

tRNomics: Genomic Organization and Processing patterns of tRNAs

Von der Fakultät für Mathematik und Informatik
der Universität Leipzig
angenommene

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM

(Dr. rer. nat.)

im Fachgebiet Informatik

vorgelegt

von M. Sc. Clara Isabel Bermudez Santana

geboren am 11. Juni 1973 in Chiquinquirá, Kolumbien

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Peter F. Stadler (Leipzig, Deutschland)
2. Professor Dr. Eric Westhof (Straßbourg, Frankreich)

Die Verleihung des akademischen Grades erfolgt mit Bestehen der Verteidigung am 13.09.2010 mit dem Gesamtprädikat *magna cum laude*.

To
Tomás and Gustavo
Your love makes me at once keep going through

Acknowledgments

- I am heartily thankful to my supervisor, Peter F. Stadler, whose encouragement, guidance, goodwill and support from the initial to the final level of this thesis enabled me to develop an understanding of the subject. To him for being Peter and helping me to go deeper in many aspects of my life and my scientific career.
- I owe my deepest gratitude to Jens and Petra for supporting me and my family in every aspect of my time in Leipzig.
- I am certain to say I am lucky for having a nice time not only 3 mts around the office 320.2 but also at the third floor of Härtelstrasse 16-18. Thanks to all of you who supported me and my family during these 4 years in Leipzig.
- To the people who found time at the hard starting of this thesis: Maribel, Borut, Markus, Camille, Claudia and Dominic. Thanks specially to Maribel and Borut for supporting me plenty of energy and enthusiasm.
- To Steve and David for our survey in transcriptome analysis and their disposition to work. It was a pleasure to work with them.
- To the people who helped and supported me in every aspect at the end of this project: Steve, David, Toralf, Camille, Jane, Jana, Sonja, Claudia, Hakim and Stephan.
- To Claudia S. Copeland for her friendship and finding a gap for editing the manuscript.
- My lovely Markus for being so patient with “my sorry ... Do you have a little time?” ... Thanks forever.
- To Jan whose encouragement and smiles after his harder days always make me re-think of my life.
- To Gustavo for being husband, mother and father without loosing and losing his own projects on life.
- To Tomás since October 1999 for helping me to feel every day of my life as a new amazing day.
- To Valeria, Maria and the families: Rubiano Ateorthua, Vasquez, Bermudez Santana and Santana Palacios for supporting Gustavo at home.
- To my patient parents Stella and Jaime who let me to go far away one day.
- To Carmen for the time that she believed in our friendship.
- To my Slovenian friends Borut and Bora who were my step-family during this 4 years.
- To my lovely friends Diana (my angel in Leipzig), Cata, Claudia, Maribel (my devil in Leipzig), Gloria and Jaqueline, Johan and Jaime O.

-
- To all our afterwork drinks and food with Gustavo, Maribel, Diana, Steve, Claudia, Jose and Jorge under the spell of Juanes, Celia Cruz, Linda Ronstadt and José José.
 - To the nice time and support at the Leipzig International School. Thanks for the wonderful time shared with the Gregorian, Alpers, Schenkel and Hipps Families.
 - To the Bier-Informatik mothers Bärbel, Petra, Kristin, Claudia, Manja and Jana.
 - For the exotic mixture Swiss-Algerian-Venezuelan-Austrian on winter 2009 that brought me joy.
 - To all our Fall and Winter Seminars and their organizers.
 - To all the people around Ritterstrasse 12.
 - To Prof. Dr. Martin Middendorf for supporting my scholarship extension..
 - To Prof. Dr. Jürgen Jost at the Max Planck Institute for Mathematics in the Sciences (MIS) in Leipzig.
 - Without hesitate to the Universidad Nacional de Colombia and DAAD-Alecol program for the financial support.
 - and to Arpe Caspary for designing Alecol to offer more German-Colombian academic exchange.

Preface

The RNA world hypothesis places RNA at center-stage during the origin of life, and has received support from many authors. In spite of difficulties in need of further examination, studies during the last ten years have demonstrated important aspects of RNA biology that were not previously known. These studies support the idea of a critical role for RNA in cellular function. Many catalytic functions for RNA are known, including translation by ribosomal RNA processing of pre-mRNA by nuclear ribonucleoproteins (snRNPs), RNA editing, and reverse transcription. In Eukaryotes, many other sorts of small RNAs have recently received attention as key components of regulatory systems, as well as main players in the RNA silencing mechanism.

No less important is the discovery of new functions and distributions of different sorts of ncRNAs, such as small nucleolar RNAs (snoRNAs), which were initially associated with a specific cellular compartment and were assumed to function exclusively as target ribosomal RNAs, but which are now being associated with different functions and broader locations. More intriguing have been recent findings indicating that snoRNAs can be processed to yield microRNA-like RNAs, and that there is a plausible connection between RNA silencing and snoRNA-mediated RNA processing systems.

