COMPUTATIONAL ANALYSIS OF NEXT GENERATION SEQUENCING DATA: FROM TRANSCRIPTION START SITES IN BACTERIA TO HUMAN NON-CODING RNAS

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von

HADI JORJANI

aus dem Iran

Basel, 2015

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultaet
auf Antrag von

Prof. Mihaela Zavolan and Prof. Ivo Hofacker

Members of the dissertation committee: Faculty representative, dissertation supervisor, and co-examiner

Basel, 9.12.2014

Date of approval by the Faculty

Signature of the Faculty representative

Prof. Dr. Joerg Schibler

The Dean of Faculty

Thanks for everything that you have done for me, and all that you are still doing

To my parents and my beloved wife...

Acknowledgements

First of all I am greatful to my supervisor Mihaela Zavolan for her constant support during these 4 years. I also thank Erik van Nimwegen for introducing me to Bayesian theory which changed my perspective to data analysis. I would also like to thank my best friend Andreas Gruber for his suggestions in order to improve the thesis. I am thankful to my friends Alexander Kanitz, Rafal Gumienny, Joao Guimaraes, Aaron Grandy, Wojciech Wojtas-Niziurski and Philipp Berninger for giving me insights to improve the thesis. Finally I am really happy to have met so many friends and colleagues who have made my stay in Basel productive and enjoyable.

Basel, 24 Nov 2014 J. H.

Abstract

The advent of next generation sequencing (NGS) technologies has revolutionized the field of molecular biology by providing a wealth of sequence data. “Transcriptomics”, which aims to identify and annotate the complete set of RNA molecules transcribed from a genome, is one of the main applications of these high-throughput methods. Special attention has been paid in determining the exact position of the 5’ ends of RNA transcripts, the transcription start sites (TSSs), and subsequently in identifying the regulatory motifs that are ultimately responsible for governing gene expression. Recently, a novel experimental approach termed dRNA-seq has emerged which enables TSS identification in prokaryotic genomes at a genome-wide scale. While the experimental procedure has reached a point of maturity, the computational downstream analysis of dRNA-seq data is still in its infancy. Analysis of dRNA-seq data was previously done manually, a tedious task that is prone to errors and biases. In order to automate this process we developed a computational tool for accurate and systematic analysis of dRNA-seq data to identify the TSSs genome-wide. In particular, we used a Bayesian framework for TSS calling and a Hidden Markov Model to infer the canonical motifs in the promoter regions of TSSs in order to further capture TSSs that show low evidence of expression. In a second contribution, we exploited the power of next generation sequencing to identify and characterize the expression and processing mechanisms of snoRNAs. SnoRNAs are a particular class of non-protein coding RNAs whose main function is post-transcriptional modification of other non-protein coding RNAs. SnoRNAs carry out their function as part of ribonucleoprotein complexes (RNPs). In order to gain insights into these protein-RNA interactions, we used a technique called PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) that allows the identification of protein-RNA contacts at nucleotide resolution. Using PAR-CLIP data, we were able to demonstrate that snoRNAs undergo precise processing and that many loci in the human genome generate snoRNA-like transcripts whose evolutionary conservation and expression are considerably lower than currently catalogued snoRNAs. Finally, we set out to use small RNA-seq data from the ENCODE project to construct a comprehensive catalog of genomic loci that give rise to snoRNAs. In addition we expanded the current catalog of human snoRNAs and studied the plasticity of snoRNA expression across different cell types. Our analysis confirmed prior observations that several snoRNAs show cell type specific expression, mainly in neurons. A more striking observation was that snoRNA expression appears to be strongly dysregulated in cancers which could lead to the identification of novel biomarkers.

Zusammenfassung

Das Aufkommen von “Next Generation Sequencing”-Technologien (NGS) hat das Gebiet der Molekularbiologie revolutioniert. Die enorme Fülle an Sequenzdaten, die mittels dieser Technologien geliefert werden kann. Das Forschungsgebiet der “Transcriptomics”, welches sich zum Ziel setzt alle RNA-Moleküle welche von einem Genom transkribiert werden zu identifizieren und zu annotieren, ist eine der Hauptanwendungen von NGS. Besonderes Augenmerk wurde dabei bisher auf die exakte Bestimmung der 5’-Enden und Transkriptionsstartstellen (TSS) von RNA-Transkripten gelegt, sowie auf der Identifizierung von regulatorischen Motiven die eine Rolle bei der Regulierung der Genexpression spielen. Seit kurzem liegt zwar mit dem sogenannten dRNA-seq ein experimenteller Ansatz vor, mit dem sich TSS auch in prokaryotischen Genomen bestimmen lassen. Aber auch wenn sich entsprechende Experimente nun routinemässig durchführen lassen, steckt die nachgeschaltete, computergestützte Analyse von dRNA-seq-Daten noch in ihren Anfängen. Erhobene Daten wurden vormals manuell ausgerwertet - ein aufwändiger Prozess der anfällig ist für Fehler und Verzerrungen bzw. Voreingenommenheiten. Um den Prozess der Ermittlung von bakteriellen TSS zu automatiseren, haben wir ein Programm zur präzisen und systematischen Auswertung von dRNA-seq-Daten entwickelt. Dieses verwendet einerseits ein Bayes-Verfahren zur Bestimmung von TSS. Andererseits kommt ein Hidden-Markov-Modell zur Herleitung von kanonischen Motiven in den Promoterregionen von TSS zum Einsatz, wodurch sich auch selten verwendete TSS bestimmen lassen. In einem zweiten Projekt haben wir die Stärken von NGS zur Katalogisierung von snoRNAs ausgenutzt. Neben der Identifizierung noch nicht bekannter Spezies stand dabei auch die Charakterisierung von snoRNAs im Hinblick auf Expression und Prozessierungsmechanismen im Vordergrund. SnoRNAs sind eine bestimmte Klasse von “nicht-kodierenden” RNAs (d.h. RNA-Moleküle die nicht als Blaupause für die Synthese von Proteinen dienen), deren Hauptfuntion in der post-transkriptionellen Modifizierung anderer “nicht-kodierender” RNAs besteht. Um ihre Aufgabe auszuführen, lagerns sich snoRNAs mit einer Reihe bestimmter Proteine zu RNA-Protein-Komplexen zusammen. Um Einblicke in diese Protein-RNA-Wechselwirkungen zu gewinnen, haben wir die Methode PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation) angewandt, welche die punktgenaue Bestimmung von Protein-RNA-Kontaktstellen ermöglicht. Mittels PAR-CLIP konnten wir aufzeigen dass die Prozessierung von snoRNAs präzise abläuft und dass viele Stellen des menschlichen Genoms snoRNA-ähnliche Transkripte generiert, deren Expression und Grad an evolutionärer Konservierung deutlich geringer sind als die bereits katalogisierter, herkömmlicher snoRNAs. Schliesslich haben wir Sequenzierungsdaten

kurzer RNA-Moleküle aus dem ENCODE-Projekt herangezogen, um eine umfassende Karte all der genomischen Regionen zu erstellen, welche Erbinformationen für die Synthese von snoRNAs tragen. So konnten wir den bestehenden snoRNA-Katalog deutlich erweitern und zusätzlich die Plastizität der Expression von snoRNAs in unterschiedlichen Zelltypen studieren. Anhand dieser Analyse konnten wir zeigen, dass snoRNAs - besonders in Nervenzellen - Zelltyp-spezifische Expressionsmuster aufweisen. Auffällig war ausserdem das unterschiedliche Expressionsmuster von snoRNAs in Krebszelllinien im Vergleich zu normalen Zellen. Dies veranlasste uns eine Reihe von snoRNAs zu identifizieren, deren Expression sich im besonderen Masse von der in gesunden Zellen unterschied und welche somit möglicherweise in naher Zukunft als “Biomarker” in der Krebsdiagnostik oder -therapie eingesetzt werden könnten.

Contents

Acknowledgements **i**

Abstract (English/Deutsch) **iii**

1 Introduction **1**
1.1 Thesis Outline . . . 2
1.2 High Throughput Sequencing . . . 2
1.2.1 Next Generation Sequencing as an Essential Tool in Molecular Biology Today . . . 2
1.2.2 Applications of NGS technology . . . 3
1.2.3 NGS platforms . . . 3
1.2.4 Analyzing NGS data . . . 4
1.3 The general framework of identifying transcriptional start sites . . . 5
1.3.1 dRNA-seq (differential RNA sequencing) . . . 5
1.3.2 dRNA-seq data analysis . . . 6
1.3.3 Basics of Bayesian analysis . . . 7
1.3.4 Hidden Markov Models . . . 8
1.4 Genome-wide identification of non-coding RNAs and their interaction partners 10
1.4.1 Non-coding RNAs . . . 10
1.4.2 The ENCODE project . . . 10
1.4.3 CLIP-based methods unravel protein-RNA interactions . . . 11

2 TSSer: A Computational tool to analyze dRNA-seq data **13**

3 Insights into snoRNA biogenesis and processing **29**

4 An updated human snoRNAome **45**

5 Discussion **71**

Appendices **75**

A Supplementary material of Chapter 4 **77**

Bibliography **104**

Curriculum Vitae **105**

1 Introduction

1.1 Thesis Outline

In the first chapter a brief and general introduction is given for the basic concepts behind the work presented in individual chapters. We thus describe the NGS technology, its history and applications. We also compare different platforms and discuss the downstream data analysis steps. In the next part we elucidate concepts that we use in Chapter 2 such as dRNA-seq protocol, Bayesian models and Hidden Markov Models. We conclude the introduction with a discussion of the methods that are used to map RNA-protein interactions such as the non-coding RNAs, the PAR-CLIP method, and the ENCODE project. Chapter 2 is the published paper where we have described our computational tool called "TSSer" which is designed to identify transcription start sites in prokaryotic genomes based on dRNA-seq data. Chapter 3 is the published paper in which we used the PAR-CLIP method to gain insights into snoRNAs biogenesis and processing. Chapter 4 is the draft of a manuscript in which we describe how we used the ENCODE data to expand the catalog of human snoRNAs and understand the plasticity of their expression across different cell types. The manuscript will be submitted shortly. In Chapter 5 we conclude our work and discuss the future directions.

1.2 High Throughput Sequencing

1.2.1 Next Generation Sequencing as an Essential Tool in Molecular Biology Today

In the realm of molecular biology "sequence" is defined as the exact order in which nucleotide bases appear in a DNA or RNA molecule or amino acids in a polypeptide. The order of nucleotides in a DNA molecule carries necessary information which serves as a prints for synthesis of proteins which are the fundamental components of all living cells and are responsible for diverse range of functions in the cell. Hence determining the order of bases in a DNA or RNA molecule is a crucial step towards understanding molecular functions. Furthermore identifying the sequence of DNA or RNA molecules to which specific DNA and RNA binding proteins bind enables us to understand molecular interactions and their consequences within cells. Novel sequencing technologies enabled the sequencing of enormous amounts of DNA or RNA molecules providing an unprecedented opportunity to study the genomes of a vast number of species at a level of detail that has not been matched in terms of costs and efficiency by any technique before. A big boost in the development of sequencing technologies came after the initial assembly of the human genome in 2001 [89, 157] . Sanger sequencing was the sequencing choice at the time of this huge project (International Human Genome Sequencing). Subsequently, the demand for a high-throughput, fast and low cost sequencing technology rapidly increased. Sanger sequencing is considered as the first-generation technology while the high-throughput sequencing technologies which emerged afterwards and were order of magnitudes faster and cheaper compared to Sanger sequencing are referred to as "second-generation" or "next-generation sequencing" (NGS) [87, 139] . NGS allows sequencing to be done in parallel, allowing to sequence a multitude of DNA / RNA molecules at the same

time. The low-cost production of large volumes of sequence data from NGS - currently up to one billion short reads per instrument run - is its main advantage over conventional DNA sequencing methods. This however, is achieved at the price of somewhat lower quality and read length [42, 128, 133]

1.2.2 Applications of NGS technology

High-throughput sequencing provided by NGS revolutionized the field of biology in the past decade by supporting a wide range of applications in molecular biology, evolutionary biology, functional genomics, metagenomics, microbiome research and medicine [118, 13, 126, 114, 159, 1, 41] . As mentioned above, transcriptomics - determining the sequence and abundance of different RNA species such as mRNAs, small and long non-coding RNAs - is one of the major applications of NGS[163] . Prior to NGS methods, measurements of gene expression were obtained with microarrays. The principle behind these was hybridization of DNA derived from cellular RNAs to predefined synthetic array of oligonucleotides. In contrast to microarrays, NGS does not require prior knowledge of the molecules that are to be quantified and there is no need for an organism-specific design. NGS has also improved the sensitivity, accuracy and dynamic range of gene expression analysis studies[129, 40].
An approach for determining the sequence specificity of DNA- and RNA-binding proteins consists in immunoprecipitation (IP) of the protein of interest with specific antibodies followed by the identification of the nucleic acids to which the protein binds. This can also be performed using NGS technology. IP followed by high-throughput sequencing allows the identification of genome-wide binding profiles of DNA-binding proteins (with Chromatin immunoprecipitation or ChIP-seq)[134, 121], genome-wide DNA methylation sites (methyl-seq) and DNase I hypersensitive sites (DNase-seq) [170, 9]. These, in turn, inform about the dynamics and regulation of gene expression. NGS has also been utilized to investigate RNA-protein interactions. Various variant methods have been proposed, that go by the names of CLIP-seq or HITS-CLIP, PAR-CLIP and iCLIP [55, 27, 68]. Other applications of NGS include finding genetic variants via resequencing the targeted regions of interest, de novo assemblies of bacterial genomes with low expenditure and high quality and identifying and classifying the spectrum of species that co-inhabit specific environments via metagenomics studies [123, 115, 33, 50, 82, 109, 112].

1.2.3 NGS platforms

Although available NGS technologies vary in the sequencing biochemistry, the workflow is quite similar and consists of the following steps : library preparation (isolating DNA or RNA molecules followed by random fragmentation of DNA and ligation of adaptors), template amplification (using polymerase chain reaction (PCR)), sequencing and imaging. These are followed by the computational analysis of the image data that leads to base calling and then the genome alignment of the resulting reads. The 454 from Roche, Solexa Genome Analyzer from Illumina and SOLiD from applied Biosystems were among the first NGS platforms that were broadly used in high-throughput studies. These platforms differ in many aspects including

inherent biases, accuracy, read length, sequencing depth, cost per run, initial infrastructure cost and bioinformatics tools to analyze their output data. These differences lead to each technology being used for specific suites of application. 454 from Roche outperformed initially the other technologies in terms of speed (few hours per run) and read length. Therefore, 454 was primarily used in applications where read length was the determining factor such as metagenomics or de novo genome ass embly. SOLiD had the highest accuracy, with applications in genome sequencing, transcriptomics research and targeted sequencing. The Illumina technology offered the cheapest sequencing method and the highest throughput. It has the capacity to handle sequencing of multiple libraries in a single instrument run using multiplexing technique and it is very versatile. It is used for a wide range of applications from the sequencing of bacterial DNA for genome assembly in microbiology studies to ChIP-seq in applications involving large genomes. PGM is a newer technology from Ion Torrent. It offers small instrument size as well as low cost, commonly used in identify microbial pathogens and whole genome sequencing of bacterial genomes [101, 6, 42, 107, 139, 2, 116, 119, 108, 115].
With the SMRT (Single Molecule Real-Time Sequencing) technology from Pacific Biosciences the third generation sequencing platforms has arrived. Sequencing in real time and eliminating the PCR amplification step are two major features of SMRT. It also produces much longer reads (average read length is 1300 bp) compared to any second generation method. Eliminating the PCR amplification step leads to lower sample preparation time and reduces biases and artifacts caused by amplification. However, these advantages come at the cost of lower throughput compared to second generation methods as well as relatively high error rates which make the computational analysis considerably more challenging [127, 84].This method is quite popular in microbiology studies, resequencing, as well as determination of isoforms in complex organisms[132]. NGS is rapidly improving in terms of quality, speed and cost and has become the method of choice in large-scale sequencing studies [123, 115, 33, 50, 82, 109, 112].
A big challenge today is to efficiently store and computa of these enormous volumes of data produced by high-throughput methods. In the next section we briefly talk about general steps that are involved in the analysis of NGS data.

1.2.4 Analyzing NGS data

Because NGS technologies are diverse and evolving rapidly, the bioinformatic analysis of the resulting data, including base-calling, sequence quality assessment, alignment of reads to a reference genome and de novo assembly evolves accordingly and it is therefore challenging. Base-calling is the process of inferring the individual nucleobases (A,C,G,T) from fluorescence intensity signals, yielding the actual sequences. There are variety of base-calling programs which mostly differ in their statistical framework and the way they report quality scores for the reads. The common way to report uncertainty of each base is using “Phred score” which is proportion to the negative of the log probability that the base call is erroneous. A comparison of common base-calling algorithms can be found in recent reviews [92, 138]. Sequence quality assessment methods are relevant not only for the basic analysis of the sequenced reads but also for identifying single nucleotide polymorphisms (SNPs). A main bioinformatic challenge

in dealing with NGS data is alignment (or mapping) of the reads to a reference genome. Although tools like BLAST and BLAT have been used for a relatively long time, they do not scale to the size of the data sets that come out of deep sequencing studies [113, 77, 4]. Thus, a new series of alignment tools have been developed recently. They differ in terms of speed, space and memory usage, the way they handle insertions/deletions and in the capacity to perform spliced alignment. The most widely used programs for the alignment of short reads to the genome are Bowtie [91, 90], BWA citeLi2009-ve,Lam2009-ys, segemehl [65] and STAR [34]. However, many other alignment programs are available such as SOAP [96, 100], GMAP and GSNAP [172, 171], Bfast [66], subread[97], CUSHAW [101], GEM [106, 46],ZOOM [99], GNUMAP [23], Maq [93], and Top Hat [152]. The accuracy, speed and general performance of these programs has been assessed recently [94, 130, 135, 47, 48]. Depending on the nature of the data, different types of analyses are performed following the mapping step. Normalization is a necessary step, while differential gene expression analysis or peak calling are specific to individual applications.

1.3 The general framework of identifying transcriptional start sites

1.3.1 dRNA-seq (differential RNA sequencing)

One of the main challenges in transcriptomics was to determine the exact locus on the genome where transcription initiates. Genome-wide studies of transcription start sites (TSS) were initially carried out in eukaryotes using a method known as cap analysis gene expression (CAGE) [58, 88, 30, 141]. Because prokaryotic RNAs lack a 5' cap structure, feature which is exploited in CAGE, the capture of TSSs in prokaryotes required the development of another technique, which came in the form of differential RNA sequencing (dRNA-seq). Limited-scope methods that have been used previously to identify TSSs of individual genes were 5' rapid amplification of clone ends (RACE), primer extension and S1 protection [14, 151, 12, 7, 160]. Here we briefly introduce the dRNA-seq method and its application in microbial transcriptomics studies.
General-purpose RNA-seq approaches can not distinguish between primary transcripts (RNAs with triphosphate at their 5' ends) and processed fragments (RNAs with monophosphate or hydroxyl group at their 5' ends). Thus, to obtain bacterial TSSs, the 5' end of transcripts that carry triphosphates needed to be captured [24]. The RNA-seq approach specifically depletes processed fragments, thereby enriching primary transcripts. Upon treatment of the sample with a 5' phosphate-dependent exonuclease (TEX), an enzyme that specifically degrades transcripts having a 5' monophosphate, processed fragments as well as the vast majority of ribosomal RNAs (rRNA) and transfer RNAs (tRNAs) that have monophosphates at their 5' ends, are specifically degraded [137].
The new approach is called dRNA-seq (differential RNA-sequencing) and is based on comparing two cDNA libraries obtained from TEX-treated and untreated samples. The RNA obtained from bacterial cells maintained in a specific condition is divided into two parts: one half is treated with TEX enzyme which specifically degrades 5' monophosphate (denoted as 5'P) RNAs and the other half is left untreated, leading to the capture of both 5' triphosphate (de-

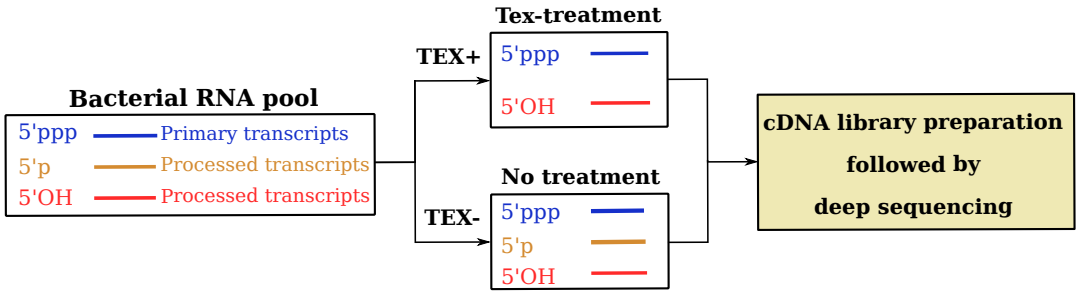


Figure 1.1 – Schematic view of dRNA-seq protocol

noted as 5'PPP) and 5'p RNAs. Then tobacco acid pyrophosphatase (TAP) is used to convert 5'PPP ends into 5'P to allow the ligation of RNA linkers. A poly(A) tail is then added to the RNA, the cDNA is then synthesized, amplified using PCR and is sequenced with high-throughput methods.
dRNA-seq was used for the time to determine the primary transcriptome of the gastric pathogen Helicobacter pylori in 2010 [136] . Since then, the dRNA-seq approach was utilized for transcriptome analysis of several organisms including several bacterial and a couple of archeal species [150, 117, 85, 175, 131, 158, 168, 161, 173, 35]. dRNA-seq enables identification of TSSs at single nucleotide resolution on a genome-wide scale and its application demonstrated that many small RNAs coding for short peptides, small non-coding RNAs and antisense transcription near TSS site [142, 136, 67, 26]are transcribed in bacteria. Identifying the exact position of TSSs is also an essential step towards investigating gene regulatory networks because it allows the focused search for transcription regulatory motifs which are present in promoter region. Accurate TSS mapping further enables the study of of 5' untranslated regions (5' UTR), which are important for translation regulation in bacteria [144]. 5' UTRs usually carry a ribosome binding site (RBS) - known as the Shine-Dalgarno (SD) sequence (AGGAGG) which is generally located around 8 bases upstream of the start codon - where the ribosome binds to initiate protein synthesis from mRNA [105, 19]. Genes that are leaderless and are translated via different mechanisms are also known[16, 25]
Illumina, 454 and SOLID sequencing have all been used to map bacterial TSSs genome-wide, though Illumina is the most popular platform [2, 42, 108, 124, 156]. For TSS identification and gene expression analysis the sequencing depth is the determining factor.

1.3.2 dRNA-seq data analysis

TSS annotation based on dRNA-seq data used to be a tedious task starting from visualizing the read profiles in a genome browser followed by manual inspection to look for any enrichment pattern of the expressed reads in TEX + (TEX treated) versus TEX - (untreated) samples. This procedure is not only laborious but also prone to errors, and thus not practical for the analysis of multiple samples and large data sets. An automated method to analyze dRNA-seq data was therefore in demand among experimental microbiologists. To fill this gap we have developed

the "TSSer" tool which enables identification of TSSs genome-wide in prokaryotic organisms systematically and in a precise way [72]. In Chapter 2 we describe our model in detail. TSSer turns out to be one of the automated methods for dRNA-seq data analysis that have been developed in the past couple of years. Other computational tools for dRNA-seq data analysis were developed more or less at the same time as TSSer [5, 35, 62]. These methods use statistical functions (e.g. Poisson distribution) to model the expression profile of reads in a defined window length [5] or considering multiple genome alignment of different strains combined with a simple peak calling strategy (lacking a statistical model)[35, 62]. Multiple genome alignment of different strains is not directly related to TSS identification based on dRNA-seq data and in fact can be used as a separate source of information to be used in conjunction with any TSS finding tool for the determination of TSSs. A rigorous benchmarking of these methods is a difficult task as an exact definition of a real TSS is not in hand. For TSSs which are highly expressed and show clear enrichment almost all these methods can capture them easily but the difference arises for TSSs which exhibit low expression and are not significantly enriched. To overcome this problem TSSer models the underlying distribution of read counts in a Bayesian framework in order to subsequently calculate the enrichment in a probabilistic manner. The HMM trained over *bona fide* TSSs also helps to recover majority of TSSs which are missed in the first round due to exhibiting low expression evidence. In Chapter 2 we have shown that TSSer achieves high consistency in TSS identification compared to manual approach and it can also detect as many TSSs which could not be captured by manual inspection of read profiles. All these methods are based on some user-defined cut-offs on their parameters and still need supervision to some extent but they facilitate TSS calling to a great extent compared to manual annotation.

1.3.3 Basics of Bayesian analysis

In TSSer we use notions of Bayesian probability theory, and we therefore give a brief introduction to these notions here. In what is called orthodox or frequentist statistics, one aims to zoom on to the correct model of the data by testing various possible models. These are denoted as "hypotheses" (H) and the data is denoted by D. To evaluate the model usually a quantity called p-value is calculated which is basically the probability of obtaining a result at least as extreme as the one that is actually observed, assuming that the "null hypothesis" (also known as counter-hypothesis) is true. lf the p-value is lower than a given significance level (e.g. $P(D|H_{null}) < 0.05$) then the null hypothesis is rejected and the alternative hypothesis is accepted. The most important point about p-value calculation or more generally the orthodox paradigm is that, p-value does not give us the probability of hypothesis or in other words $P(D|H) \neq P(H|D)$. In contrast, the Bayesian approach allows one to assign probabilities to hypotheses, empowers one to treat the model parameters as random variables and allows to infer the posterior probability of a model based on a given data i.e. calculating $P(H|D)$:

$$P(H|D) = \frac{P(H)P(D|H)}{\sum_H P(H)P(D|H)}$$

where $P(H|D)$ is called the "posterior probability" of the model given the data. $P(D|H) = L(H)$ is the likelihood function or probability of the data given the model and P(H) is the prior probability of the model before seeing the data which is usually assumed to be uninformative in case we know nothing about the model ab initio. The denominator is called marginal likelihood and is actually a normalizing factor for the density of posterior probability. The density of posterior probability is proportional to the likelihood times the prior [70].
In the Bayesian approach the probability of a model can be precisely calculated by integrating (or summing in case of dealing with discrete variables) over all possible values of parameters. Bayesian probability provides a framework for model selection - by simply calculating the probability of each model given the data - and parameter estimation - choosing the parameters set which maximizes the probability of the data - in a logical way. In chapter 2 we have used the Bayesian analysis to infer the posterior probability distribution of 5' ends of transcripts based on the observed counts and consequently we used this posterior probability to calculate the enrichment of 5' ends in dRNA-seq data.

1.3.4 Hidden Markov Models

A Hidden Markov Model (HMM) is a general probabilistic model to assign probability distributions to a sequence of observations [49]. HMM is a commonly used tool for modeling DNA and proteins sequences In the field of computational biology [36]. HMM is composed of two main components, a set of states and a set of symbols that are emitted from each state. HMM is in principle a sequence generator, it emits symbols as it passes through its states. Transitions from one state to another are associated with defined transition probabilities and in each state of HMM one symbol is emitted based on the defined emission probabilities for that state. It is called a "Markov model" because the sequence of underlying states have the "Markovian property" i.e. the next state is determined merely based on the current state and is not dependent on previous states that the HMM has passed through. It is called "Hidden" because usually the underlying sequence of states is not known and has to be inferred from the observed sequence of emitted symbols. In modeling sequence data, the emission probabilities simply define the base composition that we expect to see in that state. For example if we would like to model a state corresponding to an "A/T" rich region then the symbols "A" and "T" are emitted with higher probabilities compared to symbols "C" and "G". A schematic view of a HMM which can distinguish between A/T rich and C/G rich regions in a sequence is illustrated in Figure 2.
The probability of a sequence given the model is calculated by multiplying all emission and transition probabilities along the path which has generated that sequence (as this product is usually a small number, it is common to work with the logarithm of this product). If a sequence can be produced from alternative paths then the sum of probabilities over these paths gives one the probability of observing the sequence. To calculated this sum, algorithms known as "Forward and Backward" have been developed which enable efficient calculation of this sum using dynamic programming techniques. If we are interested to infer the most probable state path which generates the observed sequence we can use another dynamic

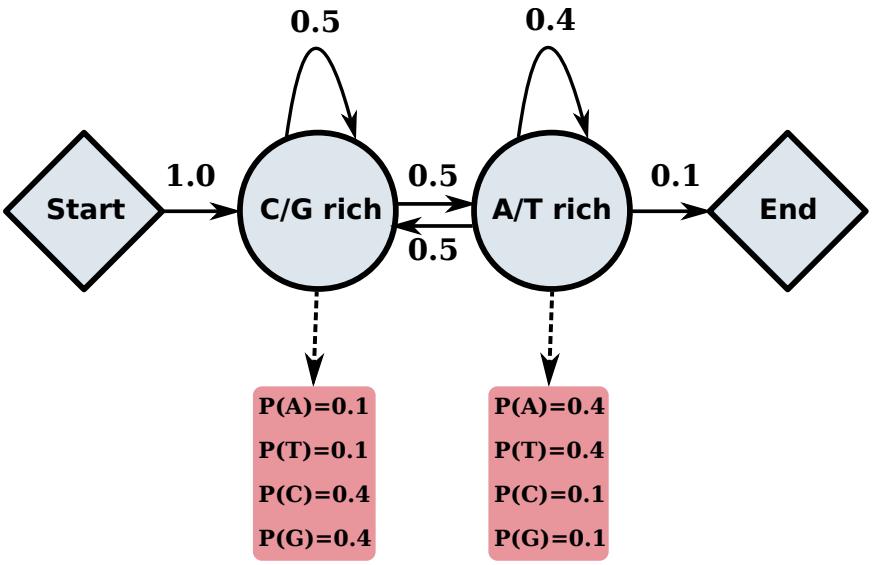


Figure 1.2 – A simple Hidden Markov Model with two states to distinguish between C/G rich and A/T rich regions in a sequence

programming-based algorithm called "Viterbi" [29]. There is another algorithm known as "posterior decoding" which finds the most probable state from which a given symbol in the sequence is emitted. Posterior decoding is based on a mixture of Forward and Backward algorithms. Another interesting problem regarding HMM is to estimate the parameters of the model based on a set of observed sequences. This can be achieved by using expectation maximization algorithms [31]. These algorithms usually start with some initial parameter set and then calculate the probability of the model based on the observed sequences (calculating likelihood). Then they update the parameters and keep on repeating these two steps until converging the likelihood values. These algorithms are discussed in detail in [36].
An application of Hidden Markov Models in computational biology first introduced in late 1980 for analysis of DNA sequences[22] and later for prediction of protein structures [146, 165]. Since then Hidden Markov Models have been used in different areas of bioinformatics such as sequence alignment [86, 10, 38, 143, 103], protein structural modeling and homology detection[37, 75, 76] and gene finding [15, 11, 83, 3]. In summary HMM has proved itself as a powerful tool to analyse the sequence data in the field of molecular biology.
In Chapter 2 we have designed a Hidden Markov Model to detect promoter regions in bacterial genomes. HMM states are corresponding to the consensus elements of bacterial promoters such as -35 and -10 motifs, spacer and discriminator regions [57]. The model was trained over a set of *bona fide* promoters and then the trained model (fitted transition probabilities and emission probabilities) was used to predict the putative promoter regions in the bacterial genome and assigning probability values to each putative promoter site in the genome. This model proved to be efficient in identifying the promoters which show low expression evidence due to condition specificity of their expression or do not exhibit sufficient enrichment due to inefficiency of the dRNA-seq experiment.

1.4 Genome-wide identification of non-coding RNAs and their interaction partners

1.4.1 Non-coding RNAs

Non-coding RNAs (ncRNAs) form a heterogeneous class of RNA molecules that do not encode information for protein production. Thus, they are not translated into proteins, but rather perform other cellular functions, being involved in a variety of processes including transcription, chromatin remodeling, RNA splicing and editing and translation[39, 21, 111].Dysregulated expression of non-coding RNAs has been observed in several diseases including cancer [95, 147], Alzheimer's disease [45] and Prader–Willi syndrome [17] [74].Highly abundant RNAs that are involved in translation and protein synthesis such as transfer RNAs (tRNAs) and ribosomal RNAs (rRNAs) constitute a big fraction of the total expressed non-coding RNAs. Other important sub-groups of non-coding RNAs are the microRNAs (miRNAs) [59],the Piwi protein-interacting RNAs (piRNAs) [162, 164]and small interfering RNA (siRNAs)[43],that are involved in gene regulation, long non-coding RNAs (lncRNAs) [79], long intergenic non-coding RNAs (lincRNAs) [102], and antisense RNAs (asRNAs) [122]. Some non-coding RNAs guide the post-transcriptional modification of other RNA species. These include the small nuclear RNAs (snRNAs) that are involved in pre-mRNA splicing [154], the small nucleolar RNAs that primarily guide methylation and pseudo-uridylation of ribosomal RNAs (snoRNAs) [81], the small Cajal body specific RNA (scaRNA) [28] and telomerase RNA component (TERC)[149] . Most ncRNAs exert their function within RNA-protein complexes (ribonucleoprotein or RNP) such as ribosomal RNAs in the ribosome, snoRNAs in the snoRNPs, miRNAs in RNA-induced silencing complex, snRNAs in snRNPs and telomerase RNAs in telomerase. The different classes of non-coding RNAs and their corresponding functions have been surveyed in a recent review [18]. Non-coding RNAs also appear to be good biomarkers for diseases and cell differentiation states [104, 166, 20]. Therefore the expression profiling of non-coding RNAs is a crucial step towards understanding their regulatory functions. High-throughput sequencing technologies have also contributed to an improved understanding of the biogenesis and functions of non-coding RNAs in the recent years. Part of the work that was carried out for this thesis has focused on the snoRNA subset of non-coding RNAs. In Chapter 4 we describe our analysis of the large data set generated by the ENCODE project towards the discovery, characterization and expression profiling of snoRNAs.

1.4.2 The ENCODE project

The ENCODE project (ENCyclopedia of DNA elements) was launched by National Human Genome Research Institute (NHGRI) to harness the power of next generation sequencing methods towards characterization of all functional elements in the human genome[148]. A large international consortium of scientists from around the globe applied state of the art experimental and computational approaches to build a comprehensive catalog of functional elements that are encoded in human genome including protein-coding and non-coding

genes, transcriptional regulatory regions (promoters, enhancers, silencers), along with their associated chromatin states and DNA methylation patterns [44, 148]. A future aim of the ENCODE project is to provide accurate annotations of transcription start sites, introns and exon boundaries, and 3' polyadenylation sites, thereby expanding our understanding of RNA processing and alternative splicing. ENCODE generated high-throughput data for a range of normal and malignant cell types, as well as for different subcellular compartments such as nucleus or cytosol. From each subcellular compartment, both long (> 200) and short (<200) RNAs were sequenced. This data set thereby provided the opportunity to identify various types of non-coding RNAs such as miRNAs and snoRNAs.[148]. UCSC ENCODE genome browser and the ENSEMBL browser made the annotation of functional elements discovered by ENCODE project available to the general scientific community.

1.4.3 CLIP-based methods unravel protein-RNA interactions

Identifying the interactions of proteins with DNA or RNA molecules is essential for our understanding of the networks which govern gene expression in individual cell types. The high-throughput experimental methods that have been developed to capture the DNA or RNA targets of individual proteins of interest are based on crosslinking the proteins to DNA using UV light and then immunoprecipitating the protein (together with its bound target sequences) with a specific antibody (Immunoprecipitation or "IP"). NGS technologies provided the necessary throughput to explore DNA/RNA-protein interactions at a genome-wide scale. ChIP-seq (Chromatin immunoprecipitation followed by high-throughput sequencing) was one of the first applications that used the above-mentioned principles [155, 73]. After successful application of this method in genome-wide studies mainly to find the binding sites of transcription factors (TFs) - the main class of regulators of gene expression on transcription level - scientists set out to apply this method to characterize binding specificity of various RNA-binding proteins. This led to the so-called "CLIP" (cross-linking immunoprecipitation) methods [71, 27]. CLIP-based protocols such as HITS-CLIP (High-throughput sequencing of RNA isolated by crosslinking immunoprecipitation), iCLIP (individual-nucleotide resolution Cross-Linking and ImmunoPrecipitation) and PAR-CLIP (Photoactivatable-Ribonucleoside-Enhanced Crosslinking Immunoprecipitation) are used for genome-wide identification of the target sites of a particular protein on RNA molecules [174, 27, 98, 145, 68, 54]. These methods can also be applied to identify the target RNAs whose interaction with a specific protein is guided by other non-coding RNAs. For instance, PAR-CLIP was applied successfully to identify the target sites of miRNA as well as snoRNAs by crosslinking of Argonaute complex and snoRNP core proteins, respectively [54, 55, 56, 80]. CLIP-based methods are making a great impact on our knowledge of post-transcriptional regulation, revealing for example, how vast the RNA-mediated interaction networks are [8]. In Chapter 3 we describe how we have utilized the PAR-CLIP method to immunoprecipitate the core proteins of snoRNP complexes as well as the Argonaute protein in order to investigate snoRNA processing [80]

2 TSSer: A Computational tool to analyze dRNA-seq data

BIOINFORMATICS ORIGINAL PAPER

Vol. 30 no. 7 2014, pages 971–974
doi:10.1093/bioinformatics/btt752

Gene expression

Advance Access publication December 25, 2013

TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data

Hadi Jorjani and Mihaela Zavolan*

Computational and Systems Biology, Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

Associate Editor: Ivo Hofacker

ABSTRACT

Motivation: Accurate identification of transcription start sites (TSSs) is an essential step in the analysis of transcription regulatory networks. In higher eukaryotes, the capped analysis of gene expression technology enabled comprehensive annotation of TSSs in genomes such as those of mice and humans. In bacteria, an equivalent approach, termed differential RNA sequencing (dRNA-seq), has recently been proposed, but the application of this approach to a large number of genomes is hindered by the paucity of computational analysis methods. With few exceptions, when the method has been used, annotation of TSSs has been largely done manually.

Results: In this work, we present a computational method called 'TSSer' that enables the automatic inference of TSSs from dRNA-seq data. The method rests on a probabilistic framework for identifying both genomic positions that are preferentially enriched in the dRNA-seq data as well as preferentially captured relative to neighboring genomic regions. Evaluating our approach for TSS calling on several publicly available datasets, we find that TSSer achieves high consistency with the curated lists of annotated TSSs, but identifies many additional TSSs. Therefore, TSSer can accelerate genome-wide identification of TSSs in bacterial genomes and can aid in further characterization of bacterial transcription regulatory networks.

