalize the rather precise tuning of alternate variants. The epigenetic explanation based on chromatin and histone modifications as regulators of alternative splicing appears to be plausible, but how histone modifications may affect the transcription of an individual exon remains an unsolved riddle. In contrast, the control by non-coding RNAs offers not only an explanation for the specificity of interaction it also allows the control of many targets by a single ncRNA similar to the miRNA translational suppression. To show experimentally that not the mRNA precursor itself, but the mRNA:ncRNA duplex is processed by the splicing machinery is a yet missing piece in the puzzle of alternate splice regulation. Our

data provide evidence that non-coding RNAs are candidates to control alternate splicing. Improved methods and detailed interpretation of the results are needed to monitor the presence of lncRNAs and their impact on the selection of alternate protein isoforms.

4.2 Generation and/or interference of activity of small ncRNAs

Besides the capability of lincRNAs to interact with protein-coding genes, it is reported that these lincRNAs further interact with noncoding RNA [9]. An interaction between a long and small ncRNA is reported to be associated with distinct functions [9]. Hence, we additionally determined the interplay of our lincRNA data with distinct types of small ncRNAs using the lincRNA reconstructions of Guttman et al. [1].

In this work, we found that many lincRNAs interplay with at least one small ncRNA: 9% (179 of 1931) in ESC, 5% (62 of 1264) in NPC and 4% (31 of 786) in MLF. lincRNAs are obviously capable to interact with distinct types of small ncRNA, such as miRNAs and snoRNAs.

There might be even a relation between type and function. This potential functionality needs further experimental examination. It is still another open question whether the lincRNA itself or the (sense) generated small ncRNA is the crucial ncRNA for alternative splicing regulation. There is a marginal evidence in our data that the lincRNA itself is responsible for the down-regulation of expression levels of their targets. The effect of lincRNAs on target isoforms' expression and events is comparable of interacting and non-interacting lincRNAs.

Part II

SNP analysis of Restless legs syndrome

5 Introduction

5.1 Motivation

Restless legs syndrome (RLS) is assumed as one of the most common neurological disorders according to Allen et al. An urge to move the legs and odd sensations in resting situations are reported symptoms of RLS. The identification of genetic loci and associated variants for the RLS phenotype is of importance (reviewed by Allen et al. [68]).

Non-synonymous single nucleotide polymorphisms (SNPs) in coding regions are crucial since the SNP in a codon alters the amino acid of the protein product. This change of an amino acid can further lead to an effect on protein function and inherited diseases. Two established approaches for the identification of disease related genetic variants are genome-wide association study (GWAS) and exome sequencing. The GWAS can just detect SNPs that are common in the population. Contrary the exome sequencing technique affords the identification of novel SNPs.

We performed an analysis for the identification and interpretation of novel RLS associated variants based on exome sequencing data. The analysis can be split in three tasks: (i) Exome sequencing (ii) Raw data processing and (iii) Qualitative modelling. This study was achieved in cooperation with Dr. Volker Stümpflen and Daniel Ellwanger (BIS group, Helmholtz Zentrum München) and Prof. Juliane Winkelmann (Neurologische Klinik und Poliklinik, Klinikum rechts der Isar). Exome sequencing data of two pedigrees known to be affected by RLS was provided by Prof. Juliane Winkelmann (NKP). The exome sequencing data was processed using an implemented NGS pipeline by our group (NGS group, Helmholtz Zentrum München) to provide lists of candidate genes and their assigned variants. Interesting candidates were modelled by a colleague of mine Daniel Ellwanger (BIS).

We found three candidate genes with novel (not in dbSNP134) and non-synonymous (deleterious) SNPs common in all eleven exomes of both pedigrees (ii Raw data processing). These three most interesting genes were primary examined by Daniel Ellwanger in detail (iii Qualitative modelling).

5.2 RLS

5.2.1 Definition and symptoms

RLS (Restless Legs Syndrome) or Ekbom's syndrome is a neurological sensory-motor disorder. The disorder is characterized by a need for movement and unpleasant sensations in the first place (reviewed by e.g. [69, 70, 71]). Symptoms similar to RLS were described for the first time by Thomas Willis, 1672 [72]. The term 'Restless legs' itself was introduced and published as doctoral thesis by the swedish neurologist Karl-Axel Ekbom, 1945 [73]. In this thesis RLS features were described, e.g. a prevalence of at least 5% and a dominant mode of inheritance [73]. Yoakum described RLS to the point as "the most common disorder you've never heard of" [74].

Clinical diagnostic criteria were later established and reviewed by the International Restless Legs Syndrome Study Group (IRLSSG) [75, 68]. The four essential RLS diagnostic criteria (updated version of 2003) are listed below, adopted from Allen et al. [68] and the review of Trenkwalder et al. [70]. All four criteria are required for a positive diagnosis of RLS [68]. Supportive and associated RLS features are listed in the Appendix B (Enumerations on Page 115).

Essential criteria

1. An urge to move the legs, usually accompanied or caused by uncomfortable and unpleasant sensations in the legs. Sometimes the urge to move is present without the uncomfortable sensations and sometimes the arms or other body parts are involved, in addition to the legs.

2. The urge to move or unpleasant sensations begin or worsen during periods of rest or inactivity such as lying down or sitting.

3. The urge to move or unpleasant sensations are partially or totally relieved by movement (walking or stretching), at least as long as the activity continues.

4. The urge to move or unpleasant sensations are worse in the evening or at night than during the day, or only occur in the evening or night. When symptoms are very severe, the worsening at night may not be noticeable but needs to have been present previously.

Symptoms of RLS are currently reported to occur most often in resting situation, worst in the evening and night (see reviews e.g. [69, 70, 71]). Primary the legs are affected, but arms are involved in up to 48.7% of patients [76]. Current estimates suggest that up to 10% of the human population is affected by RLS (at least in Europe and North

America), reviewed by Ekbom & Ulfbeg et al. [69]. For example an estimate of a prevalence of 9.7% in Orhangazi district of Bursa (Turkey) was found by a population-based study using the IRLSSG criteria for diagnosis [77].

RLS shows a wide spectrum of severity and age of onset going along with difficulties in diagnostics and medial treatment [69, 70, 71]. The severity of RLS varies across patients from a minor to major disruption. A rating scale for the evaluation of the severity of symptoms was established by IRLSSG [78]. The age of onset of RLS is reported to range from an early age, even childhood (see e.g. [79, 80]) to over 80 years of age (reviewed [70]). Most RLS patients are reported to be middle-aged or older according to Trenkwalder et al. [70]. An 'Age-at-onset' study in a large cohort of RLS patients found a large peak at 20 years of age and a smaller peak in the mid-40s [81].

Increasing age and female sex are considered as risk factors [69, 70, 71]. This consideration of these factors is based on two observations: Prevalence in the population increases with age and women are more than twice (e.g. 2.6 times more [77]) affected than men.

5.2.2 Forms

Causes of RLS are divided into primary and secondary forms. These forms are reviewed [69, 70, 71] and briefly summarized below.

1. Primary (idiopathic): In this primary form, the disease RLS is inherited. An autosomal dominant inheritance is of suggest. The identification of the responsible RLS associated genes is still a major aim in the research of RLS. This primary form of RLS is studied in our work since it is the predominant form of RLS and one research area of our cooperation partners Prof. Juliane Winkelmann at the NKP (Neurologische Klinik und Poliklinik, Klinikum rechts der Isar). Explicitly a percentage of 40-60% of cases is reported to be assigned to familial RLS. For example Juliane Winkelmann and colleagues found 77% (232 of 300) RLS patients with iRLS (idiopathic restless legs syndrome): 'positive family history' in the year 2000 [82]. Thereof 42.3% were assigned to 'definite positive' [82]. The focus of treatment is on the dopaminergic system and iron metabolism. Patients are commonly treated with levodopa and dopamine agonists in the first place (dopaminergic treatment).

   Research findings to date are briefly pointed in the following: The first genetic locus in one French Canadian pedigree was found on chromosome 12q in the year 2001 (autosomal dominant mode of inheritance) for the RLS phenotype [83].
   Desauteles et al. note in their discussion that not just this one, but rather several

genetic loci might be responsible for the RLS phenotype [83]. The gene locus on chromosome 12q was later confirmed in five more pedigrees [84]. Loci on chromosomes e.g. the following six: 12q, 14q, 9p, 2q, 20p and 16p were identified until 2009 according to the review of Ekbom et al. [69].

A search for GWAS (Genome wide association studies) hits using HuGE Navigator (version 2.0) [1] returns twelve GWAS hits (collected data since 2001). Most variants of GWAS hits are assigned to gene loci, but recent publications report variants in intergenic loci, such as lincRNAs [4]. The complete list of RLS variants and associated regions is given in Table 5.1. One intergenic RLS variant was discovered for the first time by our cooperation partner Prof. Juliane Winkelmann in the year 2011 [85].

GWAS hit	Variant	Published gene	Region
1	rs3104767	TOX3, BC034767	16q12.2
2	rs6747972	Intergenic	2p14
3	rs2300478	MEIS1	2p14
4	rs9357271	BTBD9	6p21.2
5	rs1975197	PTPRD	9p24.1
6	rs12593813	MAP2K5, SKOR1	15q23
7	rs9296249	BTBD9	6p21.2
8	rs12593813	MAP2K5, LBXCOR1	15q23
9	rs2300478	MEIS1	2p14
10	rs1975197	PTPRD	9p24.1
11	rs4626664	PTPRD	9p23
12	rs3923809	BTBD9	6p21.2

Table 5.1: List of RLS related GWAS hits, HuGE Navigator (version 2.0). Each row describes one GWAS hit. Columns represent (1-4): (1) the number of GWAS hit (2) variant (3) associated gene and (4) chromosome or region. GWAS hits 1-9 were found by Juliane Winkelmann (studies from 2007-2011).

2. Secondary (symptomatic): In the secondary form the disease RLS is the cause of a condition. Conditions can be roughly grouped into other intervening diseases and drugs. An overview of conditions is taken from the review of Byrne et al. [71] (see Figure 5.2). Contrary to the primary form a condition can be directly treated. The three most common conditions are iron deficiency, renal failure and pregnancy. It is reported that up to 25% of caucasian RLS patients are additionally affected by

[1] http://www.hugenavigator.net/HuGENavigator/downloadCenter.do

iron deficiency ('low blood' - low levels of haemoglobin) [86]. Low serum iron levels can be diagnosed: low serum ferritin, < 50 ng/ml (see associated RLS features in the Appendix B (Enumerations on Page 115)). For example iron deficiency was explicitly observed in 34% of indian RLS patients in a study in the year 2007 [86]. This percentage was increased in comparison to a control of 6% [86].

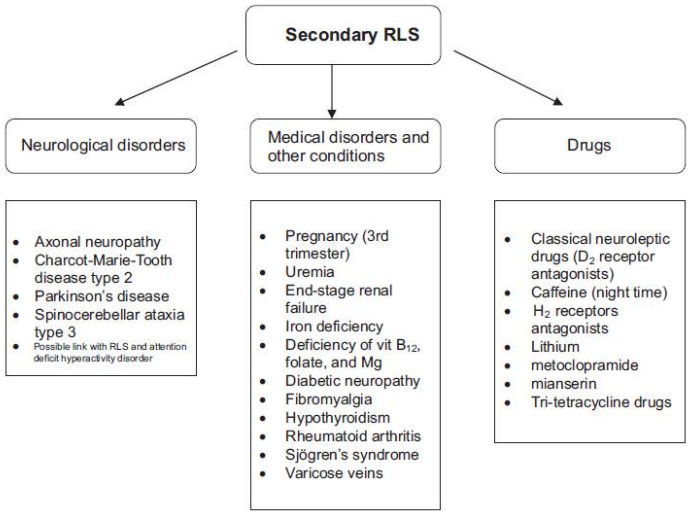


Figure 5.1: Secondary causes of RLS at a glance. This Figure is taken from the work of Byrne et al. [71]. Causes are divided into three parts: (i) Neurological disorders (ii) medical disorders and (iii) other conditions and drugs.

5.3 Exome sequencing analysis

Exome sequencing can be summarized as the use of a NGS technology to sequence the coding regions of the genome (see reviews [87, 88, 89]). Applications of exome sequencing range from the discovery of mutations and genes associated with rare disorders to clinical diagnosis. In dependence of the aim of research distinct strategies can be pursued for discovery. Sequencing and variant filtering can be applied on exomes of related or unrelated affected individuals. For example one or multiple genes with a novel variant, shared among affected individuals of a pedigree, can be identified. In addition to the discovery of mutations, exome sequencing was successfully applied to diagnose genetic diseases. In one of the first studies diagnosis of chloride-losing diarrhea was achieved by the identification of a homozygous missense variant in the disease associated gene SLC26A3 [90].

The workflow of exome sequencing is shown in Figure 5.2. This Figure is taken from the review of Ku et al. [88] and is divided into three parts: (i) Sample Preparation and sequencing, (ii) Primary Data Processing and (iii) Secondary Data processing. In the first part of the workflow genomic DNA is extracted and a sequence library is prepared. This library is further used to sequence the exome using a NGS technology. Secondly, the resulting NGS data is processed to generate raw sequence reads. These reads are mapped onto genome and PCR duplicates are removed. Finally, in the third part, variants are called and annotated. As annotation resources dbSNP or 1000GP can be used to retrieve known SNPs. It has to be mentioned that SNPs can not just be annotated as e.g. 'novel', but other filterings can be applied. For example it might be of interest whether the causal SNP is predicted to show functional effects. Causal mutations are identified and interpreted among variant lists.

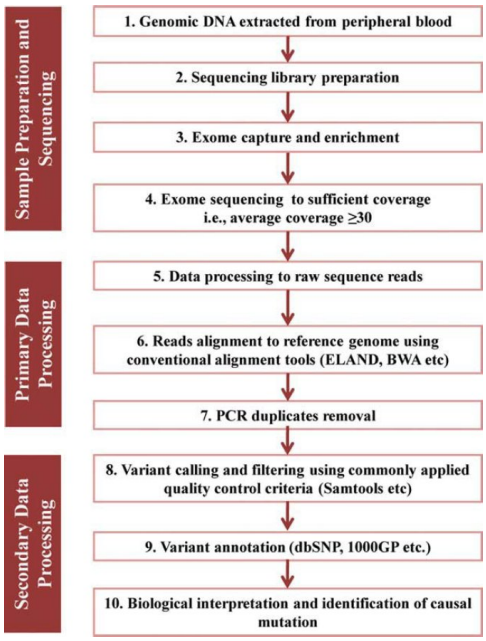


Figure 5.2: Exome sequencing analysis workflow. This Figure is taken from the work of Ku et al. [88]. The workflow is split into three parts: (i) Sample Preparation and sequencing (with steps 1-4) , (ii) Primary Data Processing (with steps 5-7) and (iii) Secondary Data processing with steps (8-10).

6 Methods

6.1 Exome sequencing data

RLS diagnosis and exome sequencing was performed by our collaboration partners, Juliane Winkelmann and colleagues at the NKP (Neurologische Klinik und Poliklinik, Klinikum rechts der Isar). The group of Prof. Juliane Winkelmann focus on the molecular and cellular processes of PD (Parkinson's disease) and RLS (Restless Legs Syndrome). Exome sequencing data was handed in for two families known to be affected by the disease: RLS (Restless Legs Syndrome). The first family comprises data of 5 exomes and the second one of 6 exomes. The generational pedigrees are available by Juliane Winkelmann on demand. The data was generated with (Illumina) sequencing and paired-end reads of 54 nucleotides [nt].

6.2 Data analysis design

One major aim is the identification and biological interpretation of novel variants in our exome sequencing data. We achieved the following analysis of the experimental data in close teamwork with Daniel Ellwanger and Volker Stümpflen from the BIS (Biological Information Systems) group at the Helmholtz Zentrum München. Hence the data analysis is subdivided into two tasks accordingly processed by our group: NGS (Next Generation Sequencing) group and the BIS group.