Recent advances in the biology of transfer RNAs (tRNAs) have not only enriched our knowledge about their functions in translation but also posit that these classical non-coding housekeeping RNAs are key components of the small RNA-mediated gene regulation system. As occurred for the understanding of snoRNA cellular location, the knowledge of tRNA cellular location has expanded. Biosynthesis of tRNA was previously thought to occur solely in the nucleus, with tRNA functioning only in the cytoplasm of eukaryotic cells. However, based on recent findings demonstrating that pre-tRNA splicing can occur in the cytoplasm, that aminoacylation is also possible in the nucleus, and that tRNA retrograde travel (from the cytoplasm to the nucleus) is possible, it is clear that tRNA will be discovered to have many unanticipated functions in diverse cellular processes. In the next decade, plenty of surprises are expected, not only with regard to the nuclear-cytoplasmic dynamics of tRNA but also for its importance in the global regulation of RNA silencing. Many respect of tRNA biology is presented on chapter 1, some issues discussed are part of the publication Tanzer, T., Riester, M., Hertel, J., **Bermudez-Santana, C**, Gorodkin, J. Hofacker, I. Stadler. P.F. *Evolutionary Genomics and Systems Biology: Chapter: Evolutionary genomics of microRNAs and their relatives. March 2010, Wiley-Blackwell.*

tRNAs are among the most ancient genes and can be traced back to the putative RNA World. They are ubiquitous in all organisms, but a comparative survey of genomic organization is not available in the literature. Although the diversity of tRNA genes in eukaryotes has been previously reported for 11 eukaryotic genomes, and a comparison of 50 genomes

from all three domains of life reveals domain-specific structural and functional features as well as a suggestive diversity of tRNA function, less is known about their specific configuration. Therefore, in chapter 2, we present a computational survey to gain insight into the genomic locations of tRNAs on a genome wide scale. The main contributions are based on the following publications: *Genomic Organization of Eukaryotic tRNAs*. **Bermudez-Santana C.I.**, Stephan-Otto, C., Kirtsten, T., Engelhardt, J., Prohaska, S., Steigele, S. and Stadler, P.F. 2010, BMC genomics. In press., *Homology-Based Annotation of Non-coding RNAs in the Genomes of Schistosoma mansoni and Schistosoma japonicum* Copeland, C., Marz, M., Dominic, R. D., Hertel, J., Brindley, P., **Bermudez-Santana, C.**, Kehr, S., Stephan, C., Stadler, P.F. BMC Genomics 2009, 10:464. and *Comparative Analysis of Non-Coding RNAs in Nematodes* Tafer, H., Rose, D., Marz, M., Hertel, J., Bartschat, S., Kehr, S., Otto, W., Donath, A., Tanzer, A., **Bermudez-Santana, C.**, Gruber, A., Juhling, F., Engelhardt, J., Busch, A., Hiller, M., Stadler, P. Dieterich, C. 2010. Submitted to Genome Consortium.

Recent findings from transcriptome data analysis regarding the processing of tRNA-derived small RNAs presented the opportunity to undertake a detailed comparison of plausible patterns of tRNAs based on the analysis of deep sequencing libraries. Since recent studies have revealed roles for tRNAs as plausible players in other diverse aspects of cellular biology, we present in chapter 3 a computational survey to identify and classify three main classes of ncRNAs from a human brain library. This chapter is based on the following publications: *Identification and Classification of Small RNAs in Transcriptome Sequence Data*. Langenberger, D., **Bermudez-Santana, C.I.**, Stadler, P.F., Hoffmann, S. Pac Symp Biocomput. 2010:80-7 and *Evidence for Human microRNA-Offset RNAs in Small RNA Sequencing Data*. Langenberger, D., **Bermudez-Santana, C.**, Hertel, J., Khaitovich, P., Hoffmann, S., Stadler, P.F. Bioinformatics. 2009 Sep 15;25(18):2298-301.

Finally, we extended our survey to analyze and classify patterns of small RNA derived from tRNA families. We have developed a new approach based on the classification of tRNA-short-read-block patterns from small RNA libraries from *Bos taurus*, *Caenorhabditis elegans*, *Canis familiaris*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Macaca mulatta*, *Mus musculus* and *Rattus norvegicus*. This study will be presented in the Fourteenth Conference on Research in Computational Molecular Biology. *Searching tRNA processing patterns in transcriptome sequencing data*. **Bermudez-Santana, C.I.**, Langenberger, D., Hoffmann, S. Stadler, P.F. RECOMB. 2010. August, Lisbon, Portugal.