Availability: TSSer is freely available under GPL license at http://www.clipz.unibas.ch/TSSer/index.php

Contact: mihaela.zavolan@unibas.ch

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

Received on June 27, 2013; revised on December 16, 2013; accepted on December 21, 2013

1 INTRODUCTION

Identification of transcription start sites (TSSs) is a key step in the study of transcription regulatory networks. It enables identification of promoter regions, and thereby the focused search for binding sites of transcription factors. Although for species such as mouse and human, methods to capture TSSs have been developed >10 years ago (Shiraki *et al.*, 2003), owing to differences in messenger RNA (mRNA) processing, these methods cannot be applied to bacteria. Recently, however, a method for genome-wide identification of bacterial TSSs has been proposed (Sharma *et al.*, 2010). The method, called differential RNA sequencing (dRNA-seq), uses the 5′ mono-phosphate-dependent terminator exonuclease (TEX) that specifically degrades 5′ mono-phosphorylated RNA species such as processed RNA, mature ribosomal RNAs and transfer RNAs, whereas primary mRNA transcripts that carry a 5′ triphosphate remain intact. This approach results in an enrichment of primary transcripts, allowing TSSs to be identified by comparison of the TEX-treated samples to control untreated ones. As an automated computational method to identify TSSs based on dRNA-seq data has not been available, TSS annotation based on dRNA-seq data required substantial effort on the part of the curators. The aim of our work was to develop an automated analysis method to support future analyses of dRNA-seq data. We here introduce a rigorous computational method that enables identification of a large proportion of *bona fide* TSSs with relative ease. The method is based on quantifying 5′ enrichment of TSSs and also the significance of their expression relative to nearby putative TSSs. Benchmarking our method on several recently published datasets, we find that the identified TSSs are in good agreement with those annotated manually, and that a relatively large number of additional TSSs that also have the expected transcription regulatory signals are identified. TSSer is freely available at http://www.clipz.unibas.ch/TSSer/index.php.

2 APPROACH

The input to TSSer is dRNA-seq data, consisting of one or more pairs of TSS-enriched (TEX-treated) and TSS-not-enriched samples. There are two main criteria that we use to define TSSs. The first criterion stems from the obvious expectation that TSSs are enriched in the TEX-treated compared with the TEX-untreated samples (Sharma *et al.*, 2010). To quantify the enrichment, we explored two methods. In one approach we calculated, for each genomic position, a '*z*-score' of the observed number of reads in the TEX-treated sample compared with number of reads in the TEX-untreated sample. The second method aims to take advantage of the information from multiple replicates: we use a Bayesian framework to quantify the probability that a genomic position is overrepresented across a number of TEX-treated samples. The second main criterion that we use to pinpoint reliable TSSs rests on the observation that in bacteria, the majority of genes have a single TSS (Cho *et al.*, 2009). Thus, we expect that in a specific sample, for each transcribed gene, there will typically be one main TSS, as opposed to multiple TSSs in relatively close vicinity. In other words, *bona fide* TSSs should exhibit a 'local enrichment' in reads compared with neighboring genomic positions. We will now describe the computation of different measures of TSS enrichment.

*To whom correspondence should be addressed.

3 METHODS

3.1 Quantifying 5′ enrichment in a TEX-treated compared with a TEX-untreated sample

In preparing the dRNA-seq sample, one captures mRNAs from bacterial cells and sequences their 5′-ends. The capture of the mRNAs could be viewed as a sampling process that gives rise to hypergeometrically distributed counts of reads from individual positions in the genome. However, given that the number of reads originating at a given genomic position is small relative to the total number of obtained reads, we can approximate the hypergeometric distribution by a binomial distribution. That is, if the total number of reads in the sample is N, and the fraction f of these corresponds to a given TSS of interest, then the probability to observe the TSS represented by n of the N reads in the sample follows a binomial distribution:

$$P(n|f,N) = \binom{N}{n} f^n (1-f)^{N-n}$$

Letting f_+ and f_- denote the frequency of reads derived from a given genomic position in the TEX-treated (TSS-enriched) and TEX-untreated (non-enriched) samples, respectively, what we would like to determine is the enrichment defined as follows:

$$P(f_+ > f_-|n_+, N_+, n_-, N_-) = P(f_+ - f_- > 0|n_+, N_+, n_-, N_-).$$

We do not know the underlying frequencies f_+ and f_-. Rather, we approximate the probability of enrichment based on observed counts as explained in the Supplementary Material. With x being the observed frequency of reads derived from a given position (i.e. $x_+ = \frac{n_+}{N_+}$ and $x_- = \frac{n_-}{N_-}$ for the TEX-enriched and not enriched samples, respectively), the probability that a genomic position has a higher expression in the TEX-treated compared with the untreated sample is given by the following equation:

$$P(f_+ - f_- > 0|n_+, N_+, n_-, N_-) = \phi\left(\frac{x_+ - x_-}{\sqrt{\frac{x_+(1-x_+)}{N_+} + \frac{x_-(1-x_-)}{N_-}}}\right)$$

where ϕ is the cumulative of Gaussian distribution (error function). In case of having multiple paired samples, the average value of $\phi(t)$ for a given genomic position would quantify the 5′ enrichment probability. We call this measure 'z-score'. Alternatively, when we have replicates of paired (TEX-treated and untreated) samples, we can calculate the 5′ enrichment λ_s for each pair separately:

$$\lambda_s = \left\langle \frac{f_+}{f_-} \right\rangle$$

Assuming that λ_s follows a normal distribution with mean μ and variance σ^2, we can calculate the probability that a TSS is enriched across a panel of k replicate paired samples:

$$P(\mu > 1|\lambda) = \frac{\int_1^\infty \left(\frac{1}{(\mu-\mu_*)^2+\sigma_*^2}\right)^{\frac{k-1}{2}} d\mu}{\int_0^\infty \left(\frac{1}{(\mu-\mu_*)^2+\sigma_*^2}\right)^{\frac{k-1}{2}} d\mu}$$

where $\lambda = (\lambda_1, \lambda_2, ..., \lambda_k)$ and μ_* and σ_* are mean and variance of λ, respectively, and k is the number of replicates (details of the derivation are given in the Supplementary Material).

3.2 Quantifying local enrichment

To quantify the local enrichment of a putative TSS, we examine the frequencies of sequenced reads in a region of length 2l centered on the putative TSS ($[x-l, x+l]$). That is, we define the local enrichment L as follows:

$$L = \frac{\sum_{i\in[x-l,x+l], n_{+,i}\le n_{+,x}} n_{+,i}}{\sum_{j\in[x-l,x+l]} n_{+,j}} \quad (1)$$

where $n_{+,i}$ is number of reads derived from position i in the TEX-treated sample. The value of L would be 1 for the position with maximum expression in the interval, corresponding to a perfect local enrichment. When replicates are available, we compute the average local enrichment over these samples. We chose l such that it covers typical 5′ UTR lengths and intergenic regions, i.e. 300 nt. This value is of course somewhat arbitrary, but we found that it allows a good selection of TSSs in practice.

3.3 Identification of TSSs

To identify TSSs, we compute these measures based on all available samples. Because we observed that the precision of start sites is not perfect but there are small variations in the position used to initiate transcription, we also apply single linkage clustering to select the representative among closely spaced (up to 10 nt) TSSs. We then select the parameters that give us the maximum number of annotated genes being associated with TSSs, restricting the total number of predicted TSSs to be in within a narrow range, ±50% of the number of annotated genes in the genome.

4 EVALUATION OF THE TSS IDENTIFICATION METHOD

To evaluate our method and verify its accuracy, we applied it to several recently published datasets [*Helicobacter pylori*, *Salmonella enterica serovar Typhimurium* (Kröger *et al.*, 2012) and *Chlamydia pneumoniae* (Albrecht *et al.*, 2009)] for which a mixture of computational analysis and manual curation was used to annotate TSSs. We here present an in-depth analysis of the TSS identification approaches for *H.pylori*. Similar analyses for the other species are given in the Supplementary Tables S4–S6.

In the *H.pylori* genome, our method identified 2366 TSSs. Of these, 1306 (55%) TSSs are in the reference set of 1893 curated TSSs reported by Sharma *et al.*, 2010, which we refer to them as 'Common' TSSs. Thus, 69% of the curated sites are included in our TSS list. A number of reasons contributed to our method failing to identify another 31% curated TSSs, which we refer to them as 'Reference only'.

- In our approach, we only use reads that were at least 18 nt in length and mapped with at most 10% error to the genome. This selection appears to have led to the loss of 187 (32%) of the 587 curated TSSs in the mapping process, before applying the TSSer inference.
- The majority of the curated sites that we did not retrieve appear to have been supported by a small number of reads. Two hundred twenty-six (38%) of the 587 curated TSSs that we did not identify were supported by less than a single read per 100 000 on average and we required that a TSS is supported by at least 1 read (see Fig. 1a).
- Finally, 174 (30% of the curated TSSs that we did not retrieve) did not pass our enrichment criteria (see Fig.1c). Accepting these TSSs as putative TSSs would have to be accompanied by the inclusion of many false positives.

In summary, 70% of the manually curated TSSs that are not in the 'TSSer' prediction set were not lost due to TSSer scoring but rather before because they had little evidence of expression, even though we mapped 70.43% of the reads to the genome, compared with 80.86% in the original analysis (Sharma *et al.*, 2010). Only 30% of the TSSs that were in the reference list were not

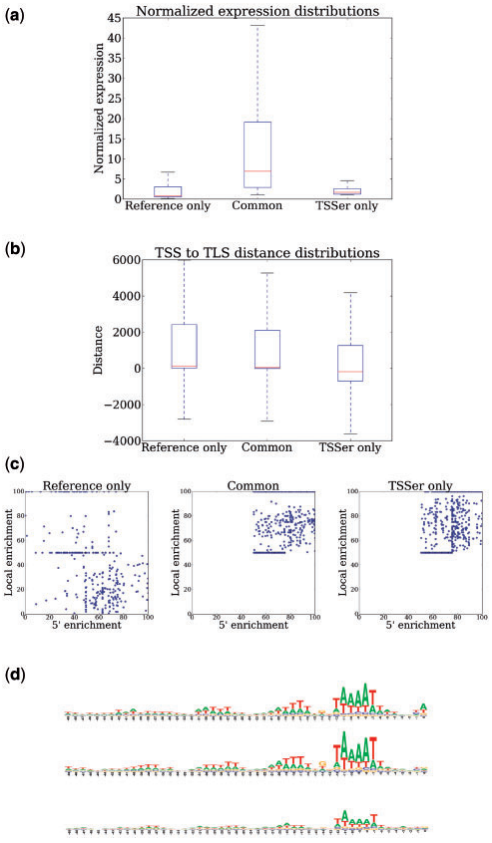


Fig. 1. Properties of TSSs that were present only in the reference list (left), both in the reference and the TSSer list (middle) or only in the TSSer list (right). (**a**) Box plot of averaged normalized expression (the boxes are drawn from the first to the third quantile and the median is shown with the red line). (**b**) Box plot of the displacement distribution relative to the start codon. (**c**) Scatterplots of 5′ versus local enrichment (both shown as percentage). (**d**) Sequence logos indicating the position-dependent (5′ → 3′ direction) frequencies of nucleotides upstream of the TSS (datasets are shown from top to bottom rather than from left to right)

present in the TSSer list because they did not satisfy our criteria for enrichment in reads. Further investigating the features [enrichment values, distance to start codon (TLS) and presence of transcriptional signals (see Supplementary Material)] of these TSSs that we did not identify, we found that a large proportion are likely to be *bona fide* TSSs, i.e. false negatives of our method.

On the other hand, we identified an even larger number of TSSs (1060) that were not present in the curated list. We refer to these as 'TSSer only'. Of these, 198 TSSs correspond to 142 genes that were not present in the reference list. Of the remaining 862 TSSs that are only identified by our method, 287 TSSs are 'Antisense' TSSs, 58 TSSs are 'Orphan' and 379 TSSs are alternative TSSs for genes that did have at least one annotated TSS in the reference set (the definition of these categories is given in Section 2.3 of Supplementary Material). These TSSs share the properties of TSSs jointly identified by our method and the manual curation (Fig. 1), indicating that they are also *bona fide* TSSs. To further support the TSSs that were identified by TSSer and were missing in the reference list, we compared these TSSs with the 'Common' category and also 'Reference only' category in the following aspects:

- Average normalized expression (Fig. 1a): 'TSSer only' TSSs have almost the same expression distribution as TSSs in 'Reference only' category and both have lower expression compared with the TSSs in the 'Common' set. This indicates that TSSs with high expression are equally well identified by the two methods, and that the difference between methods manifests itself at the level of TSSs with low expression.
- TSS to TLS distance: Figure 1b shows that TSSer identifies putative TSSs that are closer, on average, to the translation start, compared with the TSSs that were manually curated. The proportion of internal TSS identified by TSSer is also higher and it remains to be determined what proportion of these represents *bona fide* transcription initiation starts.
- Enrichment values: Figure 1c shows that TSSs identified by TSSer only have strong 5' and local enrichment, whereas those that are present in the 'Reference only' set have low local enrichment. This indicates that these sites are located in neighborhoods that give comparable initiation at spurious sites and thus these sites would be difficult to identify simply based on their expression parameters.
- Strength of transcriptional signals: Figure 1d shows that TSSs identified by TSSer share transcriptional signals such as the −10 box with the other categories of sites. The overall weaker sequence bias may indicate that a larger proportion of 'TSSer only' sites are false positives, consistent with the higher proportion of sites that TSSer identified downstream of start codons (Fig. 1a). To further investigate the transcription regulatory signals, we also implemented a hidden Markov model (HMM) that we trained on the 'Common' sites to find transcription regulatory motifs. We then applied this model to the sequences from each individual subset (see Supplementary Material for details). The results from the HMM further confirm that a large proportion of the 'TSSer only' sites have similar scores to the sites in the other two categories, indicating that TSSer captures a substantial number of *bona fide* TSSs that were not captured during manual curation.

5 DISCUSSION

Deep sequencing has truly revolutionized molecular biology. It enabled not only the assembly of the genomes of thousands of species, but also annotation of transcribed regions in these genomes and the generation of a variety of maps for DNA-binding factors, non-coding RNAs and RNA-binding factors. High-throughput studies revealed that not only eukaryotic but also

prokaryotic genomes are more complex than initially thought. In particular, bacterial genomes encode relatively large numbers of non-coding RNAs with regulatory functions (Waters and Storz, 2009) and antisense transcripts (Georg and Hess, 2011). Such transcripts are of particular interest because they are frequently produced in response to and contribute to the adaptation to specific stimuli (Repoila and Darfeuille, 2009). The availability of a large number of bacterial genomes further enables identification of regulatory elements through comparative genomics-based approaches (Arnold *et al.*, 2012). However, these methods benefit from accurate annotation of TSSs that enables a focused search for transcription factor binding sites. Although the data supporting TSS identification can be obtained with relative ease (Sharma *et al.*, 2010), the annotation of TSSs has so far been carried out manually, which is tedious and likely leads to an incomplete set of TSSs. Only recently, as our manuscript was in the review process, methods for automated annotation of TSSs based on dRNA-seq data started to emerge (Dugar *et al.*, 2013) (see also http://www.tbi.univie.ac.at/newpapers/pdfs/TBI-p-2013-4.pdf). The method that we propose here is meant to provide a starting point into the process of TSS curation. Because it uses dRNA-Seq data, it is clear that only TSSs from which there is active transcription during the experiment can be annotated. As we have determined in the benchmark against the *H.pylori*, there remain TSSs for which the expression evidence is poor, yet have the properties of *bona fide* TSSs. Additional samples, covering conditions in which these TSSs are expected to be expressed are necessary to identify them. Alternatively, they can be brought in during the process of manual curation. Nonetheless, the advantage of an unbiased automated method such as the one we propose here is that it allows the discovery of TSSs that may not be expected or easily evaluated such as those of antisense transcripts, alternative TSSs and TSSs corresponding to novel genes. Furthermore, this method can provide an initial set of high-confidence TSSs that can be used to train more complex models of transcription regulation, which could be used to iteratively identify additional TSSs, that may be supported by a small number of reads. To illustrate this point, we here used an HMM, which we trained on high-confidence TSSs from the 'Common' category, to provide an additional list of putative TSSs that appear to have appropriate transcription regulatory signals but that were not captured with high abundance or enrichment in the experiment (Supplementary Table S8). Thirty-six percent of the TSSs that were only present in the reference annotation are part of this list. More sophisticated versions of this approach could be used toward comprehensive annotation of TSSs in bacterial genomes. Finally, the method can be applied to other systems in which genomic regions give rise to an increased number of transcripts in specific conditions.

6 CONCLUSION

We have proposed an approach for genome-wide identification of TSSs in bacteria, which uses dRNA-Seq data to quantify the 5′ and local enrichment in reads at putative TSSs and their corresponding significance. The method is implemented in an automated pipeline, which we applied to several recently published dRNA-Seq datasets. A thorough benchmarking of the TSSs proposed by our method relative to manual curation indicates that the method performs well in identifying known TSSs and is able to further detect novel TSSs that have the expected properties of *bona fide* TSS. Thus, our method should enable rapid identification of TSSs in bacterial genomes starting from dRNA-Seq data.

ACKNOWLEDGEMENTS

The authors thank A. R. Gruber and A. Rzepiela for critical reading of the manuscript.

Funding: Work in the Zavolan laboratory is supported by the University of Basel and the Swiss National Science Foundation (grant number 31003A_147013).

Conflict of Interest: none declared.

REFERENCES

Albrecht,M. *et al.* (2011) The transcriptional landscape of Chlamydia pneumoniae. *Genome Biol.*, **12**, R98.

Arnold,P. *et al.* (2012) MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences. *Bioinformatics*, **28**, 487–494.

Cho,B.K. *et al.* (2003) The transcription unit architecture of the *Escherichia coli* genome. *Nat. Biotechnol.*, **27**, 1043–1049.

Dugar,G. *et al.* (2013) High-resolution transcriptome maps reveal strain-specific regulatory features of multiple Campylobacter jejuni isolates. *PLoS Genet.*, **9**, e1003495.

Georg,J. and Hess,W.R. (2011) cis-antisense RNA, another level of gene regulation in bacteria. *Microbiol. Mol. Biol. Rev.*, **75**, 286–300.

Kröger,C. *et al.* (2012) The transcriptional landscape and small RNAs of Salmonella enterica serovar Typhimurium. *Proc. Natl Acad. Sci. USA*, **109**, 1277–1286.

Repoila,F. and Darfeuille,F. (2009) Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. *Biol. Cell*, **101**, 117–131.

Sharma,C.M. *et al.* (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, **464**, 250–255.

Shiraki,T. *et al.* (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl Acad. Sci. USA*, **100**, 15776–15781.

Waters,L.S. and Storz,G. (2009) Regulatory RNAs in bacteria. *Cell*, **136**, 615–628.

Supplementary Information, Materials and Methods

TSSer: An accurate method for identifying transcription start sites in prokaryotes from next generation sequencing data

Hadi Jorjani & Mihaela Zavolan

December 16, 2013

Contents

1 Read mapping and count normalization **1**

2 TSS identification **2**
2.1 Computation of the 5' enrichment 2
2.1.1 z-score 2
2.1.2 λ-score 5
2.2 Single linkage clustering 7
2.3 Generating the list of high-confidence TSSs 7

3 Evaluation of the TSS identification method **7**
3.1 Hidden Markov Model of transcription regulatory elements 8

1 Read mapping and count normalization

We used in our study two pairs of cDNA libraries (TEX-untreated/treated) obtained from *Helicobacter pylori* cells in mid-log phase (ML-/+) or exposed to acid stress (AS-/+). These were the primary samples which were used for the initial annotation of TSSs in the *H.pylori* genome[1]. We obtained the raw data from the NCBI Short Read Archive (http://www.ncbi.nlm.nih.gov/Traces/sra), accession number SRA010186.
The raw data for Chlamydia and Salmonella can be obtained from the following links:

http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE24999
http://bioinf.gen.tcd.ie/~sathesh/Seqs/

Initial inspection of data sets generated with the dRNA-seq method revealed that a large proportion of sequences had trailing A nucleotides or nucleotides that could not be accurately called. Thus, we included in our processing procedure a 'cleaning step', in which we removed the adaptor sequence as well as trailing polyAs and polyNs (N - nucleotides that could not be accurately called). Because reads with long low complexity regions remained, we decided to map the sequences using the local sequence alignment program *BLAST* [2]. Then, for the inference of TSSs with TSSer we only considered sequences that had at least 18 nucleotides from the 5' end that were aligned to the genome, with at least 90% identity and at no more than 2 loci. In counting the reads associated with individual genomic loci, we weighted each read with $\frac{1}{\text{number of loci}}$, thus assuming that the read could have come from any of the loci to which it mapped equally well.

Supplementary Table 1: Mapping statistics for *Helicobacter pylori* samples: AS and ML stand for 'acid stress' and 'mid-log phase or control growth' respectively. + and − represent TEX-treated and TEX-untreated samples

Sample Name	Total reads	Mapped reads	Percent mapped	Percent mapped uniquely	Structural RNA content
AS+	540133	344332	63.75%	62.38%	47.77%
AS-	427455	307962	72.04%	63.47%	43.65%
ML+	528169	366775	69.44%	44.75%	66.00%
ML-	528373	406581	76.95%	27.67%	84.83%

A second observation that we made when initially inspecting the data was that the relative fraction of structural RNAs (i.e. ribosomal RNAs and tRNAs) differs dramatically between samples (see Supplementary Table 1 for the *Helicobacter* samples [1]), in a way that does not appear to be systematic. The terminator exonuclease degrades RNAs that have a 5' monophosphate group, but not those that have 5 tri-phosphate or hydroxyl. Structural RNAs such as rRNAs and tRNAs are processed (from polycistronic transcripts in the first case, by RNase P in the second case), have 5'-monophosphates and are therefore substrates of TEX. mRNAs, with 5-triphosphates, are not. We would thus expect then that TEX-treated samples are depleted in structural RNAs compared to the untreated samples, but that is not what we observed. We thus normalize the read counts to the total number of reads that map to regions other than those annotated as structural RNAs. To compare the read counts between samples we calculated the normalized count for each start site (the position to which the 5' end of a read maps) and whenever we use the term 'normalized expression' we mean relative to the total number of reads that do not map to regions annotated as structural RNAs.

2 TSS identification

We used two main criteria to automatically identify TSSs genome-wide. The first was that the putative TSS should have relatively more reads in the TEX-treated sample compared to the untreated one. We call this criterion *5' enrichment* and we quantify it via two different methods, to account for the possibility that the data includes or not replicates. In a first approach we quantify the 5' enrichment of particular genomic position in the TEX-treated compared to the untreated samples through the 'z-score'. In the second approach, we compute a probability that a genomic position is enriched across all multiple replicates of pairs of TEX-treated and untreated samples. The second criterion that we used to distinguish *bona fide* TSSs from background is based on the expectation that a real TSS is represented in a TEX-treated sample at a higher level compared to other genomic positions in relatively close vicinity. We call this criterion 'local enrichment'. Below we describe the computation of these quantities.

2.1 Computation of the 5' enrichment

2.1.1 z-score

The distribution of the number of reads associated with a specific TSS, which are derived from the mRNAs that were transcribed from that TSS, should follow a hypergeometric distribution. Because the number of reads associated with a given TSS is very small relative to the total number of reads, we approximate this hypergeometric distribution by a binomial distribution. Thus, assuming that a fraction f of the total number of mRNAs originates from a specific TSS, the probability to observe n reads from this TSS in a sample of N reads is given by

$$P(n|f,N) = \binom{N}{n} f^n (1-f)^{N-n}, \tag{1}$$

The mean and variance in the number of reads are given by $\langle n \rangle = Nf$ and $Var(n) = Nf(1-f)$, respectively. Applying Bayes' theorem, we obtain the posterior probability for f,

$$P(f|n,N) = (N+1)\binom{N}{n} f^n (1-f)^{N-n}, \tag{2}$$

with mean $\langle f \rangle = \frac{n+1}{N+1}$ and variance $Var(f) = \frac{(n+1)(n+2)}{(N+2)(N+3)}$. Having the posterior probability distribution for f we can define the enrichment at a particular genomic position as

$$P(f_+ > f_-|n_+, N_+, n_-, N_-) \tag{3}$$

We can write this equation in these two different forms, namely

$$P\left(f_+ > f_-|n_+, N_+, n_-, N_-\right) = P\left(f_+ - f_- > 0|n_+, N_+, n_-, N_-\right),$$

or

$$P\left(f_+ > f_-|n_+, N_+, n_-, N_-\right) = P\left(\frac{f_+}{f_-} > 1|n_+, N_+, n_-, N_-\right).$$

For the first form,

$$\begin{aligned} P\left(f_+ - f_- > 0|n_+, N_+, n_-, N_-\right) &= \int_0^1 \int_{f_-}^1 P\left(f_+, f_-|n_+, N_+, n_-, N_-\right) df_+ df_- \\ &= \int_0^1 \int_{f_-}^1 P\left(f_+|n_+, N_+\right) P\left(f_-|n_-, N_-\right) df_+ df_- \end{aligned} \tag{4}$$

Substituting Eq.2, the enrichment probability takes the form of an integral of an 'incomplete Beta function' which we cannot solve analytically.

$$\int_0^1 \int_{f_-}^1 (N_+ + 1)\binom{N_+}{n_+} {f_+}^{n_+} (1-f_+)^{N_+ - n_+} (N_- + 1)\binom{N_-}{n_-} {f_-}^{n_-} (1-f_-)^{N_- - n_-} df_+ df_-. \tag{5}$$

However, we can derive a Gaussian approximation as follows. Let us write the log-likelihood

$$log\left(P(f|n,N)\right) = G(f).$$

Expanding around the peak, which occurs at $a = \frac{n}{N}$, we have

$$G(f) = G(a) + \frac{\partial G}{\partial f}|_{f=a} \frac{(f-a)}{1!} + \frac{\partial^2 G}{\partial f^2}|_{f=a} \frac{(f-a)^2}{2!} + \dots \tag{6}$$

Considering that at the peak $\frac{\partial G}{\partial f} = 0$, we have

$$G(f) = G(a) + \frac{\partial^2 G}{\partial f^2}|_{f=a} \frac{(f-a)^2}{2!} + \dots \tag{7}$$

and

$$\begin{aligned} P(f|n,N) &= e^{G(f)} \\ &= e^{G(a) + \frac{\partial^2 G}{\partial f^2}|_{f=a} \frac{(f-a)^2}{2!}} \\ &= e^{G(a)} e^{\frac{\partial^2 G}{\partial f^2}|_{f=a} \frac{(f-a)^2}{2}}. \end{aligned}$$

We now calculate $\frac{\partial^2 G}{\partial f^2}|_{f=\frac{n}{N}}$:

$$\begin{aligned}
\frac{\partial^2 G}{\partial f^2} &= \frac{\partial^2 log[P(f|n,N)]}{\partial f^2} \\
&= \frac{\partial \frac{\partial log[P(f|n,N)]}{\partial f}}{\partial f} \\
&= \frac{\partial \frac{\partial log[(N+1)\binom{N}{n} f^n (1-f)^{N-n}]}{\partial f}}{\partial f} \\
&= \frac{\partial \frac{\partial [\log(N+1)+\log\binom{N}{n}+\log f^n+\log(1-f)^{N-n}]}{\partial f}}{\partial f} \\
&= \frac{\partial \frac{\partial [\log(N+1)+\log\binom{N}{n}+n\log f+(N-n)\log(1-f)]}{\partial f}}{\partial f} \\
&= \frac{\partial [\frac{n}{f} - \frac{N-n}{(1-f)}]}{\partial f} \\
&= -\frac{n}{f^2} + \frac{N-n}{(1-f)^2}
\end{aligned}$$

whose value at $f = \frac{n}{N}$ is given by:

$$\begin{aligned}
-\frac{n}{f^2} + \frac{N-n}{(1-f)^2}|_{f=\frac{n}{N}} &= -\frac{n}{(\frac{n}{N})^2} + \frac{N-n}{(1-\frac{n}{N})^2} \\
&= -\frac{N^2(N-2n)}{n(N-n)} \\
&=\approx -\frac{N^3}{n(N-n)}.
\end{aligned}$$

Thus, letting $\mu_f = \frac{n}{N}$ and $\sigma_f^2 = \frac{n(N-n)}{N^3}$, we have that

$$\begin{aligned}
P(f|n,N) &\approx e^{G(a)} e^{\frac{\partial^2 G}{\partial f^2}|_{f=a} \frac{(f-a)^2}{2}} \\
&\approx e^{G(a)} e^{-\frac{(f-\mu_f)^2}{2\sigma_f^2}} \\
&\approx \mathcal{N}(\frac{n}{N}, \frac{n(N-n)}{N^3})
\end{aligned}$$

Thus, we find that we can approximate $P\left(f|n,N\right)$ as a Gaussian. We can now derive a closed form for $P(f_+ - f_-|n_+, N_+, n_-, N_-)$ as it is the difference of two independent Gaussian distributions:

$$P(f_+ - f_-|n_+, N_+, n_-, N_-) \approx \mathcal{N}(x_+ - x_-, \frac{x_+(1-x_+)}{N_+} + \frac{x_-(1-x_-)}{N_-}) \tag{8}$$

with $x = \frac{n}{N}$ being the proportion of reads associated with the putative TSS. Moreover, $P(f_+ - f_-|n_+, N_+, n_-, N_-)$ is essentially the probability distribution of the standard score

$$P(f_+ - f_-|n_+, N_+, n_-, N_-) \approx \phi(\frac{x_+ - x_-}{\sqrt{\frac{x_+(1-x_+)}{N_+} + \frac{x_-(1-x_-)}{N_-}}}) \tag{9}$$

which we use to quantify the enrichment of the TSS in the TEX-treated compared to TEX-untreated sample.

2.1.2 λ-score

When we have multiple paired samples we can calculate the 5' enrichment for each genomic position in each TEX-treated compared to untreated sample and then evaluate the posterior probability that the mean of this distribution is greater than 1. Let us call λ_s the ratio of the normalized number of reads associated with a TSS in the TEX-treated compared to the untreated sample.

Assuming that the enrichment ratios λ_s have a Gaussian distribution across replicates (with mean and standard deviation μ and σ, respectively) we can calculate the posterior probability of the mean of this distribution being greater than one, which would correspond to the putative TSS being enriched, taking into account the evidence from all the pairs of TEX-treated and untreated samples. That is, the probability of the data, meaning the vector $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$ is given by

$$P(\lambda|\mu,\sigma) = \prod_{s=1}^{n} \frac{1}{\sigma\sqrt{2\pi}} e^{\frac{-(\lambda_s-\mu)^2}{2\sigma^2}}$$

Applying Bayes' theorem, we have that

$$P(\mu,\sigma|\lambda) = cP(\lambda|\mu,\sigma) = c\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{s=1}^{n}(\lambda_s-\mu)^2} \tag{10}$$

with c a constant.
The values of μ and σ that maximize $P(\mu,\sigma|\lambda)$ can be derived by solving $\frac{\partial P(\mu,\sigma|\lambda)}{\partial \mu} = 0$ and $\frac{\partial P(\mu,\sigma|\lambda)}{\partial \sigma} = 0$ respectively, and are

$$\mu_* = \langle \lambda_s \rangle \tag{11}$$

$$\sigma_*^2 = \langle (\lambda_s - \langle \lambda_s \rangle)^2 \rangle \tag{12}$$

To calculate $P(\mu > 1|\lambda)$ we must first determine the posterior probability of μ which we do by integrating over σ in Eq. 10:

$$\begin{aligned} P(\mu|\lambda) &= \int_0^\infty P(\mu,\sigma|\lambda)d\sigma \\ &= \int_0^\infty c\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{s=1}^{n}(\lambda_s-\mu)^2} d\sigma \\ &= c(2\pi)^{-\frac{n}{2}} \int_0^\infty \left(\frac{1}{\sigma}\right)^n e^{\frac{-c_\mu}{2\sigma^2}} d\sigma, \end{aligned}$$

with $c_\mu = \sum_{s=1}^{n}(\lambda_s - \mu)^2$. Performing the integral of the Gamma function we obtain

$$\begin{aligned} P(\mu|\lambda) &= c(2\pi)^{-\frac{n}{2}} \left[2^{\frac{n-3}{2}} c_\mu^{\frac{1-n}{2}} \Gamma\left(\frac{n-1}{2}\right)\right] \\ &= k c_\mu^{\frac{1-n}{2}} \end{aligned} \tag{13}$$

where $k = c2^{\frac{-3}{2}}\pi^{-\frac{n}{2}}\Gamma(\frac{n-1}{2})$ is a constant.

Now we can calculate $P(\mu > \mu_0|\lambda)$ given Eq. 13 as follows:

$$\begin{aligned}
P(\mu > \mu_0|\lambda) &= \int_{\mu_0}^{\infty} P(\mu|\lambda)d\mu \\
&= \int_{\mu_0}^{\infty} kc_{\mu}{}^{\frac{1-n}{2}} d\mu \\
&= k\int_{\mu_0}^{\infty} (\sum_{s=1}^{n}(\lambda_s - \mu)^2)^{\frac{1-n}{2}} d\mu \\
&= k\int_{\mu_0}^{\infty} [\sum_{s=1}^{n}(\lambda_s^2 - 2\mu\lambda_s + \mu^2)]^{\frac{1-n}{2}} d\mu \\
&= k\int_{\mu_0}^{\infty} [n\sum_{s=1}^{n}\frac{(\lambda_s^2 - 2\mu\lambda_s + \mu^2)}{n}]^{\frac{1-n}{2}} d\mu \\
&= kn\int_{\mu_0}^{\infty} [\langle\lambda_s^2\rangle - 2\mu\langle\lambda_s\rangle + \mu^2]^{\frac{1-n}{2}} d\mu \\
&= kn\int_{\mu_0}^{\infty} [\langle\lambda_s^2\rangle - \langle\lambda_s\rangle^2 + \langle\lambda_s\rangle^2 - 2\mu\langle\lambda_s\rangle + \mu^2]^{\frac{1-n}{2}} d\mu \\
&= kn\int_{\mu_0}^{\infty} [\sigma_*^2 + (\mu - \mu_*)^2]^{\frac{1-n}{2}} d\mu
\end{aligned}$$

Thus,

$$P(\mu > \mu_0|\lambda) = K\int_{\mu_0}^{\infty} [\sigma_*^2 + (\mu - \mu_*)^2]^{\frac{1-n}{2}} d\mu \tag{14}$$

where K is a constant which can be calculated from the constraint that

$$P(\mu > 0|\lambda) = K\int_{0}^{\infty} [\sigma_*^2 + (\mu - \mu_*)^2]^{\frac{1-n}{2}} d\mu = 1 \tag{15}$$

Therefore $K = \frac{1}{\int_0^{\infty} [\sigma_*^2 + (\mu - \mu_*)^2]^{\frac{1-n}{2}} d\mu}$ and considering Eq. 14 we will have

$$P(\mu > \mu_0|\lambda) = \frac{\int_{\mu_0}^{\infty} [\sigma_*^2 + (\mu - \mu_*)^2]^{\frac{1-n}{2}} d\mu}{\int_{0}^{\infty} [\sigma_*^2 + (\mu - \mu_*)^2]^{\frac{1-n}{2}} d\mu} \tag{16}$$

Finally, the quantity in which we are interested:

$$P(\mu > 1|\lambda) = \frac{\int_{1}^{\infty} (\frac{1}{(\mu - \mu_*)^2 + \sigma_*^2})^{\frac{n-1}{2}} d\mu}{\int_{0}^{\infty} (\frac{1}{(\mu - \mu_*)^2 + \sigma_*^2})^{\frac{n-1}{2}} d\mu} \tag{17}$$

The expression depends on the enrichment factors λ. Rather than using the maximum likelihood values of f_+ and f_-, we compute the *expected value* of the ratio of these two frequencies. This can be shown to take the value

$$\begin{aligned}
\lambda_s &= \langle\frac{f_+}{f_-}\rangle \\
&= \frac{\frac{n_+ + 1}{N_+ + 2}}{\frac{n_-}{N_- + 1}}
\end{aligned}$$

which can be approximated as $\lambda_s = \frac{x_+}{x_-}$, with $x_+ = \frac{n_+}{N_+}$ and $x_- = \frac{n_-}{N_-}$.

2.2 Single linkage clustering

Observing that many of the well-represented TSSs were associated with reads that started at closely-spaced positions, we applied single-linkage clustering to the set of putative TSSs before generating our list of high-confidence TSSs. The selected distance for clustering should be large enough to cluster the putative start sites in close vicinity of each other which are results of imprecise transcription initiation and should be small enough not to cluster alternative transcription start sites. Here we used 10 nucleotides as the single-linkage clustering distance. From a single-linkage cluster we reported the site with the highest average expression in the TEX-treated samples.

2.3 Generating the list of high-confidence TSSs

To define a list of high-confidence TSSs, we selected cut-off values for our parameters that allowed inclusion of most annotated genes while keeping the total number of TSSs close to total number of annotated genes ($0.5\times$ number of genes $<$ total TSS $<$ $1.5\times$ number of genes). For the *Helicobacter* genome, this selection corresponded to values of 50% minimum local and 5' enrichment and 1.0 for average normalized expression. A list of different cut-off values for the TSSer parameters and their associated number of identified TSSs is given in Supplementary Table 7. We obtained a total of 2366 predicted TSSs, classified as follows:

Supplementary Table 2: Representation of various types of TSSs in the *Helicobacter pylori* dRNA-seq data

Total number of TSSs	Primary	Antisense	Internal	Orphan
2366	984	751	602	129

The annotated TSSs were grouped hierarchically into one of these four categories according to their relative position to the closest annotated gene. Primary TSSs are defined to be those within a distance of $\leq$ 300 nucleotides upstream of an annotated open reading frame (ORF) or up to $\leq$ 100 nucleotides downstream from the start codon. Antisense TSSs are those situated inside or within $\leq$ 100 nucleotides of an annotated ORF on the opposite strand. Internal TSSs are defined to be those within an annotated ORF on the sense strand. Finally, orphan TSSs are those that have no annotated ORF in close proximity.

3 Evaluation of the TSS identification method

To evaluate the accuracy of our TSS identification method, we benchmarked it against the manually-constructed TSS map of *Helicobacter pylori*. After removing TSSs corresponding to structural RNAs, 1893 manually curated TSSs remained. We referred to these as the 'Reference' set. Considering two TSSs which are at most 5 nucleotides away from each other as shared, we found that 1306 (69%) of the TSSs on the reference list are also identified by our method. The other 31% of TSSs in the reference list were not present in our list. On the other hand, we identified 1060 TSSs that were not present in the reference set.