The workflow is shown in Figure 6.1. In a first task we processed the exome sequencing data across the eleven exomes of patients using our implemented semi-automatic NGS pipeline. The NGS pipeline was applied to identify gene loci carrying novel non-synonymous (deleterious) SNPs. A list of gene candidates carrying these SNPs was subsequently examined by Daniel Ellwanger in a second task. In this second task large scale networks were created, based on a bioinformatics pipeline analysing the impact of these candidates on the function of the expressed protein.

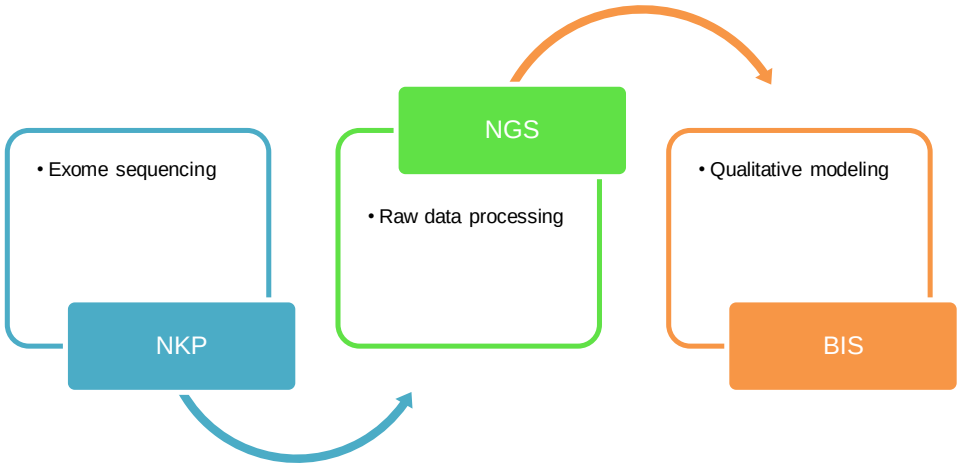


Figure 6.1: Illustration of the data analysis design. The experimental data of Winkelmann et al. was processed by the NGS group. Interesting candidates were modelled by the BIS group.

6.3 Part NGS - Raw data processing

6.3.1 Analysis workflow

Briefly the NGS pipeline can be summarized in the following five steps, see Figure 6.2. In a first step the reads were mapped onto genome with an unspliced aligner. Next, variants were determined. These variants were filtered to identify novel non-synonymous (deleterious) SNPs. These potential novel SNPs were assigned to gene loci. Gene loci could be further classified. This processing procedure was run for each exome separately with the RNA-seq reads as input resulting in a gene list as output. Interesting candidates were identified and further validated. We mainly go into detail of our part in the following (NGS Raw data processing).

It has to be noted that the reads were already pre-aligned (assembly hg19) with the (unspliced) aligner BWA [91] and SNPs were called using samtools [55] by the group of Prof. Juliane Winkelmann. $> 99\%$ of reads could be mapped to the human reference genome in all exomes of the two pedigrees arguing for the correctness of their paired-end alignments. The list of called SNPs were the foundation for our subsequent analysis (skipping the first two steps (1) and (2) of our NGS pipeline). For each patient, we analyzed the list of variants separately in the three (3-5) steps.

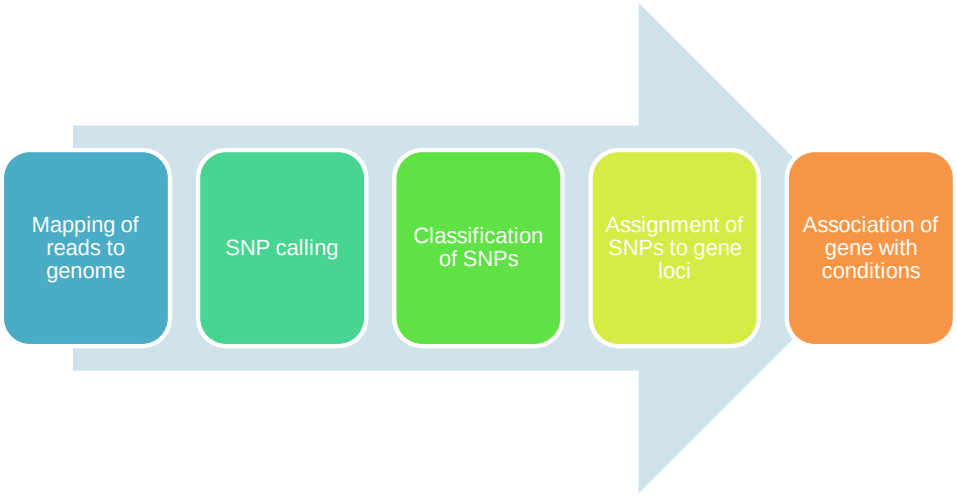


Figure 6.2: Illustration of the raw data processing. The experimental data of Winkelmann et al. was processed by the NGS group in five sequential steps (each step in one rectangle): (1) Reads were mapped onto genome with an aligner (2) Variants were identified (3) SNPs were assigned to categories (e.g. non-synonymous) (4) and gene loci (5) Gene loci were categorized (e.g. novel deleterious non-synonymous SNPs)

6.3.2 Filtering of SNPs

We used the tool ANNOVAR (Function III: Filter based annotation) [92] - version 0800 (download: February 20, 2011) (step 3 of our NGS pipeline) for filtering of variants. The filterings were applied as described online[1]. Variants were scanned against the AVSIFT data set to identify non-synonymous SNPs (assembly hg19 and threshold for SIFT 0.00). Since we are interested in the disease RLS we restricted our following analysis to non-synonymous SNPs. The SIFT score ranges from 0 to 1 (threshold 0.05): Prediction of Amino Acid change, damaging or deleterious with $<= 0.05$ and neural or tolerated with > 0.05. Non-synonymous were ranked according to the SIFT score in deleterious and neutral (see [93] for SIFT). Each non-synonymous SNP was additionally checked for an annotation in dbSNP131 (assembly hg19). The data sets of AVSIFT and dbSNP131 are available online[2].

6.3.3 Assignment of SNPs to gene loci

In a further step (step 4), each non-synonymous SNP was assigned to a gene locus. We used the RefSeq gene track provided by UCSC genome browser[3], track hg19 referring to

[1] http://www.openbioinformatics.org/annovar/annovar_filter.html
[2] http://www.openbioinformatics.org/annovar/annovar_download.html
[3] http://genome.ucsc.edu/cgi-bin/hgTables?command=start

the latest RefSeq annotation version - 37 (download: October 6, 2011). This resulted in a gene list for each patient including the information about its non-synonymous SNPs.

6.3.4 Association of identified loci with PD and RLS relevance

In addition to the filtering procedure using ANNOVAR [92] for SNPs we incorporated reference gene lists provided by Daniel Ellwanger (step 5). In detail, we used one reference of 392 genes potentially playing a role in the PD (Parkinson's disease) from HuGE Navigator[4] - version 2.0. Winkelmann et al. further provided a list of 7 genes directly associated with the progression of RLS [94, 95, 96]. Moreover they previously identified 103 genes playing a hypothetical role in RLS in other studies. These 109 RLS candidates were considered in our analysis, as well. Hence each gene was additionally classified in the following categories: at least one novel variant (no annotation in dbSNP131) (i) association to PD (ii) previous candidate for RLS (iii).

[4]http://www.hugenavigator.net/HuGENavigator/downloadCenter.do

7 Results

7.1 Overview of statistics of variant lists for each patient

As explained in the methods we used previously called SNPs provided for two families (including in total 11 patients affected by RLS) by Winkelmann and colleagues. We applied distinct filterings on these variants for each exome. Two filterings for SNPs were used: (i) dbSNP131 and (ii) AVSIFT as described in the methods. The frequency of SNPs for the distinct filterings are listed in Table 7.1 for each exome of family 1 and in Table 7.2 for family 2 to give a first overview of our data. For example the exome of patient: 27997 of family 1 harbours 44000 variants. 7533 of 44000 variants are non-synonymous (in AVSIFT). 4312 of 44000 are not annotated in dbSNP (not in dbSNP131). We used the stringent set of non-synonymous variants for further analysis as explained in the methods. Next, we created a gene list per exome. The frequency of genes is additionally listed in Tables 7.1 and 7.2, with the restriction to: (i) genes with non-synonymous SNPs regardless of an annotation in dbSNP (in AVSIFT) and (ii) thereof with potential novel SNPs (not in dbSNP131).

Frequency of called SNPs	27997	23352	25279	23347	25406
no filtering	44000	52101	44548	44607	44776
in AVSIFT (non-synonymous)	7533	7870	7537	7561	7697
not in AVSIFT	36467	45231	37011	37046	37079
in dbSNP131	39688	47332	39603	39836	40380
not in dbSNP131 (candiate for novel variant)	4312	5769	4945	4771	4396
Frequency of genes	27997	23352	25279	23347	25406
with non-synonymous SNPs	4733	4800	4672	4691	4728
with novel non-synonymous SNPs	407	423	416	419	396

Table 7.1: Overview of the statistics of the list of variants for each exome (of one patient) of family 1. Each patient has an abbreviation according to the group of Winkelmann (columns (2-6) in row 1). Each row describes the frequencies of called SNPs for one filtering criterion and each of the 5 patients. Column 1 lists the distinct filterings set by the tool ANNOVAR [92] for SNPs: AVSIFT (SIFT score >= 0) and dbSNP131, other conditions for gene lists: >= 1 non-synonymous (novel) SNP.

Frequency of called SNPs	45475	22141	22123	22132	22137	22166
no filtering	45472	45793	45247	65177	46177	45652
in AVSIFT (non-synonymous)	7520	7859	7583	7211	9321	7581
not in AVSIFT	36520	36622	36155	54435	52325	36733
in dbSNP131	39672	40174	39365	54788	40342	40031
not in dbSNP131 (candiate for novel variant)	4368	4307	4373	6858	4385	4283
Frequency of genes	45475	22141	22123	22132	22137	22166
with non-synonymous SNPs	4687	4804	4658	5517	4737	4689
with novel non-synonymous SNPs	377	402	368	601	362	364

Table 7.2: Overview of the statistics of the list of variants for each exome (of one patient) of family 2. Each patient has an abbreviation according to the group of Winkelmann (columns (2-7) in row 1). Each row describes the frequencies of called SNPs or genes for one filtering criterion and each of the 6 patients. Column 1 lists the distinct filterings set by the tool ANNOVAR [92] for SNPs: AVSIFT (SIFT score >= 0) and dbSNP131, other conditions for gene lists: >= 1 non-synonymous (novel) SNP.

7.2 Subset of exomes carrying potential gene candidates

As explained, we determined for each exome of one patient a list of genes. This gene is according to our NGS pipeline associated with at least one non-synonymous SNP. The SNP (deleterious) of a gene might be causal for the disease phenotype.

The proportion of patients carrying such a gene was calculated, as subsets of exomes (of RLS patients): n out of 11 exomes, with n as the number of exomes. The values or subset sizes range from n = 1 to 11 exomes. We adopted the definition of subsets in context of exome sequencing from the work of Ng et al. [97]. As addressed in the motivation of the introduction, exome sequencing was not performed for a control group by Juliane Winkelmann and colleagues. It would be necessary to examine the proportion of affected and healthy individuals (control group) carrying a variant of a gene (gene X for simplification) to determine the penetrance. Further family background should be considered in the analysis using knowledge about the relations and symptoms between individuals of a pedigree. Subsets give us a slight evidence for the 'penetrance' of gene X without a control. A low penetrance might refer e.g. to a subset size of one (n = 1) or the minority of RLS patients carrying Gene X. Contrary the majority or even all patients of both pedigrees could carry Gene X (n = 11), high penetrance or highly heritable.

A gene of high penetrance should be of most importance for the monogenic form of RLS (especially with subset sizes n = 11, 10, 9 in our data). The distribution of the frequency of genes per subset size is listed in Table 7.3, column 2 with any condition.

One of our issues is the detection of novel (deleterious) non-synonymous SNPs (i). Furthermore we are interested for associations between the diseases RLS and PD (ii): to which degree can be PD related genes found in our RLS exome sequencing data? Candidate genes for RLS were previously identified by other studies (unpublished) and were checked for confirmation in our data (iii). We used the reference gene lists of 392 and 109 genes reported to be associated with the diseases PD and RLS for the last two issues (ii,iii) (provided by Daniel Ellwanger). Each gene with non-synonymous SNPs was assigned to an association: novel variant, PD, RLS (i,ii,iii). The frequencies are additionally listed in Table 7.3, columns 3-5.

For example all patients share 2035 genes with at least one non-synonymous SNP (subset size n = 11). Thereof 8 genes harbour at least one novel non-synonymous variant, 39 might be related to the disease PD and 24 of previous found RLS genes could be confirmed with the exome sequencing approach.

Subset size: n of 11	Frequency of genes, non-synonymous SNPs			
		Thereof association with		
	any condition	novel SNP	PD	RLS
11	2035	8	39	24
10	2662	12	53	30
9	3204	23	65	35
8	3653	27	70	37
7	4093	40	80	38
6	4560	59	86	40
5	5072	130	96	50
4	5658	272	107	60
3	6302	558	113	72
2	7095	1031	128	74
1	8383	2375	150	77

Table 7.3: Overview of the frequency of genes per subset size. We selected genes with non-synonymous SNPs. We calculated the proportion of patients carrying a candidate gene, as subsets: n out of 11 exomes, with n as the number of exomes. The values range from n = 1 to 11 exomes. We adopted the definition of subsets in context of exome sequencing from the work of Ng et al. [97]. Each row describes the frequency of genes in one subset size. Column 1 lists the subset size n of exomes. Columns (2-5) show the frequencies of genes common in the subset of patients: (2) all genes with non-synonymous SNPs. These frequencies of all genes are subdivided in (3) with at least one novel non-synonymous SNP (4) related to PD (5) previous candidates for RLS.

7.2.1 Novel RLS variants detected in candidates

As mentioned 8 genes are shared among all eleven affected individuals in our data and are the most reliable candidates (see Table 7.3). Thereof just 4 genes harbour SNPs that are not annotated in dbSNP131 (and not in the preview version of dbSNP134, marked with a star*). To date just about 7 genes are reported to be associated with RLS [94, 95, 96]. The discovery of further novel variants is important.

It has to be remarked that these genes have a subset size n = 11, but a part of associated novel SNPs might slightly vary across exomes (especially between the families). For example, one variant (position) can be absent in one exome, but present in the majority of the other exomes in the worst case. In the other case, the fluctuation of a SNP could be just marginal e.g. in the nucleotide change or expression level (read coverage) of the position on genome.

The candidate genes and potential novel variant coordinates are listed in the following Table 7.4 for the maximal subset size n = 11. We report the statistics of a SNP (including for example the nucleotide change) according to merged SNP calls and filtering runs (reported in the majority of exomes). Additionally to the subset size, non-synonymous SNPs were ranked based on the SIFT score. We further report genes and their SNPs for subset sizes n = 10 and n = 9, with the restriction to one further association to a disease (RLS or PD).

We could identify a gene PLXNA2 associated to PD carrying a novel deleterious SNP (score = 0.00). This might be another interesting candidate. The 109 RLS genes of the reference list have to be carefully considered since they are just candidates. These candidates still need further examination and confirmation. The gene FLNB could be confirmed in our exome sequencing data (subset size n = 9), accordingly with a novel variant. This variant might give a further hint to the cause of RLS. This gene is already reported in context of other severe disorders apart from PD: such as dyspalsia boomerang and Larsen syndrome (20301736).