In summary, these chapters include findings of three novel aspects of tRNA biology: genome organization, preliminary transcriptome data analysis, and the classification of a novel class of tRNA-derived small RNAs from transcriptome data. In general, this survey concluded that the genomic organization of tRNA is characterized by complex, lineage-specific patterns with extensive variability that is in striking contrast to the extreme levels of sequence-conservation in the tRNA genes themselves. Our comprehensive analysis of eukaryotic tRNA gene distributions provides a basis for further studies into the interplay of tRNA gene arrangements and genome organization in general. This tRNA processing survey illustrated that patterns are generally conserved across species but that some superfamilies are outliers. The analysis suggests that every tRNA has a specific pattern and thus undergoes a characteristic maturation. The mechanism underlying the processing of these tRNA shreds remains to be clarified, as do any related functional implications.

Abstract

Surprisingly little is known about the organization and distribution of tRNAs and tRNA-related sequences on a genome-wide scale. While tRNA complements are usually reported in passing as part of genome annotation efforts, and peculiar features such as the tandem arrangements of tRNAs in *Entamoeba histolytica* have been described in some detail, comparative studies are rare. We therefore set out to systematically survey the genomic arrangement of tRNAs in a wide range of eukaryotes to identify common patterns and taxon-specific peculiarities. We found that tRNA complements evolve rapidly and that tRNA locations are subject to rapid turnover. At the phylum level, distributions of tRNA numbers are very broad, with standard deviations on the order of the mean. Even within fairly closely related species, we observe dramatic changes in local organization. Consistent with this variability, syntenic conservation of tRNAs is also poor in general, with turn-over rates comparable to those of unconstrained sequence elements. We conclude that the genomic organization of tRNAs shows complex, lineage-specific patterns characterized by extensive variability, and that this variability is in striking contrast to the extreme levels of sequence-conservation of the tRNA genes themselves. Our comprehensive analysis of eukaryotic tRNA distributions provides a basis for further studies into the interplay between tRNA gene arrangements and genome organization in general.

Secondly, we focused on the investigation of small non-coding RNAs (ncRNAs) from whole transcriptome data. Since ncRNAs constitute a significant part of the transcriptome, we explore this data to detect and classify patterns derived from transcriptome-associated loci. We selected three distinct ncRNA classes: microRNAs, snoRNAs and tRNAs, all of which undergo maturation processes that lead to the production of shorter RNAs. After mapping the sequences to the reference genome, specific patterns of short reads were observed. These read patterns appeared to reflect RNA processing and, if so, should specify the RNA transcripts from which they are derived. In order to investigate whether the short read patterns carry information on the particular ncRNA class from which they originate, we performed a random forest classification on the three distinct ncRNA classes listed above. Then, after exploring the potential classification of general groups of ncRNAs, we focused on the identification of small RNA fragments derived from tRNAs. After mapping transcriptome sequence data to reference genomes, we searched for specific short read patterns reflecting tRNA processing. In this context, we devised a common tRNA coordinate system based on conservation and secondary structure information that allows vector representation of processing products and thus comparison of different tRNAs by anticodon and amino acid. We report patterns of tRNA processing that seem to be conserved across species. Though the mechanisms and functional implications underlying these patterns remain to be clarified, our analysis suggests that each type of tRNA exhibits a specific pattern and thus appears to undergo a characteristic maturation process.

Contents

1	tRNA world and its link to RNAi pathway	3
1.1	tRNA biology	3
1.2	Linking transcriptome analysis and HTS technology	7
1.3	Computational identification of tRNA candidates	12
2	Genomic Organization of Eukaryotic tRNAs	15
2.1	Introduction	15
2.2	Methodology	17
2.3	Results and Discussions	20
2.4	Conclusions	32
3	non-coding RNA identification from transcriptome data	35
3.1	Introduction	35
3.2	Methodology	36
3.3	Results	40
3.4	Discussion	50
3.5	Conclusions	52
4	Computational analysis of tRNA-derived small RNAs	53
4.1	Introduction	53
4.2	Methodology	54
4.3	Results	58
4.4	Discussion	63
4.5	Conclusions	67
5	Conclusions and outlook	68
5.1	Conclusions	68
5.2	Outlook	69
	List of Figures	75
	List of Tables	76
	Bibliography	76
	Curriculum Vitae	91