We defined the following categories of TSSs:

- Those only present on the reference list (587 TSSs), which we refer to as 'Reference only'.
- Those identified both by our method and also through manual curation (1306 TSSs). We refer to this set as the 'Common' set.
- Those identified only by our method (1060 TSSs), to which we refer as 'TSSer only'.

We then compared the properties of these categories of TSSs. The results, summarized in Figure 1 of the main manuscript, indicate that TSSer indeed identifies a large number of TSSs that are not present in the reference list, yet whose properties are very similar to those of high-confidence TSSs. Namely, panel (a) of the figure indicates that TSSs that were only identified by TSSer have higher expression compared to those that were only on the reference list, panel (b) indicates that they are located closer to the translation start, and panel (c) indicates that TSSs identified by TSSer have stronger enrichment (particularly local enrichment) compared to TSSs which are only present in the reference set.

For *Helicobacter*, the number of dRNA-seq samples was rather large, covering a few conditions with distinct expression patterns. Other data sets are typically smaller, so we would not expect to get a total number of TSSs that is in the range of the number of genes. Nonetheless, enrichment thresholds of 40 to 60 percent appear to give good results on data from at least two other species, as summarized below.

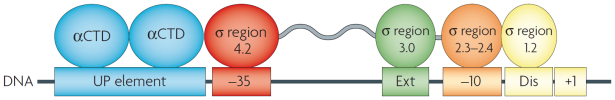
Supplementary Table 3: Information related to investigated organisms

	Helicobacter	*Salmonella*	*Chlamydia*
5' enrichment cut-off	50%	50%	50%
Local enrichment cut-off	50%	50%	60%
Normalized expression cut-off	1.0	1.0	1.0
Single-linkage distance (nt)	10	10	10
Total identified TSSs	2366	1574	1234
Common TSSs	1306	826	262
Reference only	587	992	272
TSSer only	1060	748	972

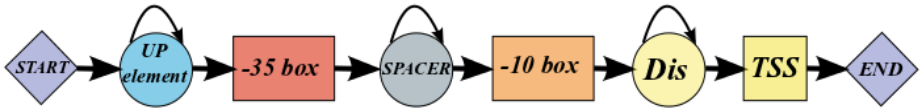
3.1 Hidden Markov Model of transcription regulatory elements

To uncover additional evidence for the putative TSSs identified by our method being *bona fide* TSS, we modeled the transcriptional signals that are known to be present upstream of the TSS in bacteria. In particular, bacterial transcription appears to be dependent on motifs that are located at 35 and 10 nucleotides upstream of the TSS, which are called '-35 box' and 'Pribnow box' motifs [3]. We thus trained a Hidden Markov Model (HMM) with the structure shown in Supplementary Figure 1(b) on the set of common TSSs and applied this model to all putative TSSs, either generated by our method or present on the reference list, to compute the posterior score for each sequence upstream of a putative TSS. To train the HMM we applied the *Baum-Welch algorithm* and to calculate the probability of each sequence we used the *Forward* algorithm (see ref.[4] chapter 3). The results are summarized in Supplementary Figure 2. As is apparent from Supplementary Figure 1(c), the Pribnow box motif is very clear but the -35 motif is not, in line with the results reported by Sharma et al. [1]. Nonetheless, the HMM captures the A/T-rich bias of the region upstream of the Pribnow box, as reported by Sharma et al. [1]. The "TSSer only" category is almost twice the size and more heterogeneous than the "Reference only" category (Supplementary Figure 1(b)). If we select the 587 TSSs with the highest HMM score (the same number as contained in the "Reference only" data set), these TSSs are very similar to those in the "Reference only" set (Supplementary Figure 3). This suggests that TSSer identifies a large number of *bona fide* TSSs that were not present in the reference. It further suggests a strategy to refine the TSS list. Namely, one could use the sites with the most clear 5' and local enrichment to abstract a model of the transcription regulatory signals, and then apply this model to putative TSSs that are less clear in the expression data to construct a more comprehensive TSS annotation. We used the HMM posterior score as a measure of the strength of transcriptional signals. From the putative start sites that had at least one mapped read in at least one of the TEX^+ but did not pass our initial criteria for expression or enrichment, we found an additional 1992 that

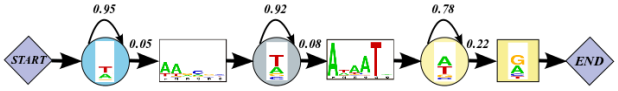
had a posterior score at least as high as the average of the "Common" set. These TSSs are listed in Supplementary Table 8, and they include a further 211 TSSs from "Reference Only" category.



(a)

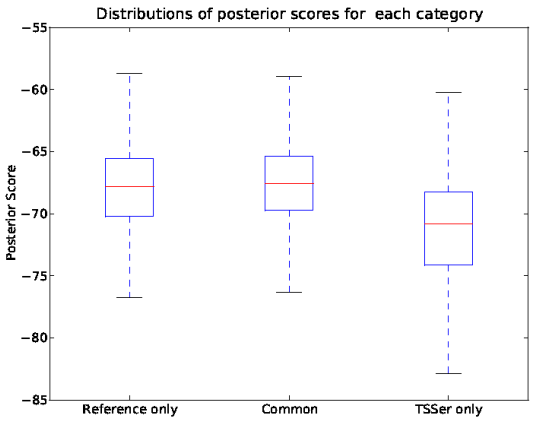


(b)

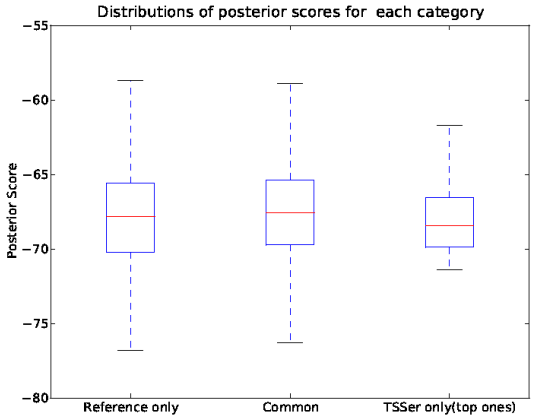


(c)

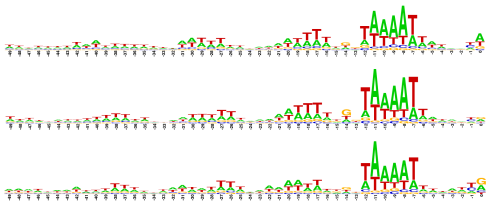
Supplementary Figure 1: (a) DNA elements and RNA polymerase modules that contribute to promoter recognition by σ^{70} [3] (b) Structure of the Hidden Markov Model to detect transcription regulatory signals. In each of '-35 box' and 'Pribnow box' states six nucleotides are emitted according to probabilities which can be summarized in weight matrices associated with these states, and in the other states only mono-nucleotides are emitted.(c) Illustration of HMM trained on the "Common" set of TSSs



Supplementary Figure 2: Distribution of sequence scores for each TSS category calculated based on trained HMM.



(a)



(b)

Supplementary Figure 3: Properties of TSSs that were present only in the reference list (left panels), both in the reference and the TSSer list (middle panels), or the top ones from the TSSer only category (right panels). (a). Box plots representing HMM posterior score distributions for each category. (b). Sequence logos indicating the position-dependent (5'→3' direction) frequencies of nucleotides upstream of the TSS (data sets are shown from top to bottom ("Reference only", "Common", "TSSer only") rather than from left to right).

References

[1] Sharma CM, Hoffmann S, Darfeuille F, Reignier J, Findeiss S, Sittka A, Chabas S, Reiche K, Hackermuller J, Reinhardt R, Stadler PF, Vogel J: **The primary transcriptome of the major human pathogen Helicobacter pylori**. *Nature* 2010, **464**(7286):250–255.

[2] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403–410.

[3] Haugen SP, Ross W, Gourse RL: **Advances in bacterial promoter recognition and its control by factors that do not bind DNA**. *Nat Rev Microbiol* 2008, **6**(7):507–519.

[4] Durbin R, Eddy S, Krogh A, Mitchison G: **Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge Univ** 1998.

3 Insights into snoRNA biogenesis and processing

Kishore *et al. Genome Biology* 2013, **14**:R45
http://genomebiology.com/2013/14/5/R45

Genome Biology

RESEARCH Open Access

Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing

Shivendra Kishore†, Andreas R Gruber†, Dominik J Jedlinski, Afzal P Syed, Hadi Jorjani and Mihaela Zavolan*

Abstract

Background: In recent years, a variety of small RNAs derived from other RNAs with well-known functions such as tRNAs and snoRNAs, have been identified. The functional relevance of these RNAs is largely unknown. To gain insight into the complexity of snoRNA processing and the functional relevance of snoRNA-derived small RNAs, we sequence long and short RNAs, small RNAs that co-precipitate with the Argonaute 2 protein and RNA fragments obtained in photoreactive nucleotide-enhanced crosslinking and immunoprecipitation (PAR-CLIP) of core snoRNA-associated proteins.

Results: Analysis of these data sets reveals that many loci in the human genome reproducibly give rise to C/D box-like snoRNAs, whose expression and evolutionary conservation are typically less pronounced relative to the snoRNAs that are currently cataloged. We further find that virtually all C/D box snoRNAs are specifically processed inside the regions of terminal complementarity, retaining in the mature form only 4-5 nucleotides upstream of the C box and 2-5 nucleotides downstream of the D box. Sequencing of the total and Argonaute 2-associated populations of small RNAs reveals that despite their cellular abundance, C/D box-derived small RNAs are not efficiently incorporated into the Ago2 protein.

Conclusions: We conclude that the human genome encodes a large number of snoRNAs that are processed along the canonical pathway and expressed at relatively low levels. Generation of snoRNA-derived processing products with alternative, particularly miRNA-like, functions appears to be uncommon.

Background

Small nucleolar RNAs (snoRNAs) are a specific class of small non-protein coding RNAs that are best known for their function as guides of modifications (2'-O-methylation and pseudouridylation) of other non-protein coding RNAs such as ribosomal, small nuclear and transfer RNAs (rRNAs, snRNAs and tRNAs, respectively) [1-3]. Based on sequence and structural features, snoRNAs are divided into two classes. C/D box snoRNAs share the consensus C (RUGAUGA, R = A or G) and D (CUGA) box motifs, which are brought into close proximity by short regions of complementarity between the snoRNA 5' and 3' ends [4,5] and are bound by the four core proteins of the small ribonucleoprotein complex (snoRNP), namely 15.5K, NOP56, NOP58 and Fibrillarin (FBL) [6-8] during snoRNA maturation. Fibrillarin is the methyltransferase that catalyzes the 2'-O-methylation of the ribose in target RNAs [9]. Most C/D box snoRNAs also contain additional conserved C' and D' motifs located in the central region of the snoRNA. The other class of snoRNAs is defined by a double-hairpin structure with two single-stranded H (ANANNA, N = A, C, G or U) and ACA box domains [10], and are therefore called H/ACA box snoRNAs. They associate with four conserved proteins, Dyskerin (DKC1), Nhp2, Nop10 and Gar1, to form snoRNPs that are functionally active in pseudouridylation. Although all four H/ACA proteins are necessary for efficient pseudouridylation [10], it is Dyskerin that provides the pseudouridine synthase activity [11]. While H/ACA and C/D box snoRNAs accumulate in the nucleolus, some snoRNAs reside in the nucleoplasmic Cajal bodies (CBs) where they guide modifications of snRNAs [2] and are called small Cajal body-specific RNAs

* Correspondence: mihaela.zavolan@unibas.ch
† Contributed equally
Computational and Systems Biology, Biozentrum, University of Basel, Klingelbergstrasse 50-70, 4056 Basel, Switzerland

(scaRNAs). In addition to the typical H/ACA snoRNA features, vertebrate H/ACA box scaRNAs carry a CB localization signal called CAB box (UGAG) in the loop of their 5' and/or 3' hairpins [12].

Immediately upstream of the D box and/or the D' box, C/D box snoRNAs contain 10 to 21 nucleotide-long antisense elements that are complementary to sites in their target RNAs [13-15]. The nucleotide in the target RNA that is complementary to the fifth nucleotide upstream from the D/D' box of the snoRNA is targeted for 2'-O-methylation by the snoRNP [14,15]. H/ACA box snoRNAs contain two antisense elements termed pseudouridylation pockets, located in the 5' and 3' hairpin domains of the snoRNA [16,17]. Substrate uridines are selected through base-pairing interactions between the pseudouridylation pocket and target RNA sequences that flank the targeted uridine.

Deep-sequencing studies revealed a surprising diversity of small RNAs derived from non-coding RNAs (ncRNAs) known as small derived RNAs (sdRNAs) with well-established functions such as tRNAs [18,19], Y RNAs [20], vault RNAs [21], ribosomal RNAs [22], spliceosomal RNAs [23] and snoRNAs [24-26]. In fact, the profile of sequenced reads observed for some of these small RNA species are very characteristic and have even been used for ncRNA gene finding based on sequencing data [27,28]. The majority of C/D box and H/ACA snoRNAs seems to be extensively processed, producing stable small RNAs from the termini of the mature snoRNA [29] and the processing pattern is conserved across cell types [30]. Thus, it appears that snoRNAs are versatile molecules that give rise to snoRNA-derived miRNAs [24,31], other small RNAs [25,29] or longer processing fragments [32].

To gain insight into the complexity of snoRNA processing and the functional relevance of the derived sdRNAs, we undertook a comprehensive characterization of products generated from snoRNA loci, combining high-throughput sequencing of long and short RNA fragments with photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation (PAR-CLIP) of core snoRNA-associated proteins and with data from Argonaute 2 (Ago2) immunoprecipitation sequencing (IP-seq) experiments. We found that many loci in the human genome can give rise to C/D box-like snoRNAs. Among the novel snoRNAs that we identified are very short sequences, extending little beyond the C and D boxes, which are essential for the binding of core snoRNA proteins. Compared to the snoRNAs that are already known, the novel snoRNA candidates exhibit a lower level of evolutionary conservation and a lower expression level. These findings indicate that the C/D box snoRNA structure evolves relatively easily and that C/D box snoRNA-like molecules are produced from many more genomic loci than are currently annotated. We further found that C/D box snoRNAs are very specifically processed inside the regions of terminal complementarity, retaining in the mature form only four to five nucleotides upstream of the C box and two to five nucleotides downstream of the D box. Sequencing of the small RNA population as well as of the small RNAs isolated after Ago2 immunoprecipitation revealed that despite their cellular abundance, C/D box-derived small RNAs are not efficiently incorporated into the Ago2 protein. Our extensive data thus indicate that, contrary to previous suggestions [25,33], snoRNA-derived small RNAs that carry out non-canonical, particularly miRNA-like, functions are rare.

Results

PAR-CLIP of C/D box and H/ACA box snoRNP core proteins identifies their RNA binding partners

To investigate the RNA population comprehensively that associates with both C/D box and H/ACA box small nucleolar ribonucleoproteins we performed PAR-CLIP as previously described [34] with antibodies against the endogenous Fibrillarin (FBL), NOP58 and Dyskerin (DKC1) proteins, in HEK293 cells (for details see Materials and methods). For NOP56 we used a stable cell line expressing FLAG-tagged NOP56 and anti-FLAG antibodies. Because we recently found that the choice of the ribonuclease and reaction conditions influences the set of binding sites obtained through cross-linking and immunoprecipitation (CLIP) [35], we also generated a Fibrillarin PAR-CLIP library employing partial digestion with micrococcal nuclease (MNase) instead of RNase T1. PAR-CLIP libraries were sequenced on Illumina sequencers, mapped and annotated through the CLIPZ web server [36]. The obtained libraries were comparable to those from previous PAR-CLIP studies in terms of size, rates of mapping to genome and proportion of cross-link-indicative T→C mutations (Table 1). The DKC1 PAR-CLIP library shows a lower frequency of T→C mutations compared to all other libraries, but T→C mutations were still the most frequent in this library as well (data not shown).

Compared to the libraries that we previously generated for HuR and Ago2 [35], two proteins whose primary targets are mRNAs, we found that snoRNAs, rRNAs and snRNAs were strongly enriched in PAR-CLIP libraries generated for the snoRNP core proteins (Table 1). The fact that not only snoRNAs but also the primary targets of snoRNAs, namely ribosomal RNAs and small nuclear RNAs, are enriched in these samples suggests that like Ago2 cross-linking, which captures both miRNAs and their targets [34,35], cross-linking of core snoRNPs efficiently captures both snoRNAs and targets. To quantify the specificity of our PAR-CLIP libraries, we intersected the 200 clusters with the highest read density per nucleotide from each library with curated snoRNA gene annotations based on snoRNA-LBME-db [37] (Table 2). Currently, snoRNA-LBME-db lists about

Table 1 Summary of CLIPZ mapping statistics and annotation categories for PAR-CLIP samples.

Feature	FBL	FBL (MNase)	NOP56	NOP58 rep A	NOP58 rep B	DKC1	Ago2 rep A	HuR rep A
Mapping rate	60.47%	73.3%	26.6%	41.4%	46.6%	47.5%	67.9%	72.4%
Library size	3,755,090	7,396,138	2,789,209	3,678,032	3,798,895	7,727,966	5,899,130	5,491,479
T→C mutations among all observed mutations	64.8%	57.7%	48.6%	67.9%	73.0%	19.7%	55.8%	58.8%
snoRNAs	33.79%	31.55%	29.95%	39.05%	44.10%	13.13%	0.18%	0.01%
snRNAs	20.87%	33.17%	15.45%	22.36%	25.60%	10.18%	0.28%	0.02%
rRNAs	18.64%	13.83%	8.12%	7.42%	7.16%	15.53%	1.07%	0.17%
mRNAs	14.47%	11.61%	22.27%	19.42%	15.14%	17.40%	50.07%	47.87%
Repeats	6.42%	1.60%	15.51%	6.08%	3.36%	18.39%	11.29%	42.08%
tRNAs	1.57%	2.67%	2.44%	0.99%	0.57%	5.10%	0.75%	0.14%
miRNAs	0.07%	0.18%	0.02%	0.01%	0.01%	0.05%	20.41%	00.00%
Other Categories	2.74%	3.66%	3.01%	2.98%	2.78%	2.80%	3.86%	1.99%
No annotation	1.43%	1.74%	3.21%	1.69%	1.27%	17.43%	12.10%	7.71%

Ago2: Argonaute 2; DKC1: Dyskerin; FBL: Fibrillarin; miRNA: micro RNA; MNase: micrococcal nuclease; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; rRNA: ribosomal RNA; snoRNA: small nucleolar RNA; snRNA: small nuclear RNA; tRNA: transfer RNA

153 human C/D box snoRNA loci and 108 human H/ACA box snoRNA loci that are known to be ubiquitously expressed. For each of the C/D box specific PAR-CLIP libraries, more than 100 of the top 200 clusters could be assigned to C/D box snoRNAs indicating the specificity of our CLIP experiments and the broad coverage of the snoRNA genes by the sequencing reads obtained from HEK293 cells. The Dyskerin PAR-CLIP data set showed a weaker enrichment in snoRNAs compared to the data sets for the core C/D box-specific proteins, with 57% of all known H/ACA box snoRNAs being represented among the 200 top-ranking clusters. scaRNAs were detected in both H/ACA box and C/D box specific libraries, as expected because many scaRNAs have both C/D box and H/ACA box elements. Finally, minor fractions of H/ACA box snoRNAs were also found in PAR-CLIP libraries of the C/D box-specific proteins, and *vice versa*. This could be caused by the close spatial arrangement of snoRNPs on the target molecule, or could indicate that H/ACA box snoRNAs and C/D box snoRNAs guide modifications on each other.

Binding patterns of core proteins on snoRNAs

As mentioned in the introduction, both C/D box and H/ACA box snoRNAs carry very specific functional sequence and structure elements, which are recognized by the snoRNP core proteins. We thus asked whether different C/D box core proteins have distinct preferences in binding different regions of the C/D box snoRNAs. Figure 1A depicts PAR-CLIP read profiles along selected snoRNA genes (profiles for all scaRNA and snoRNA genes are in Additional file 1). Both C/D box core proteins as well as the H/ACA box specific Dyskerin bind to SCARNA6, which has a hybrid structure composed of both C/D box and H/ACA box elements. However, while the CLIP reads from the Fibrillarin, NOP56 and NOP58 samples cover the C and D box motifs, Dyskerin was preferentially cross-linked to the H-box motif and to the 5' end of the first H/ACA box stem. For the C/D box snoRNAs, different snoRNA core proteins gave very similar cross-linking patterns (Figure 1B), which we quantified through the correlation coefficient between read densities obtained along

Table 2 Annotation summary of the top 200 clusters inferred from PAR-CLIP experiments with snoRNA core proteins.

PAR-CLIP library	C/D box snoRNAs	H/ACA box snoRNAs	scaRNAs	mRNA exons	Other
FBL	123 (61.5%)	9 (4.5%)	10 (5.0%)	5 (2.5%)	53 (26.5%)
FBL (MNase)	106 (53.0%)	16 (8.0%)	10 (5.0%)	26 (13.0%)	42 (21.0%)
NOP56	115 (57.5%)	28 (14.0%)	15 (7.5%)	2 (1.0%)	40 (20.0%)
NOP58 rep A	114 (57.0%)	14 (7.0%)	10 (5.0%)	9 (4.5%)	52 (26.0%)
NOP58 rep B	125 (62.5%)	4 (2.0%)	10 (5.0%)	9 (4.5%)	52 (26.0%)
DKC1	11 (5.5%)	62 (32.0%)	18 (9.0%)	7 (3.5%)	102 (51.0%)
Ago2 rep A	0 (0.0%)	0 (0.0%)	1 (0.5%)	59 (29.5%)	140 (70.0%)
HuR rep A	0 (0.0%)	0 (0.0%)	0 (0.0%)	117 (58.5%)	83 (41.5%)

Ago2: Argonaute 2; DKC1: Dyskerin; FBL: Fibrillarin; MNase: micrococcal nuclease; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; scaRNA: small Cajal body-specific RNA; snoRNA: small nucleolar RNA;

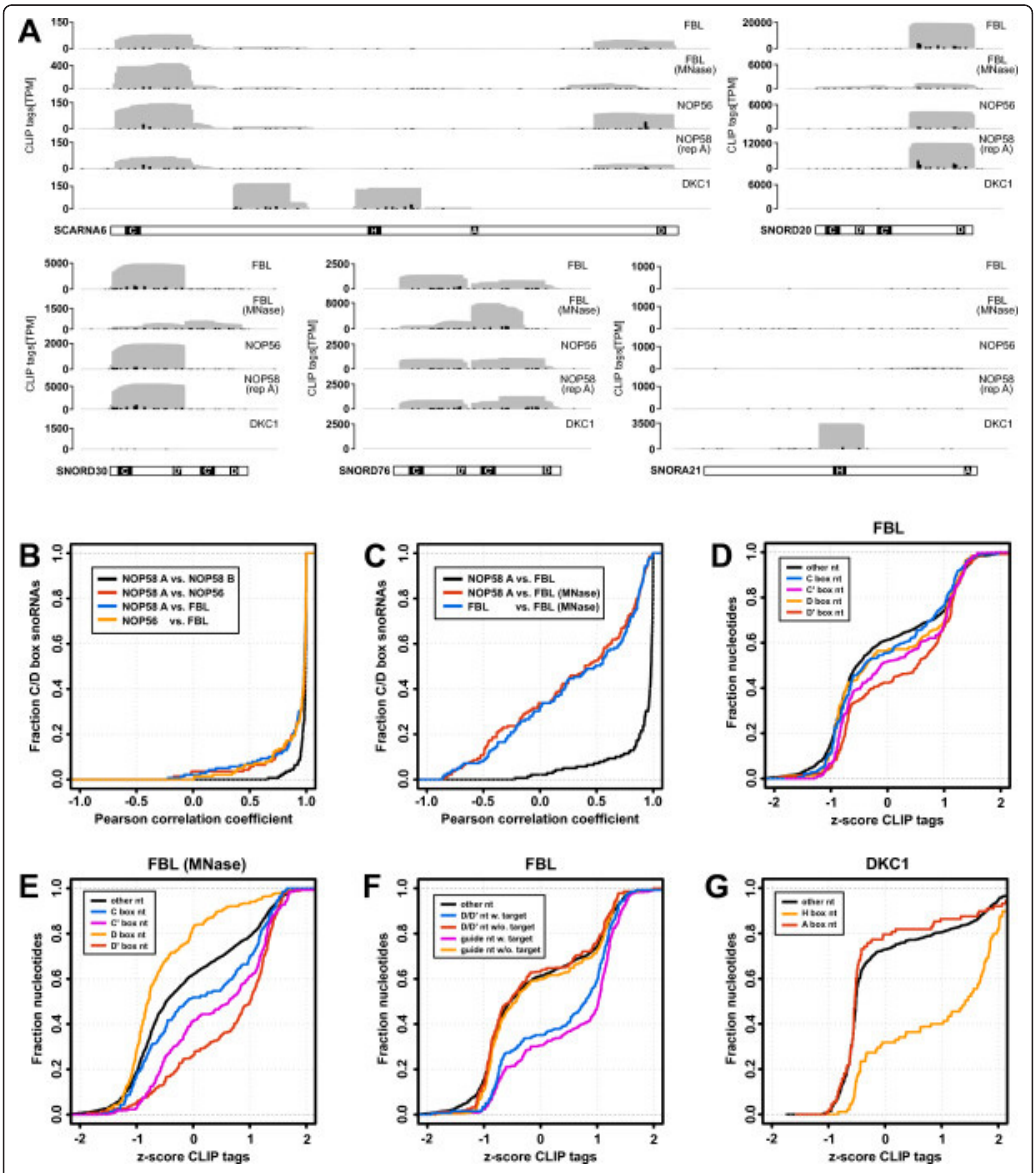


Figure 1 Summary of PAR-CLIP data of snoRNP core proteins. **(A)** Profiles of sequencing reads obtained from PAR-CLIP experiments for selected snoRNAs. Black bars in the profiles indicate the number of T→C mutations observed in PAR-CLIP reads at a particular nucleotide. **(B)** Similarity of binding profiles of core proteins that associate with C/D box snoRNAs. **(C)** Comparison of protein binding profiles as inferred from RNase T1-treated and MNase-treated PAR-CLIP samples. **(D, E)** Preferential binding of Fibrillarin to box elements as inferred from PAR-CLIP samples prepared with T1 (D) and MNase ribonucleases (E). **(F)** Comparison of binding preferences at D'/D box elements and guide regions for snoRNAs with and without a known target. **(G)** Analysis of binding preferences of Dyskerin for H/ACA box snoRNA-specific elements. D, E, F and G show the cumulative distributions of CLIP read coverage *z*-scores for nucleotides located in various regions of the snoRNA relative to the overall coverage of the snoRNA. CLIP: cross-linking and immunoprecipitation; MNase: micrococcal nuclease; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; snoRNA: small nucleolar RNA; snoRNP: small nucleolar ribonucleoprotein.

individual snoRNAs in pairs of samples. Comparing NOP58 to Fibrillarin and NOP56 we found that 109 (78%) and 111 (80%) snoRNA genes had a correlation coefficient of at least 0.9. To put this in perspective, between biological replicates of NOP58, 130 out of 139 snoRNAs investigated have a correlation coefficient of at least 0.9. This indicates that Fibrillarin, NOP56 and NOP58 form a tight complex that contacts the snoRNA. As noticed before, however [35], the nuclease treatment has a strong influence on the relative number of tags obtained from different positions along a snoRNA (Figure 1C). Only 19 snoRNA genes (14%) show a correlation ≥ 0.90 in their tag profiles obtained with RNase T1- and MNase-treated Fibrillarin PAR-CLIP samples, reflecting the fact that T1 nuclease is more efficient and generates a more biased position-dependent distribution of reads than MNase (Figure 1A). Figures 1D and Figure 1E summarize these results, showing that nucleotides in D' boxes are most frequently cross-linked, followed by nucleotides in the C' and C boxes, and then by nucleotides in the D box and in the rest of the snoRNA. MNase treatment in particular results in very poor coverage of the D box. On the other hand, we observed gene-specific differences in the binding of the core proteins. For example, SNORD20 only shows a peak of CLIP reads at the D box, SNORD30 only at the C box, while SNORD76 has peaks at both C and D boxes (Figure 1A).

We further asked whether the binding pattern of Fibrillarin reflected in the abundance of CLIP reads differs between guide regions of the snoRNAs that have a target annotated in snoRNA-LBME-db and orphan guide regions. For guide regions, we took the nine nucleotides upstream of the D and D' boxes and as a reference we compared the coverage of the D and D' boxes themselves (Figure 1F). We found that guide regions with a known target and their associated D/D' boxes generally have a higher coverage compared to those that are orphan (70% compared to 40% positive *z*-scores of the average coverage per position in the guide region relative to the entire snoRNA, Figure 1G). This could indicate that the binding to the target renders the snoRNA-core protein complex more accessible to cross-linking.

For H/ACA box snoRNAs we found that Dyskerin strongly prefers the H box nucleotides (Figure 1G), which in 70% of the snoRNAs have a positive *z*-score for coverage compared to the entire snoRNA. This is expected because these snoRNAs are highly structured, with most nucleotides being engaged in base pairs in the two hairpin stems and a few nucleotides are free to interact with the proteins.

Identification of novel snoRNA genes from PAR-CLIP and small RNA sequencing

We screened the top 500 clusters from each PAR-CLIP library that did not overlap with known ncRNAs, mRNAs or repeat elements for potentially novel snoRNA genes. To identify H/ACA box genes we employed the SnoReport program [38], while for C/D box snoRNA detection we applied a custom approach searching for a C box motif (RUGAUGA, R = A or G; allowing one mismatch) at the 5' end and a D box motif (MUGA, M = A or C) at the 3' end, requiring that a terminal stem of at least four canonical base pairs can be formed by the nucleotides flanking the C and D boxes. We combined these computational screens with isolation and sequencing of the 20 to 200 nucleotide RNA fraction from HEK293 cells, which provides evidence for expression of the predicted snoRNAs. Requiring a minimal average coverage per nucleotide of at least 1 tag per million (TPM) in least one type-specific CLIP library as well as in the small RNA-seq library, we identified 77 and 20 putative C/D and H/ACA box snoRNAs, respectively (Additional files 2 and 3). We additionally screened 14 distinct small RNA sequence libraries from the recently released ENCODE data [39] and found that more than 75% of our putative C/D box snoRNAs were detected in at least one cell type other than HEK293 (see Additional file 4). We further tested the expression of the 20 most abundantly sequenced candidate snoRNAs by Northern blotting (see Additional file 5). Nine of the twenty candidates were also detectable in this assay, while an additional nine C/D box snoRNAs are supported by the ENCODE data (see Additional file 4).

To determine whether the candidates we identified as described are entirely novel snoRNA genes or so far undescribed homologs of known snoRNAs, we performed a BLAST search against the snoRNA genes from snoRNA-LBME-db (requiring an *E*-value ≤ 10^{-3}). We further compared the loci of the putative snoRNAs with the snoRNA annotation available in ENSEMBL release 65 [40], which is based on automatic annotation with sequence/structure models available in the Rfam database [41]. Out of the 20 H/ACA box snoRNA candidates, 18 show sequence or structural homology to known snoRNAs, while candidates ZL4 (annotated as nc053 in [42], but not classified as a snoRNA by the authors) and ZL36 appear to be novel H/ACA box snoRNAs without a known homolog. The homology search additionally revealed that ZL4 is conserved until *Xenopus tropicalis*.

Of the 77 C/D box snoRNAs, only seven showed sequence homology to known C/D box snoRNA genes, but in one case (ZL1) the homology consisted solely of a long GU-rich region. The evolutionary conservation of the guide regions of five of these snoRNAs (ZL11, ZL109, ZL126, ZL127 and ZL132) suggests that they target the same nucleotides on ribosomal RNA as their homologs. A sixth snoRNA, ZL142, appears to be a human homolog of the GGN68 snoRNA of chickens [43,44]. An additional comparison with the results of another large snoRNA analysis [45], revealed that ZL2 and ZL107 have been

previously described as SNORD41B and Z39, respectively. In order to further characterize the 69 potentially novel C/D box snoRNAs (including ZL1, which only had homology with a known snoRNA in a GU-rich region), we first asked whether their C and D boxes are evolutionarily conserved (Additional file 1). To this end, we computed their average position-wise phastCons scores [46], which we obtained from the UCSC genome browser. Five candidates including ZL1 showed an average phastCons score per nucleotide higher than 0.25 for C and D box nucleotides. A comprehensive homology search of vertebrate genomes allowed us to trace the evolutionary origin of these snoRNAs and to annotate C' and D' boxes as well as putative guide regions based on sequence conservation. ZL1 is highly conserved in vertebrates including *Petromyzon marinus*, while for ZL5, ZL6, ZL8 and ZL24 we were not able to retrieve any homologs outside of mammals.

The remaining C/D box snoRNAs show overall weak conservation in mammals and in primates (Additional file 1). The C' and D' box elements of these snoRNAs, which are typically more variable in sequence, were particularly difficult to annotate without supporting evidence from evolutionary conservation. Because it is not clear that these snoRNAs have a C-D'-C'-D box architecture, we refer to them as C/D box-like. The small RNA sequence data indicates that these C/D box-like snoRNAs are only weakly expressed (Additional file 6). Interestingly, while the shortest C/D box snoRNA that has been characterized so far is SNORD49B, which has 48 nucleotides, 23 of our C/D box-like snoRNAs are even shorter. Figure 2 depicts PAR-CLIP tags and small RNA-seq reads for four of these snoRNAs which we called mini-snoRNAs. ZL77 is among the shortest, with 27 nucleotides in length, and only 7 nucleotides available as a potential guide region between the C and D boxes, while ZL49 and ZL103 are slightly longer (14 and 15 nucleotides between the C and D boxes). Another mini-snoRNA, ZL63, generated a considerable number of reads in all the CLIP libraries as well as in the RNA sequence data.

Our screen could further identify a snoRNA with mixed C/D box and H/ACA box structure. SCARNA21, a computationally predicted H/ACA box snoRNA [47], is surrounded by conserved C and D box elements enclosed by a terminal stem structure (Additional file 7). Northern blot analysis revealed that the prevalent form in the cells is the one that contains the C/D box elements and not the short form, which would be the single H/ACA box snoRNA.

Target prediction for newly identified snoRNA genes

To gain insight into the function of the novel snoRNAs that we identified, we sought to determine whether they have canonical targets. We employed the programs RNAsnoop and PLEXY to predict targets of H/ACA box and C/D box snoRNAs, respectively [48,49]. As potential target sequences we considered ribosomal and spliceosomal RNAs obtained from snoRNA-LBME-db. Indeed, for the highly conserved C/D box snoRNAs ZL1, ZL5 and ZL6 (which share the guide region), as well as for the H/ACA box snoRNA ZL4, we could identify canonical targets (Figure 3). ZL1 and ZL4 are both predicted to guide modifications on the U2 snRNA, 2'-O-methylation of U47 and pseudouridylation of U15, respectively. The pseudouridylation of U2 snRNA at U15 has already been described, but the guiding snoRNA was not known [50]. With primer extension assays we could further validate the U47 modification (see Additional File 8). SnRNA modifications are known to occur in Cajal bodies. Consistent with ZL4 H/ACA box snoRNA being a scaRNA that is recruited to Cajal bodies, is the presence of the

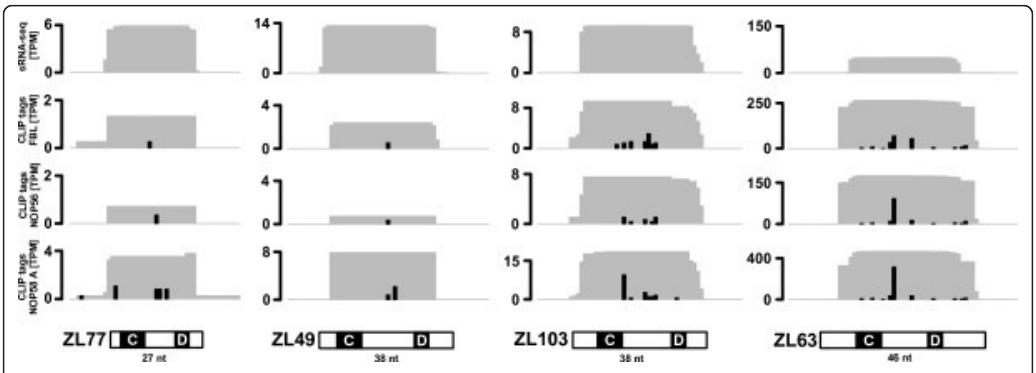


Figure 2 Small RNA-seq and PAR-CLIP reads mapping to mini-snoRNAs. Mini-snoRNAs ZL77, ZL49, ZL103 and ZL63 are shown. Black bars in the panels corresponding to PAR-CLIP libraries indicate the number of T→C mutations observed at individual nucleotides. CLIP: cross-linking and immunoprecipitation; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; snoRNA: small nucleolar RNA.

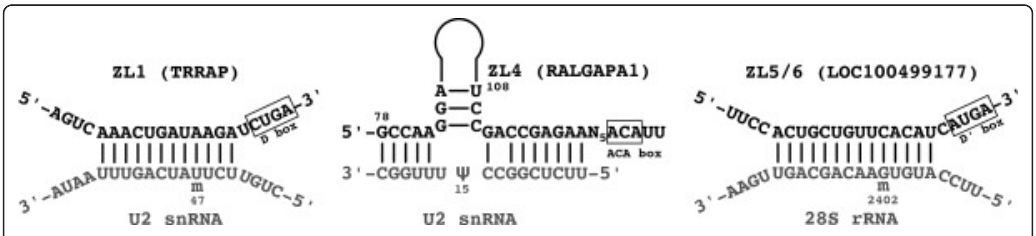


Figure 3 Predicted structure of hybrids between novel snoRNAs and target RNAs. The snoRNAs are given at the top of each panel together with the symbol of the host gene in which the snoRNA resides (in parentheses). The targets are indicated at the bottom of the panels. rRNA: ribosomal RNA; snoRNA: small nucleolar RNA; snRNA: small nuclear RNA

CAB box motif (UGAG), known to mediate this transport [12], in the hairpin loops. For the C/D box snoRNA ZL1 targeting U2 snRNA we could not identify an H/ACA box-like structural domain with a CAB box. Interestingly, however, this snoRNA candidate contains a long GU repeat, a feature shared by SCARNA9, the only Cajal body-associated snoRNA that lacks H/ACA and CAB boxes. This suggests that the GU element serves as an import signal into Cajal bodies. For ZL5/6, the predicted modification site on the 28S rRNA is in fact a known modification site for which the guide was so far unknown. We could not predict a target for the newly identified C/D box domain of SCARNA21.