Genes with novel (not in dbSNP134) non-synonymous (deleterious) SNPs are highlighted in orange and examined by Daniel Ellwanger in detail.

Gene name	Chr	Pos	SIFT score	Association (PD or RLS)	Nucleotides	Zygosity
n = 11						
KIAA1267	chr17	44144993	0.05	PD	C/G	hom.
PRSS1	chr7	142460335*	0.47	-	A/G	het.
		142460369	1.00	-	G/A	het.
SELRC1	chr1	53158524*	0.51	-	A/C	het.
CYP2A7	chr19	41383849	1.00	-	C/G	het.
ABCC6	chr16	16271357	0.37	-	T/C	hom.
RETSAT	chr2	85571228*	0.07	-	G/C	het.
CDC27	chr17	45234303*	0.04	-	G/C	het.
		45234417*	0.01	-	A/G	het.
FAM135A	chr6	71187020*	0.00	-	A/C	het.
n = 10						
PLXNA2	chr1	208272313*	0.00	PD	A/C	het.
n = 9						
FLNB	chr3	58145363*	0.08	RLS	A/C	het.

Table 7.4: Overview of the statistics of the list of gene candidates with novel SNPs. Each row describes one gene with novel non-synonymous SNPs and the information about further association to PD or RLS. We calculated the proportion of patients carrying a candidate gene, as subsets: n out of 11 exomes, with n as the number of exomes. The values range from n = 1 to 11 exomes. We adopted the definition of subsets in context of exome sequencing from the work of Ng et al. [97]. For n = 11 all candidate genes are listed with novel non-synonymous SNPs, n = 10 and n = 9 just the additional genes to n = 11 and the restriction to have at least one further association. Columns 1-7: (1) Name of the gene (2) Chromosome of the gene (3) Position of the novel SNP (no annotation in dbSNP131 - in dbSNP134 marked with a star) on genome positions of distinct SNPs are separated by newline (just an extract of all novel ones spread in more than half of all patients) (4) SIFT (Sorting Tolerant From Intolerant) score of AVSIFT, ranging from 0 (deleterious) to 1 (neutral) with <= 0.05 damaging or deleterious and > 0.05 tolerated (0 applied default threshold for ANNOVAR to get all non-synonymous SNPs), slightly above threshold could be considered as deleterious according to the SIFT manual (5) Any other associations to disease (PD or RLS), (6) Nucleotides (7) Zygosity (homozygous or heterozygous.) The most interesting candidate genes with potential novel, deleterious non-synonymous SNPs are highlighted in orange.

7.2.2 Parkinson's disease relevant genes in RLS patients identified

The overlap of the reference lists of PD and RLS associated genes is zero. Although we obviously don't have any overlap we found several several Parkinson's disease (PD) associated genes in our data. 39 of 8383 genes (with non-synonymous SNPs) are common in all exomes and reported to be involved in PD (subset size n = 11). The frequency of 39 is quite high in relation to 392 included PD relevant genes (reference list of Daniel Ellwanger). The frequency increases (maximal with 150) with decreasing subset size as expected (subset size n = 1). This observation implies that about 10% (39 of 392) up to 38% (150 of 392) of PD associated genes are found in RLS exome sequencing data.

For example the PD candidate gene KIAA1267 is abundant in all eleven patients of both families (subset size n = 11). This gene harbours one non-synonymous SNP among other variants: rs144838667, chr17:44144993 (already annotated in dbSNP134). This might be an interesting discovery, as well, since this SNP seems to be a quite novel variant and of relevance for RLS. The gene KIAA1267 is of unknown function, but interestingly reported to potentially play a role in Parkinson's disease [98]. There is one recent publication indicating that several genes, namely MAPT, STH, and KIAA1267 were significantly increased in expression in PD brains [98]. This might be an indication for associations between PD and RLS (common genes and treatment with Dopamin?). Another example is the gene LRRK2 (Leucine-Rich Repeat Kinase 2), reported as one idiopathic (monogenic) form of Parkinson's disease [99]. There is one study describing that three out of 138 probands having the LRRK2 mutation G2019S analyzed in context of a study about PD are affected by RLS [100] (common SNPs?).

An overlap would be significant under the assumption that co-morbidity is caused by the same locus. LRRK2 is included in our exome data and interestingly in addition to KIAA1267 spread in all eleven patients, with already annotated non-synonymous SNPs.

8 Discussion

The determination of the causes inducing the disease RLS is one important research issue. We analyzed the idiopathic form of RLS in our exome sequencing data. Gene loci with novel non-synonymous SNPs could be identified. Three of these loci are spread across all eleven exomes of RLS affected patients in both families. We were not capable to determine the significance of our findings since a control group of healthy people is missing in the provided exome sequencing data. Potential candiate genes were analyzed in detail by a colleague of mine: Daniel Ellwanger.

According to reported findings it is supposable that not just a single gene is the cause for the disease RLS, rather other multiple genes and other factors might intervene. We focus on transcriptomics - NGS sequencing in this part of the dissertation. The inclusion of other 'omics' approaches is important to gain more comprehensive knowledge about diseases.

One disadvantage of the exome sequencing approach itself lies in the restriction of the sequencing and discovery of disease relevant SNPs on coding regions of the genome. Relevant variants in noncoding and intergenic loci are not considered in this technique. The first and main part of our work focus on lincRNAs (large intergenic ncRNAs). There is increasing evidence that these lincRNAs harbour SNPs related to disease phenotypes. Cabili et al. found several hundreds of these lincRNAs in disease associated regions (GWAS catalogue) [4]. One example of a lincRNA is PRNCR1 (prostate cancer non-coding RNA 1) [5]. This lincRNA is located in a so called 'gene-desert' region on chr8 [5] and harbours a variant. The expression of PRNCR1 is recently observed to be up-regulated in PC (prostate cancer) cells and to be associated with PC susceptibility [5].

Conclusion, contribution and outlook

Improvement in the research of the development of sequencing technologies make it possible to sequence and investigate whole genomes, exomes and transcriptomes with the NGS technique. RNA-seq provides an unprecedented insight into the complex network of interacting coding and noncoding RNA in a cell line or tissue. This network of interactions can be used for detailed bioinformatics analysis facing new biological challenges and questions. In this doctoral thesis, we lay the focus on two applications of NGS. The first application of RNA-seq and main part of this thesis is an analysis of the functionality of lincRNAs. The second application of Exome-seq is about the detection of novel RLS variants.

In the first part of this doctoral thesis we analysed recently published RNA-seq data in three mouse cell lines including novel lincRNA reconstructions [1]. lincRNA (large intergenic ncRNA) is a subclass of lncRNA (long ncRNA) and is predominant among ncRNA types in the GENCODE7 data [8]. This subclass is associated with a variety of regulatory actions in recent publications including the investigated actions, alternative splicing regulation and the generation of small ncRNAs in our work [15, 16, 9].

Potential interactions between lincRNAs and coding, noncoding RNA partners were identified with sequence similarity searches. In the following we highlight the influence of lincRNAs on the regulation of their target alternative transcripts' expression and events since the results of this analysis are the most promising findings and contributions of the first part of this thesis. Target and target site candidates of high sequence similarity were searched for the detection of lincRNA:mRNA duplexes in the analysis workflow. We calculated a set of features for identified sequence regions, such as GO enrichment and GC content. The change in expression of splice variants was calculated and alternate splice events predicted for target genes for investigation of their regulatory role in splicing regulation.

We identified a set of potential lincRNA:mRNA duplexes, with an enrichment at 5' introns. The expression levels of nearby coding exons are significantly decreased. Accordingly to the enrichment at 5' introns, the strongest regulatory influence of lincRNAs was found on 5' exons: (i) strongest down-regulation of the expression and (ii) a sig-

nificant increased frequency of the splice events AFE and ALE (Alternative First and Last Exon). These findings show supportive evidence that lincRNA:mRNA duplexes present a new regulatory mechanism of alternative splicing. Alternate splice variants were demonstrated to be regulated via duplexes with recent RNA-seq data in three cell lines for the first time in this doctoral thesis.

Ongoing research in RNA-seq and the detection of lincRNAs among transcriptome reconstructions will lead to new publications with increased data amounts and cell lines. The contribution of this thesis is a framework for the generation of advanced computational pipelines, the incorporation of new ideas and the applications of the pipeline on new RNA-seq data sets to confirm and continue this work on regulatory actions of lincRNAs. This work mainly focused on two actions. Since there is sound evidence that there are further actions, these can be additionally incorporated in the improved analysis in an equivalent fashion as one idea. In addition, single steps in the analysis workflow for the action (Regulation of their target alternative transcripts' expression and events) can be improved. For example, the sequence similarity search can be modified to check whether the interaction is specific or just restricted to the segment for the removal of FP's (False Positives). Likewise to the tool BLAST [101] a hit can be taken and adjacent sequences inspected to check whether an alignment (extension) is possible. In addition, quantitative data concerning the total score/length of the homolog segment can be extracted. Further, parameter settings of applied tools, like Blat can be modified or entirely exchanged by other tools. Single cases of candidate lincRNA:mRNA duplexes that were found with our bioinformatics analysis should be taken for detailed biological experiments to reconfirm their influence on the regulation of alternate splicing by Biologists.

In the second part we analysed exome sequencing data of individuals of two pedigrees that are affected by the disease RLS (Restless legs syndrome) in cooperation with Juliane Winkelmann (Neurologische Klinik und Poliklinik, Klinikum rechts der Isar) and Daniel Ellwanger (BIS group, Helmholtz Zentrum München). RLS is a neurological disorder with a prevalence of at least 5%. There is evidence for an autosomal dominant inheritance [73]. The detection and interpretation of gene loci associated with the RLS phenotype is still an important challenge in the research of RLS.

The analysis workflow for this challenge was split in three tasks. Our task (NGS) is the raw data processing of the provided exome sequencing data. It has to be mentioned, that the steps, mapping of reads onto genome and SNP calling, were already achieved by our cooperation partners. We continued with the classification of SNPs. SNPs were annotated in e.g. known and novel. Further annotated SNPs were assigned to gene loci.

Associated genes could be further inspected in association with conditions, such as other disease. We could create a list of candidate genes with novel and non-synonymous SNPs in agreement with Juliane Winkelmann. These candidate genes can be used for further research.

Since the provided exome sequencing data lack a control group of healthy individuals the significance of our identified variants remains unclear. The findings have to be confirmed with a control group and further investigated with Qualitative modeling by Daniel Ellwanger in his task. The analysis workflow of our NGS task can be modified and extended. Tools for e.g. the alignment of reads can be exchanged or adjusted. Information about the relation of individuals within a pedigree is necessary to adjust and improve our analysis workflow. Another idea is the incorporation of parent-child trios in the analysis design to identify de novo mutations.

Bibliography

[1] Mitchell Guttman, Manuel Garber, Joshua Z Levin, Julie Donaghey, James Robinson, Xian Adiconis, Lin Fan, Magdalena J Koziol, Andreas Gnirke, Chad Nusbaum, John L Rinn, Eric S Lander, and Aviv Regev. Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincrnas. *Nat Biotechnol*, 28(5):503–510, May 2010.

[2] Taishin Kin, Kouichirou Yamada, Goro Terai, Hiroaki Okida, Yasuhiko Yoshinari, Yukiteru Ono, Aya Kojima, Yuki Kimura, Takashi Komori, and Kiyoshi Asai. frnadb: a platform for mining/annotating functional rna candidates from non-coding rna sequences. *Nucleic Acids Res*, 35(Database issue):D145–D148, Jan 2007.

[3] Philipp Kapranov, Jill Cheng, Sujit Dike, David A Nix, Radharani Duttagupta, Aarron T Willingham, Peter F Stadler, Jana Hertel, Jörg Hackermüller, Ivo L Hofacker, Ian Bell, Evelyn Cheung, Jorg Drenkow, Erica Dumais, Sandeep Patel, Gregg Helt, Madhavan Ganesh, Srinka Ghosh, Antonio Piccolboni, Victor Sementchenko, Hari Tammana, and Thomas R Gingeras. Rna maps reveal new rna classes and a possible function for pervasive transcription. *Science*, 316(5830):1484–1488, Jun 2007.

[4] Moran N Cabili, Cole Trapnell, Loyal Goff, Magdalena Koziol, Barbara Tazon-Vega, Aviv Regev, and John L Rinn. Integrative annotation of human large intergenic noncoding rnas reveals global properties and specific subclasses. *Genes Dev*, 25(18):1915–1927, Sep 2011.

[5] Suyoun Chung, Hidewaki Nakagawa, Motohide Uemura, Lianhua Piao, Kyota Ashikawa, Naoya Hosono, Ryo Takata, Shusuke Akamatsu, Takahisa Kawaguchi, Takashi Morizono, Tatsuhiko Tsunoda, Yataro Daigo, Koichi Matsuda, Naoyuki Kamatani, Yusuke Nakamura, and Michiaki Kubo. Association of a novel long non-coding rna in 8q24 with prostate cancer susceptibility. *Cancer Sci*, 102(1):245–252, Jan 2011.

[6] Ewan A Gibb, Carolyn J Brown, and Wan L Lam. The functional role of long non-coding rna in human carcinomas. *Mol Cancer*, 10:38, 2011.

[7] Brian J Haas and Michael C Zody. Advancing rna-seq analysis. *Nat Biotechnol*, 28(5):421–423, May 2010.

[8] Monya Baker. Long noncoding rnas: the search for function. *Nat Methods*, 8:379–383, Apr 2011.

[9] Jeremy E Wilusz, Hongjae Sunwoo, and David L Spector. Long noncoding rnas: functional surprises from the rna world. *Genes Dev*, 23(13):1494–1504, Jul 2009.

[10] Reini F Luco, Mariano Allo, Ignacio E Schor, Alberto R Kornblihtt, and Tom Misteli. Epigenetics in alternative pre-mrna splicing. *Cell*, 144(1):16–26, Jan 2011.

[11] Eric T Wang, Rickard Sandberg, Shujun Luo, Irina Khrebtukova, Lu Zhang, Christine Mayr, Stephen F Kingsmore, Gary P Schroth, and Christopher B Burge. Alternative isoform regulation in human tissue transcriptomes. *Nature*, 456(7221):470–476, Nov 2008.

[12] Lawrence A Chasin. Searching for splicing motifs. *Adv Exp Med Biol*, 623:85–106, 2007.

[13] Jennifer C Long and Javier F Caceres. The sr protein family of splicing factors: master regulators of gene expression. *Biochem J*, 417(1):15–27, Jan 2009.

[14] Siew Ping Han, Yue Hang Tang, and Ross Smith. Functional diversity of the hnrnps: past, present and perspectives. *Biochem J*, 430(3):379–392, Aug 2010.

[15] Tim R Mercer, Marcel E Dinger, and John S Mattick. Long non-coding rnas: insights into functions. *Nat Rev Genet*, 10(3):155–159, Mar 2009.

[16] Xiangting Wang, Xiaoyuan Song, Christopher K Glass, and Michael G Rosenfeld. The long arm of long noncoding rnas: Roles as sensors regulating gene transcriptional programs. *Cold Spring Harb Perspect Biol*, Jun 2010.

[17] John L Rinn, Michael Kertesz, Jordon K Wang, Sharon L Squazzo, Xiao Xu, Samantha A Brugmann, L. Henry Goodnough, Jill A Helms, Peggy J Farnham, Eran Segal, and Howard Y Chang. Functional demarcation of active and silent chromatin domains in human hox loci by noncoding rnas. *Cell*, 129(7):1311–1323, Jun 2007.

[18] A. T. Willingham, A. P. Orth, S. Batalov, E. C. Peters, B. G. Wen, P. Aza-Blanc, J. B. Hogenesch, and P. G. Schultz. A strategy for probing the function of noncoding rnas finds a repressor of nfat. *Science*, 309(5740):1570–1573, Sep 2005.