Chapter 1

tRNA world and its link to RNAi pathway

1.1 tRNA biology

RNAs have been found to participate in an ever-increasing number of pathways of cellular functionality. It is well accepted that most RNA function relies on specific patterns of base pairing and molecular interactions within a single RNA molecule or among sets of interacting molecules, either RNA-RNA, RNA-protein or RNA-cofactors. However, RNA structure can be considered at a number of different levels. Structured RNA can achieve its function at the level of single-stranded ribonucleic sequences (small RNAs or long RNAs), patterns of double-helical stretches interspersed with loops (RNA secondary structure), or complex interactions between secondary structure elements forming three-dimensional functional units (tertiary structure). Functional RNA molecules (tRNAs, rRNAs, snoRNAs, microRNAs, etc.) usually have characteristic spatial structures [101]. During the last decade, both experimental and computational research approaches based on these structures have shown that many important aspects of cell biology are dependent on structured RNAs. Transfer RNAs (tRNAs) are among the most ancient of RNA genes. They can be traced back to the putative RNA World [54], before the separation of the three Domains of Life. There is clear evidence, furthermore, that all tRNA genes are homologs, deriving from an ancestral “proto-tRNA” [42], which in turn may have emerged from even smaller components, see e.g. [41, 147, 33, 51, 34]. tRNAs are essential molecules for protein biosynthesis that couple specific amino acids with corresponding codons. Thus, tRNA functions as an adapter molecule that converts the genetic information stored in the genomic nucleotide sequence into amino acid sequences [129]. This process is achieved through molecular interaction between tRNA and aminoacyl-tRNA synthetases (aaRSs). This interaction assures translation fidelity through accurate recognition of aaRSs and discrimination between cognate and non-cognate tRNAs by aaRSs. Recognition motifs, structural features required for tRNA aminoacylation by aaRSs, play a major role in maintaining tRNA amino acid specificity [74, 29, 152, 73]. Later, recognition motifs are also required for highly coordinated interactions with the ribosome.

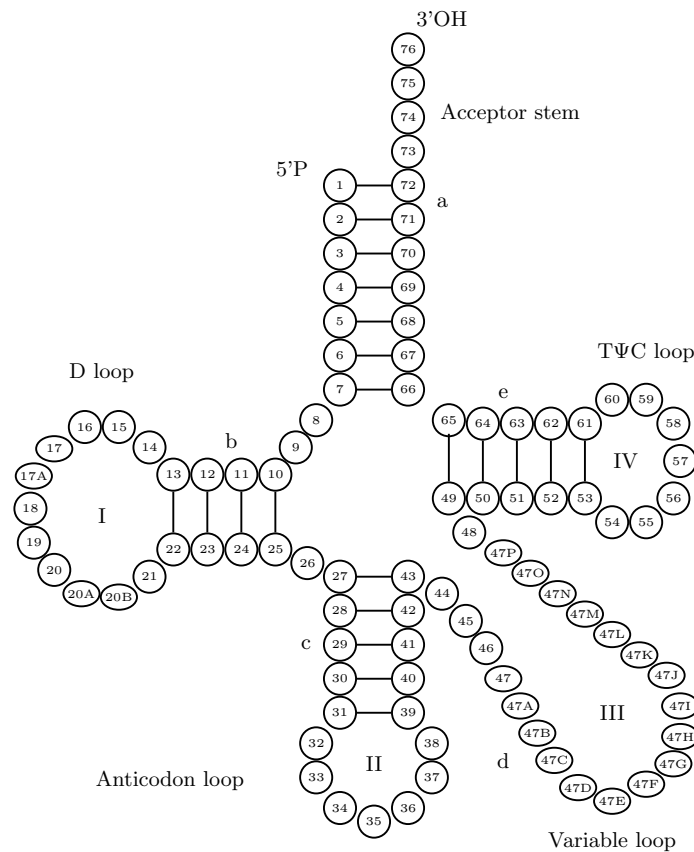


Figure 1.1: tRNA secondary structure. This structure is composed of four base-paired stems (acceptor, anticodon, T Ψ C and the D stem) and four non-base-paired loops: D, anticodon, T Ψ C and variable loop. Each structural element has specific functions in shaping the three dimensional structure. The acceptor stem and T Ψ C-arm stack together to form a continuous alpha-helix, while the D-arm and anticodon arm stack to form another continuous helix. The numbering correspond to the canonical representation by Sprinzl. [164]. Circles for nucleotide positions always conserved, in ovals non-conserved. I, II, III and IV for loops. Letters a, b, c, d and e correspond to regions that shape double helices.

Structural elements are also of fundamental importance to tRNA biology. tRNAs are highly differentiated nucleic acids comprised of 74-95 nucleotides that are folded into a bi-dimensional pattern, known as a tRNA cloverleaf, which is believed to accommodate most known tRNA sequences (see Fig. 1.1). This structure is composed of four base-paired stems (acceptor, anticodon, T Ψ C and the D stem) and four non-base-paired loops: D, anticodon, T Ψ C and variable loop. Each structural element has specific functions in shaping the three dimensional structure. The acceptor stem and T Ψ C-arm stack together to form a continuous alpha-helix, while the D-arm and anticodon arm stack to form another continuous helix. Two RNA double helices cross by 90° to form a characteristic L-shaped tertiary structure. In addition to these interactions, nucleotide modifications are essential for the tRNA to maintain its canonical L-shaped structure [129].