We were especially interested to find out whether the non-conserved C/D box-like snoRNAs and in particular the mini-snoRNAs, could guide 2'-O-methylations. To this end, we took a simple approach searching for 8-mer Watson-Crick complementarity between the putative guide regions upstream of the D boxes to ribosomal and spliceosomal RNAs. We did indeed identify seven putative interaction sites, but none of these are known modification sites (Additional file 2). Thus, the targets of these C/D box-like snoRNAs remain to be identified.

Non-canonical RNA partners of core snoRNA proteins

Although snoRNAs are best known for guiding modifications of rRNAs, snRNAs and tRNAs [1-3], some evidence has emerged for the involvement of full-length mature snoRNAs also in other biological processes such as alternative splicing [51]. To investigate this possibility, we searched our PAR-CLIP data sets for RNAs that were abundantly cross-linked, yet not known to associate with the core snoRNA proteins. In contrast to the HuR PAR-CLIP that we performed before [35], the PAR-CLIP experiments conducted with C/D box snoRNP core proteins repeatedly identified several non-coding RNAs including vault RNA 1-2, 7SK RNA and 7SL RNA as well as H/ACA box snoRNAs. Similarly, in the Dyskerin PAR-CLIP we observed cross-linking of several C/D box snoRNAs.

We performed primer extension experiments to determine potential sites for 2'-O-methyl and pseudouridine modification in prominent ncRNAs such as 7SK RNA, 7SL RNA and vault RNA 1-2 (see Additional file 9 for primer extension assays and Additional file 10 for a catalog of identified modifications sites and target predictions). Indeed, we found that all three of these RNA species carry modifications. Vault RNA 1-2 contains four 2'-O-methyl sites, 7SK RNA carries at least six 2'-O-methyl sites and one pseudouridylation site, and 7SL RNA contains several sites of pseudouridylation. Additionally, we sought to determine whether C/D box and H/ACA box snoRNAs guide modifications on each other. We thus performed 2'-O-methylation primer extension assays on SNORA61 and pseudouridylation assays on SNORD16 and SNORD35A. We found that SNORA61 potentially carries one 2'-O-methylation, while SNORD16 and SNORD35A carry two and six pseudouridylated residues, respectively. To identify C/D box snoRNAs that could guide the observed 2'-O-methylations, we searched for 8-mer complementarity upstream of D and D' boxes of C/D box and C/D box-like snoRNAs, but we did not find sequences complementary to the modification sites. To predict guiding H/ACA box snoRNAs we employed the program RNAsnoop using stringent filtering criteria. We identified potential guiding H/ACA box snoRNAs for 7SK RNA residue Ψ250 and 7SL RNA residue Ψ226.

Previous studies reported that snoRNAs may function in alternative splicing [32,51] and we also repeatedly observed cross-linking of C/D box core proteins to regions that are annotated as exons of protein coding genes. To determine whether these mRNA regions are targeted by snoRNAs, we selected, from the top 1,000 clusters located in mRNA exons in NOP58 libraries, the 157 that were present in both NOP58 replicates and a third CLIP library with at least 10 TPM per nucleotide (Additional file 11). We identified complementarities to the 8-mer guide regions of snoRNAs in 79 of these clusters. In contrast, in shuffled CLIPed regions we only found 60 complementarities to snoRNA guide regions (average of 100 simulations on

shuffled sequences). Thus, the mRNA sequences that we isolated in the CLIP experiments are consistent with the possibility that snoRNAs act as guides in some steps of mRNA processing.

snoRNA processing patterns

It has become apparent that many ncRNAs such as tRNAs, snRNAs, rRNAs and snoRNAs are extensively processed into small, stable RNA fragments originating mainly from the termini of the mature ncRNA [29], which in some cases are incorporated in the Argonaute proteins to function as microRNAs [24]. To identify snoRNA-derived small RNAs that could potentially act as miRNAs comprehensively, we isolated and sequenced the RNA fraction of 18 to 30 nucleotides from HEK293 cells. Small RNAs derived from C/D box snoRNAs constitute about 1.7% of the small RNA pool in this size range in HEK293 cells (Table 3). Consistent with the results of Li and colleagues [29], we found that most of the 513,339 reads overlapping with C/D box snoRNA genes originate from the 5' or 3' ends (38.7% and 46.0%, respectively). Visual inspection of the alignment of these reads to the snoRNAs revealed, however, that start and end positions of the reads do not generally coincide with the annotated snoRNA termini, which were inferred based on the characteristic C/D box snoRNA terminal stem (Figure 4A). Instead, the reads that we obtained indicate specific trimming that generates sharp 5' ends for 5'-end-derived reads and sharp 3' ends for 3'-end-derived reads. To determine whether this trimming may occur in the process of generating small RNAs from mature C/D box snoRNAs, we isolated small RNAs of length 20 to 200 nucleotides that presumably included the full-length, mature snoRNAs (average C/D box snoRNA length is 70 to 90 nucleotides) and performed a 150-cycle sequencing run. Figure 4A depicts the alignment of reads obtained in the small RNA fraction and the reads obtained in the 150-cycle sequencing run for three selected C/D box snoRNAs. Strikingly, the sharp ends of C/D box snoRNA-derived small RNAs coincide with the 5' and 3' ends of the mature form. More generally, we found that for 84% and 70% of the top 50 expressed C/D box snoRNAs, the most prominent start and end positions, respectively, obtained from long sequencing reads coincided with the most prominent start and end positions obtained from small RNA sequencing. This suggests that the observed trimming of the terminal closing stem occurs during the excision of the snoRNA from the intron and is not specific to the processing of the mature snoRNA form into smaller fragments. Furthermore, we found that it is the distance to the C or D boxes that seems to determine the observed ends of the snoRNAs rather than the length of the terminal closing stem (Figure 4B). The 5' end is sharply defined four to five nucleotides upstream of the C box, while the 3' end is more variably located two to five nucleotides downstream of the D box. In most cases this will leave mature C/D box snoRNAs with a terminal 5' overhang compared to the 3' end. This suggests that, similar to other small RNAs [52,53], snoRNAs are trimmed presumably by exonucleases, to boundaries that are determined by the proteins with which these small RNAs are complexed.

Small RNAs derived from C/D box snoRNA termini appear to be abundant in the cells, and can be incorporated into Argonaute proteins to act as miRNAs [31]. To determine the relative participation of various small RNA classes in the Argonaute-dependent gene silencing, we immunopurified Ago2 from HeLa cells and sequenced the associated small RNA fraction. We found that, as expected, miRNAs constitute the most abundant RNA class that associates with Ago2 (approximately 90%), while C/D box snoRNAs account only for 0.005% of the IP-seq reads (Table 3). Assuming that overall proportions of small RNAs derived from tRNAs and snoRNAs are fairly constant across cell types, we can estimate the efficiency with which small RNAs (from the total small RNA pool) are incorporated in the Argonaute proteins. We found, for example, that although small RNAs derived from tRNAs are 5.6 times more abundant than C/D box

Table 3 Functional annotation of sequencing reads obtained in sRNA sequencing and HeLa Ago2 IP sequencing.

RNA class	HEK293 sRNA sequencing (18 to 30 nucleotides)	HeLa Ago2 immunoprecipitation sequencing (asynchronous cells)	HeLa Ago2 immunoprecipitation sequencing (mitotic cells)
microRNAs	18.304%	89.750%	82.237%
tRNAs	9.694%	0.204%	0.298%
snRNAs	5.275%	0.029%	0.071%
C/D box snoRNAs	1.751%	0.005%	0.054%
H/ACA box snoRNAs	0.318%	0.026%	0.046%
No annotation	64.658%	9.985%	17.293%

Ago2: Argonaute 2; IP: immunoprecipitation; snRNA: small nuclear RNA; snoRNA: small nucleolar RNA; sRNA: small RNA; tRNA: transfer RNA

Kishore *et al. Genome Biology* 2013, **14**:R45
http://genomebiology.com/2013/14/5/R45

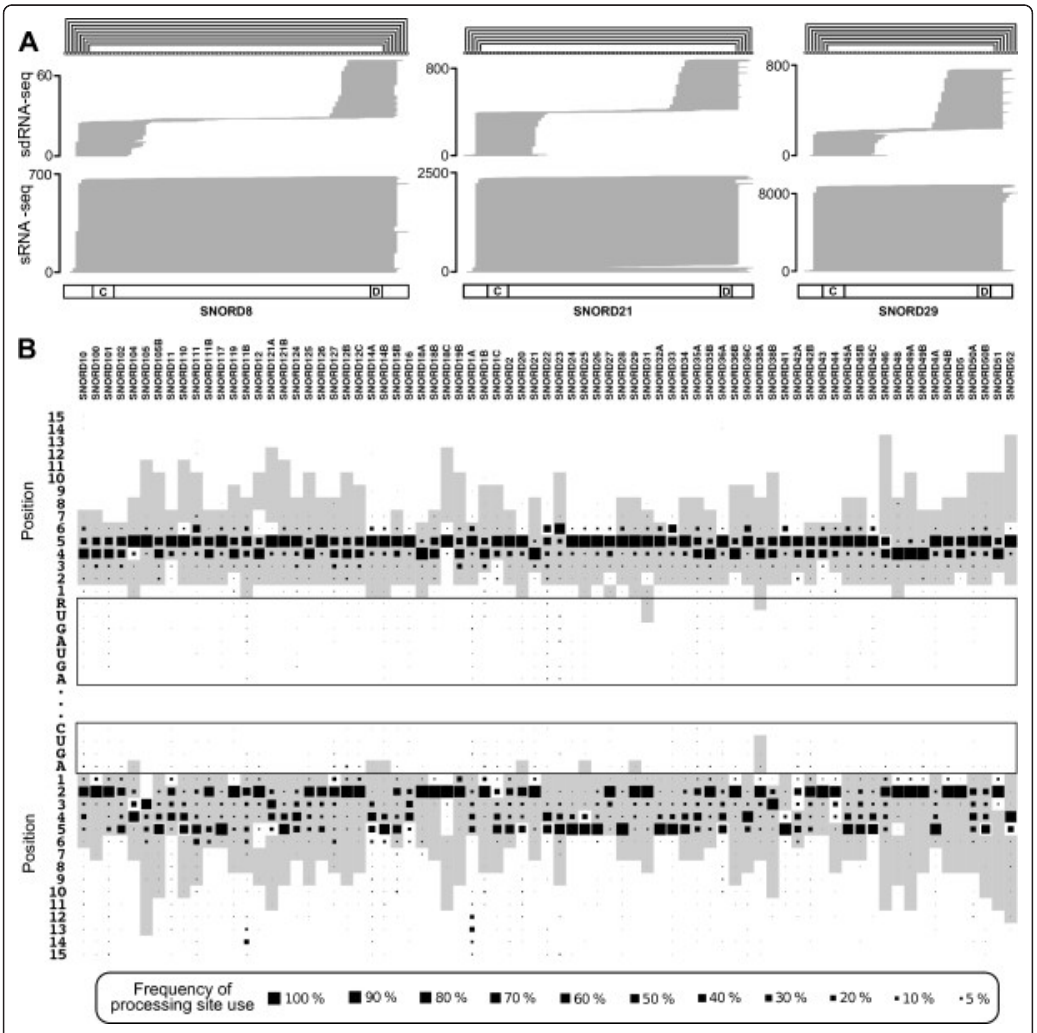


Figure 4 Terminal processing of C/D box snoRNAs. **(A)** Profiles of sequencing reads obtained from two small RNA seq libraries for three selected C/D box snoRNAs (SNORD8, SNORD21 and SNORD29). Upper: sdRNA sequencing, 18 to 30 nucleotides. Lower: sRNA sequencing, 20 to 200 nucleotides. Secondary structure annotation of the terminal closing stem is given on the top of the figure, while the locations of C and D motifs are shown on the bottom. **(B)** Detailed analysis of terminal stem processing for C/D box snoRNA expressed in HEK293 cells. The y-axis indicates individual nucleotides, with their specific identity for the nucleotides in C/D boxes and position relative to the boxes for the flanking nucleotides. Each column corresponds to a snoRNA, whose identity is shown at the top of the panel. Grey boxes indicate nucleotides that are predicted to be paired in the terminal stem. The size of black boxes is proportional to the number of sRNA sequencing reads that start (5′ end) or end (3′ end) at a particular nucleotide. See Additional File 16 for analysis of all C/D box snoRNAs expressed in HEK293 cells. sdRNA: small derived RNA; snoRNA: small nucleolar RNA; sRNA: small RNA

derived snoRNAs, tRNA fragments are 40 times more abundant in the Ago2-associated fraction. Thus, tRNA-derived small RNAs appear to be more efficiently incorporated in Ago2 than C/D box snoRNA fragments. This is consistent with observations that tRNAs are cleaved by nucleases such as Angiogenin and even Dicer to generate processing fragments that are active in translation regulation [54,55]. Similarly, small RNAs derived from H/ACA

box snoRNAs are 5.5 times less abundant than small RNAs derived from C/D box snoRNAs in the total RNA fraction, but are 5.2 times more efficiently picked up by Ago2. The H/ACA box snoRNA SCARNA15, which has been shown to be processed into smaller fragments that act as microRNAs [24], is represented in this library with 3,636 reads, 29% of all reads mapped to H/ACA box snoRNA loci (see Additional file 12 for a full listing of all snoRNAs). The C/D box snoRNA with the highest number of reads in the Ago2 IP library is SNORD1A with 1,140 reads, but the majority of C/D box snoRNAs are represented by less than 50 reads.

Of all categories of small RNAs, C/D box snoRNA fragments are those that show the strongest nuclear retention, and are found in the cytoplasm with only low frequency [56]. Thus, this physical separation could account for the low frequency of association between C/D box snoRNA-derived RNAs and Ago2. We therefore wondered whether the association of this abundant class of RNA fragments with Ago2 increases in the mitotic phase of the cell cycle, when the nuclear membrane is dissolved. We collected HeLa cells that were in the mitotic phase through mitotic shake off, immunopurified Ago2 and again sequenced the Ago2-associated small RNA fraction. We found that, indeed, the relative abundance of C/D box-derived fragments in Argonaute increased in this condition (Table 3), to 0.054% relative to 0.005%. Nonetheless, these results indicate that C/D box snoRNAs do not generally carry out miRNA-like functions, and that the number of H/ACA box snoRNAs with a dual function is very limited.

Discussion

To gain insight into the processing of snoRNAs and the functions of snoRNA-derived small RNAs, we performed PAR-CLIP experiments with snoRNP core proteins. Analysis of PAR-CLIP reads showed that C/D box core proteins Fibrillarin, NOP56 and NOP58 have a very similar binding pattern, overlapping with the box elements. Excluding snoRNA families SNORD113 to SNORD116, which are multi-copy families and do not have guide complementarity to rRNAs or snRNAs, snoRNA-LBME-db currently lists 153 C/D box snoRNAs, of which 40 and 78 have a guide region targeting a known modification at the D box and D' box, respectively. Evolutionary conservation profiles of the remaining putative guide regions suggest that most of them are not functional. In support of this concept, our analysis revealed that C/D box core proteins cross-linked more effectively to guide regions that are known to have a target compared to orphan guide regions.

Combining computational prediction with data from small RNA sequencing and PAR-CLIP we identified novel C/D and H/ACA box snoRNAs, and assigned guiding snoRNAs to several modifications on rRNAs and snRNAs that were previously described as orphans. In addition to these *bona fide* snoRNAs, we uncovered a group of C/D box-like snoRNAs that only have a C and a D box as opposed to the common C-D'-C'-D architecture. These C/D box-like snoRNAs are only weakly conserved and most of them are expressed at low levels. The unusual architecture and the weak evolutionary conservation are likely reasons why these RNA species have not been uncovered by computational ncRNA gene finders [57]. Some of the identified C/D box-like snoRNAs are extremely short, one being only 27 nucleotides in length, leaving hardly enough space for a guide region. The requirements for C/D box snoRNA biogenesis appear to be simply the presence of C and D boxes and a short region of complementarity flanking these boxes, leading probably to the production of many snoRNA-like molecules as the C/D box core proteins scan intronic regions of pre-mRNAs. An interesting lead to follow in further investigating the potential function of the C/D box-like snoRNAs originating in the introns of many genes comes from a recent study conducted in *Drosophila*, in which Schubert and colleagues showed that snoRNAs are required for maintenance of higher-order structures of chromatin accessibility [58].

In our PAR-CLIP experiments we also repeatedly cross-linked ncRNAs that are not usual snoRNA targets. We observed H/ACA box snoRNAs in PAR-CLIP experiments targeting the C/D box core proteins. *Vice versa*, we found C/D box snoRNAs in the PAR-CLIP targeting Dyskerin, which is an essential component of H/ACA box snoRNPs. Primer extension assays indicated that these snoRNAs carry modifications that would be expected from the protein complexes to which they were cross-linked, but we were, in general, not able to identify snoRNAs that could guide these modifications. One drawback may be that in the case of the 2'-O-methyl primer extension assays we cannot be sure that it was indeed a 2'-O-methyl modification as opposed to any other nucleoside modification that caused the stoppage of the reverse transcriptase. However, we can be fairly certain that we identified *bona fide* pseudouridylation sites. Particularly, in the case of SNORD35A we were able to identify five putative pseudouridylated residues but no convincing guiding sequence in a known H/ACA box snoRNA. This suggests either that even more snoRNAs remain to be identified or that these pseudouridylations are caused by a protein-only mechanism not requiring guidance by H/ACA box snoRNAs.

The processing patterns of snoRNAs have raised substantial interest and some controversy in recent years [30,32,59]. We strikingly found that snoRNA excision out of the intron follows a well-defined pattern leaving mature snoRNAs with four to five nucleotides upstream of the C box, and two to five nucleotides downstream of the D box, irrespective of the length of the terminal

closing stem. Our data support the observations of Darzacq and Kiss [5] that the terminal stem serves to bring the C and D box elements into close proximity so as to be more easily recognized by snoRNP proteins, which then protect the snoRNA from further trimming by the exosome, but may not be needed for the functional, mature snoRNA. This implies that the core proteins actively protect and stabilize the maturing snoRNA.

We further quantified the abundance of snoRNA-derived small RNAs in HEK293 cells, and consistent with other studies [29], we found that small RNAs derived from the ends of C/D box snoRNAs are indeed abundant. However, we did not find evidence that these sdRNAs efficiently associate with Ago2 to act as microRNAs, even in conditions when the accessibility of these sdRNAs to Ago2 should be higher, such as in mitotic cells. We thus conclude that a microRNA-like function of snoRNA-derived small RNAs is an exception rather than a rule. Most of the sdRNAs from C/D box snoRNAs originate from the termini of mature snoRNAs, and hence carry C and D box motifs. It might be that snoRNA core proteins are still attached to these fragments, protect them from total degradation, sequester them in the nucleus and prevent these sdRNAs from being loaded into Ago2.

Deep-sequencing-based studies revealed a very complex landscape of transcription and processing of RNAs. The non-canonical products identified initially in such studies raises the question of additional, yet unknown, functions of molecules that have been studied for many years. What has become apparent more recently, however, is that deep sequencing allows us to construct a very detailed picture of the kinetics of processing various classes of RNAs and of their interactions with proteins that protect them from degradation. Intersection of many data sets such as those generated in our study will eventually reveal kinetic and regulatory aspects of cellular processes at a fine level of detail.

Materials and methods

PAR-CLIP experiments

PAR-CLIP was performed with HEK293 Flp-In cells (Invitrogen). Cells were grown in thirty 15-cm cell culture plates per experiment to approximately 80% confluency. At 12 h before harvest, 4-thiouridine (Sigma) was added to the cells to a final concentration of 100 μM. PAR-CLIP was carried out as described previously [34]. For immunoprecipitation, antibodies were coupled to protein-A or protein-G Dynabeads (Invitrogen). Antibodies used against endogenous proteins were α-NOP58 (sc-23705 from Santa Cruz Biotechnology), α-Dyskerin H-300 (sc-48794, Santa Cruz Biotechnology), α-Dyskerin C-15 (sc-26982, Santa Cruz Biotechnology) and α-Fibrillarin AFB01 monoclonal antibody line 72B9, lot 011 (from Cytoskeleton, Inc, AFB01). The α-Ago2 (11A9) monoclonal antibody was a gift from Gunter Meister. For PAR-CLIP with NOP56 we used a HEK293 cell line with a stably integrated FLAG-NOP56 fusion gene and IP was done with monoclonal α-FLAG antibody M2 from Sigma. For one Fibrillarin targeted PAR-CLIP the immunoprecipitated complexes were treated with micrococcal nuclease (MNase, from New England Biolabs) for 5 min at 37°C [35]. After SDS-PAGE, gels were blotted onto nitrocellulose membranes to reduce the background from free RNAs [60]. The PAR-CLIP libraries were prepared as described in Additional file 13 and submitted to deep sequencing on an Illumina HiSeq 2000.

The reads obtained from PAR-CLIP experiments were mapped to the human genome (hg19 assembly from UCSC, February 2009) and annotated with the CLIPZ server [36]. Reads marked with the CLIPZ annotation categories 'fungal', 'bacterial,' or 'vector' were discarded and only reads that mapped uniquely to the genome were used in the analyses. The library size was scaled to 1,000,000 for all samples to obtain a normalized expression value (tags per million).

Small RNA sequencing

Small RNA sequencing libraries were prepared from size-selected RNAs of 18 to 30 nucleotides (sdRNA sequencing) and 20 to 200 nucleotides (sRNA sequencing). HEK293 total RNA was extracted and treated with DNase. Next, 20 units of T4 polynucleotide kinase and 2 μl of [γ-32P] ATP (10 μCi/μl) were used to radiolabel 10 μg of RNA at the 5'-ends. The RNA was separated together with a radiolabeled 20-nucleotide ladder on a 12% polyacrylamide gel, the bands corresponding to 18 to 30 nucleotides (for sdRNA sequencing libraries) or 20 to 200 nucleotides (for sRNA sequencing libraries) were excised, the RNA was extracted overnight in a 0.4-M NaCl solution and finally precipitated with ethanol. Small RNA libraries were prepared according to a published protocol [61] and sequenced on an Illumina HiSeq 2000 instrument, for 36 (sdRNA sequencing) and 150 cycles (sRNA sequenicng library). Adaptor removal was done with the CLIPZ server, and the mapping to the human genome was then done with the Segemehl software (v. 0.1.3) with parameters '-D 1 -A 90' [62]. The Gene Expression Omnibus (GEO) accession number for the PAR-CLIP and sRNA-seq data is GSE43666.

Identification of novel C/D snoRNAs and H/ACA snoRNAs from PAR-CLIP and small RNA sequencing data

For each PAR-CLIP library we inferred binding regions of the proteins of interest by clustering reads whose corresponding loci were at most 25 nucleotides apart. To annotate known snoRNA and scaRNA genes we first retrieved sequences from the snoRNA-LBME-db [37], mapped them to the human genome (a list of motif and

secondary structure annotated snoRNAs is available in Additional file 13). The 500 binding regions that accumulated the highest number of reads in each individual CLIP library, but did not overlap with known snoRNA or scaRNA genes, ncRNA genes or repeat elements, were screened for novel snoRNA candidates. We used SnoReport [38] to detect H/ACA box snoRNAs, while for detection of C/D box snoRNAs we searched for protein-binding regions that contained motifs corresponding to the C box (RTGATGA; allowing one mismatch) and to the two most common D box motifs (CTGA and ATGA). Sequences that contained both a C box and a D box motif were extended by ten nucleotides in order to search for a terminal closing stem. If a compact closing stem composed of at least four canonical base pairs with at least two G-C/C-G base pairs was found, the sequence was considered a snoRNA candidate. To evaluate the specificity of our C/D box snoRNA gene finding approach, we applied the same procedure to two types of clusters of PAR-CLIP reads from the NOP58 rep A sample both extended by 25 nucleotides on each side. First were the top 100 clusters (defined in terms of the number of reads associated with the cluster) that overlapped with C/D box snoRNA annotation, which served as a positive control. In this set, our program reported 80 sequences as putative snoRNAs. The second type of cluster contained the top 100 clusters that overlap with mRNA exon annotation. These should not contain snoRNAs, and indeed, we only obtained five putative C/D box snoRNAs candidates. Similarly low numbers of snoRNA candidates were obtained from randomized sequences (not shown). Altogether, these tests indicated that our method has very good specificity. In contrast, the number of predictions we obtained from CLIPed clusters without a known annotation was 11 for the top 100 such clusters.

Candidates that showed expression of at least 1 TPM per nucleotide in the 20 to 200 nucleotides small RNA sequencing run (only uniquely mapped reads that covered at least 50% of the candidate snoRNA sequence were considered), and had at least 1 TPM per nucleotide in at least one of the type-specific CLIP libraries were considered putative snoRNAs. They were consecutively numbered, and named as 'ZL#'. To further validate the newly found snoRNAs, we searched for evidence of expression in recently published small RNA-seq libraries from the ENCODE project [39]. Files with the genome coordinates of mapped reads (BAM files) were obtained from the ENCODE data coordination center at UCSC [63] and uniquely mapping reads were used for the analysis. In addition, we selected the 20 candidate C/D box snoRNAs with the highest read count in our data for validation by Northern blotting (see Additional file 13 for details on the experiment). To evaluate the evolutionary conservation of the putative snoRNAs, we carried out a homology search against the vertebrate genomes available in the UCSC genome browser. Once an initial set of homologs was identified, we built sequence/structure models and continued to search for more distant homologs with the Infernal software [64].

Detection of 2'-O-ribose-methylated and pseudouridylated residues

To identify 2'-O-methylated residues we used a reverse transcriptase-based method coupled with polyacrylamide gel analysis as described in [65]. The method is based on the observation that cDNA synthesis is noticeably impaired in the presence of a 2'-O-methyl when deoxynucleotide triphosphate fragments (dNTPs) are limiting [65,66], giving rise to a characteristic pattern of gel banding immediately preceding the 2'-O-methyls, with strong bands at low dNTP concentrations (0.004 mM) [66], becoming weaker with increasing concentrations of dNTPs.

To map pseudouridines in candidate RNAs we used a method that relies on chemical modification of RNA bases with N-cyclohexyl-N'-β (4-methyl morpholinium) -ethylcarbodiimide (CMC) [67]. The method involves carbodiimide adduct formation with U, G and pseudouridine followed by mild alkali treatment, which removes the adduct from U and G but not from the N-3 of pseudouridine. This modification results in the blockage of reverse transcription one residue 3' of the pseudouridine on the sequencing gel. For a detailed description of assays used to map 2'-O-methyls and pseudouridines see Additional file 13. As a proof of principle, we first applied these assays to the spliceosomal RNA U6, which is known to carry 2'-O-methylated and pseudouridylidated residues. In addition to the well-documented sites, we also observed novel 2'-O-methyl sites that have not been previously reported so far (Additional file 14).

To predict C/D box snoRNAs that could guide 2'-O-methylation, we searched for 8-mer complementarity (only canonical base pairs allowed) to regions immediately or one nucleotide upstream of the D and D' boxes of C/D box and C/D box-like snoRNAs. To predict H/ACA box snoRNAs that could guide pseudouridylations, we used the program RNAsnoop [48]. We first determined for each H/ACA snoRNA stem an energy cutoff value by running simulations on 1,000 random sequences of length 100. Only if an RNAsnoop prediction had an energy value lower than 90% of the random sequences, and at least three canonical base pairs on each side of the binding pocket, did we consider it as a hit.

Ago2 immunoprecipitation sequencing of asynchronous and mitotic cells

Mitotic cells were collected using mitotic shake-off [68,69], a technique based on the observation that cells become rounded and more easily detachable from the culture

vessel as they progress into metaphase during mitosis [70]. Details of the experimental setup are given in Additional file 13. To be able to confirm microscopically that we collected mitotic cells we used HeLa cells with the human histone H2B gene fused to green fluorescent protein (see Additional file 15).

Ago2 was immunoprecipitated from mitotic and asynchronous cells; the Ago2-associated RNAs were extracted and used to prepare cDNA libraries as described above [61], which were then submitted to deep sequencing. Adaptor removal was with the CLIPZ server, and reads were then mapped with Segemehl as described above. In the analysis of small RNA libraries (Ago2-IP and HEK293 sdRNA sequencing (18 to 30 nucleotides)), we considered both uniquely and multi-mapping reads that were annotated based on their mapping to genes in one of the following categories: tRNAs (from the UCSC Table Browser), microRNAs (from mirBase) and snRNAs (from ENSEMBL release 59), C/D box snoRNAs and H/ACA box snoRNAs (curated data set from this work).

Additional material

Additional file 1: Profiles of PAR-CLIPs reads obtained with various core snoRNP proteins for snoRNAs and scaRNAs. The proteins and normalized read counts are shown on the y-axis. The snoRNA and location of boxes are shown at the bottom. Red bars in the profiles indicate the number of T→C mutations observed at individual nucleotides in the PAR-CLIP reads.

Additional file 2: List of novel C/D box, C/D box-like snoRNAs and mini-snoRNAs obtained in this study.

Additional file 3: List of novel H/ACA snoRNAs or homologs of known snoRNAs (indicated in the 'BLAST hits' column) that were obtained in this study.

Additional file 4: RNA-seq read profiles from selected ENCODE small RNA-seq samples along the novel C/D box and H/ACA box snoRNA loci identified in our study.

Additional file 5: Northern blots for selected novel C/D box snoRNAs. Among the 20 most abundantly expressed (in the small RNA-seq data) novel C/D box snoRNAs we could confirm the presence of ZL1, ZL2, ZL8, ZL11, ZL63, ZL107, ZL116, ZL126 and ZL127 by Northern blotting.

Additional file 6: Expression of C/D box and C/D box-like snoRNAs in our small RNA-seq run (20 to 200 nucleotides; sequenced 150 cycles). Only reads that cover at least 50% of the snoRNA locus were considered.

Additional file 7: SCARNA21 has a C/D box H/ACA box hybrid structure. (A) Screenshot from the UCSC genome browser showing conserved C and D box elements. **(B)** Northern blot probing for H/ACA box structure only (left) and for the hybrid structure (right).

Additional file 8: Primer extension assays for U2 snRNA. Primer extension assay reveals a 2'-O-methyl modification site for nucleotide U47.

Additional file 9: Primer extension assays for non-canonical snoRNA targets. Primer extension runs reveal 2'-O-methyl (A-C) and pseudouridine (D-G) modification sites in several non-canonical RNAs. **(A)** SNORA61: G50. **(B)** VTRNA1-2: G30, U31, C33, A34. **(C)** 7SK RNA: C137, G139, C141, G148, C150, G151. **(D)** SNORD16: U52, U55. **(E)** SNORD35A: U26, U31, U37, U43, U45, U51. **(F)** 7SK RNA: U250. **(G)** 7SL RNA: U226, U233, U236, U266, U273.

Additional file 10: Summary of nucleotide modifications detected by primer extension assays and predicted guide snoRNA-target interactions.

Additional file 11: Analysis of PAR-CLIP clusters overlapping with mRNA exon annotation. Shown are genome coordinates, host transcript and exon identifier, the number of C and D boxes predicted within the genomic region, snoRNAs to whose guide regions these mRNA fragments are complementary and the number of (normalized) reads obtained from the regions in various PAR-CLIP libraries.

Additional file 12: Detailed list of reads mapping to snoRNA loci in Ago2 IP-seq libraries.

Additional file 13: Supplementary materials and methods. Detailed information about the experimental methods (PAR-CLIP library preparation, Northern blotting, primer extension assays, mitotic shake-off and Ago2 immunoprecipitation and sequencing). In addition, the annotated C/D and H/ACA snoRNAs used in this study are listed.

Additional file 14: Primer extension assays on spliceosomal RNA U6. (A) Primer extension assay on spliceosomal RNA U6 detected documented 2'-O-methylation as well as potentially novel 2'-O-methylation sites. **(B)** Primer extension assay detected documented pseudouridine sites in U6. CTRL indicates the untreated sample, +CMC the sample treated with 1-cyclohexyl-3-(2-morpholinoethyl)carbodiimide metho-p-toluenesulfonate (CMC).

Additional file 15: Asynchronous and mitotic GFP-tagged HeLa cells. Green fluorescent protein appears in green and cell boundaries in orange. **(A)** In an asynchronous cell culture only a few cells are in the mitotic phase, which can be seen from the condensed chromatin and the rounded cell morphology. **(B)** Cell obtained with mitotic shake-off. The procedure enriches for round cells containing condensed chromatin.

Additional file 16: Extended version of Figure 4B showing all snoRNA genes expressed in HEK293 cells.

Abbreviations

Ago2: Argonaute 2; CB: Cajal body; CLIP: cross-linking and immunoprecipitation; DKC1: Dyskerin; dNTP: deoxynucleotide triphosphate; FBL: Fibrillarin; IP: immunoprecipitation; IP-seq: immunoprecipitation sequencing; miRNA: micro RNA; MNase: micrococcal nuclease; ncRNA: non-coding RNA; PAR-CLIP: photoactivatable-ribonucleoside-enhanced cross-linking and immunoprecipitation; rRNA: ribosomal RNA; scaRNA: small Cajal body-specific RNA; sdRNA: small derived RNA; snoRNA: small nucleolar RNA; snoRNP: small nucleolar ribonucleoprotein; snRNA: small nuclear RNA; sRNA: small RNA; TPM: tags per million; tRNA: transfer RNA.

Authors' contributions

SK and MZ conceived the project. SK performed the experiments, with help from DJJ (Ago2 IP and primer extensions) and APS (novel snoRNA validation). ARG performed the computational analysis of the sequencing data, with help from HJ (computational prediction of snoRNA targets). ARG, DJJ, SK and MZ wrote the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors thank Erich Nigg (Biozentrum) for providing the HeLa cells in which the histone H2B gene is fused to green fluorescent protein and for advice on the isolation of mitotic cells. We are grateful to Gunter Meister for the anti-Ago2 antibody and Ina Nissen and Christian Beisel from the Quantitative Genomics Facility of the D-BSSE of ETH Zurich, for help with deep sequencing. All computations were carried out on the [BC]2 HPC infrastructure at the University of Basel. Work in the Zavolan lab is supported by the University of Basel and the Swiss National Science Foundation (grant #31003A 127307). SK acknowledges the support of the Gebert Rüf Foundation Rare Diseases Program (grant GRS-046/11).