[19] Vidisha Tripathi, Jonathan D Ellis, Zhen Shen, David Y Song, Qun Pan, Andrew T Watt, Susan M Freier, C. Frank Bennett, Alok Sharma, Paula A Bubulya, Benjamin J Blencowe, Supriya G Prasanth, and Kannanganattu V Prasanth. The nuclear-retained noncoding rna malat1 regulates alternative splicing by modulating sr splicing factor phosphorylation. *Mol Cell*, 39(6):925–938, Sep 2010.

[20] Ming-De Yan, Chih-Chen Hong, Gi-Ming Lai, Ann-Lii Cheng, Ya-Wen Lin, and Shuang-En Chuang. Identification and characterization of a novel gene saf transcribed from the opposite strand of fas. *Hum Mol Genet*, 14(11):1465–1474, Jun 2005.

[21] Manuel Beltran, Isabel Puig, Cristina Peña, José Miguel García, Ana Belén Alvarez, Raúl Peña, Félix Bonilla, and Antonio García de Herreros. A natural antisense transcript regulates zeb2/sip1 gene expression during snail1-induced epithelial-mesenchymal transition. *Genes Dev*, 22(6):756–769, Mar 2008.

[22] Shivendra Kishore and Stefan Stamm. The snorna hbii-52 regulates alternative splicing of the serotonin receptor 2c. *Science*, 311(5758):230–232, Jan 2006.

[23] Yarden Katz, Eric T Wang, Edoardo M Airoldi, and Christopher B Burge. Analysis and design of rna sequencing experiments for identifying isoform regulation. *Nat Methods*, 7(12):1009–1015, Dec 2010.

[24] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature*, 431(7011):931–945, Oct 2004.

[25] E. N. C. O. D. E. Project Consortium, Ewan Birney, John A Stamatoyannopoulos, Anindya Dutta, Roderic Guigó, Thomas R Gingeras, Elliott H Margulies, Zhiping Weng, Michael Snyder, Emmanouil T Dermitzakis, Robert E Thurman, Michael S Kuehn, Christopher M Taylor, Shane Neph, Christoph M Koch, Saurabh Asthana, Ankit Malhotra, Ivan Adzhubei, Jason A Greenbaum, Robert M Andrews, Paul Flicek, Patrick J Boyle, Hua Cao, Nigel P Carter, Gayle K Clelland, Sean Davis, Nathan Day, Pawandeep Dhami, Shane C Dillon, Michael O Dorschner, Heike Fiegler, Paul G Giresi, Jeff Goldy, Michael Hawrylycz, Andrew Haydock, Richard Humbert, Keith D James, Brett E Johnson, Ericka M Johnson, Tristan T Frum, Elizabeth R Rosenzweig, Neerja Karnani, Kirsten Lee, Gregory C Lefebvre, Patrick A Navas, Fidencio Neri, Stephen C J Parker, Peter J Sabo, Richard Sandstrom, Anthony Shafer, David Vetrie, Molly Weaver, Sarah Wilcox, Man Yu, Francis S Collins, Job Dekker, Jason D Lieb, Thomas D Tullius, Gregory E Crawford, Shamil Sunyaev, William S Noble, Ian Dunham, France Denoeud, Alexandre Reymond, Philipp Kapranov, Joel Rozowsky, Deyou Zheng, Robert

Castelo, Adam Frankish, Jennifer Harrow, Srinka Ghosh, Albin Sandelin, Ivo L Hofacker, Robert Baertsch, Damian Keefe, Sujit Dike, Jill Cheng, Heather A Hirsch, Edward A Sekinger, Julien Lagarde, Josep F Abril, Atif Shahab, Christoph Flamm, Claudia Fried, Jörg Hackermüller, Jana Hertel, Manja Lindemeyer, Kristin Missal, Andrea Tanzer, Stefan Washietl, Jan Korbel, Olof Emanuelsson, Jakob S Pedersen, Nancy Holroyd, Ruth Taylor, David Swarbreck, Nicholas Matthews, Mark C Dickson, Daryl J Thomas, Matthew T Weirauch, James Gilbert, Jorg Drenkow, Ian Bell, XiaoDong Zhao, K. G. Srinivasan, Wing-Kin Sung, Hong Sain Ooi, Kuo Ping Chiu, Sylvain Foissac, Tyler Alioto, Michael Brent, Lior Pachter, Michael L Tress, Alfonso Valencia, Siew Woh Choo, Chiou Yu Choo, Catherine Ucla, Caroline Manzano, Carine Wyss, Evelyn Cheung, Taane G Clark, James B Brown, Madhavan Ganesh, Sandeep Patel, Hari Tammana, Jacqueline Chrast, Charlotte N Henrichsen, Chikatoshi Kai, Jun Kawai, Ugrappa Nagalakshmi, Jiaqian Wu, Zheng Lian, Jin Lian, Peter Newburger, Xueqing Zhang, Peter Bickel, John S Mattick, Piero Carninci, Yoshihide Hayashizaki, Sherman Weissman, Tim Hubbard, Richard M Myers, Jane Rogers, Peter F Stadler, Todd M Lowe, Chia-Lin Wei, Yijun Ruan, Kevin Struhl, Mark Gerstein, Stylianos E Antonarakis, Yutao Fu, Eric D Green, Ulaş Karaöz, Adam Siepel, James Taylor, Laura A Liefer, Kris A Wetterstrand, Peter J Good, Elise A Feingold, Mark S Guyer, Gregory M Cooper, George Asimenos, Colin N Dewey, Minmei Hou, Sergey Nikolaev, Juan I Montoya-Burgos, Ari Löytynoja, Simon Whelan, Fabio Pardi, Tim Massingham, Haiyan Huang, Nancy R Zhang, Ian Holmes, James C Mullikin, Abel Ureta-Vidal, Benedict Paten, Michael Seringhaus, Deanna Church, Kate Rosenbloom, W. James Kent, Eric A Stone, N. I. S. C. Comparative Sequencing Program, Baylor College of Medicine Human Genome Sequencing Center, Washington University Genome Sequencing Center, Broad Institute, Children's Hospital Oakland Research Institute, Serafim Batzoglou, Nick Goldman, Ross C Hardison, David Haussler, Webb Miller, Arend Sidow, Nathan D Trinklein, Zhengdong D Zhang, Leah Barrera, Rhona Stuart, David C King, Adam Ameur, Stefan Enroth, Mark C Bieda, Jonghwan Kim, Akshay A Bhinge, Nan Jiang, Jun Liu, Fei Yao, Vinsensius B Vega, Charlie W H Lee, Patrick Ng, Atif Shahab, Annie Yang, Zarmik Moqtaderi, Zhou Zhu, Xiaoqin Xu, Sharon Squazzo, Matthew J Oberley, David Inman, Michael A Singer, Todd A Richmond, Kyle J Munn, Alvaro Rada-Iglesias, Ola Wallerman, Jan Komorowski, Joanna C Fowler, Phillippe Couttet, Alexander W Bruce, Oliver M Dovey, Peter D Ellis, Cordelia F Langford, David A Nix, Ghia Euskirchen, Stephen Hartman, Alexander E Urban, Peter Kraus, Sara Van Calcar, Nate Heintzman, Tae Hoon Kim, Kun Wang, Chunxu Qu, Gary Hon, Rosa Luna, Christopher K Glass, M. Geoff Rosenfeld, Shelley Force Aldred, Sara J Cooper, Anason Halees, Jane M Lin, Hennady P Shulha, Xiaoling Zhang, Mousheng Xu, Jaafar N S Haidar, Yong Yu,

Yijun Ruan, Vishwanath R Iyer, Roland D Green, Claes Wadelius, Peggy J Farnham, Bing Ren, Rachel A Harte, Angie S Hinrichs, Heather Trumbower, Hiram Clawson, Jennifer Hillman-Jackson, Ann S Zweig, Kayla Smith, Archana Thakkapallayil, Galt Barber, Robert M Kuhn, Donna Karolchik, Lluis Armengol, Christine P Bird, Paul I W de Bakker, Andrew D Kern, Nuria Lopez-Bigas, Joel D Martin, Barbara E Stranger, Abigail Woodroffe, Eugene Davydov, Antigone Dimas, Eduardo Eyras, Ingileif B Hallgrímsdóttir, Julian Huppert, Michael C Zody, Gonçalo R Abecasis, Xavier Estivill, Gerard G Bouffard, Xiaobin Guan, Nancy F Hansen, Jacquelyn R Idol, Valerie V B Maduro, Baishali Maskeri, Jennifer C McDowell, Morgan Park, Pamela J Thomas, Alice C Young, Robert W Blakesley, Donna M Muzny, Erica Sodergren, David A Wheeler, Kim C Worley, Huaiyang Jiang, George M Weinstock, Richard A Gibbs, Tina Graves. Identification and analysis of functional elements in 1genome by the encode pilot project. *Nature*, 447(7146):799–816, Jun 2007.

[26] Alka Saxena and Piero Carninci. Long non-coding rna modifies chromatin: Epigenetic silencing by long non-coding rnas. *Bioessays*, Sep 2011.

[27] Miao-Chih Tsai, Robert C Spitale, and Howard Y Chang. Long intergenic noncoding rnas: new links in cancer progression. *Cancer Res*, 71(1):3–7, Jan 2011.

[28] C. I. Brannan, E. C. Dees, R. S. Ingram, and S. M. Tilghman. The product of the h19 gene may function as an rna. *Mol Cell Biol*, 10(1):28–36, Jan 1990.

[29] Mitchell Guttman, Ido Amit, Manuel Garber, Courtney French, Michael F Lin, David Feldser, Maite Huarte, Or Zuk, Bryce W Carey, John P Cassady, Moran N Cabili, Rudolf Jaenisch, Tarjei S Mikkelsen, Tyler Jacks, Nir Hacohen, Bradley E Bernstein, Manolis Kellis, Aviv Regev, John L Rinn, and Eric S Lander. Chromatin signature reveals over a thousand highly conserved large non-coding rnas in mammals. *Nature*, 458(7235):223–227, Mar 2009.

[30] Paulo P Amaral, Michael B Clark, Dennis K Gascoigne, Marcel E Dinger, and John S Mattick. lncrnadb: a reference database for long noncoding rnas. *Nucleic Acids Res*, 39(Database issue):D146–D151, Jan 2011.

[31] Chandra Shekhar Pareek, Rafal Smoczynski, and Andrzej Tretyn. Sequencing technologies and genome sequencing. *J Appl Genet*, 52(4):413–435, Nov 2011.

[32] Michael L Metzker. Sequencing technologies - the next generation. *Nat Rev Genet*, 11(1):31–46, Jan 2010.

[33] Manuel Garber, Manfred G Grabherr, Mitchell Guttman, and Cole Trapnell. Computational methods for transcriptome annotation and quantification using rna-seq. *Nat Methods*, 8(6):469–477, Jun 2011.

[34] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, Jan 2009.

[35] Cole Trapnell, Lior Pachter, and Steven L Salzberg. Tophat: discovering splice junctions with rna-seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[36] Cole Trapnell, Brian A Williams, Geo Pertea, Ali Mortazavi, Gordon Kwan, Marijke J van Baren, Steven L Salzberg, Barbara J Wold, and Lior Pachter. Transcript assembly and quantification by rna-seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat Biotechnol*, 28(5):511–515, May 2010.

[37] Xuezhong Cai and Bryan R Cullen. The imprinted h19 noncoding rna is a primary microrna precursor. *RNA*, 13(3):313–316, Mar 2007.

[38] Margaret S Ebert, Joel R Neilson, and Phillip A Sharp. Microrna sponges: competitive inhibitors of small rnas in mammalian cells. *Nat Methods*, 4(9):721–726, Sep 2007.

[39] Rajnish A Gupta, Nilay Shah, Kevin C Wang, Jeewon Kim, Hugo M Horlings, David J Wong, Miao-Chih Tsai, Tiffany Hung, Pedram Argani, John L Rinn, Yulei Wang, Pius Brzoska, Benjamin Kong, Rui Li, Robert B West, Marc J van de Vijver, Saraswati Sukumar, and Howard Y Chang. Long non-coding rna hotair reprograms chromatin state to promote cancer metastasis. *Nature*, 464(7291):1071–1076, Apr 2010.

[40] Wing Pui Tsang, Enders K O Ng, Simon S M Ng, Hongchuan Jin, Jun Yu, Joseph J Y Sung, and Tim Tak Kwok. Oncofetal h19-derived mir-675 regulates tumor suppressor rb in human colorectal cancer. *Carcinogenesis*, 31(3):350–358, Mar 2010.

[41] C. Joel McManus and Brenton R Graveley. Rna structure and the mechanisms of alternative splicing. *Curr Opin Genet Dev*, 21(4):373–379, Aug 2011.

[42] Douglas L Black. Mechanisms of alternative pre-messenger rna splicing. *Annu Rev Biochem*, 72:291–336, 2003.

[43] Barmak Modrek and Christopher Lee. A genomic view of alternative splicing. *Nat Genet*, 30(1):13–19, Jan 2002.

[44] E. S. Lander, L. M. Linton, B. Birren, C. Nusbaum, M. C. Zody, J. Baldwin, K. Devon, K. Dewar, M. Doyle, W. FitzHugh, R. Funke, D. Gage, K. Harris, A. Heaford, J. Howland, L. Kann, J. Lehoczky, R. LeVine, P. McEwan, K. McKernan, J. Meldrim, J. P. Mesirov, C. Miranda, W. Morris, J. Naylor, C. Raymond, M. Rosetti, R. Santos, A. Sheridan, C. Sougnez, N. Stange-Thomann, N. Stojanovic, A. Subramanian, D. Wyman, J. Rogers, J. Sulston, R. Ainscough, S. Beck, D. Bentley, J. Burton, C. Clee, N. Carter, A. Coulson, R. Deadman, P. Deloukas, A. Dunham, I. Dunham, R. Durbin, L. French, D. Grafham, S. Gregory, T. Hubbard, S. Humphray, A. Hunt, M. Jones, C. Lloyd, A. McMurray, L. Matthews, S. Mercer, S. Milne, J. C. Mullikin, A. Mungall, R. Plumb, M. Ross, R. Shownkeen, S. Sims, R. H. Waterston, R. K. Wilson, L. W. Hillier, J. D. McPherson, M. A. Marra, E. R. Mardis, L. A. Fulton, A. T. Chinwalla, K. H. Pepin, W. R. Gish, S. L. Chissoe, M. C. Wendl, K. D. Delehaunty, T. L. Miner, A. Delehaunty, J. B. Kramer, L. L. Cook, R. S. Fulton, D. L. Johnson, P. J. Minx, S. W. Clifton, T. Hawkins, E. Branscomb, P. Predki, P. Richardson, S. Wenning, T. Slezak, N. Doggett, J. F. Cheng, A. Olsen, S. Lucas, C. Elkin, E. Uberbacher, M. Frazier, R. A. Gibbs, D. M. Muzny, S. E. Scherer, J. B. Bouck, E. J. Sodergren, K. C. Worley, C. M. Rives, J. H. Gorrell, M. L. Metzker, S. L. Naylor, R. S. Kucherlapati, D. L. Nelson, G. M. Weinstock, Y. Sakaki, A. Fujiyama, M. Hattori, T. Yada, A. Toyoda, T. Itoh, C. Kawagoe, H. Watanabe, Y. Totoki, T. Taylor, J. Weissenbach, R. Heilig, W. Saurin, F. Artiguenave, P. Brottier, T. Bruls, E. Pelletier, C. Robert, P. Wincker, D. R. Smith, L. Doucette-Stamm, M. Rubenfield, K. Weinstock, H. M. Lee, J. Dubois, A. Rosenthal, M. Platzer, G. Nyakatura, S. Taudien, A. Rump, H. Yang, J. Yu, J. Wang, G. Huang, J. Gu, L. Hood, L. Rowen, A. Madan, S. Qin, R. W. Davis, N. A. Federspiel, A. P. Abola, M. J. Proctor, R. M. Myers, J. Schmutz, M. Dickson, J. Grimwood, D. R. Cox, M. V. Olson, R. Kaul, C. Raymond, N. Shimizu, K. Kawasaki, S. Minoshima, G. A. Evans, M. Athanasiou, R. Schultz, B. A. Roe, F. Chen, H. Pan, J. Ramser, H. Lehrach, R. Reinhardt, W. R. McCombie, M. de la Bastide, N. Dedhia, H. Blöcker, K. Hornischer, G. Nordsiek, R. Agarwala, L. Aravind, J. A. Bailey, A. Bateman, S. Batzoglou, E. Birney, P. Bork, D. G. Brown, C. B. Burge, L. Cerutti, H. C. Chen, D. Church, M. Clamp, R. R. Copley, T. Doerks, S. R. Eddy, E. E. Eichler, T. S. Furey, J. Galagan, J. G. Gilbert, C. Harmon, Y. Hayashizaki, D. Haussler, H. Hermjakob, K. Hokamp, W. Jang, L. S. Johnson, T. A. Jones, S. Kasif, A. Kaspryzk, S. Kennedy, W. J. Kent, P. Kitts, E. V. Koonin, I. Korf, D. Kulp, D. Lancet, T. M. Lowe, A. McLysaght, T. Mikkelsen, J. V. Moran, N. Mulder, V. J. Pollara, C. P. Ponting, G. Schuler, J. Schultz, G. Slater, A. F. Smit, E. Stupka, J. Szustakowski, D. Thierry-Mieg, J. Thierry-Mieg, L. Wagner, J. Wallis, R. Wheeler, A. Williams, Y. I. Wolf, K. H. Wolfe, S. P. Yang, R. F. Yeh, F. Collins, M. S. Guyer, J. Pe-

terson, A. Felsenfeld, K. A. Wetterstrand, A. Patrinos, M. J. Morgan, P. de Jong, J. J. Catanese, K. Osoegawa, H. Shizuya, S. Choi, Y. J. Chen, J. Szustakowki, and International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409(6822):860–921, Feb 2001.