In addition to their current functionality, certain aspects of these structural elements

point to one of the more surprising steps in the origin, diversification and maintenance of life. The acceptor stem includes the 5' and 3' ends of the tRNA and the 3' end harbors a 3'CCA motive that is aminoacylated with a specific amino acid by cognate aaRSs. The anticodon stem harbors the anticodon loop where the anticodon triplet is located. This triplet, through the mediation of the ribosome and other enzymes, facilitates subsequent decoding of the genetic code by inducing binding of the tRNA to its complementary anticodon sequences on the mRNA [74]. This deciphering of the primary genetic code establishes the crucial role of tRNA structure in the decoding of genetic information [74]. However, the existence of a second genetic code, written into the structure of the tRNA and the aaRSs, has also been widely documented. This code presumably recognizes determinants of tRNA identity hidden in a highly conserved and compact common structure with L shaped architecture [74, 29]. Many efforts have been made to characterize the interaction between this sequence-structure and aminoacylation steps [152]. Although the diversity and conservation of the tRNA world is well documented [142], several fundamental biological questions remain open. These include questions regarding aspects of tRNA identity, its role in the evolution of the genetic code, and its role in nucleotide modifications, as well as the effect of structural deviation of tRNA on aminoacylation [60] [73].

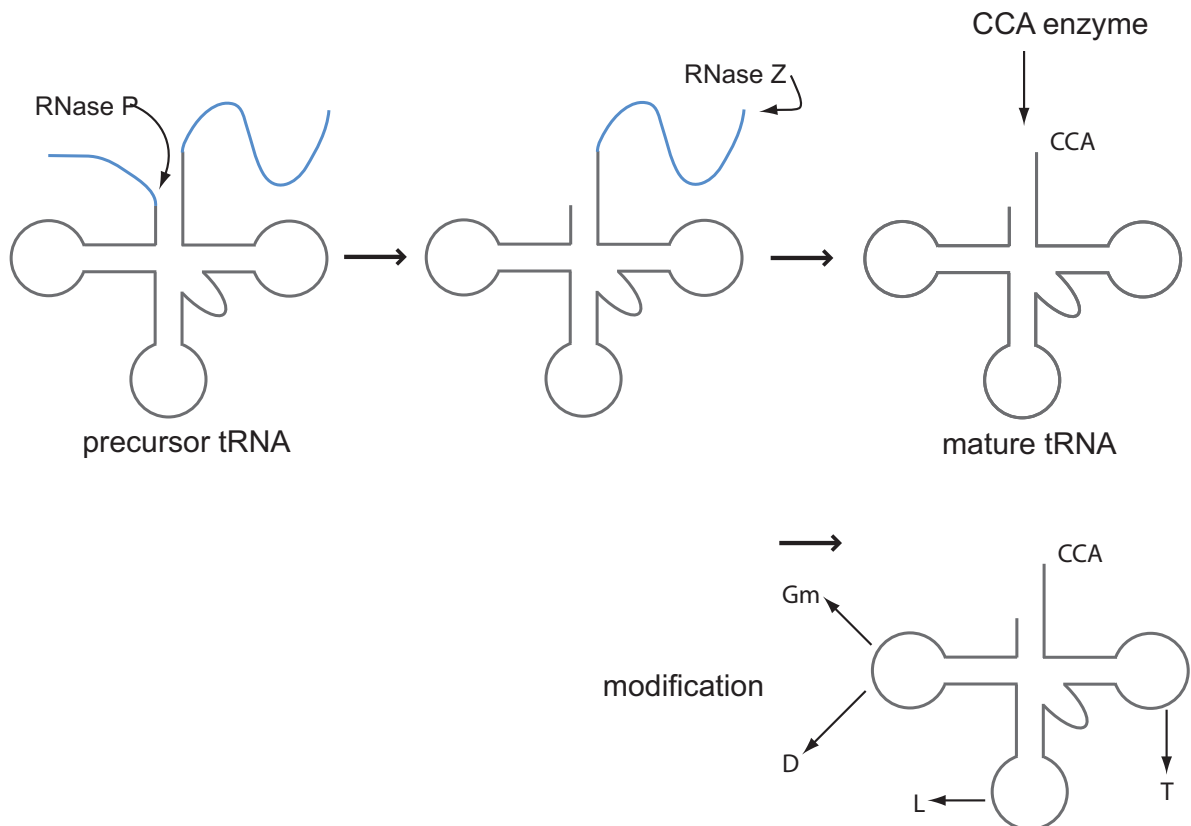


Figure 1.2: Summary of the more important steps of the biogenesis of functional mature tRNAs.