Received: 18 March 2013 Revised: 15 May 2013
Accepted: 26 May 2013 Published: 26 May 2013

References

1. Decatur W, Fournier M: **rRNA modifications and ribosome function.** *Trends Biochem Sci* 2002, **27**:344-51.
2. Darzacq X, Jády B, Verheggen C, Kiss A, Bertrand E, Kiss T: **Cajal body-specific small nuclear RNAs: a novel class of 2′-O-methylation and pseudouridylation guide RNAs.** *EMBO J* 2002, **21**:2746-56.
3. Clouet d'Orval B, Bortolin ML, Gaspin C, Bachellerie JP: **Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp.** *Nucleic Acids Res* 2001, **29**:4518-29.
4. Tollervey D, Kiss T: **Function and synthesis of small nucleolar RNAs.** *Curr Opin Cell Biol* 1997, **9**:337-42.
5. Darzacq X, Kiss T: **Processing of intron-encoded box C/D small nucleolar RNAs lacking a 5′,3′-terminal stem structure.** *Mol Cell Biol* 2000, **20**:4522-31.
6. Brown JW, Echeverria M, Qu LH: **Plant snoRNAs: functional evolution and new modes of gene expression.** *Trends Plant Sci* 2003, **8**:42-9.
7. Kiss T: **Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs.** *EMBO J* 2001, **20**:3617-22.
8. McKeegan KS, Debieux CM, Boulon S, Bertrand E, Watkins NJ: **A dynamic scaffold of pre-snoRNP factors facilitates human box C/D snoRNP assembly.** *Mol Cell Biol* 2007, **27**:6782-93.
9. Tollervey D, Lehtonen H, Jansen R, Kern H, Hurt EC: **Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly.** *Cell* 1993, **72**:443-57.
10. Kiss T, Fayet-Lebaron E, Jády BE: **Box H/ACA small ribonucleoproteins.** *Mol Cell* 2010, **37**:597-606.
11. Lafontaine DL, Bousquet-Antonelli C, Henry Y, Caizergues-Ferrer M, Tollervey D: **The box H + ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase.** *Genes Dev* 1998, **12**:527-37.
12. Richard P, Darzacq X, Bertrand E, Jády BE, Verheggen C, Kiss T: **A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs.** *EMBO J* 2003, **22**:4283-93.
13. Nicoloso M, Qu LH, Michot B, Bachellerie JP: **Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2′-O-ribose methylation of rRNAs.** *J Mol Biol* 1996, **260**:178-95.
14. Kiss-László Z, Henry Y, Bachellerie JP, Caizergues-Ferrer M, Kiss T: **Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs.** *Cell* 1996, **85**:1077-88.
15. Cavaillé J, Nicoloso M, Bachellerie JP: **Targeted ribose methylation of RNA in vivo directed by tailored antisense RNA guides.** *Nature* 1996, **383**:732-5.
16. Ganot P, Bortolin ML, Kiss T: **Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs.** *Cell* 1997, **89**:799-809.
17. Bortolin ML, Ganot P, Kiss T: **Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs.** *EMBO J* 1999, **18**:457-69.
18. Lee Y, Shibata Y, Malhotra A, Dutta A: **A novel class of small RNAs: tRNA-derived RNA fragments (tRFs).** *Genes Dev* 2009, **23**:2639-49.
19. Haussecker D, Huang Y, Lau A, Parameswaran P, Fire A, Kay M: **Human tRNA-derived small RNAs in the global regulation of RNA silencing.** *RNA* 2010, **16**:673-95.
20. Nicolas F, Hall A, Csorba T, Turnbull C, Dalmay T: **Biogenesis of Y RNA-derived small RNAs is independent of the microRNA pathway.** *FEBS Lett* 2012, **586**:1226-30.
21. Persson H, Kvist A, Vallon-Christersson J, Medstrand P, Borg A, Rovira C: **The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs.** *Nat Cell Biol* 2009, **11**:1268-71.
22. Zywicki M, Bakowska-Zywicka K, Polacek N: **Revealing stable processing products from ribosome-associated small RNAs by deep-sequencing data analysis.** *Nucleic Acids Res* 2012, **40**:4013-24.
23. Kawaji H, Nakamura M, Takahashi Y, Sandelin A, Katayama S, Fukuda S, Daub C, Kai C, Kawai J, Yasuda J, Carninci P, Hayashizaki Y: **Hidden layers of human small RNAs.** *BMC Genomics* 2008, **9**:157.
24. Ender C, Krek A, Friedlander M, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, Meister G: **A human snoRNA with microRNA-like functions.** *Mol Cell* 2008, **32**:519-28.
25. Taft R, Glazov E, Lassmann T, Hayashizaki Y, Carninci P, Mattick J: **Small RNAs derived from snoRNAs.** *RNA* 2009, **15**:1233-40.
26. Shen M, Eyras E, Wu J, Khanna A, Josiah S, Rederstorff M, Zhang M, Stamm S: **Direct cloning of double-stranded RNAs from RNase protection analysis reveals processing patterns of C/D box snoRNAs and provides evidence for widespread antisense transcript expression.** *Nucleic Acids Res* 2011, **39**:9720-30.
27. Jung C, Hansen M, Makunin I, Korbie D, Mattick J: **Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data.** *BMC Genomics* 2010, **11**:77.
28. Langenberger D, Pundhir S, Ekstrøm C, Stadler P, Hoffmann S, Gorodkin J: **deepBlockAlign: a tool for aligning RNA-seq profiles of read block patterns.** *Bioinformatics* 2012, **28**:17-24.
29. Li Z, Ender C, Meister G, Moore P, Chang Y, John B: **Extensive terminal and asymmetric processing of small RNAs from rRNAs, snoRNAs, snRNAs, and tRNAs.** *Nucleic Acids Res* 2012, **40**:6787-99.
30. Scott M, Ono M, Yamada K, Endo A, Barton G, Lamond A: **Human box C/D snoRNA processing conservation across multiple cell types.** *Nucleic Acids Res* 2012, **40**:3676-88.
31. Li W, Saraiya A, Wang C: **The profile of snoRNA-derived microRNAs that regulate expression of variant surface proteins in *Giardia lamblia*.** *Cell Microbiol* 2012, **14**:1455-73.
32. Kishore S, Khanna A, Zhang Z, Hui J, Balwierz P, Stefan M, Beach C, Nicholls R, Zavolan M, Stamm S: **The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing.** *Hum Mol Genet* 2010, **19**:1153-64.
33. Brameier M, Herwig A, Reinhardt R, Walter L, Gruber J: **Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs.** *Nucleic Acids Res* 2011, **39**:675-86.
34. Hafner M, Landthaler M, Burger L, Khorshid M, Hausser J, Berninger P, Rothballer A, Ascano M Jr, Jungkamp A, Munschauer M, Ulrich A, Wardle G, Dewell S, Zavolan M, Tuschl T: **Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.** *Cell* 2010, **141**:129-41.
35. Kishore S, Jaskiewicz L, Burger L, Hausser J, Khorshid M, Zavolan M: **A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins.** *Nat Methods* 2011, **8**:559-64.
36. Khorshid M, Rodak C, Zavolan M: **CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins.** *Nucleic Acids Res* 2011, **39**:D245-52.
37. Lestrade L, Weber M: **snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs.** *Nucleic Acids Res* 2006, **34**:D158-62.
38. Hertel J, Hofacker I, Stadler P: **SnoReport: computational identification of snoRNAs with unknown targets.** *Bioinformatics* 2008, **24**:158-64.
39. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, Xue C, Marinov GK, Khatun J, Williams BA, Zaleski C, Rozowsky J, Roder M, Kokocinski F, Abdelhamid RF, Alioto T, Antoshechkin I, Baer MT, Bar NS, Batut P, Bell K, Bell I, Chakrabortty S, Chen X, Chrast J, Curado J, *et al*: **Landscape of transcription in human cells.** *Nature* 2012, **489**:101-8.
40. **ENSEMBL release 65..** [http://www.ensembl.org].
41. Burge S, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki E, Eddy S, Gardner P, Bateman A: **Rfam 11.0: 10 years of RNA families.** *Nucleic Acids Res* 2013, **41**:D226-32.
42. Yan D, He D, He S, Chen X, Fan Z, Chen R: **Identification and analysis of intermediate size noncoding RNAs in the human fetal brain.** *PLoS One* 2011, **6**:e21652.
43. Zhang Y, Wang J, Huang S, Zhu X, Liu J, Yang N, Song D, Wu R, Deng W, Skogerbo G, Wang XJ, Chen R, Zhu D: **Systematic identification and characterization of chicken (*Gallus gallus*) ncRNAs.** *Nucleic Acids Res* 2009, **37**:6562-74.
44. Marz M, Gruber AR, Höner Zu Siederdissen C, Amman F, Badelt S, Bartschat S, Bernhart SH, Beyer W, Kehr S, Lorenz R, Tanzer A, Yusuf D, Tafer H, Hofacker IL, Stadler PF: **Animal snoRNAs and scaRNAs with exceptional structures.** *RNA Biol* 2011, **8**:938-46.
45. Yang J, Zhang X, Huang Z, Zhou H, Huang M, Zhang S, Chen Y, Qu L: **snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome.** *Nucleic Acids Res* 2006, **34**:5112-23.
46. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, Clawson H, Spieth J, Hillier LW, Richards S, Weinstock GM, Wilson RK, Gibbs RA, Kent WJ, Miller W, Haussler D: **Evolutionarily conserved**

elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* 2005, **15**:1034-50.
47. Schattner P, Barberan-Soler S, Lowe T: **A computational screen for mammalian pseudouridylation guide H/ACA RNAs.** *RNA* 2006, **12**:15-25.
48. Tafer H, Kehr S, Hertel J, Hofacker IL, Stadler PF: **RNAsnoop: efficient target prediction for H/ACA snoRNAs.** *Bioinformatics* 2010, **26**:610-16.
49. Kehr S, Bartschat S, Stadler PF, Tafer H: **PLEXY: efficient target prediction for box C/D snoRNAs.** *Bioinformatics* 2011, **27**:279-80.
50. Karijolich J, Yu Y: **Spliceosomal snRNA modifications and their function.** *RNA Biol* 2010, **7**:192-204.
51. Kishore S, Stamm S: **The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C.** *Science* 2006, **311**:230-2.
52. Berninger P, Jaskiewicz L, Khorshid M, Zavolan M: **Conserved generation of short products at piRNA loci.** *BMC Genomics* 2011, **12**:46.
53. Valen E, Preker P, Andersen PR, Zhao X, Chen Y, Ender C, Dueck A, Meister G, Sandelin A, Jensen TH: **Biogenic mechanisms and utilization of small RNAs derived from human protein-coding genes.** *Nat Struct Mol Biol* 2011, **18**:1075-82.
54. Cole C, Sobala A, Lu C, Thatcher SR, Bowman A, Brown JW, Green PJ, Barton GJ, Hutvagner G: **Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs.** *RNA* 2009, **15**:2147-60.
55. Yamasaki S, Ivanov P, Hu GF, Anderson P: **Angiogenin cleaves tRNA and promotes stress-induced translational repression.** *J Cell Biol* 2009, **185**:35-42.
56. Liao J, Ma L, Guo Y, Zhang Y, Zhou H, Shao P, Chen Y, Qu L: **Deep sequencing of human nuclear and cytoplasmic small RNAs reveals an unexpectedly complex subcellular distribution of miRNAs and tRNA 3' trailers.** *PLoS One* 2010, **5**:e10563.
57. Bernhart SH, Hofacker IL: **From consensus structure prediction to RNA gene finding.** *Brief Funct Genomic Proteomic* 2009, **8**:461-71.
58. Schubert T, Pusch M, Diermeier S, Benes V, Kremmer E, Imhof A, Längst G: **Df31 protein and snoRNAs maintain accessible higher-order structures of chromatin.** *Mol Cell* 2012, **48**:434-44.
59. Scott MS, Ono M: **From snoRNA to miRNA: dual function regulatory non-coding RNAs.** *Biochimie* 2011, **93**:1987-92.
60. Ule J, Ule A, Spencer J, Williams A, Hu J, Cline M, Wang H, Clark T, Fraser C, Ruggiu M, Zeeberg B, Kane D, Weinstein J, Blume J, Darnell R: **Nova regulates brain-specific splicing to shape the synapse.** *Nat Genet* 2005, **37**:844-52.
61. Hafner M, Landgraf P, Ludwig J, Rice A, Ojo T, Lin C, Holoch D, Lim C, Tuschl T: **Identification of microRNAs and other small regulatory RNAs using cDNA library sequencing.** *Methods* 2008, **44**:3-12.
62. Hoffmann S, Otto C, Kurtz S, Sharma C, Khaitovich P, Vogel J, Stadler P, Hackermüller J: **Fast mapping of short sequences with mismatches, insertions and deletions using index structures.** *PLoS Comput Biol* 2009, **5**: e1000502.
63. **ENCODE data coordination center at UCSC..** [http://genome.ucsc.edu/ENCODE/downloads.html].
64. Nawrocki EP, Kolbe DL, Eddy SR: **Infernal 1.0: inference of RNA alignments.** *Bioinformatics* 2009, **25**:1335-7.
65. Maden BE, Corbett ME, Heeney PA, Pugh K, Ajuh PM: **Classical and novel approaches to the detection and localization of the numerous modified nucleotides in eukaryotic ribosomal RNA.** *Biochimie* 1995, **77**:22-9.
66. Maden BE: **Mapping 2'-O-methyl groups in ribosomal RNA.** *Methods* 2001, **25**:374-82.
67. Ofengand J, Del Campo M, Kaya Y: **Mapping pseudouridines in RNA molecules.** *Methods* 2001, **25**:365-73.
68. Morla AO, Draetta G, Beach D, Wang JY: **Reversible tyrosine phosphorylation of cdc2: dephosphorylation accompanies activation during entry into mitosis.** *Cell* 1989, **58**:193-203.
69. Pines J, Hunter T: **Isolation of a human cyclin cDNA: evidence for cyclin mRNA and protein regulation in the cell cycle and for interaction with p34cdc2.** *Cell* 1989, **58**:833-46.
70. Elvin P, Evans CW: **Cell adhesiveness and the cell cycle: correlation in synchronized Balb/c 3T3 cells.** *Biol Cell* 1983, **48**:1-9.

doi:10.1186/gb-2013-14-5-r45
Cite this article as: Kishore *et al.*: Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biology* 2013 **14**:R45.

4 An updated human snoRNAome

An updated human snoRNAome

Hadi Jorjani[1], Stephanie Kehr[2], Jana Hertel[2], Peter F. Stadler[2], Mihaela Zavolan[1], Andreas R. Gruber[1]

[1] Computational and Systems Biology, Biozentrum, University of Basel and Swiss Institute of Bioinformatics, Basel CH-4056, Switzerland.
[2] Bioinformatics Group, Department of Computer Science, and Interdisciplinary Center for Bioinformatics, University of Leipzig, D-04107 Leipzig, Germany.

Keywords: C/D box snoRNA, H/ACA box snoRNA, scaRNA, non-coding RNAs, snoRNA-derived fragments, gene annotation

Abstract

Small nucleolar RNAs (snoRNAs) are a class of non-coding RNAs that guide the post-transcriptional processing of other non-coding RNAs, mostly ribosomal RNAs. Recently, snoRNAs have been implicated in several other processes ranging from microRNA-like silencing to alternative splicing. A comprehensive catalog of these molecules, their processing products and expression profiles is essential for studying their functions. Here we have constructed an up-to-date catalog of human snoRNAs by combining data from various databases with de novo prediction and extensive literature review to provide curated genomic coordinates for the mature snoRNAs. By analysing small RNA-seq data from the ENCODE project we characterize the plasticity of snoRNA gene expression as well as their processing patterns. Finally, we identify snoRNAs whose expression is most strongly and reproducibly dysregulated in cancer cell lines.

Introduction

SnoRNAs are a specific class of small (from 60 to -- with a few exceptions --160 nucleotides) non-protein coding RNAs that are best known for guiding post-transcriptional modification of other non-protein coding RNAs such as ribosomal, small nuclear and transfer RNAs (rRNAs, snRNAs and tRNAs, respectively) [1–6]. Based on defined sequence motifs and secondary structure elements, snoRNAs are classified as either C/D box or H/ACA box. The two classes guide 2'-O-methylation and pseudouridylation of nucleotides on the target molecules, respectively. In C/D box snoRNAs the C box (RUGAUGA, R = A or G) and D box (CUGA) are brought into close proximity when the 5' and 3' ends of the snoRNA form a stem structure [7,8]. Most C/D box snoRNAs have additional, less conserved, C and D box motifs in the central region of the snoRNA which are termed C' and D' boxes. To carry out their function snoRNAs form ribonucleoprotein (RNP) complexes with the 15.5K, NOP56, NOP58, and fibrillarin proteins [9,10]. In these complexes, fibrillarin catalyses the 2'-O-methylation of the ribose in target RNAs [11]. The nucleotide undergoing the modification is determined by the complementarity to the 10 to 21 nucleotides (nt) guide region that is located upstream of the D or D' box. The fifth nucleotide upstream of the D/D' box will undergo the 2'-O-methylation [12–14].

H/ACA box snoRNAs adopt a well defined secondary structure consisting of two hairpins that are joined by a single-stranded region termed the H box (ANANNA, N = A, C, G or U) and further have an ACA box (AYA, Y = C or U) motif at the 3' end [15,16]. Similar to C/D box snoRNAs, H/ACA snoRNAs form RNP complexes with a set of four proteins, Dyskerin, Nhp2, Nop10 and Gar1. This RNP is active in pseudouridylation, with Dyskerin acting as the pseudouridine synthase [17]. Target recognition by H/ACA box snoRNAs also involves RNA-RNA interactions, of the single-stranded region in the snoRNA hairpin structures with the target RNA [18,19].

Canonical snoRNAs accumulate in the nucleolus, the primary site of ribosome synthesis, where they carry out their functions. ScaRNAs (small Cajal body-specific RNAs) are a specific subset of snoRNAs that guide modifications of spliceosomal RNAs and hence are found specifically enriched in Cajal bodies, the primary site of spliceosomal RNAs biogenesis [2]. The import of snoRNAs into Cajal bodies requires the presence of special sequence motifs. H/ACA box snoRNAs have the CAB boxes (UGAG) located in the hairpin loops of the two stem structures [20], while the import of C/D box snoRNAs seems to be dependent on a long UG dinucleotide repeat element [21]. There is evidence that both motifs are recognized by WDR79 which facilitates transport to Cajal bodies [21,22]. Beyond these snoRNAs with canonical structures some long scaRNAs with hybrid structures that are able to function in both methylation and pseudouridylation have been characterized [2,23]. Moreover, the primate specific Alu repeat elements can give rise to H/ACA box like snoRNAs termed AluACA RNAs that also seem to accumulate in Cajal bodies [24].

Interestingly, it appears that snoRNAs can guide other types of RNA processing, beyond methylation and pseudouridylation (see ref. 25 for a recent review). For example, SNORD22, SNORD14, SNORD3 and SNORD118 are involved in the processing of ribosomal RNA precursors [26]. Even though these RNA molecules have C and D box motifs, it seems that they do not show terminal end trimming C/D

box snoRNAs are usually subject to [27]. This likely suggests that these snoRNAs are in complex with additional proteins that assist in executing their function and prevent the usual C/D box specific trimming. Some evidence suggests that the brain-specific C/D box SNORD115 family regulates the alternative splicing of the serotonin receptor 5-HT(2C) mRNA [28,29]. Many C/D box as well as H/ACA box snoRNAs seem to undergo some kind of processing, yielding smaller fragments whose function remains elusive [27,30]. SCARNA15 provides a well documented example of an H/ACA box snoRNA that has a microRNA-like function [31]. Whether this function can be more generally carried out by other snoRNAs remains unknown. To add to the complexity of this class of RNAs, recent high-throughput sequencing-based studies identified C/D box-like snoRNAs as short as 27 nucleotides [27], that barely could host an antisense region, and as long as a few thousand nucleotides. The latter have been termed long non-coding snoRNAs (sno-lncRNAs) [32]. A summary of the currently known snoRNA classes is shown in Figure 1.

Despite a few recent genome-wide surveys for detection of novel snoRNAs, recent studies [21,27] have clearly demonstrated that our catalog of human snoRNA loci is far from complete. The data resources on snoRNAs [33,34] that have become standard in the field have either ceased to exist or to be updated. Furthermore, the focus of the research community has moved towards characterization of snoRNA genes in species other than human [35–39]. A recent attempt to improve the accuracy of snoRNA gene annotation [40] clearly demonstrates that a well designed, uniform analysis strategy is needed in trying to expand the catalog of snoRNAs while maintaining annotation accuracy. Here we sought to fill these gaps by providing an up-to-date catalog of snoRNA loci in the human genome, their processing patterns and expression profiles across tissues.

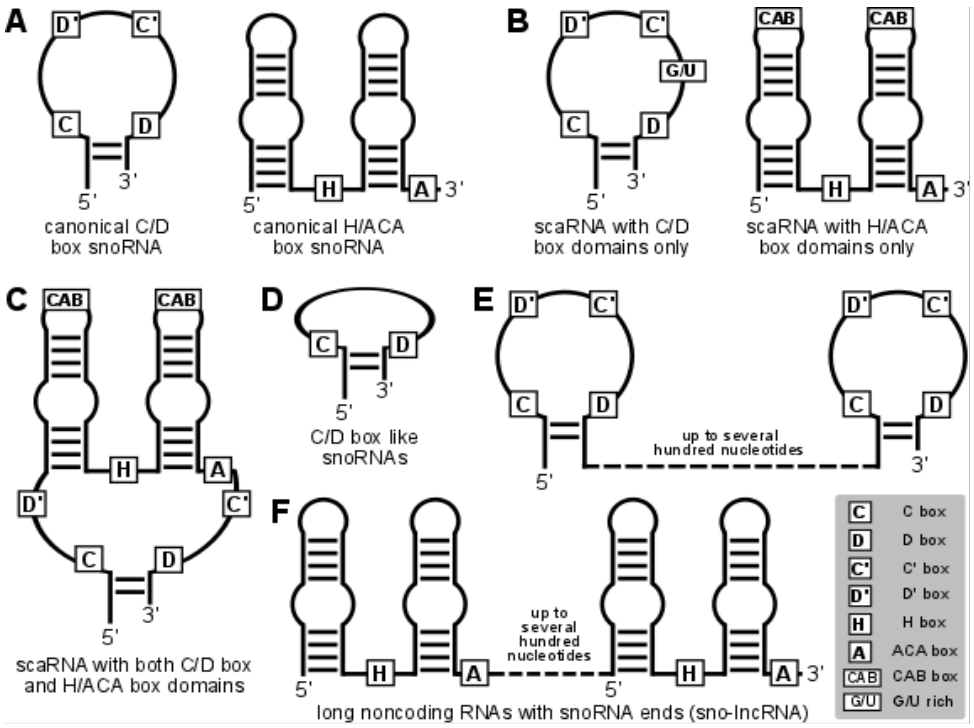


Figure 1. Schematic overview of different structural snoRNA classes. (**A**) Canonical C/D box snoRNAs have a C box and D box motif located close to the terminal stem, and additional boxes termed C' and D' box internally. Canonical H/ACA box snoRNAs are composed to two stem structures with an internal H box motif and an ACA box motif a the 3' end.

(**B**) SnoRNAs that execute their function in Cajal bodies additionally have specific import motifs termed CAB box in the case of H/ACA box snoRNAs, or a G/U rich sequence in the case of C/D box snoRNAs. (**C**) Several hybrid snoRNAs that consist of both a C/D box and an H/ACA box domain have been identified. Recent studies, have also uncovered extremely short C/D box like snoRNAs (**D**) as well as long noncoding RNAs with snoRNA ends that cover several hundred nucleotides (**E,F**).

Results

Curation of known snoRNA gene loci

In contrast to other types of molecules such as mRNAs or microRNAs, fewer studies attempted to sequence the full complement of mature human snoRNAs. Thus, the annotation of human snoRNA genes frequently started from computational predictions. Especially in the case of C/D box snoRNAs a consistent procedure for defining the 5' and 3' ends of their mature forms is lacking, and different pragmatic definitions such as the longest terminal stem, the longest evolutionarily conserved terminal stem, or the experimentally determined ends were used in different studies. However, the sequencing data that we obtained in a recent study indicated C/D box snoRNAs undergo uniform trimming at both the 5' and the 3' end [27], irrespective of the length of the terminal stem. In this work we use this observation to provide a unified catalog of mature human snoRNAs, with their 5' and 3' ends defined based on their coverage in small RNA sequencing data sets.

We retrieved 272 C/D box snoRNA, 108 H/ACA box snoRNA and 24 scaRNA that are currently annotated by the HUGO Gene Nomenclature Committee (HGNC) and mapped them to the human genome (hg19). We further obtained the genomic coordinates of small RNA sequencing reads from 114 data sets that were generated by the ENCODE consortium [41]. Intersecting the loci of sequenced small RNAs with those of the known snoRNAs, we identified, for each known snoRNA, the 5' and 3' ends that were most represented among the small RNA sequencing (sRNA-seq) reads (see Methods for details). As reported previously [27,4227,] the ends of C/D box snoRNAs undergo precise processing: the 5' end is located 4-5 nt upstream of the C box motif and 3' end is located up to 5 nt downstream of the D box motif. The same processing pattern is also observed here based on curated coordinates (see Supplementary Figure S1). The curated loci of the known, mature snoRNAs, are compiled in Supplementary File 1. For some snoRNAs e.g. SCARNA21, SNORD11B, or SNORA58 the sequence inferred from the small RNA sequencing data differed considerably from the sequence that is current deposited in HUGO. Supplementary File 2 shows a visualization of snoRNA loci including the HUGO sequence, the sRNA-seq read profile along these loci and the 5' and 3' ends that were inferred based on the sRNA-seq data.

An updated catalog of human snoRNA genes

To provide an up-to-date catalog of human snoRNAs, we integrated data from several sources, including a *de novo* genome-wide search. Our strategy is outlined in Figure 2. Specifically, we collected snoRNAs from the recently published literature, from RFAM-based predictions that were generated by the GENCODE consortium [43], from deepbase [44], and from our in-house snoRNA database at the University of Leipzig. To these we added genome-wide *de novo* predictions obtained with the following workflow, which is summarized schematically in Figure 2: Starting from genomic regions that gave rise to at least 5 reads in the entire sRNA-seq data set generated by the ENCODE consortium,

we extracted regions extending 20 nt upstream and 100 nt downstream of the start and end of the read cluster respectively. We used the snoReport [45] and snoSeeker [46] software to carry out snoRNA gene predictions. Additionally, we searched for cases in which degenerate C box and D box motifs with at most 100 length define potential C/D box-like snoRNA transcripts [27] (see Materials and Methods for a detailed description of the algorithm). We consolidate these initial candidates to a non-redundant set of putative snoRNA loci, excluding those that overlapped with repeat-annotated genomic regions. To generate a high-confidence set of snoRNA loci, we defined a set of strict rules to identify snoRNA candidates whose expression as mature forms was strongly supported by the sRNA-seq data (see Materials and Methods). This analysis yielded over 200 human snoRNAs that are currently not covered by the human gene annotation (Table 1 and Supplementary File 1). Finally, we used the the Infernal software and Rfam sequence-structure models to identify candidates which have relatively close homologs among the already known snoRNAs. We assigned each snoRNA to the family with the closest homology that had a p-value lower than 10^{-6}. Table 1 summarizes these findings.

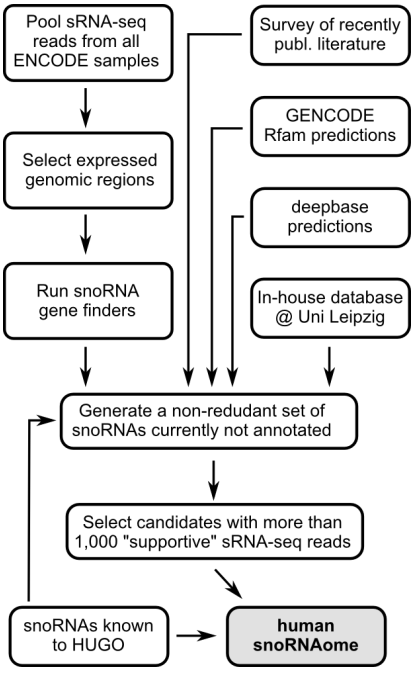


Figure 2. Outline of the snoRNA annotation strategy used in this study. We combined *de novo* search on ENCODE sRNA-seq expressed regions with snoRNA genes and predictions from various databases as well as extensive literature review. Finally, all candidate sequences were checked for a supportive sRNA-seq read pattern to identify high confidence, currently not annotated snoRNA genes.

Table 1. Overview of the snoRNAs identified in this study. Numbers in parentheses indicate snoRNAs without close homologs among the already annotated snoRNAs from the RFAM database [64,65].

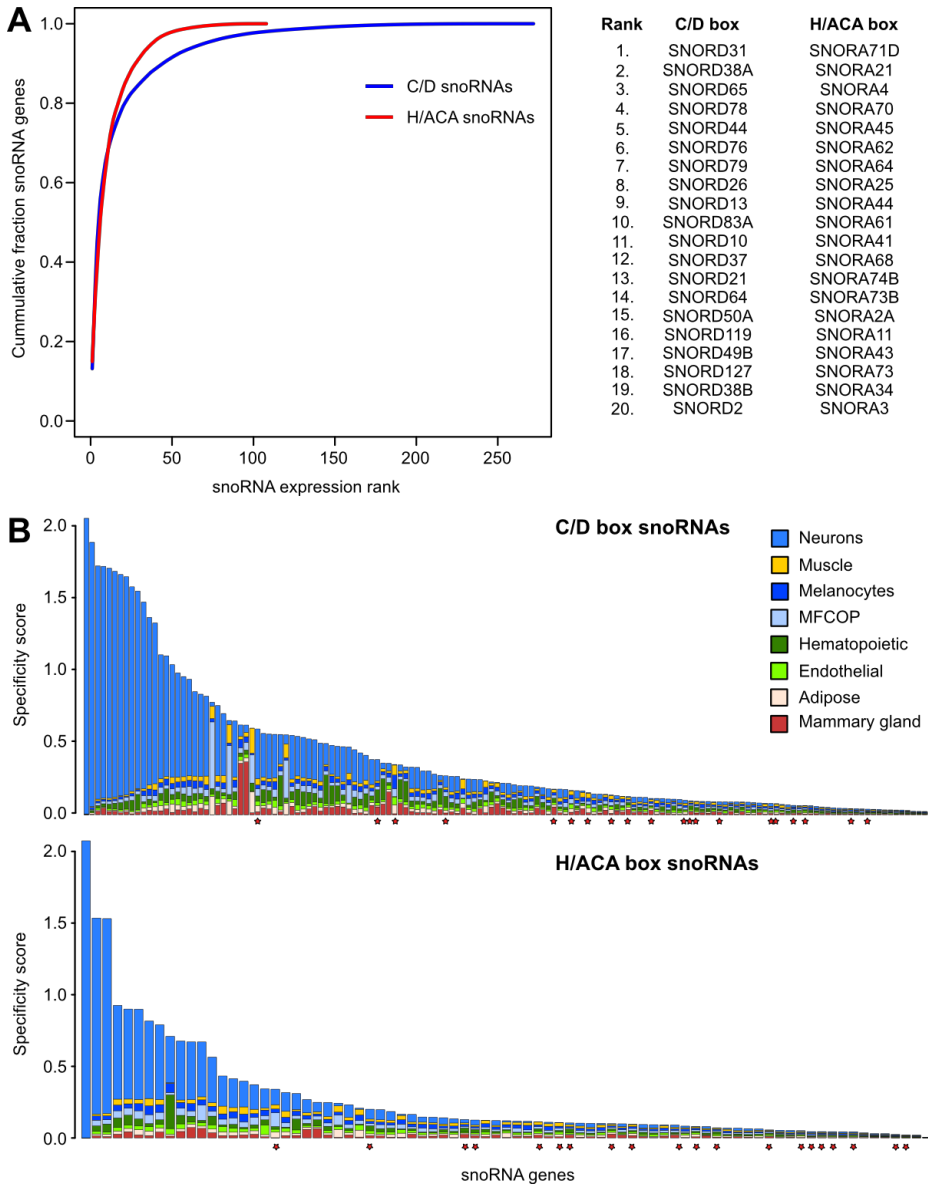
	HUGO annotation	Currently not annotated	Total
C/D box snoRNAs	272	119 (77)	391
C/D box-like snoRNAs	-	93 (92)	93
H/ACA snoRNAs	108	54 (12)	162
scaRNAs	24	2 (0)	26

Expression profiling of human snoRNAs

The plasticity of snoRNA expression across cell types has been relatively poorly studied, although changes in snoRNA expression have been observed in cancers [47]. Because the ENCODE consortium profiled noncoding RNA expression over a diverse set of normal and malignant cell types we analyzed the tissue specificity of expression of the snoRNAs in our catalog. We found that both H/ACA box and the C/D box snoRNA pools are dominated by a few abundantly expressed snoRNAs (Figure 3A). In particular, 21 and 18 C/D box and H/ACA box snoRNAs account for more than 80% of sRNA-seq reads captured for the respective snoRNA class. Of these abundantly expressed snoRNAs, only two of the C/D box family (SNORD83A and SNORD64) and only four of the H/ACA family (SNORA73B, SNORA11, SNORA73A and SNORA51) do not have well confirmed target sites on ribosomal RNAs. This indicates that abundantly expressed snoRNAs are essential for ribosome biogenesis. Consistently, these snoRNAs also show little variation in expression across cell types (Fig. 3B,C denoted by red stars; high resolution versions of these figures including gene names can be found in Supplementary Figure S2). On the other hand, some snoRNAs, belonging to both the C/D box and the H/ACA box class, do exhibit cell type-specific expression. The vast majority of these are expressed in neuronal cell types and include the well known, neuronal specific orphan SNORD115 and SNORD116 families [28,48,49] as well as snoRNAs with canonical ribosomal targets such as SNORD100 and SNORD33. The orphan H/ACA box SNORA35, which is known to be expressed in neurons [50], has the strongest cell type specificity among the H/ACA box snoRNAs. However, H/ACA box snoRNAs with canonical ribosomal targets such as SNORA54 or SNORA22 also show a cell type specific bias of expression. A comprehensive listing of snoRNAs that show cell type specific expression can be found in Table 2.

Furthermore, we performed hierarchical clustering of a subset of sRNA-seq samples that have been generated from decapped (tobacco acid phosphatase (TAP)-treated) RNAs isolated from whole cells (Supplementary Figure S3), and found a striking separation of normal and malignant cell lines with several snoRNAs being differentially expressed in all cancer cell lines compared to cells of non-malignant origin. This is consistent with the results of prior studies that identified snoRNAs as putative cancer biomarkers [51–55]. It also parallels a recent finding that a specific set of tRNAs undergoes increased expression in cancers, with possible consequences on the translational efficiency in these

cells[56]. To facilitate further investigations into these cancer-associated snoRNAs we compiled the list of snoRNAs with the most significant differential expression in cancer cell lines (Table3 and Table 4). Finally, cells of neuronal origin have a snoRNA expression profile that stands out from those other cell types, due to the relatively large number of neuron-specific snoRNAs. Other cell types show more similar profiles, although the mammary gland and hematopoietic cell types tend to cluster closer together, as do the muscle and adipose tissue. The remaining cell types (melanocytes, fibroblasts, osteoblasts, chondrocytes and placental tissue) form one big cluster with no clear boundaries.
(see supplementary Figures S3 and S4).



Rank	C/D box	H/ACA box
1.	SNORD31	SNORA71D
2.	SNORD38A	SNORA21
3.	SNORD65	SNORA4
4.	SNORD78	SNORA70
5.	SNORD44	SNORA45
6.	SNORD76	SNORA62
7.	SNORD79	SNORA64
8.	SNORD26	SNORA25
9.	SNORD13	SNORA44
10.	SNORD83A	SNORA61
11.	SNORD10	SNORA41
12.	SNORD37	SNORA68
13.	SNORD21	SNORA74B
14.	SNORD64	SNORA73B
15.	SNORD50A	SNORA2A
16.	SNORD119	SNORA11
17.	SNORD49B	SNORA43
18.	SNORD127	SNORA73
19.	SNORD38B	SNORA34
20.	SNORD2	SNORA3

Figure 3. Expression profiling of snoRNA genes in ENCODE sRNA-seq data. (**A**) The pool of human snoRNA genes is dominated by a few abundantly expressed snoRNA genes. (**B**) Evaluation of tissue specific expression of snoRNA genes. The top panel show values for C/D box snoRNAs, while the bottom panel does for H/ACA box snoRNAs. The higher the specificity score is the more biased the expression to a specific tissue or cell type is. MFOCP is an acronym for melanocytes, fibroblasts, osteoblasts, chondrocytes and placental tissue.

Limited evidence of tissue-specificity of snoRNA-derived fragments

Several previous studies described snoRNA-derived fragments and suggested that, with some exceptions, the pattern of processing is conserved across snoRNAs and tissues [30,57]. Furthermore, various groups proposed that snoRNA-derived fragments may have non-canonical functions [30,31,48,58–63]. We asked whether the relative proportion of short (less than 40 nt) snoRNA-derived fragments differs between snoRNAs and whether it differs across cell types (see Materials and Methods) for a given snoRNA. We observed that the majority of C/D box snoRNAs (75%) are found predominantly as mature forms in the data. That is, the proportion of processing products is <50% of the reads associated with the snoRNA. The cumulative distribution of this proportion is shown in Supplementary Figure S5. Furthermore, we found only minor differences in this proportion across the tissues where the snoRNAs are expressed. Notable exceptions are the SNORD115, 116, 113 and 114 families . A group of snoRNA comprising SNORD50, SNORD19, SNORD32B, SNORD123, SNORD111, SNORD72, SNORD93, SNORD23 and SNORD85, gives rise to > 90% processed fragments, yet we did not find evidence that these snoRNAs are processed into shorter forms in a cell type-dependent manner (Supplementary File S3 and Supplementary Figure S6).

Conclusions

The wide availability of deep sequencing technologies has prompted thorough investigations into the processing and expression patterns of all types of RNA molecules, including those with relatively well characterized functions such as the snoRNAs. In turn, the improved understanding of these molecules' biogenesis enables their identification in large-scale data sets with increased accuracy. Among the small RNAs, snoRNAs have a relatively long history, going back to the late 1960's (Maxwell and Fournier 1995). A comprehensive database of human C/D box and H/ACA box snoRNAs has been constructed (https://www-snorna.biotoul.fr/) [34], but unfortunately, this database has not been updated since deep sequencing studies started to uncover additional snoRNA molecules. Furthermore, the number of novel snoRNAs that emerged from these recent studies varies widely, and there is some controversy concerning the criteria that were used in defining the snoRNAs.

Here we combined known sequence and structure properties of snoRNAs with recently uncovered patterns of processing and with expression evidence to generate an updated catalog of human C/D box and H/ACA box snoRNAs. Our analysis suggests that although many genomic regions may give rise to RNAs that are processed by the snoRNA-processing machinery and even bind the core proteins of the snoRNP complex, as has been observed before [27], only a relatively small number (hundreds) of these molecules are expressed at a level that is comparable to other well-characterized snoRNAs.
Finally, our analysis indicates that snoRNA expression is not "static", but can undergo some dynamics. Although it has been long known that neurons specifically express a large number of snoRNAs, here we found a striking difference in snoRNA expression between normal and malignant cells. Whether changes in snoRNA expression are reflected in the processing of the target molecules such as rRNAs and whether this has a consequence for the mRNA translation are very interesting questions that remain to be investigated in the future. Our study facilitates these studies by providing a catalog of snoRNAs and the associated rRNA modifications that could then be studied in a targeted manner.

Materials and Methods

Curation of mature forms of known snoRNA genes

A list of snoRNA genes currently annotated by HGNC was obtained from www.genenames.org (3.3.2014) and the corresponding sequence entries were retrieved from the NCBI Nucleotide database via accession numbers as identifiers. Retrieved sequences were then mapped to the hg19 human genome with BLAT to infer their genomic loci. To annotate the genomic coordinates of mature snoRNA genes, we took advantage of the massive sRNA-seq data produced by the ENCODE Consortium [41]. We retrieved the BAM files containing the genomic loci of the reads from 114 sRNA-seq data sets (read length of 101 nt) from the UCSC ENCODE analysis hub (http://genome.ucsc.edu/ENCODE).

To select reads that could support mature snoRNA genes, we used the following criteria: First, we required that either the sRNA-seq read covers at least 75% of a snoRNA gene or the sRNA-seq read was longer than 90 nt (and the snoRNA gene was presumably too long to be covered by sRNA-seq reads). Second, we required that the first and last genomic positions where the sRNA-seq read mapped were at most 5 nt away from the start and end position of the annotated snoRNA gene to which the read mapped. After thus identifying sRNA-seq reads associated with individual snoRNA genes, we redefined the boundaries of the mature snoRNA forms as the positions where most of the sRNA-reads associated with the locus started or ended, respectively. For snoRNA loci with too few sRNA-seq supporting reads, we manually curated the genomic coordinates of the mature forms based on the sRNA-seq reads profile (see Supplementary file 1). To further validate this procedure, we examined the distance between the 5' and 3' ends and the C and D box motifs, respectively. We found that, as shown before in ref. 27, the 5' end of C/D box snoRNA was located 4-5 nt upstream of the C box motif, and the 3' end at most 5 nt downstream of the D box motif. In turn, we used this information as another indication for curating the 5' and 3' end coordinates of the mature snoRNAs for which the sRNA-seq data did not sufficiently or completely covered the loci. Annotated snoRNA with a coverage of less than 100 reads (corresponding to 0.0087 TPM) are SNORD 113-1,113-2, 116-28, 114-7, 115-45,115-47, 108, 114-2, 56B, 114-8, 114-30, 116-10 and SNORA29. It is worth noting that the majority of these come from large, repetitive families.

Identification of predicted snoRNAs with supporting expression data from the ENCODE project

To uncover additional snoRNA genes that have supporting expression evidence, we first collected predictions of two computational tools, snoSeeker [46] and snoReport [45], that have been specifically designed to predict snoRNA genes. To that end, we restricted the search space to genomic regions that were supported by at least five reads in the combined set of sRNA-seq samples and extended these loci by 20 nt from the 5' end and 100 nt from the 3' end. The predictions of snoSeeker and snoReport were pooled and candidate snoRNAs genes overlapping with already annotated snoRNA genes were removed. This step yielded 820,835 putative C/D box snoRNA loci and 316,076 H/ACA box snoRNA

loci.