[45] Heebal Kim, Robert Klein, Jacek Majewski, and Jurg Ott. Estimating rates of alternative splicing in mammals and invertebrates. *Nat Genet*, 36(9):915–6; author reply 916–7, Sep 2004.

[46] Eddo Kim, Amir Goren, and Gil Ast. Alternative splicing: current perspectives. *Bioessays*, 30(1):38–47, Jan 2008.

[47] Jamal Tazi, Nadia Bakkour, and Stefan Stamm. Alternative splicing and disease. *Biochim Biophys Acta*, 1792(1):14–26, Jan 2009.

[48] Nuno André Faustino and Thomas A Cooper. Pre-mrna splicing and human disease. *Genes Dev*, 17(4):419–437, Feb 2003.

[49] Reini F Luco and Tom Misteli. More than a splicing code: integrating the role of rna, chromatin and non-coding rna in alternative splicing regulation. *Curr Opin Genet Dev*, 21(4):366–372, Aug 2011.

[50] Delphine Bernard, Kannanganattu V Prasanth, Vidisha Tripathi, Sabrina Colasse, Tetsuya Nakamura, Zhenyu Xuan, Michael Q Zhang, Frédéric Sedel, Laurent Jourdren, Fanny Coulpier, Antoine Triller, David L Spector, and Alain Bessis. A long nuclear-retained non-coding rna regulates synaptogenesis by modulating gene expression. *EMBO J*, 29(18):3082–3093, Sep 2010.

[51] Ping Ji, Sven Diederichs, Wenbing Wang, Sebastian Böing, Ralf Metzger, Paul M Schneider, Nicola Tidow, Burkhard Brandt, Horst Buerger, Etmar Bulk, Michael Thomas, Wolfgang E Berdel, Hubert Serve, and Carsten Müller-Tidow. Malat-1, a novel noncoding rna, and thymosin beta4 predict metastasis and survival in early-stage non-small cell lung cancer. *Oncogene*, 22(39):8031–8041, Sep 2003.

[52] Shivendra Kishore, Amit Khanna, Zhaiyi Zhang, Jingyi Hui, Piotr J Balwierz, Mihaela Stefan, Carol Beach, Robert D Nicholls, Mihaela Zavolan, and Stefan Stamm. The snorna mbii-52 (snord 115) is processed into smaller rnas and regulates alternative splicing. *Hum Mol Genet*, 19(7):1153–1164, Apr 2010.

[53] W. James Kent. Blat–the blast-like alignment tool. *Genome Res*, 12(4):656–664, Apr 2002.

[54] Ali Mortazavi, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. Mapping and quantifying mammalian transcriptomes by rna-seq. *Nat Methods*, 5(7):621–628, Jul 2008.

[55] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, Aug 2009.

[56] Guojie Zhang, Guangwu Guo, Xueda Hu, Yong Zhang, Qiye Li, Ruiqiang Li, Ruhong Zhuang, Zhike Lu, Zengquan He, Xiaodong Fang, Li Chen, Wei Tian, Yong Tao, Karsten Kristiansen, Xiuqing Zhang, Songgang Li, Huanming Yang, Jian Wang, and Jun Wang. Deep rna sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*, 20(5):646–654, May 2010.

[57] Ron Edgar, Michael Domrachev, and Alex E Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Res*, 30(1):207–210, Jan 2002.

[58] Kagnew Hailesellasse Sene, Christopher J Porter, Gareth Palidwor, Carolina Perez-Iratxeta, Enrique M Muro, Pearl A Campbell, Michael A Rudnicki, and Miguel A Andrade-Navarro. Gene function in early mouse embryonic stem cell differentiation. *BMC Genomics*, 8:85, 2007.

[59] Kerstin Haase. Regulatory influence of small nuclear RNAs on splicing. *Diploma Thesis*, pages 1–96, 2011.

[60] Lioudmila V Sharova, Alexei A Sharov, Timur Nedorezov, Yulan Piao, Nabeebi Shaik, and Minoru S H Ko. Database for mrna half-life of 19 977 genes obtained by dna microarray analysis of pluripotent and differentiating mouse embryonic stem cells. *DNA Res*, 16(1):45–58, Feb 2009.

[61] Ying Huang, Beifang Niu, Ying Gao, Limin Fu, and Weizhong Li. Cd-hit suite: a web server for clustering and comparing biological sequences. *Bioinformatics*, 26(5):680–682, Mar 2010.

[62] Inbal Paz, Martin Akerman, Iris Dror, Idit Kosti, and Yael Mandel-Gutfreund. Sfmap: a web server for motif analysis and prediction of splicing factor binding sites. *Nucleic Acids Res*, 38(Web Server issue):W281–W285, Jul 2010.

[63] Martin Akerman, Hilda David-Eden, Ron Y Pinter, and Yael Mandel-Gutfreund. A computational approach for genome-wide mapping of splicing factor binding sites. *Genome Biol*, 10(3):R30, 2009.

[64] Joonhee Han, Ji Xiong, Dong Wang, and Xiang-Dong Fu. Pre-mrna splicing: where and when in the nucleus. *Trends Cell Biol*, 21(6):336–343, Jun 2011.

[65] Alicia Oshlack, Mark D Robinson, and Matthew D Young. From rna-seq reads to differential expression results. *Genome Biol*, 11(12):220, 2010.

[66] Keith R Bradnam and Ian Korf. Longer first introns are a general property of eukaryotic gene structure. *PLoS One*, 3(8):e3093, 2008.

[67] Jian-Hua Yang, Peng Shao, Hui Zhou, Yue-Qin Chen, and Liang-Hu Qu. deepbase: a database for deeply annotating and mining deep sequencing data. *Nucleic Acids Res*, 38(Database issue):D123–D130, Jan 2010.

[68] Richard P Allen, Daniel Picchietti, Wayne A Hening, Claudia Trenkwalder, Arthur S Walters, Jacques Montplaisi, Restless Legs Syndrome Diagnosis, Epidemiology workshop at the National Institutes of Health, and International Restless Legs Syndrome Study Group. Restless legs syndrome: diagnostic criteria, special considerations, and epidemiology. a report from the restless legs syndrome diagnosis and epidemiology workshop at the national institutes of health. *Sleep Med*, 4(2):101–119, Mar 2003.

[69] Karl Ekbom and J. Ulfberg. Restless legs syndrome. *J Intern Med*, 266(5):419–431, Nov 2009.

[70] Claudia Trenkwalder, Walter Paulus, and Arthur S Walters. The restless legs syndrome. *Lancet Neurol*, 4(8):465–475, Aug 2005.

[71] Ruth Byrne, Smita Sinha, and K. Ray Chaudhuri. Restless legs syndrome: diagnosis and review of management options. *Neuropsychiatr Dis Treat*, 2(2):155–164, Jun 2006.

[72] Willis Thomas. De anima brutorum. *London: Wells and Scott*, 1672.

[73] Ekbom Karl-Axel. Restless legs. *Acta Med Scand*, 158:4–124, 1945.

[74] Yoakum R. Night walkers: do your legs seem to have a life of their own? your torment has a name. *Modern Maturity*, 55:82–4, 1994.

[75] A. S. Walters. Toward a better definition of the restless legs syndrome. the international restless legs syndrome study group. *Mov Disord*, 10(5):634–642, Sep 1995.

[76] M. Michaud, A. Chabli, G. Lavigne, and J. Montplaisir. Arm restlessness in patients with restless legs syndrome. *Mov Disord*, 15(2):289–293, Mar 2000.

[77] Sevda Erer, Necdet Karli, Mehmet Zarifoglu, Alis Ozcakir, and Demet Yildiz. The prevalence and clinical features of restless legs syndrome: a door to door population study in orhangazi, bursa in turkey. *Neurol India*, 57(6):729–733, 2009.

[78] Arthur S Walters, Cheryl LeBrocq, Anjana Dhar, Wayne Hening, Ray Rosen, Richard P Allen, Claudia Trenkwalder, and International Restless Legs Syndrome Study Group. Validation of the international restless legs syndrome study group rating scale for restless legs syndrome. *Sleep Med*, 4(2):121–132, Mar 2003.

[79] Suresh Kotagal and Michael H Silber. Childhood-onset restless legs syndrome. *Ann Neurol*, 56(6):803–807, Dec 2004.

[80] A. S. Walters, D. L. Picchietti, B. L. Ehrenberg, and M. L. Wagner. Restless legs syndrome in childhood and adolescence. *Pediatr Neurol*, 11(3):241–245, Oct 1994.

[81] S. Whittom, Y. Dauvilliers, M-H. Pennestri, F. Vercauteren, N. Molinari, D. Petit, and J. Montplaisir. Age-at-onset in restless legs syndrome: a clinical and polysomnographic study. *Sleep Med*, 9(1):54–59, Dec 2007.

[82] J. Winkelmann, T. C. Wetter, V. Collado-Seidel, T. Gasser, M. Dichgans, A. Yassouridis, and C. Trenkwalder. Clinical characteristics and frequency of the hereditary restless legs syndrome in a population of 300 patients. *Sleep*, 23(5):597–602, Aug 2000.

[83] A. Desautels, G. Turecki, J. Montplaisir, A. Sequeira, A. Verner, and G. A. Rouleau. Identification of a major susceptibility locus for restless legs syndrome on chromosome 12q. *Am J Hum Genet*, 69(6):1266–1270, Dec 2001.

[84] Alex Desautels, Gustavo Turecki, Jacques Montplaisir, Lan Xiong, Arthur S Walters, Bruce L Ehrenberg, Kateri Brisebois, Amelie K Desautels, Yves Gingras, William G Johnson, Elio Lugaresi, Giorgio Coccagna, Daniel L Picchietti, Alice Lazzarini, and Guy A Rouleau. Restless legs syndrome: confirmation of linkage to chromosome 12q, genetic heterogeneity, and evidence of complexity. *Arch Neurol*, 62(4):591–596, Apr 2005.

[85] Juliane Winkelmann, Darina Czamara, Barbara Schormair, Franziska Knauf, Eva C Schulte, Claudia Trenkwalder, Yves Dauvilliers, Olli Polo, Birgit Högl, Klaus Berger, Andrea Fuhs, Nadine Gross, Karin Stiasny-Kolster, Wolfgang Oertel, Cornelius G Bachmann, Walter Paulus, Lan Xiong, Jacques Montplaisir, Guy A Rouleau, Ingo Fietze, Jana Vávrová, David Kemlink, Karel Sonka, Sona Nevsimalova, Siong-Chi Lin, Zbigniew Wszolek, Carles Vilariño-Güell, Matthew J Farrer, Viola Gschliesser, Birgit Frauscher, Tina Falkenstetter, Werner Poewe,

Richard P Allen, Christopher J Earley, William G Ondo, Wei-Dong Le, Derek Spieler, Maria Kaffe, Alexander Zimprich, Johannes Kettunen, Markus Perola, Kaisa Silander, Isabelle Cournu-Rebeix, Marcella Francavilla, Claire Fontenille, Bertrand Fontaine, Pavel Vodicka, Holger Prokisch, Peter Lichtner, Paul Peppard, Juliette Faraco, Emmanuel Mignot, Christian Gieger, Thomas Illig, H-Erich Wichmann, Bertram Müller-Myhsok, and Thomas Meitinger. Genome-wide association study identifies novel restless legs syndrome susceptibility loci on 2p14 and 16q12.1. *PLoS Genet*, 7(7):e1002171, Jul 2011.

[86] Sunad Rangarajan and George Albert D'Souza. Restless legs syndrome in indian patients having iron deficiency anemia in a tertiary care hospital. *Sleep Med*, 8(3):247–251, Apr 2007.

[87] Michael J Bamshad, Sarah B Ng, Abigail W Bigham, Holly K Tabor, Mary J Emond, Deborah A Nickerson, and Jay Shendure. Exome sequencing as a tool for mendelian disease gene discovery. *Nat Rev Genet*, 12(11):745–755, Nov 2011.

[88] Chee-Seng Ku, David N Cooper, Constantin Polychronakos, Nasheen Naidoo, Mengchu Wu, and Richie Soong. Exome sequencing: dual role as a discovery and diagnostic tool. *Ann Neurol*, 71(1):5–14, Jan 2012.

[89] Jacek Majewski, Jeremy Schwartzentruber, Emilie Lalonde, Alexandre Montpetit, and Nada Jabado. What can exome sequencing do for you? *J Med Genet*, 48(9):580–589, Sep 2011.

[90] Murim Choi, Ute I Scholl, Weizhen Ji, Tiewen Liu, Irina R Tikhonova, Paul Zumbo, Ahmet Nayir, Ayşin Bakkaloğlu, Seza Ozen, Sami Sanjad, Carol Nelson-Williams, Anita Farhi, Shrikant Mane, and Richard P Lifton. Genetic diagnosis by whole exome capture and massively parallel dna sequencing. *Proc Natl Acad Sci U S A*, 106(45):19096–19101, Nov 2009.

[91] Heng Li and Richard Durbin. Fast and accurate short read alignment with burrows-wheeler transform. *Bioinformatics*, 25(14):1754–1760, Jul 2009.

[92] Kai Wang, Mingyao Li, and Hakon Hakonarson. Annovar: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res*, 38(16):e164, Sep 2010.

[93] Pauline C Ng and Steven Henikoff. Sift: Predicting amino acid changes that affect protein function. *Nucleic Acids Res*, 31(13):3812–3814, Jul 2003.