The biogenesis of functional mature tRNAs is an amazingly complicated process [189]. The tRNA is first transcribed by RNA polymerase III (in eukaryotes) as a precursor with 5' leader and 3' trailer extensions [129]. Thus, tRNA transcripts must be processed into their standard length, but they then must also be posttranscriptionally modified. Three main steps lead their maturation: (1) removal of the 5' leader and 3' trailer extensions, (2) modification of some nucleic bases, and (3) addition of CCA to their 3' ends. For the subset of intronic tRNAs, an extra step, the splicing out of an intron (intervening sequence), is also required.

The complexity of tRNA maturation requires a chain of interactions of diverse proteins. The **La** protein autoantigen, which functions in the earliest stages of the biogenesis of many noncoding RNAs [25], is the first protein in this chain. This phosphoprotein forms a complex **La-pre-tRNA** that protects the pre-tRNA 3' end from degradation and assists in correct folding of certain pre-tRNAs. It is also the substrate of the ribonucleoprotein enzyme RNase P, which removes the 5' leader sequence through a single endonucleolytic cleavage [25].

Processing of the 3' end is more complicated. In *E. coli*, the 3' end CCA is encoded by its tRNA complement. RNase E carries out the first step of tRNA maturation through cleavage at positions usually a few nucleotides downstream of the 3' end of the tRNA, either together with or after RNase P action. Then, many other exoribonucleases, mainly RNase T, PH, D, II, etc., shorten the trailer. If the exonucleases trim the trailer sequence beyond the amino acid attachment CCA sequence, a template-independent RNA polymerase, known as CCA-adding enzyme, repairs the CCA terminus [129]. RNase Z endonucleolytically cleaves the sequence 5' of the CCA sequence, and then the CCA terminus is synthesized de novo by CCA-adding enzyme. In eukaryotes, RNase Z cleaves after discriminator nucleotides in tRNA precursors that are generally CCA-less. The addition of CCA to eukaryotic CCA-less tRNAs is catalyzed by tRNA nucleotidyl transferase [107].

In a third universal step, maturation of tRNAs is accomplished by a set of enzymes that act on multiple tRNA substrates, catalyzing the same base modification at a particular position, or a defined set of positions [141]. tRNA modifications are divided into two main categories. Modifications localized in the tRNA core region (D- and T ψ C) that contribute to the stabilization of the L-shaped tertiary structure and modifications occurring within the anticodon loop have the dual functions of precise codon pairing on the one hand and accurate recognition by the cognate aminoacyl-tRNA synthetases on the other. [129]. See Fig. 1.2. However, alterations to the mechanisms that assure tRNA structural stability and the universality of tRNA modifications, which are normally extensively and extremely stable, are also well documented [18, 25, 129]. Some tRNAs lacking specific modifications are subject to degradation pathways [141] and to rapid tRNA decay.

Besides their primary ancestral function in translation, tRNAs appear to have acquired several additional modes of employment throughout evolution. Several recent studies, for instance, have reported tRNA-derived small RNAs in different Eukaryotic clades [106, 154, 19, 23, 98], which at least in part appear to be utilized in the RNAi pathway. Furthermore, tRNA genes are a prolific source of repetitive elements (SINEs) [167], and of tRNA-derived small RNAs such as the small brain-specific non-messenger RNA BC1 RNA [148, 80] and other SINE-derived ncRNAs [132].

Multiple copies of functional tRNA genes, numerous pseudogenes, and tRNA-derived

repeats are characteristic byproducts of tRNA evolution throughout the Eukarya [49]. In general, tRNA genes appear to evolve rapidly. In *E. coli*, the rate of tRNA gene duplication/deletion events is on the order of one per million years [187], and a recent analysis of schistosome genomes revealed striking differences in the tRNA complement between the closely related platyhelminths *S. mansoni* and *S. japonicum* [26].

Although the sequence and structural evolution of tRNAs themselves has received quite a bit of attention [70, 58, 113, 114], much less is known about the genomic organization of tRNA genes. Recent evidence, however, indicates that tRNA genes play a role in eukaryotic genome organization [120], e.g. by acting as barriers that separate chromatin domains. In trypanosomes, for example, tRNA genes mostly appear at the boundaries of transcriptional units and may be involved in the deposition of special nucleosome variants in these regions [173]. Furthermore, there is a link between tRNA loci, in particular clusters of tRNA genes, and chromosomal instability [35, 93, 5, 31, 82]. A recent study showed that tRNA genes may act as barriers to the progression of the DNA replication fork [120], providing a possible mechanism for the formation of genomic fragile sites. The genomic evolution of tRNA genes thus may be linked to the evolution of genome organization. Nevertheless, reports on clade-specific features, such as the strong increase in tRNA introns seen in Thermoproteales [166], are rare.