Because the sequence and structure constraints on snoRNAs appear to be weaker compared to, for example, tRNAs, we expect a higher false-positive rate of prediction for snoRNAs compared to tRNAs. Here we used the observation that C/D box snoRNAs undergo precise processing which leaves only 4-5 nt upstream of the C box, and 2-5 nt downstream of the D box [27] to further validate the C/D box snoRNA prediction. Small RNA-seq reads that mapped to C/D box snoRNA loci were considered 'supportive' of a snoRNA mature form if the 5' end of the read was located 4-5 nt upstream of the inferred C box and the 3' end of the read was located 2-5 nt downstream of the D box. For C/D box snoRNA genes with a predicted length of more than 100 nt, we could only enforce that the 5' end is processed as expected, but we required that the sRNA-seq reads cover at least 75% of the length of the predicted snoRNA gene or are at least 90 nt in length. For H/ACA box snoRNAs, a read was labelled as supportive if the 5' end of the read was located +/- 5 nt around the predicted 5' end of the snoRNA locus, and the read either covered at least 75% of the length of the snoRNA locus or was at least 90 nt in length. 8,000 predicted C/D box snoRNAs and 7,772 predicted H/ACA box snoRNAs had at least one supportive read, but only 121 and 114, respectively, remained when we required at least 1000 supportive reads (corresponding to 0.087 TPM) in the entire data set. In the next step, candidate snoRNA loci were filtered for redundancy and loci overlapping with predictions obtained from deepBase, Leipzig, and GENCODE were removed. Finally, we removed candidate loci where more than 25% of the loci overlapped with repeat annotation and discarded those that did not have support by uniquely mapped reads. In the end, our *de novo* prediction yielded 12 and 74 H/ACA box and C/D box snoRNA loci, respectively. These putative snoRNAs can be found in Supplementary File 1, under "de novo" category.

In previous work [27], we found that core snoRNP proteins bind snoRNA-like RNAs, that we not reported in snoRNA databases. To capture these cases, we carried out a genome-wide scan for C/D box snoRNA-like molecules that are supported by sRNA-seq evidence. We started from genomic regions defined by a degenerate C box ("TGATGA", "TGGTGA", "TGATGT", "TGATGC" or "TGTTGA") and a D box (C|ATGA) separated by 10-90 nts. Applying the same filtering steps as we did for the predictions generated by snoReport/snoSeeker (see above) we identified 93 CD-box like candidates that have at least 1000 supportive reads in the sRNA-seq data. These can be found under the "snoRNA-like" category in the Supplementary File 1.

Analysis of the expression profiles of known snoRNA genes and snoRNA-derived fragment based on ENCODE

The expression level of a given snoRNA in a sample was calculated based on the total number of reads (uniquely and multi-mapping) from that sample that overlapped with the snoRNA locus. The normalization of read counts was done relative to the total number of reads obtained in the sample. The ENCODE project generated sRNA-seq samples from a range of cell types, both normal and malignant, as well as from distinct sub-cellular compartments ("Cell", "Cytosol", "Chromatin", "Nucleus" and "Nucleolus"). Furthermore, to capture various types of small RNAs, the RNA was subjected to various treatments (tobacco acid phosphatase ("TAP") to remove cap structures, calf intestinal phosphatase and

TAP ("CIP-TAP") to further remove 5' and 3' phosphates, as well as left untreated "No treatment")). Based on the calculated expression values of each snoRNA in each sample we carried out hierarchical clustering of the snoRNAs expression profiles as well as the samples based on the similarity in their corresponding snoRNA expression profiles. The results are shown in Supplementary Figure S7 for C/D and H/ACA box snoRNAs. Because samples that were prepared similarly and were generated from the same cellular compartment tended to cluster together for the expression analysis across cell types we used samples that were obtained from the same cellular compartment ("cell") and with the same treatment ("TAP"), as these covered the largest variety of cell types. Furthermore, we normalized the reads relative to the total expression of snoRNAs in the given sample, excluding other types of molecules. Because snoRNAs tend to form families of closely related sequences, we also grouped snoRNAs that were more than 80% identical over their entire length. Supplementary File S4 contains the list of snoRNAs and their corresponding cluster representatives. The expression level of a cluster representative was defined as the average expression level of all snoRNAs associated with that cluster. When replicates were available, we further averaged expression over replicates as well. Given the normalized expression levels thus calculated, we evaluated the specificity of expression or of processing of individual snoRNAs as follows. To quantify the specificity of expression, we first computed the relative frequency of each snoRNA in a given sample. Then we calculated a specificity score defined as

$$S(p_1, p_2, .., p_n) = log(n) - \sum_{i=1}^{n} p_i log(p_i)$$

where p_i is the normalized frequency of the snoRNA in sample i. The specificity score is maximal when the snoRNA is expressed in a single sample and minimal when the relative frequency of the snoRNA is the same across all samples.

snoRNAs dysregulated in cancer

To directly compare snoRNA expression between normal and malignant cells, we averaged the snoRNA expression separately over normal and malignant cell types. The ratio of these quantities gives us the fold-change of expression between normal and malignant cells.

Expression profiling of snoRNA-derived fragments

To determine whether processed fragments are generated in a cell type-specific manner, we first separated the reads into those that correspond to the mature snoRNA and to shorter processed products. Because the sRNA-seq samples should in principle contain only full-length RNAs and based on the length distribution of snoRNAs (Supplementary Figure S8), we chose a maximum length of 40 nt for a read to be considered as corresponding to a processed RNA. This is consistent with the length of snoRNA-derived fragments that was reported before [27,30,57,58]. Next, we calculated the proportion of processed reads among all reads associated with the snoRNA. Finally, we calculated a specificity score of snoRNA expression or of processing across tissues as described above for the specificity of snoRNA expression.

Table 2. Summary of snoRNAs with a highly cell type-specific expression (specificity score > 0.6). MFOCP stands for melanocytes, fibroblasts, osteoblasts, chondrocytes and placental tissue.

SnoRNA name	Cells in which it is expressed	Associated samples
SNORD115 family, SNORD116 family, SNORD100, SNORD109, SNORD107, SNORD29	Neurons	H1_neurons
SNORD33, SNORD81, SNORD105, SNORD68, SNORD11, SNORD36A, SNORD102, SNORD111, SNORD12B, SNORD30, SNORD69, SNORD32A (2), SNORD12, SNORD22, SNORD50A, SNORD11B, SNORD55, SNORD105B	Neurons and lymphoblastoid cells	H1_neurons, GM12878
SNORD11B	Neurons and pericytes	H1_neurons, HPC_PL
SNORD112	MFOCP	HCH
SNORD113-8 (7)	MFOCP	hMSC-BM
SNORD114-22 (28)	MFOCP	HPIEpC
SNORD7	Neurons and Endothelial cells	H1_neurons, HAoEC
SNORD46, SNORD42A	Mammary gland and lymphoblastoid cells	HMEpC and GM12878
SNORD125, SNORD85, SNORD91A	hematopoietic, neurons and lymphoblastoid cells	CD34+, H1_neuron, GM12878
SNORA35, SNORA36B (3)	Neurons	H1_neurons
SNORA54, SNORA22, SNORA16A (2), SNORA48, SNORA63, SNORA14B(2), SNORA5A	Neurons and lymphoblastoid cells	H1_neurons, GM12878
SNORA47	Neurons, hematopoietic and lymphoblastoid cells	H1_neurons, CD34+, GM12878

SNORA55	Neurons and pericytes	H1_neurons, HPC_PL

Table 3. SnoRNAs whose expression differs substantially (more than 5-fold) and significantly (p--value < 0.0005 in the t-test) between malignant and normal cells.

snoRNA name	Fold change (log2) (malignant vs normal cells	p-value (two sample t-test)	Expression (total reads across the 114 samples)
SNORD10	-3.79	1e-14	29256584
SNORD105B	-3.60	1.5e-5	610940
SNORD76	-3.57	1.7e-10	85151249
SNORD79	-3.32	1.1e-6	53384564
SNORD65	-3.07	2.8e-13	175648163
SNORD123	-3.07	5e-4	73633
SNORD80	-3.05	2e-6	4859473
SNORD29	-2.96	7e-10	3519100
SNORD58A	-2.94	2e-6	1797396
SNORD21	-2.74	2e-8	21310926
SNORD107	-2.56	6e-5	32073
SNORD15B	-2.56	9e-9	7858324
SNORD119	-2.32	6e-7	18167463
SNORA68	-4.10	2e-12	8591178
SNORA8	-3.51	1.3e-6	1674927
SNORA34	-3.42	4.7e-9	3559104
SNORA62	-3.36	8e-9	17837817
SNORA44	-3.10	5e-12	10314585

SNORA20	-2.68	3e-5	1280339
SNORA57	-2.66	2e-10	1857248
SNORA23	-2.63	2e-8	1262349
SNORA43	-2.54	1e-9	3794469
SNORA49	-2.49	8e-6	1770801
SNORA14B	-2.41	1e-5	398365
SNORA74B	-2.38	5e-5	6202921
SNORA60	-2.35	1e-7	77417

Table 4. snoRNAs dysregulated in different cancer types based on their expression profiles in cancerous versus normal cell lines (references are cited in case the snoRNA is found dysregulated in recent cancer studies)

	Regulation	**Comment**	**Reference**
SNORA47	**Strongly downregulated**	**All cancer types**	[53]
SNORA78	**Strongly downregulated**	**Brain Cancer**	[53]
SNORA59A	**Strongly downregulated**	**Brain and Breast Cancer**	[66]
SNORA22	**Extremely downregulated**	**Lung Cancer**	
SNORA55	**Extremely downregulated**	**Brain Cancer**	
SNORA68	**Extremely downregulated**	**All cancer types**	[53]
SNORA60 (SNORA71 cluster)	**downregulated**	**All cancer types**	[53]

SNORA44, SNORA61	**downregulated**	**All cancer types**	
SNORA62,SNORA12,SNORA52, SNORA14B, SNORA38B, SNORA84, SNORA17	**downregulated**	**Lung Cancer**	
SNORA25	**downregulated**	**Breast Cancer**	
SNORA70 cluster	**downregulated**	**All cancer types**	[53,66]
SNORA57, SNORA34, SNORA8, SNORA43, SNORA67	**downregulated**	**All cancer types**	
SNORA76	**downregulated**	**Lung Cancer**	[53]
SNORA49	**Extremely downregulated**	**Lung Cancer**	
SNORA20 , SNORA24, SNORA23, SNORA77, SNORA39 ,SNORA11	**downregulated**	**Breast Cancer**	
SNORA18, SNORA53, SNORA74B	**downregulated**	**Brain Cancer**	
SNORA50, SNORA32, SNORA36B, SNORA69, SNORA41	**Over-expressed**	**All cancer types**	
SNORA21, SNORA64	**Over-expressed**	**All cancer types**	[53]
SNORA74A, SNORA73, SNORA19, SNORA4	**Over-expressed**	**Lung Cancer**	

	Regulation	**Comment**	**Reference**
SNORD58A, SNORD107, SNORD109A, SNORD116-26 (4), SNORD116-3 (23), SNORD64, SNORD69	**extremely downregulated**	**Breast cancer**	

SNORD29	**extremely downregulated**	**Brain cancer**	
SNORD28, SNORD80, SNORD10	**downregulated**	**All cancer types (SNORD10 resides on ELF4A1 intron)**	[53]
SNORD112, SNORD113-8 (7), SNORD114-22 (28), SNORD123	**extremely downregulated**	**All cancer types (MEG3 harbors a couple of snoRNAs, including SNORD112, SNORD113, and SNORD114 and tumor suppressor miRNAs)**	[53,67,68,69,47,66]
SNORD76, SNORD83B,SNORD65	**downregulated**	**All cancer types**	[70,51]
SNORD127, SNORD119, SNORD21, snord49B,SNORD9, SNORD126, SNORD105B,SNORD65 SNORD87,SNORD58C,SNORD15B, SNORD12C, SNORD79, SNORD44	**downregulated**	**All cancer types**	
SNORD110	**downregulated**	**Lung cancer**	[51]
SNORD117, SNORD103, SNORD46, SNORD42A, SNORD71	**downregulated**	**Lung cancer**	

SNORD33	**Over-expressed**	**Lung cancer**	[70]
SNORD66, SNORD32A (2), SNORD18B (3),SNORD38A, SNORD50B, SNORD96A (2), SNORD74, SNORD36B, SNORD24, SNORD104,	**Over-expressed**	**All cancer types**	[51,70–74]
SNORD111B, SNORD85, SNORD62A(2), SNORD14E, SNORD18B (3), SNORD121B, SNORD14A(2), SNORD4A, SNORD15A, SNORD1B, SNORD77, SNORD101, SNORD63 SNORD91B, SNORD2, SNORD25, SNORD75	**Over-expressed**	**All cancer types**	
SNORD90, SNORD36A	**Over-expressed**	**Brain cancer**	
SNORD32A (2), SNORD27, SNORD30, SNORD59B, SNORD91A, SNORD86, SNORD1A, SNORD1C, SNORD93, SNORD55, SNORD72, SNORD22, SNORD100, SNORD4B, SNORD92, SNORD49A, SNORD118	**Over-expressed**	**Lung cancer**	
SNORD38B, SNORD104, SNORD24	**Over-expressed**	**Breast Cancer**	

References

1. Decatur WA, Fournier MJ. rRNA modifications and ribosome function. Trends Biochem Sci 2002; 27:344–51.

2. Darzacq X, Jády BE, Verheggen C, Kiss AM, Bertrand E, Kiss T. Cajal body-specific small nuclear RNAs: a novel class of 2'-O-methylation and pseudouridylation guide RNAs. EMBO J 2002; 21:2746–56.

3. D' Orval BC, Bortolin M-L, Gaspin C, Bachellerie J-P. Box C/D RNA guides for the ribose methylation of archaeal tRNAs. The tRNATrp intron guides the formation of two ribose-methylated nucleosides in the mature tRNATrp. Nucleic Acids Res 2001; 29:4518–29.

4. Kiss T. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. Cell 2002; 109:145–8.

5. Matera AG, Terns RM, Terns MP. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. Nat Rev Mol Cell Biol 2007; 8:209–20.

6. Bratkovič T, Rogelj B. Biology and applications of small nucleolar RNAs. Cell Mol Life Sci 2011; 68:3843–51.

7. Tollervey D, Kiss T. Function and synthesis of small nucleolar RNAs. Curr Opin Cell Biol 1997; 9:337–42.

8. Darzacq X, Kiss T. Processing of intron-encoded box C/D small nucleolar RNAs lacking a 5',3'-terminal stem structure. Mol Cell Biol 2000; 20:4522–31.

9. Kiss T. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. EMBO J 2001; 20:3617–22.

10. McKeegan KS, Debieux CM, Boulon S, Bertrand E, Watkins NJ. A Dynamic Scaffold of Pre-snoRNP Factors Facilitates Human Box C/D snoRNP Assembly. Mol Cell Biol 2007; 27:6782–93.

11. Tollervey D, Lehtonen H, Jansen R, Kern H, Hurt EC. Temperature-sensitive mutations demonstrate roles for yeast fibrillarin in pre-rRNA processing, pre-rRNA methylation, and ribosome assembly. Cell 1993; 72:443–57.

12. Nicoloso M, Qu LH, Michot B, Bachellerie JP. Intron-encoded, antisense small nucleolar RNAs: the characterization of nine novel species points to their direct role as guides for the 2'-O-ribose methylation of rRNAs. J Mol Biol 1996; 260:178–95.

13. Kiss-László Z, Henry Y, Bachellerie JP, Caizergues-Ferrer M, Kiss T. Site-specific ribose methylation of preribosomal RNA: a novel function for small nucleolar RNAs. Cell 1996; 85:1077–88.

14. Cavaillé J, Nicoloso M, Bachellerie JP. Targeted ribose methylation of RNA in vivo directed by

tailored antisense RNA guides. Nature 1996; 383:732–5.

15. Balakin AG, Smith L, Fournier MJ. The RNA world of the nucleolus: two major families of small RNAs defined by different box elements with related functions. Cell 1996; 86:823–34.

16. Ganot P, Caizergues-Ferrer M, Kiss T. The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. Genes Dev 1997; 11:941–56.

17. Lafontaine D, Bousquet-Antonelli C. The box H+ ACA snoRNAs carry Cbf5p, the putative rRNA pseudouridine synthase. Genes [Internet] 1998; Available from: http://genesdev.cshlp.org/content/12/4/527.short

18. Ganot P, Bortolin ML, Kiss T. Site-specific pseudouridine formation in preribosomal RNA is guided by small nucleolar RNAs. Cell 1997; 89:799–809.

19. Bortolin ML, Ganot P, Kiss T. Elements essential for accumulation and function of small nucleolar RNAs directing site-specific pseudouridylation of ribosomal RNAs. EMBO J 1999; 18:457–69.

20. Richard P, Darzacq X, Bertrand E, Jády BE, Verheggen C, Kiss T. A common sequence motif determines the Cajal body-specific localization of box H/ACA scaRNAs. EMBO J 2003; 22:4283–93.

21. Marnef A, Richard P, Pinzón N, Kiss T. Targeting vertebrate intron-encoded box C/D 2'-O-methylation guide RNAs into the Cajal body. Nucleic Acids Res 2014; 42:6616–29.

22. Tycowski KT, Shu M-D, Kukoyi A, Steitz JA. A conserved WD40 protein binds the Cajal body localization signal of scaRNP particles. Mol Cell 2009; 34:47–57.

23. Marz M, Gruber AR, Höner Zu Siederdissen C, Amman F, Badelt S, Bartschat S, Bernhart SH, Beyer W, Kehr S, Lorenz R, et al. Animal snoRNAs and scaRNAs with exceptional structures. RNA Biol 2011; 8:938–46.

24. Jády BE, Ketele A, Kiss T. Human intron-encoded Alu RNAs are processed and packaged into Wdr79-associated nucleoplasmic box H/ACA RNPs. Genes Dev 2012; 26:1897–910.

25. Bratkovič T, Rogelj B. The many faces of small nucleolar RNAs. Biochim Biophys Acta 2014; 1839:438–43.

26. Lafontaine DL, Tollervey D. Birth of the snoRNPs: the evolution of the modification-guide snoRNAs. Trends Biochem Sci 1998; 23:383–8.

27. Kishore S, Gruber AR, Jedlinski DJ, Syed AP, Jorjani H, Zavolan M. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. Genome Biol 2013; 14:R45.

28. Kishore S, Stamm S. The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor

2C. Science 2006; 311:230–2.

29. Doe CM, Relkovic D, Garfield AS, Dalley JW, Theobald DEH, Humby T, Wilkinson LS, Isles AR. Loss of the imprinted snoRNA mbii-52 leads to increased 5htr2c pre-RNA editing and altered 5HT2CR-mediated behaviour. Hum Mol Genet 2009; 18:2140–8.

30. Scott MS, Ono M, Yamada K, Endo A, Barton GJ, Lamond AI. Human box C/D snoRNA processing conservation across multiple cell types. Nucleic Acids Res 2012; 40:3676–88.

31. Ender C, Krek A, Friedländer MR, Beitzinger M, Weinmann L, Chen W, Pfeffer S, Rajewsky N, Meister G. A human snoRNA with microRNA-like functions. Mol Cell 2008; 32:519–28.

32. Yin Q-F, Yang L, Zhang Y, Xiang J-F, Wu Y-W, Carmichael GG, Chen L-L. Long noncoding RNAs with snoRNA ends. Mol Cell 2012; 48:219–30.

33. Xie J, Zhang M, Zhou T, Hua X, Tang L, Wu W. Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs. Nucleic Acids Res 2007; 35:D183–7.

34. Lestrade L, Weber MJ. snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs. Nucleic Acids Res 2006; 34:D158–62.

35. Ellis JC, Brown DD, Brown JW. The small nucleolar ribonucleoprotein (snoRNP) database. RNA 2010; 16:664–6.

36. Zhang Y, Liu J, Jia C, Li T, Wu R, Wang J, Chen Y, Zou X, Chen R, Wang X-J, et al. Systematic identification and evolutionary features of rhesus monkey small nucleolar RNAs. BMC Genomics 2010; 11:61.

37. Liu T-T, Zhu D, Chen W, Deng W, He H, He G, Bai B, Qi Y, Chen R, Deng XW. A global identification and analysis of small nucleolar RNAs and possible intermediate-sized non-coding RNAs in Oryza sativa. Mol Plant 2013; 6:830–46.

38. Gardner PP, Bateman A, Poole AM. SnoPatrol: how many snoRNA genes are there? J Biol 2010; 9:4.

39. Kaur D, Gupta AK, Kumari V, Sharma R, Bhattacharya A, Bhattacharya S. Computational prediction and validation of C/D, H/ACA and Eh_U3 snoRNAs of Entamoeba histolytica. BMC Genomics 2012; 13:390.

40. Makarova JA, Kramerov DA. SNOntology: Myriads of novel snoRNAs or just a mirage? BMC Genomics 2011; 12:543.

41. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, Mortazavi A, Tanzer A, Lagarde J, Lin W, Schlesinger F, et al. Landscape of transcription in human cells. Nature 2012; 489:101–8.

42. Deschamps-Francoeur G, Garneau D, Dupuis-Sandoval F, Roy A, Frappier M, Catala M, Couture S, Barbe-Marcoux M, Abou-Elela S, Scott MS. Identification of discrete classes of small nucleolar RNA featuring different ends and RNA binding protein dependency. Nucleic Acids Res

2014; 42:10073–85.

43. Derrien T, Johnson R, Bussotti G, Tanzer A, Djebali S, Tilgner H, Guernec G, Martin D, Merkel A, Knowles DG, et al. The GENCODE v7 catalog of human long noncoding RNAs: analysis of their gene structure, evolution, and expression. Genome Res 2012; 22:1775–89.

44. Yang J-H, Shao P, Zhou H, Chen Y-Q, Qu L-H. deepBase: a database for deeply annotating and mining deep sequencing data. Nucleic Acids Res 2010; 38:D123–30.

45. Hertel J, Hofacker IL, Stadler PF. SnoReport: computational identification of snoRNAs with unknown targets. Bioinformatics 2008; 24:158–64.

46. Yang J-H, Zhang X-C, Huang Z-P, Zhou H, Huang M-B, Zhang S, Chen Y-Q, Qu L-H. snoSeeker: an advanced computational package for screening of guide and orphan snoRNA genes in the human genome. Nucleic Acids Res 2006; 34:5112–23.

47. Mannoor K, Liao J, Jiang F. Small nucleolar RNAs in cancer. Biochim Biophys Acta 2012; 1826:121–8.

48. Kishore S, Khanna A, Zhang Z, Hui J, Balwierz PJ, Stefan M, Beach C, Nicholls RD, Zavolan M, Stamm S. The snoRNA MBII-52 (SNORD 115) is processed into smaller RNAs and regulates alternative splicing. Hum Mol Genet 2010; 19:1153–64.

49. Bortolin-Cavaillé M-L, Cavaillé J. The SNORD115 (H/MBII-52) and SNORD116 (H/MBII-85) gene clusters at the imprinted Prader–Willi locus generate canonical box C/D snoRNAs. Nucleic Acids Res 2012; 40:6800–7.

50. Cavaillé J, Buiting K, Kiefmann M, Lalande M, Brannan CI, Horsthemke B, Bachellerie JP, Brosius J, Hüttenhofer A. Identification of brain-specific and imprinted small nucleolar RNA genes exhibiting an unusual genomic organization. Proc Natl Acad Sci U S A 2000; 97:14311–6.

51. Mannoor K, Shen J, Liao J, Liu Z, Jiang F. Small nucleolar RNA signatures of lung tumor-initiating cells. Mol Cancer 2014; 13:104.

52. Lin Y, Li Z, Ozsolak F, Kim SW, Arango-Argoty G, Liu TT, Tenenbaum SA, Bailey T, Monaghan AP, Milos PM, et al. An in-depth map of polyadenylation sites in cancer. Nucleic Acids Res 2012; 40:8460–71.

53. Gao L, Ma J, Mannoor K, Guarnera MA, Shetty A, Zhan M, Xing L, Stass SA, Jiang F. Genome-wide small nucleolar RNA expression analysis of lung cancer by next-generation deep sequencing. Int J Cancer [Internet] 2014; Available from: http://dx.doi.org/10.1002/ijc.29169

54. Ronchetti D, Mosca L, Cutrona G, Tuana G, Gentile M, Fabris S, Agnelli L, Ciceri G, Matis S, Massucco C, et al. Small nucleolar RNAs as new biomarkers in chronic lymphocytic leukemia. BMC Med Genomics 2013; 6:27.

55. Ronchetti D, Todoerti K, Tuana G, Agnelli L, Mosca L, Lionetti M, Fabris S, Colapietro P, Miozzo M, Ferrarini M, et al. The expression pattern of small nucleolar and small Cajal body-specific RNAs characterizes distinct molecular subtypes of multiple myeloma. Blood

Cancer J 2012; 2:e96.

56. Gingold H, Tehler D, Christoffersen NR, Nielsen MM, Asmar F, Kooistra SM, Christophersen NS, Christensen LL, Borre M, Sørensen KD, et al. A dual program for translation regulation in cellular proliferation and differentiation. Cell 2014; 158:1281–92.

57. Taft RJ, Glazov EA, Lassmann T, Hayashizaki Y, Carninci P, Mattick JS. Small RNAs derived from snoRNAs. RNA 2009; 15:1233–40.

58. Falaleeva M, Stamm S. Processing of snoRNAs as a new source of regulatory non-coding RNAs: snoRNA fragments form a new class of functional RNAs. Bioessays 2013; 35:46–54.

59. Abel Y, Clerget G, Bourguignon-Igel V, Salone V, Rederstorff M. [Beyond usual functions of snoRNAs]. Med Sci 2014; 30:297–302.

60. Scott MS, Avolio F, Ono M, Lamond AI, Barton GJ. Human miRNA precursors with box H/ACA snoRNA features. PLoS Comput Biol 2009; 5:e1000507.

61. Ono M, Scott MS, Yamada K, Avolio F, Barton GJ, Lamond AI. Identification of human miRNA precursors that resemble box C/D snoRNAs. Nucleic Acids Res 2011; 39:3879–91.

62. Brameier M, Herwig A, Reinhardt R, Walter L, Gruber J. Human box C/D snoRNAs with miRNA like functions: expanding the range of regulatory RNAs. Nucleic Acids Res 2011; 39:675–86.

63. Röther S, Meister G. Small RNAs derived from longer non-coding RNAs. Biochimie 2011; 93:1905–15.

64. Burge SW, Daub J, Eberhardt R, Tate J, Barquist L, Nawrocki EP, Eddy SR, Gardner PP, Bateman A. Rfam 11.0: 10 years of RNA families. Nucleic Acids Res 2013; 41:D226–32.

65. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, Bateman A. Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 2005; 33:D121–4.

66. Ferreira HJ, Heyn H, Moutinho C, Esteller M. CpG island hypermethylation-associated silencing of small nucleolar RNAs in human cancer. RNA Biol 2012; 9:881–90.

67. Valleron W, Laprevotte E, Gautier E-F, Quelen C, Demur C, Delabesse E, Agirre X, Prósper F, Kiss T, Brousset P. Specific small nucleolar RNA expression profiles in acute leukemia. Leukemia 2012; 26:2052–60.

68. Liuksiala T, Teittinen KJ, Granberg K, Heinäniemi M, Annala M, Mäki M, Nykter M, Lohi O. Overexpression of SNORD114-3 marks acute promyelocytic leukemia. Leukemia 2014; 28:233–6.

69. Xu G, Yang F, Ding C-L, Zhao L-J, Ren H, Zhao P, Wang W, Qi Z-T. Small nucleolar RNA 113-1 suppresses tumorigenesis in hepatocellular carcinoma. Mol Cancer 2014; 13:216.

70. Liao J, Yu L, Mei Y, Guarnera M, Shen J, Li R, Liu Z, Jiang F. Small nucleolar RNA signatures

as biomarkers for non-small-cell lung cancer. Mol Cancer 2010; 9:198.

71. Michel CI, Holley CL, Scruggs BS, Sidhu R, Brookheart RT, Listenberger LL, Behlke MA, Ory DS, Schaffer JE. Small nucleolar RNAs U32a, U33, and U35a are critical mediators of metabolic stress. Cell Metab 2011; 14:33–44.

72. Dong XY, Rodriguez C, Guo P, Sun X. SnoRNA U50 is a candidate tumor-suppressor gene at 6q14. 3 with a mutation associated with clinically significant prostate cancer. Hum Mol Genet [Internet] 2008; Available from: http://hmg.oxfordjournals.org/content/17/7/1031.short

73. Dong X-Y, Guo P, Boyd J, Sun X, Li Q, Zhou W, Dong J-T. Implication of snoRNA U50 in human breast cancer. J Genet Genomics 2009; 36:447–54.

74. Tanaka R, Satoh H, Moriyama M, Satoh K, Morishita Y, Yoshida S, Watanabe T, Nakamura Y, Mori S. Intronic U50 small-nucleolar-RNA (snoRNA) host gene of no protein-coding potential is mapped at the chromosome breakpoint t(3;6)(q27;q15) of human B-cell lymphoma. Genes Cells 2000; 5:277–87.

75. Su H, Xu T, Ganapathy S, Shadfan M, Long M, Huang TH-M, Thompson I, Yuan Z-M. Elevated snoRNA biogenesis is essential in breast cancer. Oncogene 2014; 33:1348–58.

5 Discussion

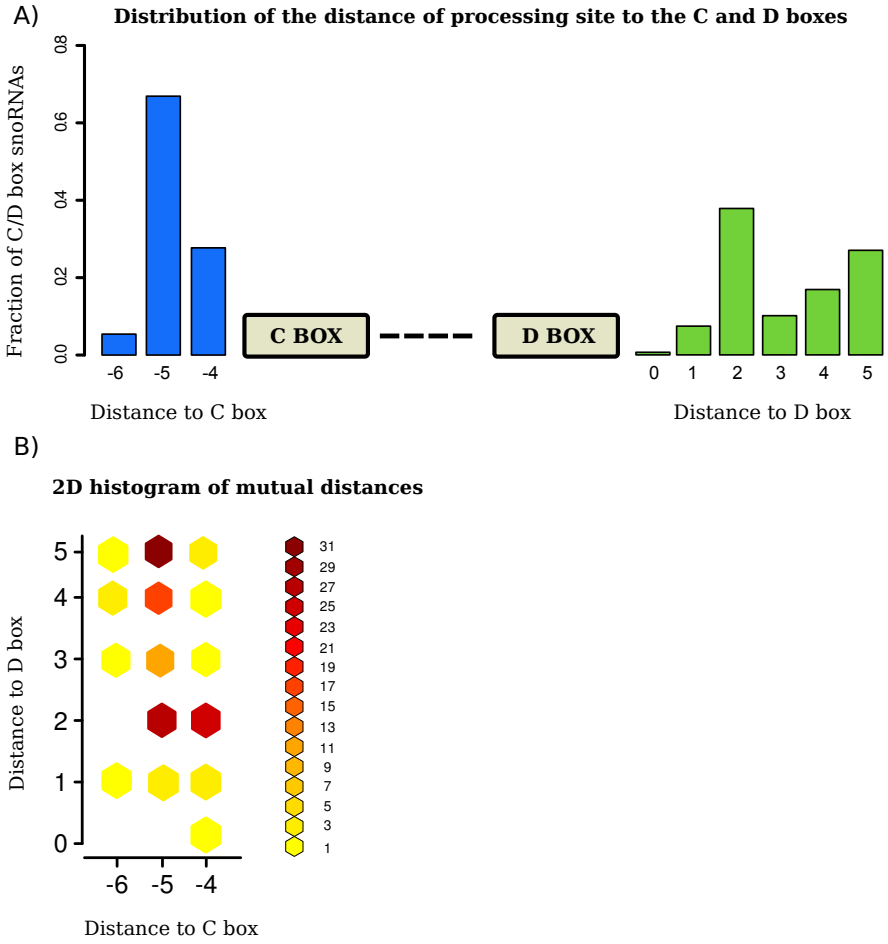
High-throughput sequencing has revolutionized the field of molecular biology. The number of applications as well as the efficiency of the technology in terms of accuracy, cost and speed is rapidly increasing. Among these applications, RNA-seq revealed evidence of pervasive transcription across the genome [69] which prompted a revision of the previously held belief that the human genome consists to a large extent of junk DNA [120]. Whether these resulting transcripts are functional or simply result from stochasticity in the activity of the transcriptional machinery (i.e. they represent transcriptional noise) is still an open question.
Over the past decade various classes of non-coding RNAs have been identified and their functions have been elucidated to a great extent [110]. It has been shown that many non-protein coding transcripts play important roles in diverse set of cellular processes[39, 111, 169]. Thus, many groups started to combine computational and experimental methods in an effort of to uncover functionally important non-coding RNAs. These studies have considered different criteria such as transcription regulatory sequence motifs, secondary structure, conservation across species and any evidence of expression (from RNA-seq, CAGE, SAGE, EST, etc) [64, 63, 125, 167, 153, 52, 53, 51, 148]. Finding non-coding RNAs is more challenging compared to protein-coding genes as they do not have a extended informative coding regions, their function being rather determined mostly by their structure. This makes the development of de novo non-coding RNA prediction algorithms more challenging. Nonetheless, the great interest that non-coding RNAs raised in the past few years resulted in a great improvement in the approaches for their identification. Next generation sequencing technologies enabled generation of vast volumes of sequences, including the genomes and transcriptomes of multiple species [148], thereby providing the material for comparative genomics approaches that could be used towards non-coding RNA identification as well.
In this work we used the NGS data to identify primary transcripts in prokaryotes and to identify novel snoRNAs in the human genome [72, 80]. In a first study we developed a mathematical model for the analysis of dRNA-seq data for identification of TSSs in bacterial genomes. Evidence from NGS has shown that the genome of prokaryotes is more complex than initially thought. Our proposed model quantifies the enrichment of a putative start site in TEX+ versus TEX- samples as well as the dominance in expression of that site relative to nearby genomic positions. The enrichment is modeled using a Bayesian probabilistic framework based on calculating the posterior probability of the underlying read count distribution. We have implemented this model using python and bash scripts as a pipeline which is publicly available and in principle can be applied to any dRNA-seq data to identify putative TSSs genome wide. Based on a set of high confidence TSSs that we derived with the above-mentioned method we trained a hidden markov model representing the consensus motifs as well as the sequence content of promoters in the species that we analyzed. We then applied this model to find additional TSSs that our TSSer model did not initially identify because their expression in the analyzed samples was too low. Alternatively, the issue of sensitivity could be addressed experimentally, by generating dRNA-seq from multiple conditions. An improved annotation of TSSs has consequences for the identification of transcription regulatory motifs and of gene expression regulatory networks, identification of 5'UTRs and characterization of translation regulatory elements therein, finding novel regulatory RNAs.Quantification of expression driven

by individual TSSs has additional application such as general analysis of gene expression and identification of transcription factors that drive gene expression in specific conditions.
The HMM that we developed is only a first step towards prediction of bacterial promoters. An improved de novo predictor may take into account the binding motifs of different sigma factors that help recruit the polymerase at transcription start sites in specific conditions, activator and repressor elements, spacing between the conserved motifs, AT richness of the given genome, distance to start codon as well as other factors that are characteristic to prokaryotic gene promoters. These models can be trained and tested based on the initial set of high confidence TSSs generated by TSSer. DRNA-seq is able to capture the 5' ends of transcripts. However, to determine the full-length bacterial transcripts, a method for mapping transcript 3' ends in prokaryotic systems is still needed. In eukaryotic systems, 3' end sequencing is a method of choice to identify 3' ends of transcripts but a counterpart method in prokaryotes is missing [140, 32]. In the second contribution, PAR-CLIP data was utilized to identify RNAs that associate with snoRNP core proteins. This study aimed to characterize in depth the processing of snoRNAs and snoRNA-derived fragments as well as finding their potential targets. We identified novel snoRNAs which could reproducibly be detected in PAR-CLIP data from different snoRNP associated proteins. We also demonstrated that stem-loops in C/D box snoRNAs undergo precise processing that leaves 4-5 nt upstream of the C box and up to 5 nt downstream of the D box. We later confirmed this processing pattern in the ENCODE data as well. We additionally found short C/D box snoRNAs (up to 28 nt) which lack C' and D' motifs but can still be incorporated into snoRNP complexes. Finally, we observed that snoRNA-derived fragments were mostly produced from snoRNA ends.
Although PAR-CLIP method has been successfully applied in several genome scale studies (such as genome-wide identification of miRNA targets [55, 56]to investigate the RNA-protein interaction, this method suffers from some drawbacks which must be improved in future. One is that the method is complex and thereby susceptible to various biases. For example, the RNase treatment that is applied during sample preparation can bias the set of identified RNAs. To identify the snoRNA targets, in this contribution we trained a biophysical model similar to one that was developed in our group for the prediction of miRNA-target interactions [78], on known snoRNA-rRNA interactions. Because the training data was very limited, we think that there is much room left for improving this model. One possible direction that could be pursued in the future is to use instead of a limited number of known snoRNA-rRNA interactions data from crosslinking and sequencing of hybrids (CLASH) experiments [60, 61]. These experiments aim to generate and capture chimeric reads that result from the ligation of hybrids that form between snoRNAs and their corresponding targets. Such methods have been used successfully to identify microRNA targets and there is great potential in applying them to finally determine the targets of the so-called orphan snoRNAs, which so far do not have any identified target.
In the third contribution we screened ENCODE expressed regions to find snoRNA genes in the human genome, hence expanding the current catalogue of known snoRNAs. The extensive amount of data provided by ENCODE project created the opportunity to validate the genomic elements (e.g. coding and non-coding transcripts) which were predicted by de novo algo-

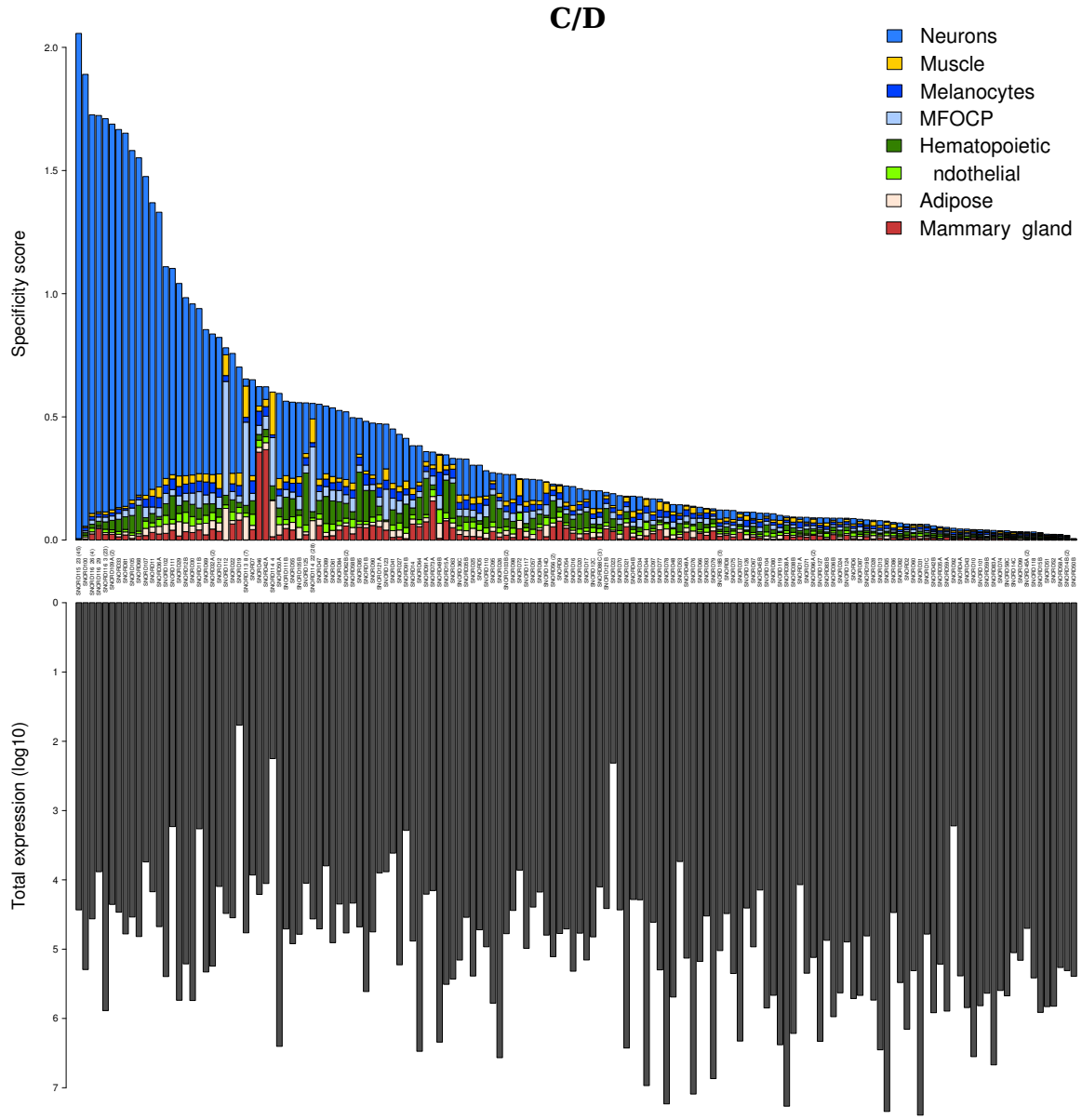
rithms. These algorithms usually have high false positive rate, hence the need for curation and experimental validation. In our contribution, small RNA sequencing data obtained by the ENCODE consortium from different cell types and tissues were used to identify novel snoRNAs and subsequently curating the coordinates of currently annotated snoRNAs as well as of novel ones. In this study the previous observation (based on PAR-CLIP-data) that C/D box mature snoRNAs undergo precise processing pattern was replicated using ENCODE data. Expression profiling of snoRNAs across a range of different tissues led to finding sample specific snoRNAs and also snoRNAs whose expression is dysregulated in cancer. snoRNA expression also exhibits apparent separation between normal and malignant cell types which emphasizes the potential role of snoRNAs as novel cancer biomarkers. We further investigated the expression pattern of snoRNA-derived fragment and found no evidence of tissue specificity in their processing across different cell types.But the functional role of this fragments compared to long form snoRNA remains to be investigated. As in this study the distinction between snoRNA expression profiles across different tissues (especially in neurons) was observed, it propounds the question that what would be the role of this snoRNAs in developmental stages and differentiation. This is an interesting question which remains to be answered in future studies. Identification of the targets of orphan snoRNAs as well as the novel ones is also a challenge which must be elucidated in future. In summary this work can serve as a reference resource for future research in snoRNA and cancer studies.