[94] Juliane Winkelmann, Peter Lichtner, Barbara Schormair, Manfred Uhr, Stephanie Hauk, Karin Stiasny-Kolster, Claudia Trenkwalder, Walter Paulus, Ines Peglau,

Ilonka Eisensehr, Thomas Illig, H-Erich Wichmann, Hildegard Pfister, Jelena Golic, Thomas Bettecken, Benno Pütz, Florian Holsboer, Thomas Meitinger, and Bertram Müller-Myhsok. Variants in the neuronal nitric oxide synthase (nnos, nos1) gene are associated with restless legs syndrome. *Mov Disord*, 23(3):350–358, Feb 2008.

[95] Barbara Schormair, David Kemlink, Darina Roeske, Gertrud Eckstein, Lan Xiong, Peter Lichtner, Stephan Ripke, Claudia Trenkwalder, Alexander Zimprich, Karin Stiasny-Kolster, Wolfgang Oertel, Cornelius G Bachmann, Walter Paulus, Birgit Högl, Birgit Frauscher, Viola Gschliesser, Werner Poewe, Ines Peglau, Pavel Vodicka, Jana Vávrová, Karel Sonka, Sona Nevsimalova, Jacques Montplaisir, Gustavo Turecki, Guy Rouleau, Christian Gieger, Thomas Illig, H-Erich Wichmann, Florian Holsboer, Bertram Müller-Myhsok, Thomas Meitinger, and Juliane Winkelmann. Ptprd (protein tyrosine phosphatase receptor type delta) is associated with restless legs syndrome. *Nat Genet*, 40(8):946–948, Aug 2008.

[96] Juliane Winkelmann, Barbara Schormair, Peter Lichtner, Stephan Ripke, Lan Xiong, Shapour Jalilzadeh, Stephany Fulda, Benno Pütz, Gertrud Eckstein, Stephanie Hauk, Claudia Trenkwalder, Alexander Zimprich, Karin Stiasny-Kolster, Wolfgang Oertel, Cornelius G Bachmann, Walter Paulus, Ines Peglau, Ilonka Eisensehr, Jacques Montplaisir, Gustavo Turecki, Guy Rouleau, Christian Gieger, Thomas Illig, H-Erich Wichmann, Florian Holsboer, Bertram Müller-Myhsok, and Thomas Meitinger. Genome-wide association study of restless legs syndrome identifies common variants in three genomic regions. *Nat Genet*, 39(8):1000–1006, Aug 2007.

[97] Sarah B Ng, Abigail W Bigham, Kati J Buckingham, Mark C Hannibal, Margaret J McMillin, Heidi I Gildersleeve, Anita E Beck, Holly K Tabor, Gregory M Cooper, Heather C Mefford, Choli Lee, Emily H Turner, Joshua D Smith, Mark J Rieder, Koh-Ichiro Yoshiura, Naomichi Matsumoto, Tohru Ohta, Norio Niikawa, Deborah A Nickerson, Michael J Bamshad, and Jay Shendure. Exome sequencing identifies mll2 mutations as a cause of kabuki syndrome. *Nat Genet*, 42(9):790–793, Sep 2010.

[98] J. E. Tobin, J. C. Latourelle, M. F. Lew, C. Klein, O. Suchowersky, H. A. Shill, L. I. Golbe, M. H. Mark, J. H. Growdon, G. F. Wooten, B. A. Racette, J. S. Perlmutter, R. Watts, M. Guttman, K. B. Baker, S. Goldwurm, G. Pezzoli, C. Singer, M. H. Saint-Hilaire, A. E. Hendricks, S. Williamson, M. W. Nagle, J. B. Wilk, T. Massood, J. M. Laramie, A. L. DeStefano, I. Litvan, G. Nicholson, A. Corbett, S. Isaacson, D. J. Burn, P. F. Chinnery, P. P. Pramstaller, S. Sherman, J. Al-hinti, E. Drasby, M. Nance, A. T. Moller, K. Ostergaard, R. Roxburgh, B. Snow, J. T.

Slevin, F. Cambi, J. F. Gusella, and R. H. Myers. Haplotypes and gene expression implicate the mapt region for parkinson disease: the genepd study. *Neurology*, 71(1):28–34, Jul 2008.

[99] Thomas Gasser. Genetics of parkinson's disease. *Curr Opin Neurol*, 18(4):363–369, Aug 2005.

[100] Joaquim J Ferreira, Leonor Correia Guedes, Mário Miguel Rosa, Miguel Coelho, Marina van Doeselaar, Dorothea Schweiger, Alessio Di Fonzo, Ben A Oostra, Cristina Sampaio, and Vincenzo Bonifati. High prevalence of lrrk2 mutations in familial and sporadic parkinson's disease in portugal. *Mov Disord*, 22(8):1194–1201, Jun 2007.

[101] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *J Mol Biol*, 215(3):403–410, Oct 1990.

[102] Stephen M Rumble, Phil Lacroute, Adrian V Dalca, Marc Fiume, Arend Sidow, and Michael Brudno. Shrimp: accurate mapping of short color-space reads. *PLoS Comput Biol*, 5(5):e1000386, May 2009.

[103] Gerton Lunter and Martin Goodson. Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Res*, 21(6):936–939, Jun 2011.

[104] Ben Langmead, Cole Trapnell, Mihai Pop, and Steven L Salzberg. Ultrafast and memory-efficient alignment of short dna sequences to the human genome. *Genome Biol*, 10(3):R25, 2009.

[105] Kai Wang, Darshan Singh, Zheng Zeng, Stephen J Coleman, Yan Huang, Gleb L Savich, Xiaping He, Piotr Mieczkowski, Sara A Grimm, Charles M Perou, James N MacLeod, Derek Y Chiang, Jan F Prins, and Jinze Liu. Mapsplice: accurate mapping of rna-seq reads for splice junction discovery. *Nucleic Acids Res*, 38(18):e178, Oct 2010.

[106] Kin Fai Au, Hui Jiang, Lan Lin, Yi Xing, and Wing Hung Wong. Detection of splice junctions from paired-end rna-seq data by splicemap. *Nucleic Acids Res*, 38(14):4570–4578, Aug 2010.

[107] Thomas D Wu and Serban Nacu. Fast and snp-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, Apr 2010.

[108] Fabio De Bona, Stephan Ossowski, Korbinian Schneeberger, and Gunnar Rätsch. Optimal spliced alignments of short sequence reads. *Bioinformatics*, 24(16):i174–i180, Aug 2008.

[109] France Denoeud, Jean-Marc Aury, Corinne Da Silva, Benjamin Noel, Odile Rogier, Massimo Delledonne, Michele Morgante, Giorgio Valle, Patrick Wincker, Claude Scarpelli, Olivier Jaillon, and François Artiguenave. Annotating genomes with massive-scale rna sequencing. *Genome Biol*, 9(12):R175, 2008.

[110] Daniel R Zerbino and Ewan Birney. Velvet: algorithms for de novo short read assembly using de bruijn graphs. *Genome Res*, 18(5):821–829, May 2008.

[111] Gordon Robertson, Jacqueline Schein, Readman Chiu, Richard Corbett, Matthew Field, Shaun D Jackman, Karen Mungall, Sam Lee, Hisanaga Mark Okada, Jenny Q Qian, Malachi Griffith, Anthony Raymond, Nina Thiessen, Timothee Cezard, Yaron S Butterfield, Richard Newsome, Simon K Chan, Rong She, Richard Varhol, Baljit Kamoh, Anna-Liisa Prabhu, Angela Tam, YongJun Zhao, Richard A Moore, Martin Hirst, Marco A Marra, Steven J M Jones, Pamela A Hoodless, and Inanc Birol. De novo assembly and analysis of rna-seq data. *Nat Methods*, 7(11):909–912, Nov 2010.

[112] Malachi Griffith, Obi L Griffith, Jill Mwenifumbo, Rodrigo Goya, A. Sorana Morrissy, Ryan D Morin, Richard Corbett, Michelle J Tang, Ying-Chen Hou, Trevor J Pugh, Gordon Robertson, Suganthi Chittaranjan, Adrian Ally, Jennifer K Asano, Susanna Y Chan, Haiyan I Li, Helen McDonald, Kevin Teague, Yongjun Zhao, Thomas Zeng, Allen Delaney, Martin Hirst, Gregg B Morin, Steven J M Jones, Isabella T Tai, and Marco A Marra. Alternative expression analysis by rna sequencing. *Nat Methods*, 7(10):843–847, Oct 2010.

[113] Soohyun Lee, Chae Hwa Seo, Byungho Lim, Jin Ok Yang, Jeongsu Oh, Minjin Kim, Sooncheol Lee, Byungwook Lee, Changwon Kang, and Sanghyuk Lee. Accurate quantification of transcriptome from rna-seq data by effective length normalization. *Nucleic Acids Res*, 39(2):e9, Jan 2011.

[114] Bo Li and Colin N Dewey. Rsem: accurate transcript quantification from rna-seq data with or without a reference genome. *BMC Bioinformatics*, 12:323, 2011.

[115] Likun Wang, Zhixing Feng, Xi Wang, Xiaowo Wang, and Xuegong Zhang. Degseq: an r package for identifying differentially expressed genes from rna-seq data. *Bioinformatics*, 26(1):136–138, Jan 2010.

[116] Mark D Robinson, Davis J McCarthy, and Gordon K Smyth. edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, Jan 2010.

[117] Simon Anders and Wolfgang Huber. Differential expression analysis for sequence count data. *Genome Biol*, 11(10):R106, 2010.

[118] Ben Langmead, Kasper D Hansen, and Jeffrey T Leek. Cloud-scale rna-sequencing differential expression analysis with myrna. *Genome Biol*, 11(8):R83, 2010.

[119] Kazumi Yamada, Junko Kano, Hajime Tsunoda, Hiroyuki Yoshikawa, Chigusa Okubo, Tadashi Ishiyama, and Masayuki Noguchi. Phenotypic characterization of endometrial stromal sarcoma of the uterus. *Cancer Sci*, 97(2):106–112, Feb 2006.

[120] R. Lin, S. Maeda, C. Liu, M. Karin, and T. S. Edgington. A large noncoding rna is a marker for murine hepatocellular carcinomas and a spectrum of human carcinomas. *Oncogene*, 26(6):851–858, Feb 2007.

[121] Jian-Hua Luo, Baoguo Ren, Sergei Keryanov, George C Tseng, Uma N M Rao, Satdarshan P Monga, Steven Strom, Anthony J Demetris, Michael Nalesnik, Yan P Yu, Sarangarajan Ranganathan, and George K Michalopoulos. Transcriptomic and genomic analysis of human hepatocellular carcinomas and hepatoblastomas. *Hepatology*, 44(4):1012–1024, Oct 2006.

[122] Fengjie Guo, Yalin Li, Yan Liu, Jiajia Wang, Yuehui Li, and Guancheng Li. Inhibition of metastasis-associated lung adenocarcinoma transcript 1 in caski human cervical cancer cells suppresses cell proliferation and invasion. *Acta Biochim Biophys Sin (Shanghai)*, 42(3):224–229, Mar 2010.

[123] Joerg Fellenberg, Ludger Bernd, Guenter Delling, Daniela Witte, and Anita Zahlten-Hinguranage. Prognostic significance of drug-regulated genes in high-grade osteosarcoma. *Mod Pathol*, 20(10):1085–1094, Oct 2007.

[124] Taka aki Koshimizu, Yoko Fujiwara, Nobuya Sakai, Katsushi Shibata, and Hiroyoshi Tsuchiya. Oxytocin stimulates expression of a noncoding rna tumor marker in a human neuroblastoma cell line. *Life Sci*, 86(11-12):455–460, Mar 2010.

[125] Katrin Panzitt, Marisa M O Tschernatsch, Christian Guelly, Tarek Moustafa, Martin Stradner, Heimo M Strohmaier, Charles R Buck, Helmut Denk, Renée Schroeder, Michael Trauner, and Kurt Zatloukal. Characterization of hulc, a novel gene with striking up-regulation in hepatocellular carcinoma, as noncoding rna. *Gastroenterology*, 132(1):330–342, Jan 2007.

[126] Imad J Matouk, Ibrahim Abbasi, Abraham Hochberg, Eithan Galun, Hassan Dweik, and Mutaz Akkawi. Highly upregulated in liver cancer noncoding rna is overexpressed in hepatic colorectal metastasis. *Eur J Gastroenterol Hepatol*, 21(6):688–692, Jun 2009.

[127] W. Chen, W. Böcker, J. Brosius, and H. Tiedge. Expression of neural bc200 rna in human tumours. *J Pathol*, 183(3):345–351, Nov 1997.

[128] Anna Iacoangeli, Yuan Lin, Eric J Morley, Ilham A Muslimov, Riccardo Bianchi, James Reilly, Jeremy Weedon, Raihanatou Diallo, Werner Böcker, and Henri Tiedge. Bc200 rna in invasive and preinvasive breast cancer. *Carcinogenesis*, 25(11):2125–2133, Nov 2004.

[129] Anne Gabory, Hélène Jammes, and Luisa Dandolo. The h19 locus: role of an imprinted non-coding rna in growth and development. *Bioessays*, 32(6):473–480, Jun 2010.

[130] K. Hibi, H. Nakamura, A. Hirai, Y. Fujikake, Y. Kasai, S. Akiyama, K. Ito, and H. Takagi. Loss of h19 imprinting in esophageal cancer. *Cancer Res*, 56(3):480–482, Feb 1996.

[131] Imad J Matouk, Nathan DeGroot, Shaul Mezan, Suhail Ayesh, Rasha Abu-lail, Abraham Hochberg, and Eithan Galun. The h19 non-coding rna is essential for human tumor growth. *PLoS One*, 2(9):e845, 2007.

[132] I. Ariel, M. Sughayer, Y. Fellig, G. Pizov, S. Ayesh, D. Podeh, B. A. Libdeh, C. Levy, T. Birman, M. L. Tykocinski, N. de Groot, and A. Hochberg. The imprinted h19 gene is a marker of early recurrence in human bladder carcinoma. *Mol Pathol*, 53(6):320–323, Dec 2000.

[133] C. M. Yballe, T. H. Vu, and A. R. Hoffman. Imprinting and expression of insulin-like growth factor-ii and h19 in normal breast tissue and breast tumor. *J Clin Endocrinol Metab*, 81(4):1607–1612, Apr 1996.

[134] V. Tanos, I. Ariel, D. Prus, N. De-Groot, and A. Hochberg. H19 and igf2 gene expression in human normal, hyperplastic, and malignant endometrium. *Int J Gynecol Cancer*, 14(3):521–525, 2004.

[135] S. Douc-Rasy, M. Barrois, S. Fogel, J. C. Ahomadegbe, D. Stéhelin, J. Coll, and G. Riou. High incidence of loss of heterozygosity and abnormal imprinting of h19 and igf2 genes in invasive cervical carcinomas. uncoupling of h19 and igf2 expression and biallelic hypomethylation of h19. *Oncogene*, 12(2):423–430, Jan 1996.

[136] T. Arima, T. Matsuda, N. Takagi, and N. Wake. Association of igf2 and h19 imprinting with choriocarcinoma development. *Cancer Genet Cytogenet*, 93(1):39–47, Jan 1997.

[137] V. Tanos, D. Prus, S. Ayesh, D. Weinstein, M. L. Tykocinski, N. De-Groot, A. Hochberg, and I. Ariel. Expression of the imprinted h19 oncofetal rna in epithelial ovarian cancer. *Eur J Obstet Gynecol Reprod Biol*, 85(1):7–11, Jul 1999.

[138] N. Berteaux, S. Lottin, E. Adriaenssens, F. Van Coppenolle, F. Van Coppennolle, X. Leroy, J. Coll, T. Dugimont, and J-J. Curgy. Hormonal regulation of h19 gene expression in prostate epithelial cells. *J Endocrinol*, 183(1):69–78, Oct 2004.