A peculiar feature of tRNA gene organization is the pattern of tRNA tandem repeats, which so far has been reported only in the protistan parasite *Entamoeba histolytica* [177, 21]. MicroRNAs derived from a precursor in which an imperfectly matched inverted repeat forms a partly double-stranded region, as observed in *Chlamydomonas* [123, 192], furthermore, suggests that head-to-head or tail-to-tail arrangements of tRNA genes might be an evolutionary source of small RNAs.

1.2 Linking transcriptome analysis and HTS technology

The transcriptome is defined as the complement of all RNA molecules, including mRNA, rRNA, tRNA, miRNA, snoRNA, snRNAs, and many other types of non-coding RNA, transcribed in one cell or a population of cells. More formal definitions include quantification of this transcribed material and its relation to differential tissue expression [185]. Efforts to understand the transcriptome have led to the identification of new regulatory elements and the deciphering of key regulatory elements in developmental and disease biology as well dramatically improved sequencing techniques. Since the initial work of Sanger [150, 149], the only sequencing method used 30 years ago, biology has benefited from huge improvements in genome analysis affecting diverse research areas and applied fields. Over the years, the increase in throughput demand for DNA sequencing has driven the development of next-generation sequencing methods that have in turn led to the development of diverse new fields within biology. Next-generation sequencing technologies are transforming the field of genomic science and indeed the field of biology in general [153]. Hundreds of papers have been published in these new fields, ranging from quantitative data on specific biological functions to characterization of the full genetic potential for the sustainable use of ecosystems.

However, the goal of deciphering whole organisms' genomes necessitated a dramatic

increase in DNA sequencing throughput ability, and this requirement was met through the impact of automated capillary electrophoresis. A major revolution in sequencing appeared in 2005 with the publication of shotgun sequencing and de novo assembly of the *Mycoplasma genitalium* genome using open micro-fabricated high density picoliter reactors [116]. This highly parallel sequencing system, developed by 454 Life Sciences, presented biologists with a new technology to greatly reduce reaction volume requirements. Then, also in 2005, George Church lab reported the development of the multiplex polony sequencing protocol, used for the first time to resequence an evolved strain of *Escherichia coli* with an accuracy of less than one error per million consensus bases [155]. Some other biological studies utilize massive parallel sequencing systems like the Solexa 1G sequencing technology, developed to map histone modifications in the human genome [11]. To map protein-DNA interactions across entire mammalian genomes, another group developed the large-scale chromatin immunoprecipitation assay (ChIPSeq), based on direct ultrahigh-throughput DNA sequencing [83]. RNA sequencing (RNA-seq), a revolutionary tool for deciphering the complexity of cell biology expression, combines HTS technology and quantitative measurement of transcripts and their isoforms (for a review see [185]). The RNA-seq approach has enabled the mapping and quantification of mammalian transcriptomes [124], characterization of the transcriptomes of stem cells, profiling of the HeLa S3 transcriptome [22, 117], understanding of the transcriptional landscape of the yeast genome [127], development of highly integrated single-base resolution maps of the epigenome in Arabidopsis [109], transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line [191], and characterization of transcriptome-based aspects of many other human diseases. In addition, this tool has enabled the emergence of one of the major areas for modern genetics: the study of alternative promoters used for transcript production [165].

Mapping problem By the time of the appearance of the new sequencing methods, most criticisms were mainly focused on the size of the read product; first read sizes rarely exceeded 25 bp. This was followed by skepticism founded on the needs to handle large volumes of data, and is being addressed by the development of new sequencing hardware and computational models, still an ongoing process. Thus, biologists and computer scientists continue to face new challenges to transform biology. In the words of Schuster [153], *“this goal will only be achieved through the development of structurally, biochemically, and biophysically detailed computational models based directly on experimental data. Once developed, these models can be simulated, analyzed, and understood through application of modern engineering and computational approaches, and the knowledge gained from these analyses can be applied to the design of additional experiments”*. With the availability of more non-Sanger sequencing methods, it is now becoming possible to assess both next-generation sequencing accuracy and the curation of the vast majority of Sanger-based referenced sequences in the public databases [153]. This will allow exhaustive exploration of the vast information derived from transcriptome data and the development of new computational approaches.

Once the size of the reads produced by HTS increased to more than 400 bp, very different error models were also faced for the new sequencing technology. For example in Solexa, the more frequent read-error types are mismatches, whereas those of 454's GS FLX are insertions and deletions [72].