Appendices

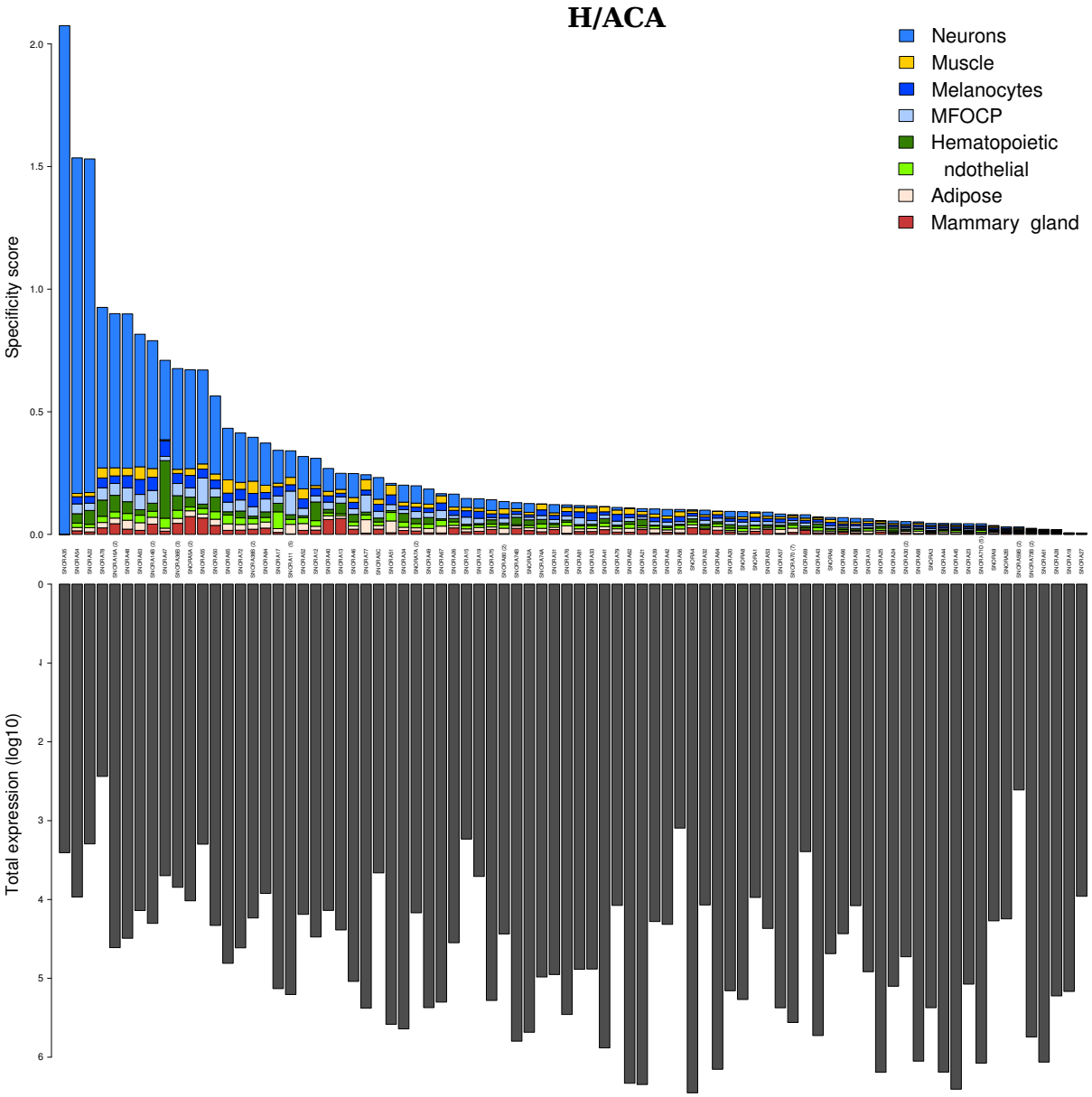
A Supplementary material of Chapter 4



Supplementary Figure S1. (**A**) Distribution of the distance of most frequent processing site to C and D boxes for C/D box snoRNAs. (**B**) 2D histogram of mutual distances of processing sites to C and D boxes

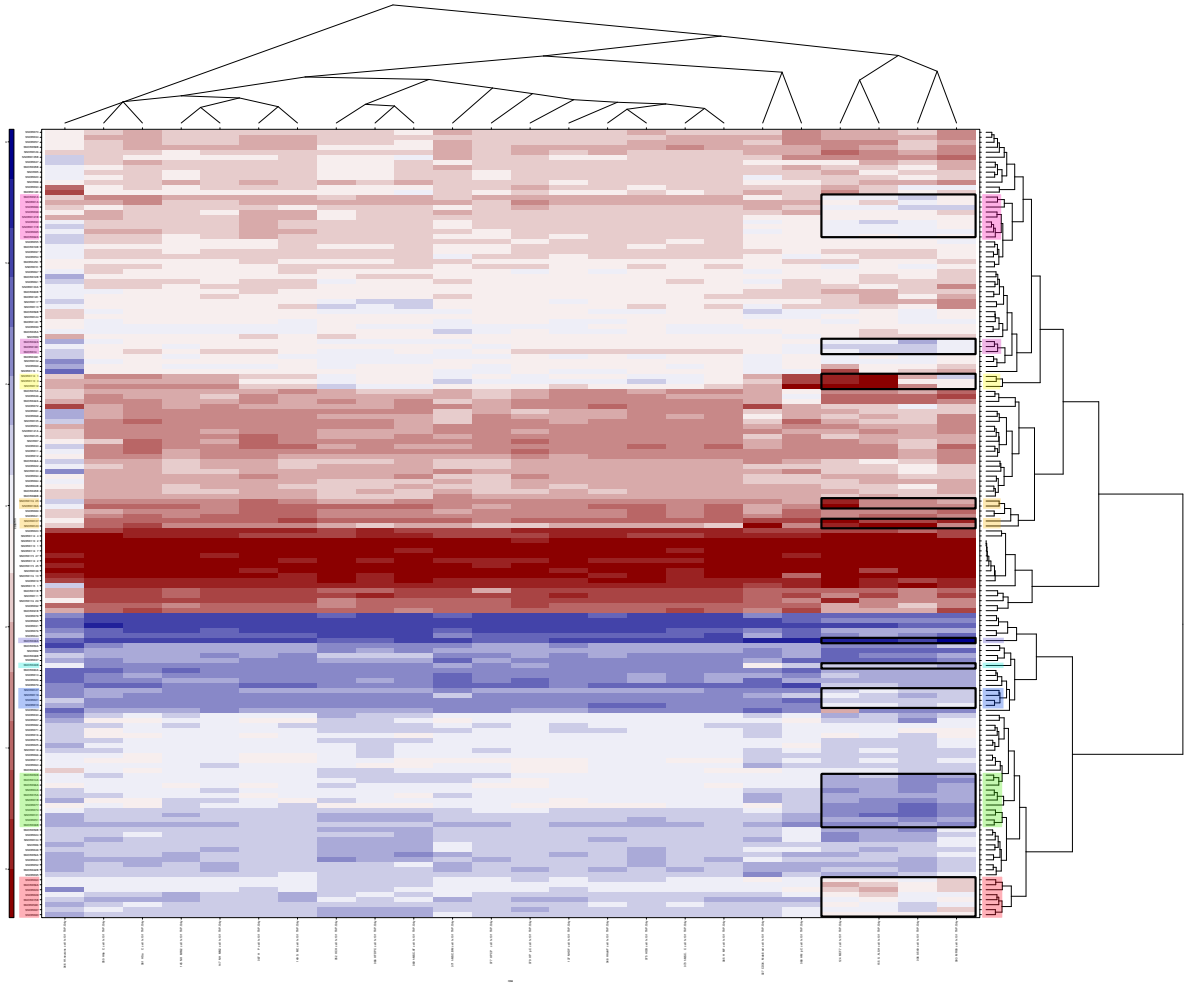


Supplementary Figure S2 (A). Barplot of specificity score of C/D box snoRNAs expression along with the total expression values across samples

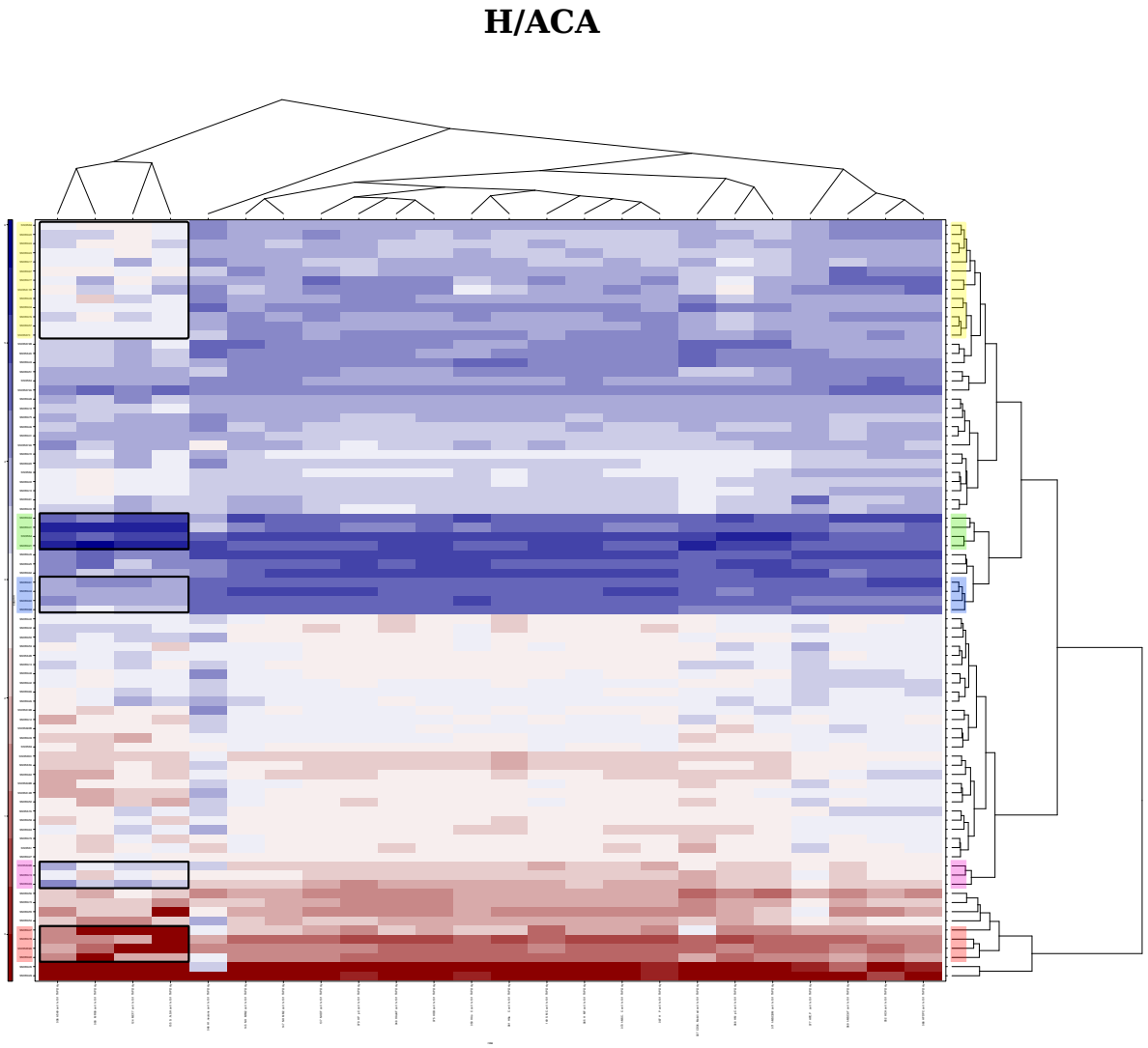


Supplementary Figure S2 (B). Barplot of specificity score of H/ACA box snoRNAs expression along with the total expression values across samples

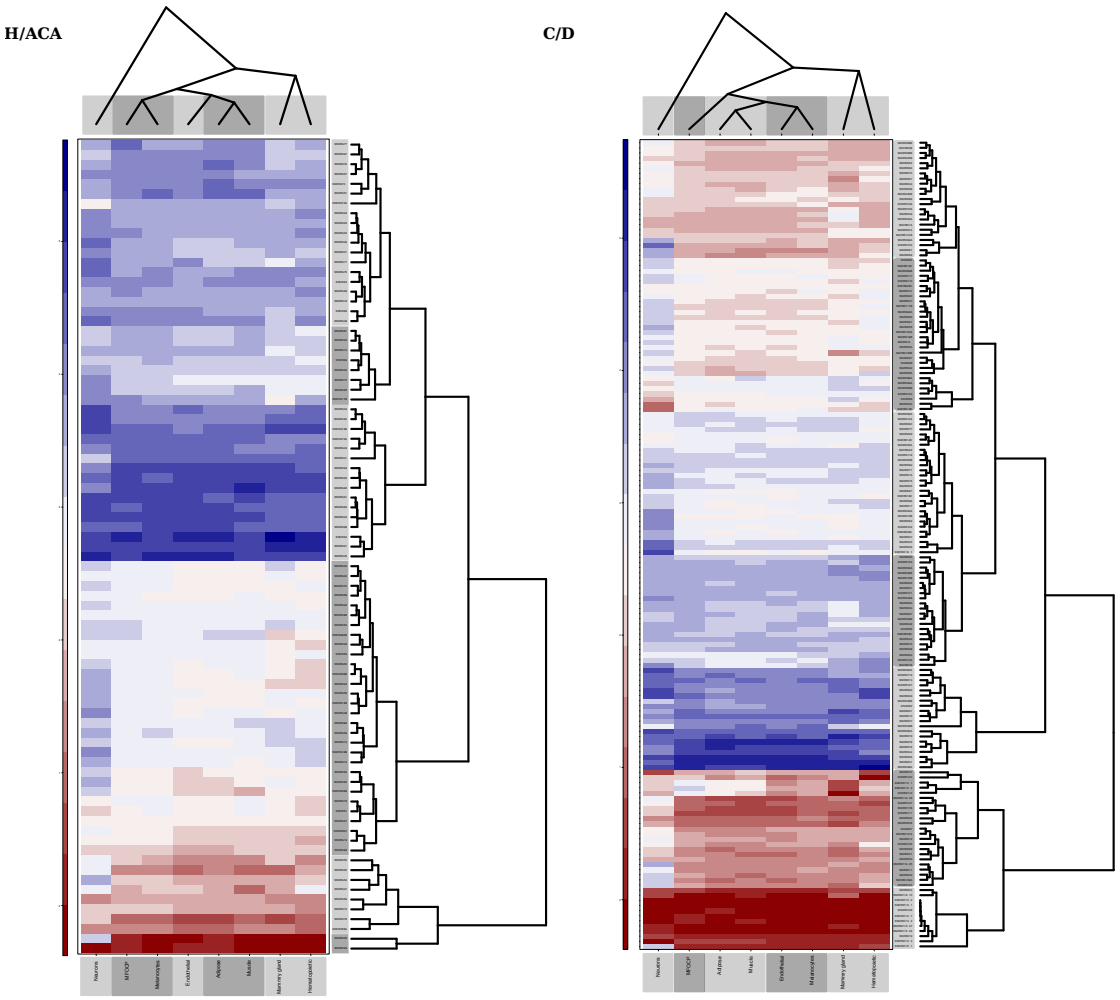
C/D



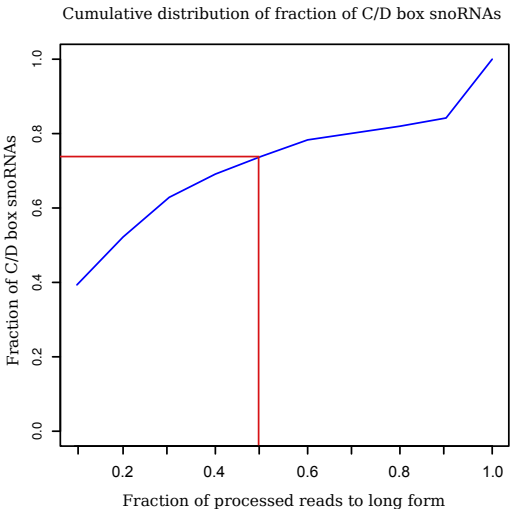
Supplementary Figure S3 (A). Hierarchical clustering of a subset of sRNA-seq samples that have been generated from decapped (tobacco acid phosphatase (TAP)-treated) RNAs isolated from whole cells. The snoRNA clusters which are dysregulated in cancer are highlited



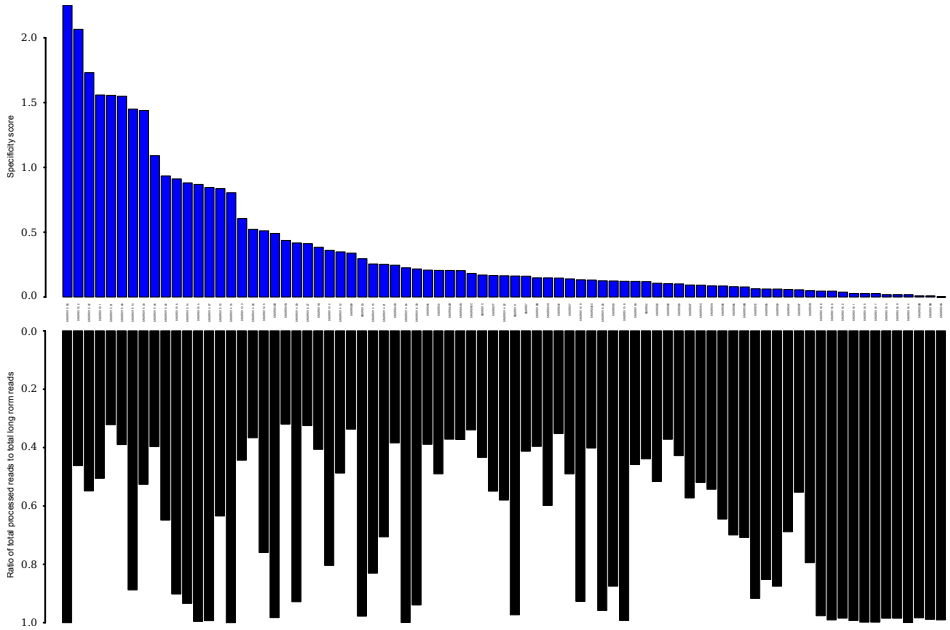
Supplementary Figure S3 (B). Hierarchical clustering of a subset of sRNA-seq samples that have been generated from decapped (tobacco acid phosphatase (TAP)-treated) RNAs isolated from whole cells. The snoRNA clusters which are dysregulated in cancer are highlited



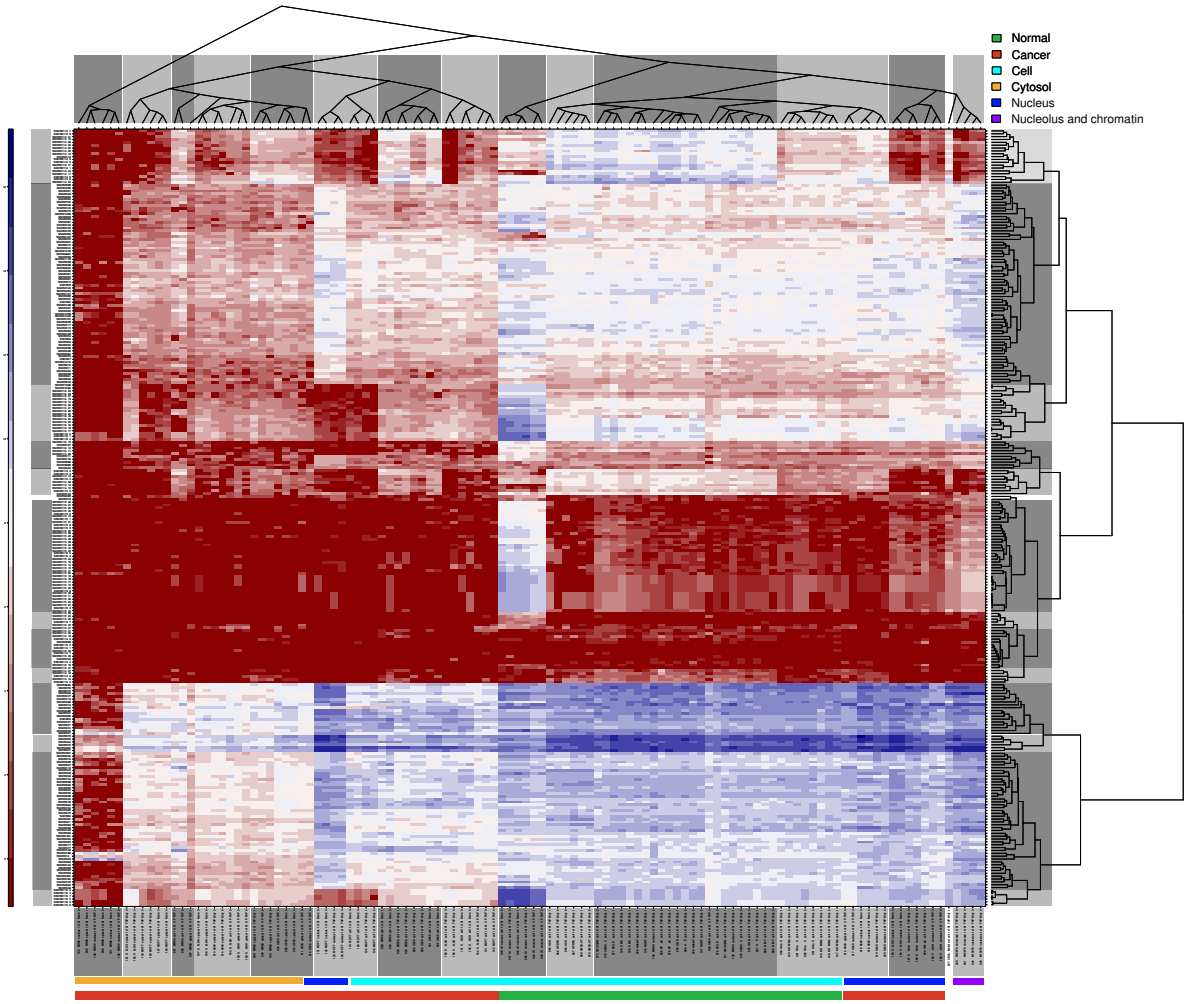
Supplementary Figure S4. Hierarchical clustering of snoRNA expression profiles based on tissue types (excluding cancerous cell types). MFOCP stands for melanocytes, fibroblasts, osteoblasts, chondrocytes and placental tissue.



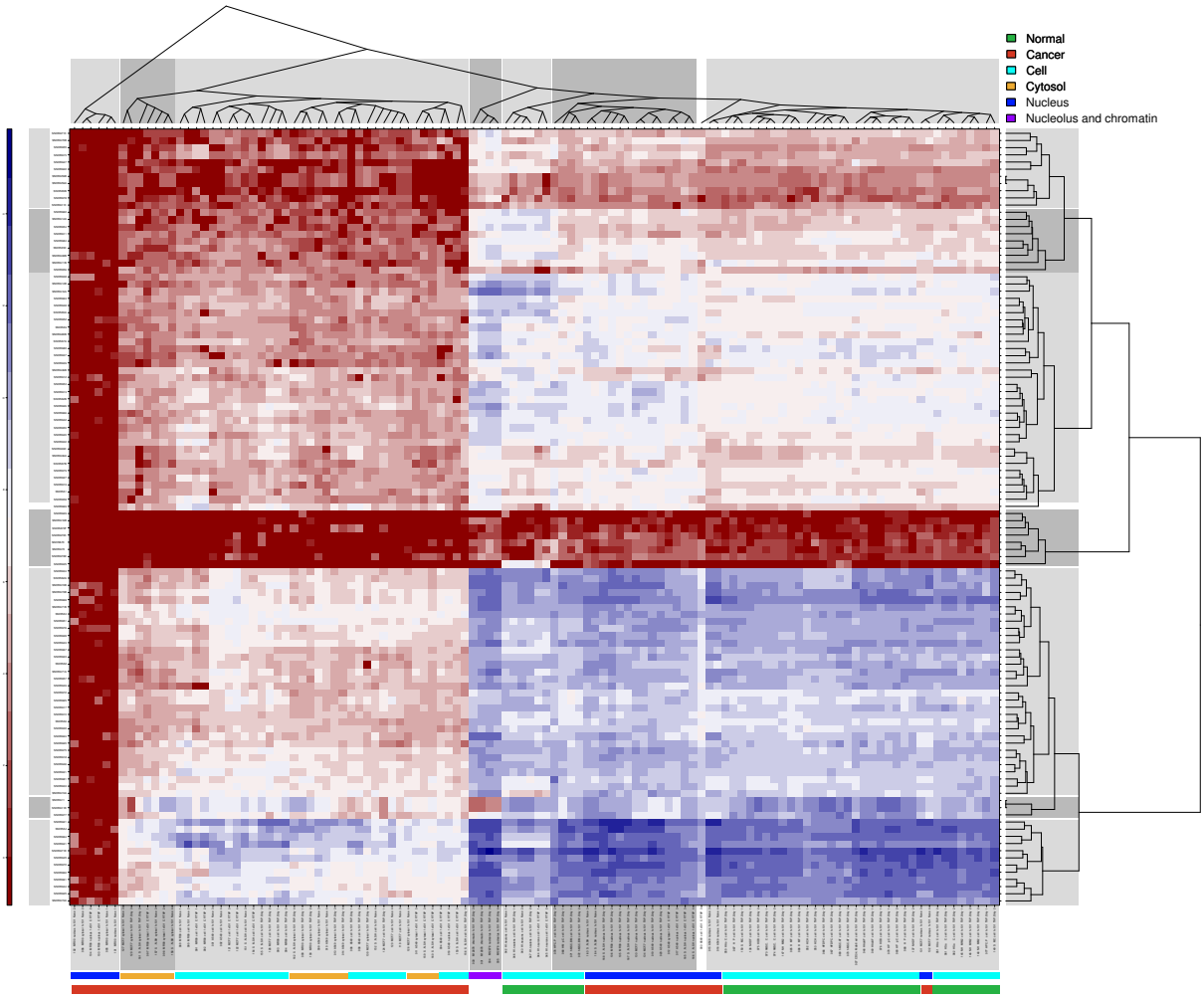
Supplementary Figure S5. Cumulative distribution of fraction of C/D box snoRNAs at different processing ratios (total ratio of processed reads to the reads which cover the majority of snoRNA gene) .



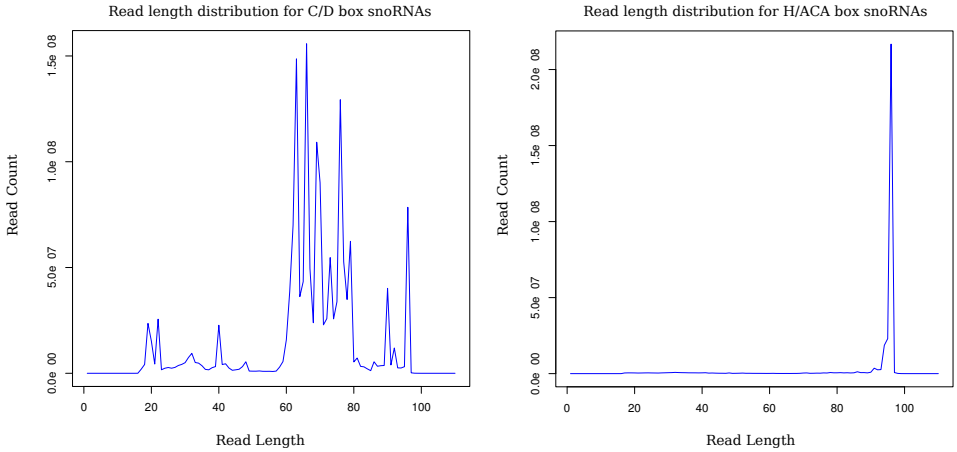
Supplementary Figure S6. Barplot of specificity score for the processing ratio of snoRNAs (ratio of processed to long form) across samples along with the ratio of total processed reads to total long from reads for each snoRNA.



Supplementary Figure S7 (A). Hierarchical clustering of all sRNA-seq samples that have been used in this study based on their C/D box snoRNAs expression profiles. Separation of normal and malignant cell types as well as different compartments of the cell and different tissues can be observed specially for C/D box class of snoRNAs



Supplementary Figure S7 (B). Hierarchical clustering of all sRNA-seq samples that have been used in this study based on their H/ACA box snoRNAs expression profiles.



Supplementary Figure S8. Read length distribution for the reads which mapped to snoRNAs loci

Bibliography

[1] C Alexander Valencia, M Ali Pervaiz, Ammar Husami, Yaping Qian, and Kejian Zhang. Application of Next-Generation-Sequencing to the diagnosis of genetic disorders: A brief overview. In *Next Generation Sequencing Technologies in Medical Genetics*, SpringerBriefs in Genetics, pages 35–43. Springer New York, 1 January 2013.

[2] C Alexander Valencia, M Ali Pervaiz, Ammar Husami, Yaping Qian, and Kejian Zhang. A survey of Next-Generation-Sequencing technologies. In *Next Generation Sequencing Technologies in Medical Genetics*, SpringerBriefs in Genetics, pages 13–24. Springer New York, 1 January 2013.

[3] Marina Alexandersson, Simon Cawley, and Lior Pachter. SLAM: cross-species gene finding and alignment with a generalized pair hidden markov model. *Genome Res.*, 13(3):496–502, March 2003.

[4] 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, 5 October 1990.

[5] Fabian Amman, Michael T Wolfinger, Ronny Lorenz, Ivo L Hofacker, Peter F Stadler, and Sven Findeiß. TSSAR: TSS annotation regime for dRNA-seq data. *BMC Bioinformatics*, 15:89, 27 March 2014.

[6] Wilhelm J Ansorge. Next-generation DNA sequencing techniques. *N. Biotechnol.*, 25(4):195–203, April 2009.

[7] L Argaman, R Hershberg, J Vogel, G Bejerano, E G Wagner, H Margalit, and S Altuvia. Novel small RNA-encoding genes in the intergenic regions of escherichia coli. *Curr. Biol.*, 11(12):941–950, 26 June 2001.

[8] Manuel Ascano, Markus Hafner, Pavol Cekan, Stefanie Gerstberger, and Thomas Tuschl. Identification of RNA-protein interaction networks using PAR-CLIP. *Wiley Interdiscip. Rev. RNA*, 3(2):159–177, March 2012.

[9] El Mustapha Bahassi and Peter J Stambrook. Next-generation sequencing technologies: breaking the sound barrier of human genetics. *Mutagenesis*, 29(5):303–310, September 2014.

[10] P Baldi, Y Chauvin, T Hunkapiller, and M A McClure. Hidden markov models of biological primary sequence information. *Proc. Natl. Acad. Sci. U. S. A.*, 91(3):1059–1063, 1 February 1994.

[11] S Batzoglou, L Pachter, J P Mesirov, B Berger, and E S Lander. Human and mouse gene structure: comparative analysis and application to exon prediction. *Genome Res.*, 10(7):950–958, July 2000.

[12] B A Bensing, B J Meyer, and G M Dunny. Sensitive detection of bacterial transcription initiation sites and differentiation from RNA processing sites in the pheromone-induced plasmid transfer system of enterococcus faecalis. *Proc. Natl. Acad. Sci. U. S. A.*, 93(15):7794–7799, 23 July 1996.

[13] Eva C Berglund, Anna Kiialainen, and Ann-Christine Syv̀anen. Next-generation sequencing technologies and applications for human genetic history and forensics. *Investig. Genet.*, 2:23, 24 November 2011.

[14] A J Berk and P A Sharp. Sizing and mapping of early adenovirus mRNAs by gel electrophoresis of S1 endonuclease-digested hybrids. *Cell*, 12(3):721–732, November 1977.

[15] C Burge and S Karlin. Prediction of complete gene structures in human genomic DNA. *J. Mol. Biol.*, 268(1):78–94, 25 April 1997.

[16] Konstantin Byrgazov, Oliver Vesper, and Isabella Moll. Ribosome heterogeneity: another level of complexity in bacterial translation regulation. *Curr. Opin. Microbiol.*, 16(2):133–139, April 2013.

[17] Jérôme Cavaillé, Hervé Seitz, Martina Paulsen, Anne C Ferguson-Smith, and Jean-Pierre Bachellerie. Identification of tandemly-repeated C/D snoRNA genes at the imprinted human 14q32 domain reminiscent of those at the Prader-Willi/Angelman syndrome region. *Hum. Mol. Genet.*, 11(13):1527–1538, 15 June 2002.

[18] Thomas R Cech and Joan A Steitz. The noncoding RNA revolution-trashing old rules to forge new ones. *Cell*, 157(1):77–94, 27 March 2014.

[19] H Chen, M Bjerknes, R Kumar, and E Jay. Determination of the optimal aligned spacing between the Shine-Dalgarno sequence and the translation initiation codon of escherichia coli mRNAs. *Nucleic Acids Res.*, 22(23):4953–4957, 25 November 1994.

[20] Wen-Dan Chen and Xiao-Feng Zhu. Small nucleolar RNAs (snoRNAs) as potential non-invasive biomarkers for early cancer detection. *Chin. J. Cancer*, 32(2):99–101, February 2013.

[21] L D Chong, L B Ray, and N R Gough. Coding and noncoding RNA: An expanding RNA world. *Sci. Signal.*, 2002(133), 2002.

[22] G A Churchill. Stochastic models for heterogeneous DNA sequences. *Bull. Math. Biol.*, 51(1):79–94, 1989.

[23] Nathan L Clement, Quinn Snell, Mark J Clement, Peter C Hollenhorst, Jahnvi Purwar, Barbara J Graves, Bradley R Cairns, and W Evan Johnson. The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. *Bioinformatics*, 26(1):38–45, 1 January 2010.

[24] C Condon. *Molecular Biology of RNA Processing and Decay in Prokaryotes*. PMBTS/Progress in Molecular Biology and Translational Science Series. Elsevier Science, 2009.

[25] Teresa Cortes, Olga T Schubert, Graham Rose, Kristine B Arnvig, Iñaki Comas, Ruedi Aebersold, and Douglas B Young. Genome-wide mapping of transcriptional start sites defines an extensive leaderless transcriptome in mycobacterium tuberculosis. *Cell Rep.*, 5(4):1121–1131, 27 November 2013.

[26] Nicholas J Croucher and Nicholas R Thomson. Studying bacterial transcriptomes using RNA-seq. *Curr. Opin. Microbiol.*, 13(5):619–624, October 2010.

[27] Robert B Darnell. HITS-CLIP: panoramic views of protein-RNA regulation in living cells. *Wiley Interdiscip. Rev. RNA*, 1(2):266–286, September 2010.