[139] M. Kondo and T. Takahashi. [altered genomic imprinting in the igf2 and h19 genes in human lung cancer]. *Nihon Rinsho*, 54(2):492–496, Feb 1996.

[140] Peggy S Eis, Wayne Tam, Liping Sun, Amy Chadburn, Zongdong Li, Mario F Gomez, Elsebet Lund, and James E Dahlberg. Accumulation of mir-155 and bic rna in human b cell lymphomas. *Proc Natl Acad Sci U S A*, 102(10):3627–3632, Mar 2005.

[141] Ivan Pasic, Adam Shlien, Adam D Durbin, Dimitrios J Stavropoulos, Berivan Baskin, Peter N Ray, Ana Novokmet, and David Malkin. Recurrent focal copy-number changes and loss of heterozygosity implicate two noncoding rnas and one tumor suppressor gene at chromosome 3q13.31 in osteosarcoma. *Cancer Res*, 70(1):160–171, Jan 2010.

[142] Gyorgy Petrovics, Wei Zhang, Mazen Makarem, Jesse P Street, Roger Connelly, Leon Sun, Isabell A Sesterhenn, Vasantha Srikantan, Judd W Moul, and Shiv Srivastava. Elevated expression of pcgem1, a prostate-specific gene with cell growth-promoting function, is associated with high-risk prostate cancer patients. *Oncogene*, 23(2):605–611, Jan 2004.

[143] V. Srikantan, Z. Zou, G. Petrovics, L. Xu, M. Augustus, L. Davis, J. R. Livezey, T. Connell, I. A. Sesterhenn, K. Yoshino, G. S. Buzard, F. K. Mostofi, D. G. McLeod, J. W. Moul, and S. Srivastava. Pcgem1, a prostate-specific gene, is overexpressed in prostate cancer. *Proc Natl Acad Sci U S A*, 97(22):12216–12221, Oct 2000.

[144] Xiaoqin Fu, Lakshmi Ravindranath, Nicholas Tran, Gyorgy Petrovics, and Shiv Srivastava. Regulation of apoptosis by a prostate-specific and prostate cancer-associated noncoding gene, pcgem1. *DNA Cell Biol*, 25(3):135–141, Mar 2006.

[145] Xiao-Song Wang, Zheng Zhang, Hong-Cheng Wang, Jian-Liang Cai, Qing-Wen Xu, Meng-Qiang Li, Yi-Cheng Chen, Xiao-Ping Qian, Tian-Jing Lu, Li-Zhang Yu, Yu Zhang, Dian-Qi Xin, Yan-Qun Na, and Wei-Feng Chen. Rapid identification of uca1 as a very sensitive and specific unique marker for human bladder carcinoma. *Clin Cancer Res*, 12(16):4851–4858, Aug 2006.

[146] Fan Wang, Xu Li, XiaoJuan Xie, Le Zhao, and Wei Chen. Uca1, a non-protein-coding rna up-regulated in bladder carcinoma and embryo, influencing cell growth and promoting invasion. *FEBS Lett*, 582(13):1919–1927, Jun 2008.

[147] Wing Pui Tsang, Timothy W L Wong, Albert H H Cheung, Chloe N N Co, and Tim Tak Kwok. Induction of drug resistance and transformation in human cancer cells by the noncoding rna cudr. *RNA*, 13(6):890–898, Jun 2007.

[148] M. J. Bussemakers, A. van Bokhoven, G. W. Verhaegh, F. P. Smit, H. F. Karthaus, J. A. Schalken, F. M. Debruyne, N. Ru, and W. B. Isaacs. Dd3: a new prostate-specific gene, highly overexpressed in prostate cancer. *Cancer Res*, 59(23):5975–5979, Dec 1999.

[149] Jacques B de Kok, Gerald W Verhaegh, Rian W Roelofs, Daphne Hessels, Lambertus A Kiemeney, Tilly W Aalders, Dorine W Swinkels, and Jack A Schalken. Dd3(pca3), a very sensitive and specific marker to detect prostate tumors. *Cancer Res*, 62(9):2695–2698, May 2002.

[150] Sergei A Korneev, Elena I Korneeva, Marya A Lagarkova, Sergei L Kiselev, Giles Critchley, and Michael O'Shea. Novel noncoding antisense rna transcribed from human anti-nos2a locus is differentially regulated during neuronal differentiation of embryonic stem cells. *RNA*, 14(10):2030–2037, Oct 2008.

[151] George A Calin, Chang gong Liu, Manuela Ferracin, Terry Hyslop, Riccardo Spizzo, Cinzia Sevignani, Muller Fabbri, Amelia Cimmino, Eun Joo Lee, Sylwia E Wojcik, Masayoshi Shimizu, Esmerina Tili, Simona Rossi, Cristian Taccioli, Flavia Pichiorri, Xiuping Liu, Simona Zupo, Vlad Herlea, Laura Gramantieri, Giovanni Lanza, Hansjuerg Alder, Laura Rassenti, Stefano Volinia, Thomas D Schmittgen, Thomas J Kipps, Massimo Negrini, and Carlo M Croce. Ultraconserved regions encoding ncrnas are altered in human leukemias and carcinomas. *Cancer Cell*, 12(3):215–229, Sep 2007.

[152] Chiara Braconi, Nicola Valeri, Takayuki Kogure, Pierluigi Gasparini, Nianyuan Huang, Gerard J Nuovo, Luigi Terracciano, Carlo M Croce, and Tushar Patel. Expression and functional role of a transcribed noncoding rna with an ultraconserved element in hepatocellular carcinoma. *Proc Natl Acad Sci U S A*, 108(2):786–791, Jan 2011.

[153] Wenqiang Yu, David Gius, Patrick Onyango, Kristi Muldoon-Jacobs, Judith Karp, Andrew P Feinberg, and Hengmi Cui. Epigenetic silencing of tumour suppressor gene p15 by its antisense rna. *Nature*, 451(7175):202–206, Jan 2008.

[154] Lasse Folkersen, Theodosios Kyriakou, Anuj Goel, John Peden, Anders Mälarstig, Gabrielle Paulsson-Berne, Anders Hamsten, Hugh Watkins, Anders Franco-Cereceda, Anders Gabrielsen, Per Eriksson, and P. R. O. C. A. R. D. I. S. consortia. Relationship between cad risk genotype in the chromosome 9p21 locus and gene

expression. identification of eight new anril splice variants. *PLoS One*, 4(11):e7677, 2009.

[155] Kyoko L Yap, Side Li, Ana M Muñoz-Cabello, Selina Raguz, Lei Zeng, Shiraz Mujtaba, Jesús Gil, Martin J Walsh, and Ming-Ming Zhou. Molecular interplay of the noncoding rna anril and methylated histone h3 lysine 27 by polycomb cbx7 in transcriptional silencing of ink4a. *Mol Cell*, 38(5):662–674, Jun 2010.

[156] Eric Pasmant, Ingrid Laurendeau, Delphine Héron, Michel Vidaud, Dominique Vidaud, and Ivan Bièche. Characterization of a germ-line deletion, including the entire ink4/arf locus, in a melanoma-neural system tumor family: identification of anril, an antisense noncoding rna whose expression coclusters with arf. *Cancer Res*, 67(8):3963–3969, Apr 2007.

[157] N. Miyoshi, H. Wagatsuma, S. Wakana, T. Shiroishi, M. Nomura, K. Aisaka, T. Kohda, M. A. Surani, T. Kaneko-Ishino, and F. Ishino. Identification of an imprinted gene, meg3/gtl2 and its human homologue meg3, first mapped on mouse distal chromosome 12 and human chromosome 14q. *Genes Cells*, 5(3):211–220, Mar 2000.

[158] Xun Zhang, Kimberley Rice, Yingying Wang, Wendy Chen, Ying Zhong, Yuki Nakayama, Yunli Zhou, and Anne Klibanski. Maternally expressed gene 3 (meg3) noncoding ribonucleic acid: isoform structure, expression, and functions. *Endocrinology*, 151(3):939–947, Mar 2010.

[159] Yunli Zhou, Ying Zhong, Yingying Wang, Xun Zhang, Dalia L Batista, Roger Gejman, Peter J Ansell, Jing Zhao, Catherine Weng, and Anne Klibanski. Activation of p53 by meg3 non-coding rna. *J Biol Chem*, 282(34):24731–24742, Aug 2007.

[160] M. Mourtada-Maarabouni, M. R. Pickard, V. L. Hedge, F. Farzaneh, and G. T. Williams. Gas5, a non-protein-coding rna, controls apoptosis and is downregulated in breast cancer. *Oncogene*, 28(2):195–208, Jan 2009.

[161] Etienne Leygue. Steroid receptor rna activator (sra1): unusual bifaceted gene products with suspected relevance to breast cancer. *Nucl Recept Signal*, 5:e006, 2007.

[162] Shilpa Chooniedass-Kothari, Mohammad Kariminia Hamedani, Sandy Troup, Florent Hubé, and Etienne Leygue. The steroid receptor rna activator protein is expressed in breast tumor tissues. *Int J Cancer*, 118(4):1054–1059, Feb 2006.

[163] Laura Poliseno, Leonardo Salmena, Jiangwen Zhang, Brett Carver, William J Haveman, and Pier Paolo Pandolfi. A coding-independent function of gene and

pseudogene mrnas regulates tumour biology. *Nature*, 465(7301):1033–1038, Jun 2010.

[164] Andrea Alimonti, Arkaitz Carracedo, John G Clohessy, Lloyd C Trotman, Caterina Nardella, Ainara Egia, Leonardo Salmena, Katia Sampieri, William J Haveman, Edi Brogi, Andrea L Richardson, Jiangwen Zhang, and Pier Paolo Pandolfi. Subtle variations in pten dose determine cancer susceptibility. *Nat Genet*, 42(5):454–458, May 2010.

[165] Yuyan Zhu, Meng Yu, Zhenhua Li, Chuize Kong, Jianbin Bi, Jun Li, Zeshou Gao, and Zeliang Li. ncran, a newly identified long noncoding rna, enhances human bladder tumor growth, invasion, and survival. *Urology*, 77(2):510.e1–510.e5, Feb 2011.

[166] Meng Yu, Miki Ohira, Yuanyuan Li, Hidetaka Niizuma, Myat Lin Oo, Yuyan Zhu, Toshinori Ozaki, Eriko Isogai, Yohko Nakamura, Tadayuki Koda, Shigeyuki Oba, Bingzhi Yu, and Akira Nakagawara. High expression of ncran, a novel non-coding rna mapped to chromosome 17q25.1, is associated with poor prognosis in neuroblastoma. *Int J Oncol*, 34(4):931–938, Apr 2009.

Appendix A

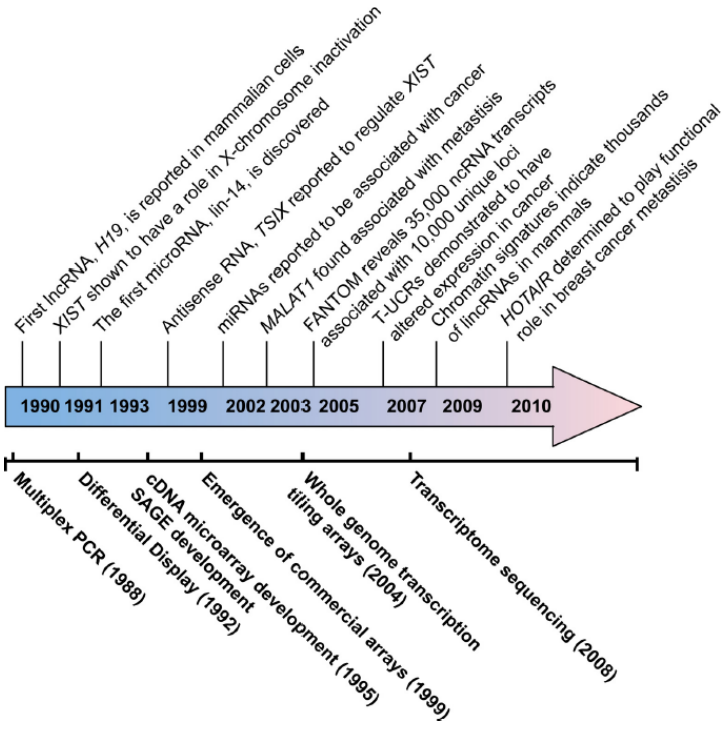


Figure A.1: Time scale of cancer-associated ncRNA in relation to technologies. This Figure is taken from the review of Gibb et al. [6]

Class	Category	Package
Read mapping		
Unspliced aligners	Seed methods	SHRiMP[102]
		Stampy[103]
	Burrows-Wheeler	Bowtie[104]
	transform methods	BWA[91]
Spliced aligners	Exon-first methods	MapSplice[105]
		SpliceMap[106]
		TopHat[35]
	Seed-extend methods	GSNAP[107]
		QPALMA[108]
Transcriptome reconstruction		
Genome-guided	Exon identification	G-Mo.R-Se[109]
reconstruction	Genome-guided	Scripture[1]
	assembly	Cufflinks[36]
Genome-	Genome-independent	Velvet[110]
idependent	assembly	TransABySS[111]
reconstruction		
Expression quantification		
Expression	Gene quantification	ALEXA-seq[112]
quantification		ERANGE[54]
		NEUMA[113]
	Isoform quantification	Cufflinks [36]
		MISO[23]
		Rsem[114]
Differential expression		Cuffdiff[36]
		Degseq[115]
		Edger[116]
		DESeq[117]
		Myrna[118]

Table A.1: Overview of RNA-seq analysis tools. This Table (Columns 1-3) is taken (slightly modified) from the review of Garber et al. [33]. As addressed by Garber et al. the RNA-seq analysis programs in this Table are not exhaustive.

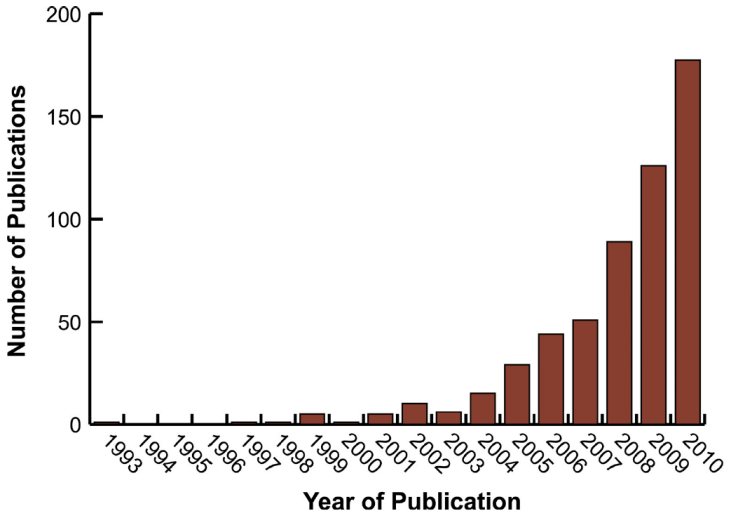


Figure A.2: Number of publications in relation to year of publications. This Figure is taken from the review of Gibb et al. [6]

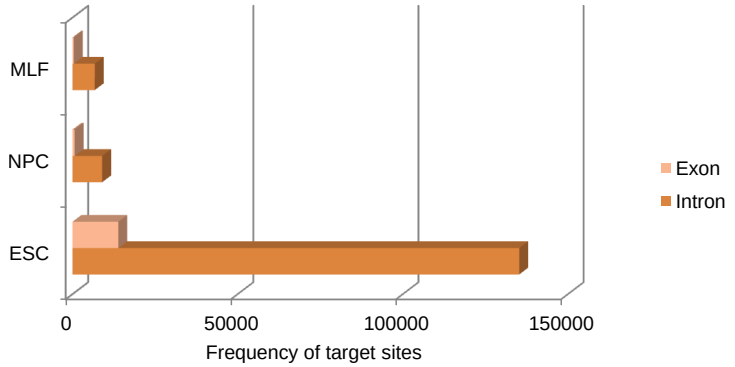


Figure A.3: Distribution of the frequency of lincRNA target sites in relation to the location on pre-mRNA along coding transcripts. We compared the frequencies of target sites located within introns to sites located within exons. Target sites and potential duplexes are enriched at introns of protein-coding genes.

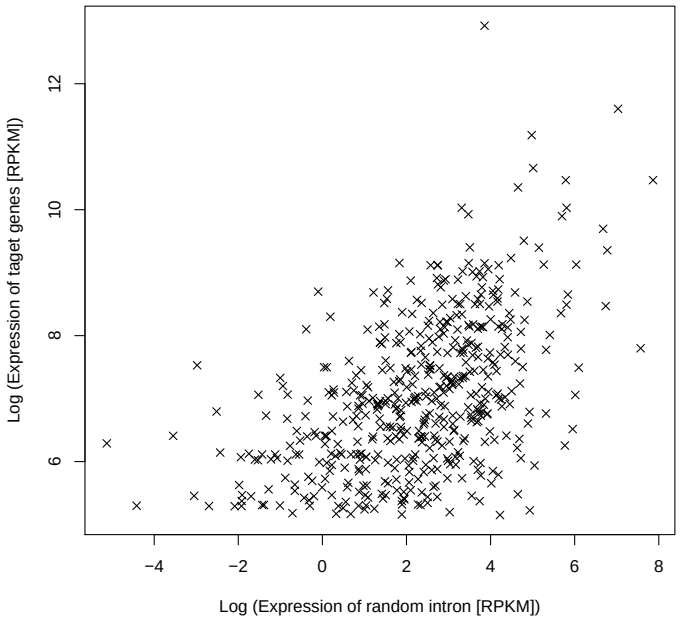
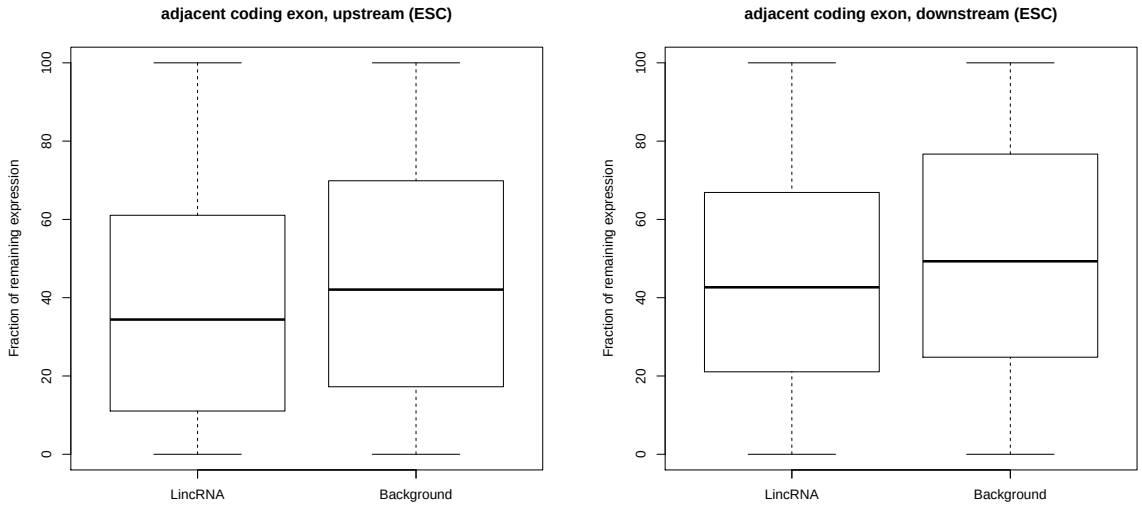


Figure A.4: Comparison of the expression of 1000 randomly sampled introns with pre-mRNA. Each data point describes one intron with the expression of the intron on the x-axis and the expression of pre-mRNA on the y-axis. As pre-mRNA we selected the maximal expressed coding exon of the associated non-targeted transcript. The expression as RPKM of 1000 randomly sampled introns is slightly correlated with the pre-mRNA. The Pearson correlation coefficient is 0.5.



(a) (b)

Figure A.5: Comparison of the fraction of remaining expression of adjacent exons of target sites with random introns (shown for the ESC cell line and stratified intron positions). The expression of adjacent coding exons of target sites (lincRNA) is significantly decreased. The comparison is shown for (a) upstream. (b) downstream exons. The p-value of the Wilcoxon rank sum test for upstream and downstream in ESC is $< 2.2e\text{-}16$. The median is shown as black line in each boxplot.

LncRNA	Size	Cytoband	Cancer types	References
HOTAIR	2158 nt	12q13.13	Breast	[17, 39]
MALAT1/α/ NEAT2	7,5 kb	11q13.1	Breast, lung, uterus, pancreas, colon, prostate, liver, cervix[1], neuroblastoma[1], osteosarcoma	[51, 119, 120, 121, 122, 123, 124]
HULC	500 nt	6p24.3	Liver, hepatic colorectal metastasis	[125, 126]
BC200	200 nt	2p21	Breast, cervix, esophagus, lung, ovary, parotid, tongue,	[127, 128]
H19	2.3 kb	11p15.5	Bladder, lung, liver, breast, endometrial, cervix, esophagus, ovary, prostate, choricarcinoma, colorectal	[28, 129, 130, 131, 37, 40, 132, 133, 134, 135, 136, 137, 138, 139]
BIC/MIRHG155/ MIRHG2	1.6 kb	21q11.2	B-cell lymphoma	[140]
PRNCR1	13kb	8q24.2	Prostate	[5]
LOC285194	2105 nt	3q13.31	Osteosarcoma	[141]
PCGEM1	1643 nt	2q32.2	Prostate	[142, 143, 144]
UCA1/CUDR	1.4 kb, 2.2 kb, 2.7 kb,	19p13.12	Bladder, colon, cervix, lung, thyroid, liver, breast, esophagus, stomach	[145, 146, 147]
DD3/PCA3	0.6 kb, 2 kb, 4 kb	9q21.22	Prostate	[148, 149]
anti-NOS2A	∼ 1.9 kb	17q23.2	Brain[1]	[150]
uc.73A	201 nt	2q22.3	Colon	[151]
TUC338 (encodes uc.338)	590 nt	12q13.13	Liver	[152]
ANRIL/p15AS/ CDK2BAS	34.8 kb & splice variants	9p21.3	Prostate, leukemia	[153, 154, 155, 156]
MEG3	1.6 kb & splicing isoforms	14q32.2	Brain (down-regulated)	[157, 158, 159]
GAS5/SNHG2	Multiple isoforms &	1q25.1	Breast (down-regulated)	[160]
SRA-1/SRA	1965 nt	5q31.3	Breast, uterus, ovary (hormone responsive tissue)	[161, 162]
PETENP1	∼ 3.9 kb	9p13.3	Prostate	[163, 164]
ncRAN	2186 nt, 2087 nt	17q25.1	Bladder, neuroblastoma	[165, 166]

Table A.2: Ovieview of reported cancer-associated lncRNAs. This Table is taken from the review of Gibb et al. [6]. [1]cell lines.

Abbreviation	Full name
easRNA	exon-associated small RNA
rasRNA	repeat-associated small RNA
pasRNA	promoter-associated small RNA
nasRNA	ncRNA-associated small RNAs
snoRNA	small nucleolar RNA
snRNA	small nuclear RNA
piRNA	piwi interacting RNA
miRNA	micro RNA

Table A.3: Overview of distinct types of small ncRNAs. Each row lists the information of one type of small ncRNA included in our work. This information is taken from DeepBase[1] and fRNAdb[2]. Column 1 represents the abbreviation and 2 the full name of a type.

Splicing factor	Motif
SF2ASF	crsmsgw, ugrwgvh
9G8	acgagagay, wggacra
SC35	gryymcyr, ugcygyy
Tra2alpha	gaagaggaag
Tra2beta	gaagaa, ghvvganr, aaguguu
SRp20	cuckucy, wcwwc
SRp40	yywcwsg
SRp55	yrcrkm
hnRNPA1	uagaca, uagagu, uagggw
hnRNPAB	auagca
hnRNPH/F	ggcgg, gggug, uguggg, uugggu
MBNL	ygcuky
NOVA1	ycay
PTB	cucucu, ucuu
CUG-BP	ugcug
YB1	caaccacaa
FOX1	ugcaug

Table A.4: Overview of the splicing factors and motifs searched by the tool SFmap [62]. The list of SF-binding motifs was previously described and taken from the work of [63]. Each row describes one type of splicing factor. Column 1 represents the abbreviation and 2 the sequence motif of a factor.

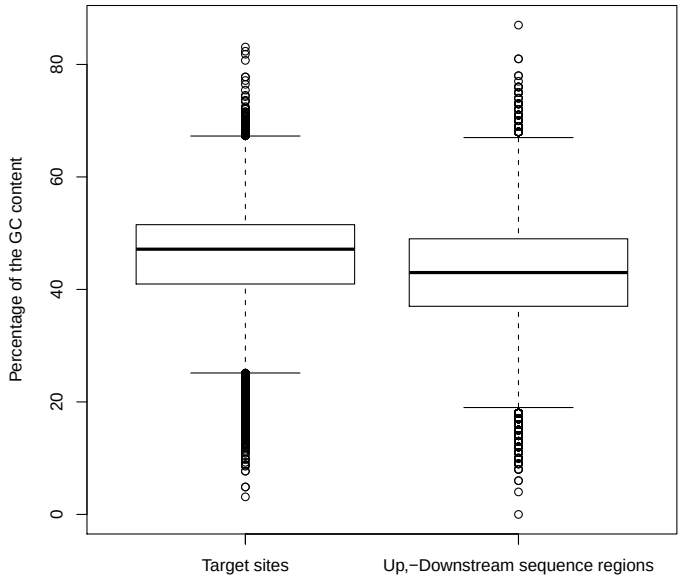
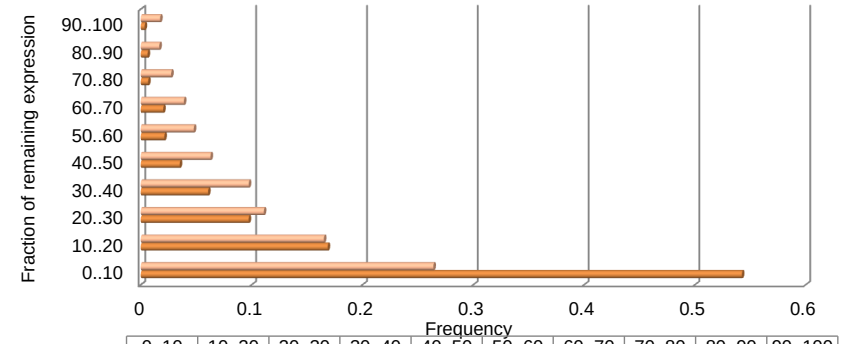


Figure A.6: Comparison of the percentage of the GC content of target sites with surrounding sequence regions. The GC content is significantly increased within target sites. The p-value of the Wilcoxon rank sum test for upstream and downstream in ESC is < 2.2e-16. The median is shown as black line in each boxplot.



	0..10	10..20	20..30	30..40	40..50	50..60	60..70	70..80	80..90	90..100
Background	0.2641	0.1653	0.1111	0.0975	0.063	0.0479	0.039	0.0275	0.0168	0.0175
LincRNAs	0.5425	0.1687	0.0974	0.0608	0.0352	0.0213	0.0203	0.0067	0.0062	0.0035

Figure A.7: Distribution of the frequency of lincRNAs and background split in non-overlapping windows of the fractions of remaining expressions. The frequencies are shown for target sites and randomly sampled introns located at the 5' intron position of coding genes. The frequencies are separated in windows of the fractions of remaining expression. In this Figure we show the results for adjacent exons upstream (corresponding to the 5' exon of protein-coding genes). We used non-overlapping window sizes of ten percent ([0..10[,..,[90..100[). The frequencies are normalized by the total frequency of target sites and random introns within the 5' intron in each window. lincRNAs most significantly down-regulate the expression of the 5' exon of their targets in comparison to random sampling. The maximal difference is 0.28 with the maximal excess of lincRNAs in the bin of 'fraction of remaining expression' [0..10[. This notes that the 5' exon is down-regulated to a remaining expression level of <10 %.

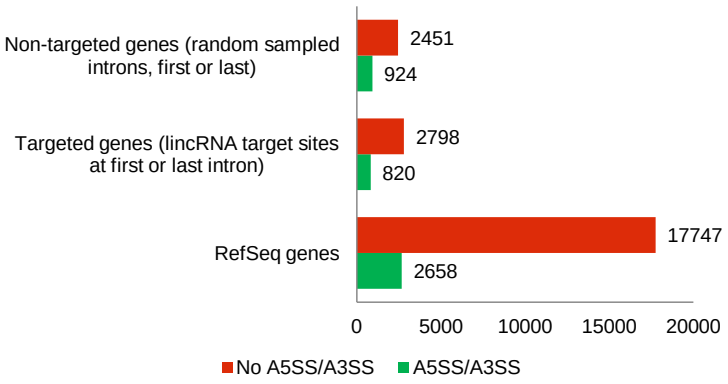


Figure A.8: Distribution of the frequencies of genes undergoing an alternative transcript event. In this Figure the results are shown for the merged event: A5SS and A3SS and the ESC cell line. We calculated the distribution for three distinct sets of genes: targeted, non-targeted genes and all RefSeq genes. We selected targets with target sites assembled at first and last introns of all sites for this analysis. The same restriction was set for random sampling. The frequencies of genes are split in undergoing an event and no event. This event is not significantly fostered by lincRNAs.

small ncRNA type	Frequency of lincRNAs		Average frequency of small ncRNAs	
	S	A	S	A
snoRNA[1]	1	1	1	1
miRNA[1]	27	30	2	2
pasRNA[1]	1	31	1	3
easRNA[1]	110	59	16	6
rasRNA[1]	130	63	26	42
miRNA[2]	8	-	1	-
piRNA[2]	18	22	2	2
snoRNA[2]	53	47	1	1

Table A.5: Statistics of the interaction between our lincRNA data set with distinct types of small ncRNAs. The number of lincRNAs with sequence similarity to a small ncRNA type are listed in this table. One lincRNA can interact with multiple distinct types of small ncRNAs. Each row describes the statistics of the interaction of lincRNAs (ESC) with one type of small ncRNA. Column 1 lists the distinct types of small ncRNAs, with small ncRNAs of DeepBase[1] and fRNAdb[2]. Columns 2-4 show the statistics of lincRNA:one type of small ncRNAs interactions: (2) the frequency of lincRNAs with sequence similarity to at least one small ncRNA (expressed interaction sites), (3) average frequency of small ncRNAs a lincRNA is interacting with. The frequencies are separated in sense: S and antisense: A interactions (strand orientation + and -).

Appendix B

Supportive criteria

1. Positive response to dopaminergic treatment.
2. Periodic limb movements (during wakefulness or sleep).
3. Positive family history of the restless legs syndrome suggestive of an autosomal dominant mode of inheritance.

Associated criteria

1. Natural clinical course of the disorder. Can begin at any stage, but most patients seen in clinical practice are middle-aged or older. Most patients seen in the clinic have a progressive clinical course, but static clinical course is sometimes seen. Remissions of a month or more are sometimes reported.
2. Sleep disturbance. The leg discomfort and the need to move result in insomnia.
3. Medical investigation/neurological examination. A neurological examination is usual in idiopathic and familial forms of the syndrome. Peripheral neuropathy or radiculopathy are sometines carried out in the non-familial form of the syndrome. A low serum ferritin may be found in the syndrome.