[28] Xavier Darzacq, Beáta E Jády, Céline Verheggen, Arnold M Kiss, Edouard Bertrand, and Tamás Kiss. Cajal body-specific small nuclear RNAs: a novel class of 2'-o-methylation and pseudouridylation guide RNAs. *EMBO J.*, 21(11):2746–2756, 3 June 2002.

[29] G David Forney, Jr. The viterbi algorithm: A personal history. 6 April 2005.

[30] Michiel de Hoon and Yoshihide Hayashizaki. Deep cap analysis gene expression (CAGE): genome-wide identification of promoters, quantification of their expression, and network inference. *Biotechniques*, 44(5):627–8, 630, 632, April 2008.

[31] A P Dempster, N M Laird, and D B Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. Series B Stat. Methodol.*, 39(1):1–38, 1 January 1977.

[32] Adnan Derti, Philip Garrett-Engele, Kenzie D Macisaac, Richard C Stevens, Shreedharan Sriram, Ronghua Chen, Carol A Rohl, Jason M Johnson, and Tomas Babak. A quantitative atlas of polyadenylation in five mammals. *Genome Res.*, 22(6):1173–1183, June 2012.

[33] Julia M Di Bella, Yige Bao, Gregory B Gloor, Jeremy P Burton, and Gregor Reid. High throughput sequencing methods and analysis for microbiome research. *J. Microbiol. Methods*, 95(3):401–414, December 2013.

[34] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. STAR: ultrafast universal RNA-seq aligner. *Bioinformatics*, 29(1):15–21, 1 January 2013.

[35] Gaurav Dugar, Alexander Herbig, Konrad U F̀orstner, Nadja Heidrich, Richard Reinhardt, Kay Nieselt, and Cynthia M Sharma. High-resolution transcriptome maps reveal strain-specific regulatory features of multiple campylobacter jejuni isolates. *PLoS Genet.*, 9(5):e1003495, May 2013.

[36] R Durbin. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, 1998.

[37] S R Eddy. Profile hidden markov models. *Bioinformatics*, 14(9):755–763, 1998.

[38] S R Eddy, G Mitchison, and R Durbin. Maximum discrimination hidden markov models of sequence consensus. *J. Comput. Biol.*, 2(1):9–23, 1995.

[39] Sean R Eddy. Non-coding RNA genes and the modern RNA world. *Nat. Rev. Genet.*, 2(12):919–929, 1 December 2001.

[40] C Edeki. Comparative study of microarray and next generation sequencing technologies. *IJCSMC*, 2012.

[41] Ashley N Egan, Jessica Schlueter, and David M Spooner. Applications of next-generation sequencing in plant biology. *Am. J. Bot.*, 99(2):175–185, February 2012.

[42] Sara El-Metwally, Osama M Ouda, and Mohamed Helmy. Next-Generation sequencing platforms. In *Next Generation Sequencing Technologies and Challenges in Sequence Assembly*, SpringerBriefs in Systems Biology, pages 37–44. Springer New York, 1 January 2014.

[43] S M Elbashir, J Harborth, W Lendeckel, A Yalcin, K Weber, and T Tuschl. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature*, 411(6836):494–498, 24 May 2001.

[44] ENCODE Project Consortium. The ENCODE (ENCyclopedia of DNA elements) project. *Science*, 306(5696):636–640, 22 October 2004.

[45] Mohammad Ali Faghihi, Farzaneh Modarresi, Ahmad M Khalil, Douglas E Wood, Barbara G Sahagan, Todd E Morgan, Caleb E Finch, Georges St Laurent, 3rd, Paul J Kenny, and Claes Wahlestedt. Expression of a noncoding RNA is elevated in alzheimer's disease and drives rapid feed-forward regulation of beta-secretase. *Nat. Med.*, 14(7):723–730, July 2008.

[46] Gregory G Faust and Ira M Hall. GEM: crystal-clear DNA alignment. *Nat. Methods*, 9(12):1159–1161, December 2012.

[47] Paul Flicek and Ewan Birney. Sense from sequence reads: methods for alignment and assembly. *Nat. Methods*, 6(11 Suppl):S6–S12, November 2009.

[48] Nuno A Fonseca, Johan Rung, Alvis Brazma, and John C Marioni. Tools for mapping high-throughput sequencing data. *Bioinformatics*, 28(24):3169–3177, 15 December 2012.

[49] Zoubin Ghahramani. AN INTRODUCTION TO HIDDEN MARKOV MODELS AND BAYESIAN NETWORKS. *Int. J. Pattern Recognit Artif Intell.*, 15(01):9–42, 2001.

[50] Ayman Grada and Kate Weinbrecht. Next-generation sequencing: methodology and application. *J. Invest. Dermatol.*, 133(8):e11, August 2013.

[51] Sam Griffiths-Jones. Annotating noncoding RNA genes. *Annu. Rev. Genomics Hum. Genet.*, 8:279–298, 2007.

[52] Sam Griffiths-Jones, Alex Bateman, Mhairi Marshall, Ajay Khanna, and Sean R Eddy. Rfam: an RNA family database. *Nucleic Acids Res.*, 31(1):439–441, 1 January 2003.

[53] Sam Griffiths-Jones, Simon Moxon, Mhairi Marshall, Ajay Khanna, Sean R Eddy, and Alex Bateman. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.*, 33(Database issue):D121–4, 1 January 2005.

[54] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. PAR-CliP–a method to identify transcriptome-wide the binding sites of RNA binding proteins. *J. Vis. Exp.*, (41), 2 July 2010.

[55] Markus Hafner, Markus Landthaler, Lukas Burger, Mohsen Khorshid, Jean Hausser, Philipp Berninger, Andrea Rothballer, Manuel Ascano, Jr., Anna-Carina Jungkamp, Mathias Munschauer, Alexander Ulrich, Greg S Wardle, Scott Dewell, Mihaela Zavolan, and Thomas Tuschl. Transcriptome-wide identification of RNA-Binding protein and MicroRNA target sites by PAR-CLIP. *Cell*, 141(1):129–141, 2 April 2010.

[56] Markus Hafner, Steve Lianoglou, Thomas Tuschl, and Doron Betel. Genome-wide identification of miRNA targets by PAR-CLIP. *Methods*, 58(2):94–105, October 2012.

[57] Shanil P Haugen, Wilma Ross, and Richard L Gourse. Advances in bacterial promoter recognition and its control by factors that do not bind DNA. *Nat. Rev. Microbiol.*, 6(7):507–519, July 2008.

[58] Y Hayashizaki. Cap analysis gene expression (CAGE). In *Cap-Analysis Gene Expression (CAGE)*, chapter 1, pages 1–5.

[59] Lin He and Gregory J Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, 5(7):522–531, July 2004.

[60] Aleksandra Helwak, Grzegorz Kudla, Tatiana Dudnakova, and David Tollervey. Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, 153(3):654–665, 25 April 2013.

[61] Aleksandra Helwak and David Tollervey. Mapping the miRNA interactome by cross-linking ligation and sequencing of hybrids (CLASH). *Nat. Protoc.*, 9(3):711–728, March 2014.

[62] Alexander Herbig, Cynthia Sharma, and Kay Nieselt. Automated transcription start site prediction for comparative transcriptomics using the SuperGenome. *EMBnet j.*, 19(A):19, 8 April 2013.

[63] I Hofacker and P F Stadler. RNAz 2.0: improved noncoding RNA detection. *Pac. Symp. Biocomput.*, 2010.

[64] I L Hofacker, B Priwitzer, and P F Stadler. Prediction of locally stable RNA secondary structures for genome-wide surveys. *Bioinformatics*, 20(2):186–190, 22 January 2004.

[65] Steve Hoffmann, Christian Otto, Stefan Kurtz, Cynthia M Sharma, Philipp Khaitovich, J̀org Vogel, Peter F Stadler, and J̀org Hackerm ́uller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comput. Biol.*, 5(9):e1000502, September 2009.

[66] N Homer. Bfast: Blat-like fast accurate search tool. 2009.

[67] Zhi-Wei Hou, Yun Wang, Hong Gao, and Sheng-Wei Hou. The principle of dRNA-seq and its applications in prokaryotic transcriptome analyses. *Hereditas*, 35(8):983–991, 30 September 2013.

[68] Ina Huppertz, Jan Attig, Andrea D'Ambrogio, Laura E Easton, Christopher R Sibley, Yoichiro Sugimoto, Mojca Tajnik, Julian K̀onig, and Jernej Ule. iCLIP: Protein-RNA interactions at nucleotide resolution. *Methods*, 65(3):274–287, February 2014.

[69] Alain Jacquier. The complex eukaryotic transcriptome: unexpected pervasive transcription and novel small RNAs. *Nat. Rev. Genet.*, 10(12):833–844, December 2009.

[70] E T Jaynes and G L Bretthorst. *Probability Theory: The Logic of Science*. Cambridge University Press, 2003.

[71] Kirk B Jensen and Robert B Darnell. CLIP: crosslinking and immunoprecipitation of in vivo RNA targets of RNA-binding proteins. *Methods Mol. Biol.*, 488:85–98, 2008.

[72] Hadi Jorjani and Mihaela Zavolan. TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. *Bioinformatics*, 30(7):971–974, 1 April 2014.

[73] Raja Jothi, Suresh Cuddapah, Artem Barski, Kairong Cui, and Keji Zhao. Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res.*, 36(16):5221–5231, 1 September 2008.

[74] Beena Kadakkuzha. Role of noncoding RNAs in diseases. *RNA & DISEASE*, 1(1), 21 October 2014.

[75] K Karplus, C Barrett, M Cline, M Diekhans, L Grate, and R Hughey. Predicting protein structure using only sequence information. *Proteins*, Suppl 3:121–125, 1999.

[76] K Karplus, K Sj̀olander, C Barrett, M Cline, D Haussler, R Hughey, L Holm, and C Sander. Predicting protein structure using hidden markov models. *Proteins*, Suppl 1:134–139, 1997.

[77] W James Kent. BLAT-The BLAST-Like alignment tool. *Genome Res.*, 12(4):656–664, 1 April 2002.

[78] Mohsen Khorshid, Jean Hausser, Mihaela Zavolan, and Erik van Nimwegen. A biophysical miRNA-mRNA interaction model infers canonical and noncanonical targets. *Nat. Methods*, 10(3):253–255, March 2013.

[79] Eun-Deok Kim and Sibum Sung. Long noncoding RNA: unveiling hidden layer of gene regulatory networks. *Trends Plant Sci.*, 17(1):16–21, January 2012.

[80] Shivendra Kishore, Andreas R Gruber, Dominik J Jedlinski, Afzal P Syed, Hadi Jorjani, and Mihaela Zavolan. Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol.*, 14(5):R45, 9 September 2013.

[81] T Kiss. Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs. *EMBO J.*, 20(14):3617–3622, 16 July 2001.

[82] Jan O Korbel, Alexander Eckehart Urban, Jason P Affourtit, Brian Godwin, Fabian Grubert, Jan Fredrik Simons, Philip M Kim, Dean Palejev, Nicholas J Carriero, Lei Du, Bruce E Taillon, Zhoutao Chen, Andrea Tanzer, A C Eugenia Saunders, Jianxiang Chi, Fengtang Yang, Nigel P Carter, Matthew E Hurles, Sherman M Weissman, Timothy T Harkins, Mark B Gerstein, Michael Egholm, and Michael Snyder. Paired-end mapping reveals extensive structural variation in the human genome. *Science*, 318(5849):420–426, 19 October 2007.

[83] I Korf, P Flicek, D Duan, and M R Brent. Integrating genomic homology into gene structure prediction. *Bioinformatics*, 17 Suppl 1:S140–8, 2001.

[84] J Korlach and P Biosciences. Understanding accuracy in SMRT® sequencing. *hpccisj.pacb.com*.

[85] Carsten Kr̀oger, Shane C Dillon, Andrew D S Cameron, Kai Papenfort, Sathesh K Sivasankaran, Karsten Hokamp, Yanjie Chao, Alexandra Sittka, Magali Hébrard, Kristian H́andler, Aoife Colgan, Pimlapas Leekitcharoenphon, Gemma C Langridge, Amanda J Lohan, Brendan Loftus, Sacha Lucchini, David W Ussery, Charles J Dorman, Nicholas R Thomson, J̀org Vogel, and Jay C D Hinton. The transcriptional landscape and small RNAs of salmonella enterica serovar typhimurium. *Proc. Natl. Acad. Sci. U. S. A.*, 109(20):E1277–86, 15 May 2012.

[86] Anders Krogh, Michael Brown, I Saira Mian, Kimmen Sj̀olander, and David Haussler. Hidden markov models in computational biology: Applications to protein modeling. *J. Mol. Biol.*, 235(5):1501–1531, 3 February 1994.

[87] Chee-Seng Ku, Yudi Pawitan, Mengchu Wu, Dimitrios H Roukos, and David N Cooper. The evolution of High-Throughput sequencing technologies: From sanger to Single-Molecule sequencing. In *Next Generation Sequencing in Cancer Research*, pages 1–30. Springer New York, 1 January 2013.

[88] Junpei Kurosawa, Hiromi Nishiyori, and Yoshihide Hayashizaki. Deep cap analysis of gene expression. *Methods Mol. Biol.*, 687:147–163, 2011.

[89] 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 Bl̀ocker, 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 Peterson, 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, 15 February 2001.

[90] Ben Langmead and Steven L Salzberg. Fast gapped-read alignment with bowtie 2. *Nat. Methods*, 9(4):357–359, April 2012.

[91] 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, 4 March 2009.

[92] Christian Ledergerber and Christophe Dessimoz. Base-calling for next-generation sequencing platforms. *Brief. Bioinform.*, 12(5):489–497, September 2011.

[93] H Li, J Ruan, and R Durbin. Maq: mapping and assembly with qualities. *Version 0.6*, 2008.

[94] Heng Li and Nils Homer. A survey of sequence alignment algorithms for next-generation sequencing. *Brief. Bioinform.*, 11(5):473–483, September 2010.

[95] Pei-Fei Li, Sheng-Can Chen, Tian Xia, Xiao-Ming Jiang, Yong-Fu Shao, Bing-Xiu Xiao, and Jun-Ming Guo. Non-coding RNAs and gastric cancer. *World J. Gastroenterol.*, 20(18):5411–5419, 14 May 2014.

[96] Ruiqiang Li, Chang Yu, Yingrui Li, Tak-Wah Lam, Siu-Ming Yiu, Karsten Kristiansen, and Jun Wang. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics*, 25(15):1966–1967, 1 August 2009.

[97] Yang Liao, Gordon K Smyth, and Wei Shi. The subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Res.*, 41(10):e108, 1 May 2013.

[98] Donny D Licatalosi, Aldo Mele, John J Fak, Jernej Ule, Melis Kayikci, Sung Wook Chi, Tyson A Clark, Anthony C Schweitzer, John E Blume, Xuning Wang, Jennifer C Darnell, and Robert B Darnell. HITS-CLIP yields genome-wide insights into brain alternative RNA processing. *Nature*, 456(7221):464–469, 27 November 2008.

[99] Hao Lin, Zefeng Zhang, Michael Q Zhang, Bin Ma, and Ming Li. ZOOM! zillions of oligos mapped. *Bioinformatics*, 24(21):2431–2437, 1 November 2008.

[100] Chi-Man Liu, Thomas Wong, Edward Wu, Ruibang Luo, Siu-Ming Yiu, Yingrui Li, Bingqiang Wang, Chang Yu, Xiaowen Chu, Kaiyong Zhao, Ruiqiang Li, and Tak-Wah Lam. SOAP3: ultra-fast GPU-based parallel alignment tool for short reads. *Bioinformatics*, 28(6):878–879, 15 March 2012.

[101] Yongchao Liu, Bertil Schmidt, and Douglas L Maskell. CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows-Wheeler transform. *Bioinformatics*, 28(14):1830–1837, 15 July 2012.

[102] Sabine Loewer, Moran N Cabili, Mitchell Guttman, Yuin-Han Loh, Kelly Thomas, In Hyun Park, Manuel Garber, Matthew Curran, Tamer Onder, Suneet Agarwal, Philip D Manos, Sumon Datta, Eric S Lander, Thorsten M Schlaeger, George Q Daley, and John L Rinn. Large intergenic non-coding RNA-RoR modulates reprogramming of human induced pluripotent stem cells. *Nat. Genet.*, 42(12):1113–1117, December 2010.

[103] Ari L̀oytynoja and Michel C Milinkovitch. A hidden markov model for progressive multiple alignment. *Bioinformatics*, 19(12):1505–1513, 12 August 2003.

[104] Jun Lu, Gad Getz, Eric A Miska, Ezequiel Alvarez-Saavedra, Justin Lamb, David Peck, Alejandro Sweet-Cordero, Benjamin L Ebert, Raymond H Mak, Adolfo A Ferrando, James R Downing, Tyler Jacks, H Robert Horvitz, and Todd R Golub. MicroRNA expression profiles classify human cancers. *Nature*, 435(7043):834–838, 9 June 2005.

[105] Jiong Ma, Allan Campbell, and Samuel Karlin. Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures. *J. Bacteriol.*, 184(20):5733–5745, October 2002.

[106] Santiago Marco-Sola, Michael Sammeth, Roderic Guigó, and Paolo Ribeca. The GEM mapper: fast, accurate and versatile alignment by filtration. *Nat. Methods*, 9(12):1185–1188, December 2012.

[107] Elaine R Mardis. Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum. Genet.*, 9:387–402, 2008.

[108] Elaine R Mardis. Next-generation sequencing platforms. *Annu. Rev. Anal. Chem.*, 6:287–303, 2013.

[109] Samuel Marguerat, Brian T Wilhelm, and J́urg B̀ahler. Next-generation sequencing: applications beyond genomes. *Biochem. Soc. Trans.*, 36(5):1091, 1 October 2008.

[110] A Gregory Matera, Rebecca M Terns, and Michael P Terns. Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs. *Nat. Rev. Mol. Cell Biol.*, 8(3):209–220, March 2007.

[111] John S Mattick and Igor V Makunin. Non-coding RNA. *Hum. Mol. Genet.*, 15(suppl 1):R17–R29, 15 April 2006.

[112] Giuseppe Matullo, Cornelia Di Gaetano, and Simonetta Guarrera. Next generation sequencing and rare genetic variants: from human population studies to medical genetics. *Environ. Mol. Mutagen.*, 54(7):518–532, August 2013.

[113] Scott McGinnis and Thomas L Madden. BLAST: at the core of a powerful and diverse set of sequence analysis tools. *Nucleic Acids Res.*, 32(Web Server issue):W20–5, 1 July 2004.

[114] John D McPherson. Clinical application of DNA sequencing: Sanger and Next-Generation platforms. In *Molecular Testing in Cancer*, pages 81–85. Springer New York, 1 January 2014.

[115] M L Metzker. Sequencing technologies-the next generation. *Nat. Rev. Genet.*, 2009.

[116] Riten Mitra, Ryan Gill, Susmita Datta, and Somnath Datta. Statistical analyses of next generation sequencing data: An overview. In *Statistical Analysis of Next Generation Sequencing Data*, Frontiers in Probability and the Statistical Sciences, pages 1–24. Springer International Publishing, 1 January 2014.

[117] Jan Mitschke, Jens Georg, Ingeborg Scholz, Cynthia M Sharma, Dennis Dienst, Jens Bantscheff, Bj̀orn Voss, Claudia Steglich, Annegret Wilde, J́org Vogel, and Wolfgang R Hess. An experimentally anchored map of transcriptional start sites in the model cyanobacterium synechocystis sp. PCC6803. *Proc. Natl. Acad. Sci. U. S. A.*, 108(5):2124–2129, 1 February 2011.

[118] Olena Morozova and Marco A Marra. Applications of next-generation sequencing technologies in functional genomics. *Genomics*, 92(5):255–264, November 2008.

[119] Samuel Myllykangas, Jason Buenrostro, and Hanlee P Ji. Overview of sequencing technology platforms. In *Bioinformatics for High Throughput Sequencing*, pages 11–25. Springer New York, 1 January 2012.

[120] S Ohno. So much“ junk” DNA in our genome. *Brookhaven Symp. Biol.*, 1972.

[121] Peter J Park. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.*, 10(10):669–680, October 2009.

[122] Vicent Pelechano and Lars M Steinmetz. Gene regulation by antisense transcription. *Nat. Rev. Genet.*, 14(12):880–893, December 2013.

[123] Joseph F Petrosino, Sarah Highlander, Ruth Ann Luna, Richard A Gibbs, and James Versalovic. Metagenomic pyrosequencing and microbial identification. *Clin. Chem.*, 55(5):856–866, May 2009.

[124] Michael A Quail, Miriam Smith, Paul Coupland, Thomas D Otto, Simon R Harris, Thomas R Connor, Anna Bertoni, Harold P Swerdlow, and Yong Gu. A tale of three next generation sequencing platforms: comparison of ion torrent, pacific biosciences and illumina MiSeq sequencers. *BMC Genomics*, 13:341, 24 July 2012.

[125] E Rivas and S R Eddy. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8, 10 October 2001.

[126] Jason M Rizzo and Michael J Buck. Key principles and clinical applications of “next-generation” DNA sequencing. *Cancer Prev. Res.*, 5(7):887–900, July 2012.

[127] Richard J Roberts, Mauricio O Carneiro, and Michael C Schatz. The advantages of SMRT sequencing. *Genome Biol.*, 14(7):405, 3 July 2013.

[128] Yu-Hui Rogers and J Craig Venter. Genomics: massively parallel sequencing. *Nature*, 437(7057):326–327, 15 September 2005.

[129] Seong Woon Roh, Guy C J Abell, Kyoung-Ho Kim, Young-Do Nam, and Jin-Woo Bae. Comparing microarrays and next-generation sequencing technologies for microbial ecology research. *Trends Biotechnol.*, 28(6):291–299, June 2010.

[130] Matthew Ruffalo, Thomas LaFramboise, and Mehmet Koyut̀urk. Comparative analysis of algorithms for next-generation sequencing read alignment. *Bioinformatics*, 27(20):2790–2796, 15 October 2011.

[131] Cornelius Schmidtke, Sven Findeiss, Cynthia M Sharma, Juliane Kuhfuss, Steve Hoffmann, J̀org Vogel, Peter F Stadler, and Ulla Bonas. Genome-wide transcriptome analysis of the plant pathogen xanthomonas identifies sRNAs with putative virulence functions. *Nucleic Acids Res.*, 40(5):2020–2031, March 2012.

[132] Dietmar Schreiner, Thi-Minh Nguyen, Giancarlo Russo, Steffen Heber, Andrea Patrignani, Erik Ahrné, and Peter Scheiffele. Targeted combinatorial alternative splicing generates brain Region-Specific repertoires of neurexins. *Neuron*, 1 October 2014.

[133] Stephan C Schuster. Next-generation sequencing transforms today's biology. *Nat. Methods*, 5(1):16–18, January 2008.

[134] Anjali Shah. Chromatin immunoprecipitation sequencing (ChIP-Seq) on the SOLiD™ system. *Nat. Methods*, 6(4), 1 April 2009.

[135] Jing Shang, Fei Zhu, Wanwipa Vongsangnak, Yifei Tang, Wenyu Zhang, and Bairong Shen. Evaluation and comparison of multiple aligners for next-generation sequencing data analysis. *Biomed Res. Int.*, 2014:309650, 23 March 2014.

[136] Cynthia M Sharma, Steve Hoffmann, Fabien Darfeuille, Jérémy Reignier, Sven Findeiss, Alexandra Sittka, Sandrine Chabas, Kristin Reiche, J́org Hackerm̀uller, Richard Reinhardt, Peter F Stadler, and J̀org Vogel. The primary transcriptome of the major human pathogen helicobacter pylori. *Nature*, 464(7286):250–255, 11 March 2010.

[137] Cynthia M Sharma and J̀org Vogel. Differential RNA-seq: the approach behind and the biological insight gained. *Curr. Opin. Microbiol.*, 19:97–105, June 2014.

[138] Mona A Sheikh and Yaniv Erlich. Base-Calling for bioinformaticians. In *Bioinformatics for High Throughput Sequencing*, pages 67–83. Springer New York, 1 January 2012.

[139] Jay Shendure and Hanlee Ji. Next-generation DNA sequencing. *Nat. Biotechnol.*, 26(10):1135–1145, October 2008.

[140] Peter J Shepard, Eun-A Choi, Jente Lu, Lisa A Flanagan, Klemens J Hertel, and Yongsheng Shi. Complex and dynamic landscape of RNA polyadenylation revealed by PAS-Seq. *RNA*, 17(4):761–772, April 2011.

[141] Toshiyuki Shiraki, Shinji Kondo, Shintaro Katayama, Kazunori Waki, Takeya Kasukawa, Hideya Kawaji, Rimantas Kodzius, Akira Watahiki, Mari Nakamura, Takahiro Arakawa,

Shiro Fukuda, Daisuke Sasaki, Anna Podhajska, Matthias Harbers, Jun Kawai, Piero Carninci, and Yoshihide Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, 100(26):15776–15781, 23 December 2003.

[142] Navjot Singh and Joseph T Wade. Identification of regulatory RNA in bacterial genomes by genome-scale mapping of transcription start sites. *Methods Mol. Biol.*, 1103:1–10, 2014.

[143] L Smith, L Yeganova, and W J Wilbur. Hidden markov models and optimized sequence alignments. *Comput. Biol. Chem.*, 27(1):77–84, February 2003.

[144] Rotem Sorek and Pascale Cossart. Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat. Rev. Genet.*, 11(1):9–16, January 2010.

[145] Jessica Spitzer, Markus Hafner, Markus Landthaler, Manuel Ascano, Thalia Farazi, Greg Wardle, Jeff Nusbaum, Mohsen Khorshid, Lukas Burger, Mihaela Zavolan, and Thomas Tuschl. PAR-CLIP (photoactivatable Ribonucleoside-Enhanced crosslinking and immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins. *Methods Enzymol.*, 539:113–161, 2014.

[146] C M Stultz, J V White, and T F Smith. Structural analysis based on state-space modeling. *Protein Sci.*, 2(3):305–314, March 1993.

[147] H Su, T Xu, S Ganapathy, M Shadfan, M Long, T H-M Huang, I Thompson, and Z-M Yuan. Elevated snoRNA biogenesis is essential in breast cancer. *Oncogene*, 33(11):1348–1358, 13 March 2014.

[148] The ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, 9(4):e1001046, 19 April 2011.

[149] Carla A Theimer and Juli Feigon. Structure and function of telomerase RNA. *Curr. Opin. Struct. Biol.*, 16(3):307–318, June 2006.

[150] Maureen K Thomason, Thorsten Bischler, Sara K Eisenbart, Konrad U F̀orstner, Aixia Zhang, Alexander Herbig, Kay Nieselt, Cynthia M Sharma, and Gisela Storz. Global transcriptional start site mapping using dRNA-seq reveals novel antisense RNAs in escherichia coli. *J. Bacteriol.*, 29 September 2014.

[151] J A Thompson, M F Radonovich, and N P Salzman. Characterization of the 5'-terminal structure of simian virus 40 early mRNA's. *J. Virol.*, 31(2):437–446, August 1979.

[152] Cole Trapnell, Lior Pachter, and Steven L Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, 1 May 2009.

[153] Andrew V Uzilov, Joshua M Keegan, and David H Mathews. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics*, 7:173, 27 March 2006.

[154] Saba Valadkhan and Lalith S Gunawardane. Role of small nuclear RNAs in eukaryotic gene expression. *Essays Biochem.*, 54:79–90, 2013.

[155] Anton Valouev, David S Johnson, Andreas Sundquist, Catherine Medina, Elizabeth Anton, Serafim Batzoglou, Richard M Myers, and Arend Sidow. Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods*, 5(9):829–834, September 2008.

[156] Arnoud H M van Vliet. Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol. Lett.*, 302(1):1–7, January 2010.

[157] J C Venter, M D Adams, E W Myers, P W Li, R J Mural, G G Sutton, H O Smith, M Yandell, C A Evans, R A Holt, J D Gocayne, P Amanatides, R M Ballew, D H Huson, J R Wortman, Q Zhang, C D Kodira, X H Zheng, L Chen, M Skupski, G Subramanian, P D Thomas, J Zhang, G L Gabor Miklos, C Nelson, S Broder, A G Clark, J Nadeau, V A McKusick, N Zinder, A J Levine, R J Roberts, M Simon, C Slayman, M Hunkapiller, R Bolanos, A Delcher, I Dew, D Fasulo, M Flanigan, L Florea, A Halpern, S Hannenhalli, S Kravitz, S Levy, C Mobarry, K Reinert, K Remington, J Abu-Threideh, E Beasley, K Biddick, V Bonazzi, R Brandon, M Cargill, I Chandramouliswaran, R Charlab, K Chaturvedi, Z Deng, V Di Francesco, P Dunn, K Eilbeck, C Evangelista, A E Gabrielian, W Gan, W Ge, F Gong, Z Gu, P Guan, T J Heiman, M E Higgins, R R Ji, Z Ke, K A Ketchum, Z Lai, Y Lei, Z Li, J Li, Y Liang, X Lin, F Lu, G V Merkulov, N Milshina, H M Moore, A K Naik, V A Narayan, B Neelam, D Nusskern, D B Rusch, S Salzberg, W Shao, B Shue, J Sun, Z Wang, A Wang, X Wang, J Wang, M Wei, R Wides, C Xiao, C Yan, A Yao, J Ye, M Zhan, W Zhang, H Zhang, Q Zhao, L Zheng, F Zhong, W Zhong, S Zhu, S Zhao, D Gilbert, S Baumhueter, G Spier, C Carter, A Cravchik, T Woodage, F Ali, H An, A Awe, D Baldwin, H Baden, M Barnstead, I Barrow, K Beeson, D Busam, A Carver, A Center, M L Cheng, L Curry, S Danaher, L Davenport, R Desilets, S Dietz, K Dodson, L Doup, S Ferriera, N Garg, A Gluecksmann, B Hart, J Haynes, C Haynes, C Heiner, S Hladun, D Hostin, J Houck, T Howland, C Ibegwam, J Johnson, F Kalush, L Kline, S Koduru, A Love, F Mann, D May, S McCawley, T McIntosh, I McMullen, M Moy, L Moy, B Murphy, K Nelson, C Pfannkoch, E Pratts, V Puri, H Qureshi, M Reardon, R Rodriguez, Y H Rogers, D Romblad, B Ruhfel, R Scott, C Sitter, M Smallwood, E Stewart, R Strong, E Suh, R Thomas, N N Tint, S Tse, C Vech, G Wang, J Wetter, S Williams, M Williams, S Windsor, E Winn-Deen, K Wolfe, J Zaveri, K Zaveri, J F Abril, R Guigó, M J Campbell, K V Sjolander, B Karlak, A Kejariwal, H Mi, B Lazareva, T Hatton, A Narechania, K Diemer, A Muruganujan, N Guo, S Sato, V Bafna, S Istrail, R Lippert, R Schwartz, B Walenz, S Yooseph, D Allen, A Basu, J Baxendale, L Blick, M Caminha, J Carnes-Stine, P Caulk, Y H Chiang, M Coyne, C Dahlke, A Mays, M Dombroski, M Donnelly, D Ely, S Esparham, C Fosler, H Gire, S Glanowski, K Glasser, A Glodek, M Gorokhov, K Graham, B Gropman, M Harris, J Heil, S Henderson, J Hoover, D Jennings, C Jordan, J Jordan, J Kasha, L Kagan, C Kraft, A Levitsky, M Lewis, X Liu, J Lopez, D Ma, W Majoros, J McDaniel, S Murphy, M Newman, T Nguyen, N Nguyen, M Nodell, S Pan, J Peck, M Peterson, W Rowe,

R Sanders, J Scott, M Simpson, T Smith, A Sprague, T Stockwell, R Turner, E Venter, M Wang, M Wen, D Wu, M Wu, A Xia, A Zandieh, and X Zhu. The sequence of the human genome. *Science*, 291(5507):1304–1351, 16 February 2001.

[158] Michael-Paul Vockenhuber, Cynthia M Sharma, Michaela G Statt, Denis Schmidt, Zhenjiang Xu, Sascha Dietrich, Heiko Liesegang, David H Mathews, and Beatrix Suess. Deep sequencing-based identification of small non-coding RNAs in streptomyces coelicolor. *RNA Biol.*, 8(3):468–477, May 2011.

[159] Karl V Voelkerding, Shale A Dames, and Jacob D Durtschi. Next-generation sequencing: from basic research to diagnostics. *Clin. Chem.*, 55(4):641–658, April 2009.

[160] J̀org Vogel, Verena Bartels, Thean Hock Tang, Gennady Churakov, Jacoba G Slagter-J́ager, Alexander H̀uttenhofer, and E Gerhart H Wagner. RNomics in escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.*, 31(22):6435–6443, 15 November 2003.

[161] J T Wade. Where to begin? mapping transcription start sites genome-wide in escherichia coli. *J. Bacteriol.*, 20 October 2014.

[162] F Wahid, T Khan, K Hwang, and Y Kim. Piwi-interacting RNAs (piRNAs) in animals: The story so far. *Afr. J. Biotechnol.*, 8(17), 29 November 2010.

[163] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat. Rev. Genet.*, 10(1):57–63, January 2009.

[164] Toshiaki Watanabe and Haifan Lin. Posttranscriptional regulation of gene expression by piwi proteins and piRNAs. *Mol. Cell*, 56(1):18–27, 2 October 2014.

[165] J V White, C M Stultz, and T F Smith. Protein classification by stochastic modeling and optimal filtering of amino-acid sequences. *Math. Biosci.*, 119(1):35–75, January 1994.

[166] Erno Wienholds, Wigard P Kloosterman, Eric Miska, Ezequiel Alvarez-Saavedra, Eugene Berezikov, Ewart de Bruijn, H Robert Horvitz, Sakari Kauppinen, and Ronald H A Plasterk. MicroRNA expression in zebrafish embryonic development. *Science*, 309(5732):310–311, 8 July 2005.

[167] Sebastian Will, Kristin Reiche, Ivo L Hofacker, Peter F Stadler, and Rolf Backofen. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput. Biol.*, 3(4):e65, 13 April 2007.

[168] Ina Wilms, Aaron Overl̀oper, Minou Nowrousian, Cynthia M Sharma, and Franz Narberhaus. Deep sequencing uncovers numerous small RNAs on all four replicons of the plant pathogen agrobacterium tumefaciens. *RNA Biol.*, 9(4):446–457, April 2012.

[169] J E Wilusz. Noncoding RNA. In *Brenner's Encyclopedia of Genetics*, pages 84–86. Elsevier, 2013.

[170] Barbara Wold and Richard M Myers. Sequence census methods for functional genomics. *Nat. Methods*, 5(1):19–21, 19 December 2007.

[171] Thomas D Wu and Serban Nacu. Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics*, 26(7):873–881, 1 April 2010.

[172] Thomas D Wu and Colin K Watanabe. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, 21(9):1859–1875, 1 May 2005.

[173] Omri Wurtzel, Rajat Sapra, Feng Chen, Yiwen Zhu, Blake A Simmons, and Rotem Sorek. A single-base resolution map of an archaeal transcriptome. *Genome Res.*, 20(1):133–141, January 2010.

[174] Chengguo Yao, Lingjie Weng, and Yongsheng Shi. Global protein-RNA interaction mapping at single nucleotide resolution by iCLIP-seq. *Methods Mol. Biol.*, 1126:399–410, 2014.

[175] Petya Zhelyazkova, Cynthia M Sharma, Konrad U F̀orstner, Karsten Liere, J́org Vogel, and Thomas B̀orner. The primary transcriptome of barley chloroplasts: numerous noncoding RNAs and the dominating role of the plastid-encoded RNA polymerase. *Plant Cell*, 24(1):123–136, January 2012.

Hadi Jorjani

Davidsbodenstrasse 42
4056 Basel
Switzerland

Phone: (+41) 7878-32758
Email: hadi.jorjani@unibas.ch
h.jorjani@gmail.com

Personal Information

First Name: Hadi
Last Name: Jorjani
Birth Date: April 16th, 1986
Gender: Male
Marital Status: Married
Nationality: Iranian

Education

January 2011 – December 2014

Ph.D in Bioinformatics

Thesis: Computational analysis of next generation sequencing data: from transcription start sites in bacteria to human non-coding RNA

Supervised by Prof. Mihaela Zavolan
Department of Bioinformatics, Biozentrum, University of Basel, Switzerland

September 2008 – August 2010

M.Sc in Computer Engineering - Algorithms and Computations

Thesis: Transcriptional regulatory network analysis of histone post–translational modifications in computational epigenetics

Supervised by Prof. Ali Moeini
Department of Electrical and Computer Engineering, University of Tehran, Iran

September 2004 – August 2008

B.Sc in Computer Engineering

Thesis: Extraction of learning styles in an Intelligent tutoring system

Supervised by: Dr. Hasan Seydrazi
Department of Electrical and Computer Engineering, University of Tehran, Iran

September 2000 - June 2004

Diploma in Mathematics and Physics
National Organization for Development of Exceptional Talents, Gorgan, Iran

Publications

• Jorjani, H. & Zavolan, M. TSSer: an automated method to identify transcription start sites in prokaryotic genomes from differential RNA sequencing data. *Bioinformatics* **30,** 971–974 (2014).

• Kishore, S. *et al.* Insights into snoRNA biogenesis and processing from PAR-CLIP of snoRNA core proteins and small RNA sequencing. *Genome Biol.* **14,** R45 (2013).

• Jorjani, H. *et al.* An updated human snoRNAome. RNA *Biol.* To be submitted.

Research Interests

• Algorithms design. Machine learning, Bayesian data analysis
• Graph theory, Linear algebra, Combinatorics
• Stochastic modeling, Dynamical systems

Teaching Experience

Computational systems biology
Teaching Assistant
University of Basel, Department of Bioinformatics
Spring 2013

Skills

Programming Languages: C, C++, Java, Python, R, MATLAB

Languages:
Farsi: Native
English: fluent
German: basic

Honors and Awards

• Top student in sub-discipline of Information technology, 2008
• GPA qualified for studying M.Sc. at University of Tehran without entrance exam among all computer engineering students, 2008
• Bronze medal of 21th national mathematics olympiad, young scholars club, Tehran, Iran, 2003
• 9 th Place, National graduate entrance examination of Azad university in artificial intelligence field, 2008
• Top 0.1% of the nationwide university entrance exam, with nearly 500,000 participants, 2004

References

Prof. Dr. Mihaela Zavolan

Department of Bioinformatics, University of Basel
E-mail: mihaela.zavolan@unibas.ch
Tel: +41 61 267 15 77