Later, demand for the development of new mapping methods was intensified, and newly available methods are designed specifically to allow mismatch matching of short reads. Some of them only fill the necessities of specific technologies; for instances SOAP [104] and Maq [102] to map Solexa or SOLiD reads, and longer reads cannot be mapped. SOAP can perform both ungapped and gapped alignments, and has special modules for the alignment of pair-end small RNA and mRNA tag sequences. In Maq, furthermore, the problem of one or two mutations or sequencing errors in a short read leading to its mapping to the wrong location is solved by keeping all of the reads that can be mapped and evaluating the likelihood of incorrect positioning for each of them. With this approach, poor alignments can still be discarded later. Other software, such as [108], are based on extension of the multiple-spaced seeds method, in which different seeds are designed on different positions of a read. The seeds designed only to address the issue of mismatches were demonstrated to also have very high sensitivity when indels are present. Finally, `segemehl` is one approach dealing simultaneously with insertions, deletions, and mismatches. This matching method uses a variation of enhanced suffix arrays to allow a coherent treatment of these read error sources through the modification of matching statistics in such a way that all of the error sources are evaluated efficiently [72].

HTS and its importance to the study of the RNAi pathway The new DNA sequencing technologies have significantly improved throughput and dramatically reduced cost compared with capillary-based electrophoresis systems [155] and have significantly influenced today's biology. On the other hand, though the main problem of HTS was solved, the new technologies have created new challenges for the study of the transcriptome, particularly the small RNA fraction. As a consequence, a new abundant class of small noncoding RNAs was discovered. Thus, frontiers of knowledge have been expanded not only to increase our comprehension of expression and regulation, but also to allow the discovery of a tiny world that re-evaluates our comprehension of cellular functionality in many respects.

In transcriptome analysis, much attention is currently focused on understanding the role of alternative promoters in generating transcript diversity, both for protein-encoding RNAs (traditionally thought of as gene transcripts) and non-protein-encoding RNAs (ncRNAs). Whole-transcriptome analysis of many species and cell types reveals massive expression of ncRNAs. It is widely believed that ncRNAs act as regulators of transcription and translation. Recent investigations of whole RNA cDNA libraries generated by high throughput sequencing (HTS) have shown that these libraries contain both primary and processed transcripts. Further, recent mapping of functional sequence elements in the human genome has corroborated the cDNA-based finding that ncRNAs compose a significant portion of the transcriptome. Finally, numerous studies have illustrated how some mutant RNAs can cause disease [134]. In response to these findings, comprehensive computational approaches are emerging to characterize the vast repertoire of ncRNA expression.

One of the major impacts of HTS on transcriptome studies is the characterization of RNA interference (RNAi). This system, based mainly on three classes of small RNA molecules, microRNA (miRNA), small interfering RNA (siRNA) and piwiRNA (piRNA), was discovered in 1986 and 1990 through observations of transcriptional inhibition by antisense RNA expressed in transgenic plants [38], and reports of unexpected outcomes in experiments performed by plant scientists in the early 1990s [130]. Although the discov-

ery of RNAi preceded HTS technology, today's in depth studies of the roles of new small RNAs in mammalian RNAi-related gene silencing pathways [67] are only possible due to this technology.

Originally, RNAi described a variety of gene silencing processes which require small RNA to mediate site specificity. The pathways including this small RNA machinery appear to overlap to a certain extent. While they use distinct core proteins, they share several components. There are three different mechanisms of regulation of gene expression:

1. Translational inhibition. Classes: miRNA. This small RNA binds to an mRNA and causes translational inhibition. The degree of base-pairing between the miRNA and the target sequence, together with protein components in related miRNPs (Ago1), determines the mode of function. The so-called seed region (7nt on 5end of the miRNA) mediates sequence specificity. RNA degradation requires almost perfect complementarity, whereas translational inhibition allows a certain number of unpaired bases.
2. RNAi: mRNA degradation. Classes: miRNA, siRNA, tasiRNA, natsiRNA, piRNA. In contrast to translational repression, RNAi causes degradation of the target by the RNA induced silencing complex (RISC). Two factors determine this mode: the composition of the RISC complex and the small RNA:mRNA binding pattern. RNAi requires the presence of Ago2 and nearly perfect complementarity between the small RNA and its target. Whereas metazoan miRNAs target the 3 end of the mRNA and by some not yet fully understood mechanism cause blocking of translation, miRNAs in plants target the coding region and cause degradation by an siRNA- like pathway (reviewed in [47]).
3. Transcriptional gene silencing and Imprinting. Class: miRNA, rasiRNA, piRNA. Small RNAs have been shown to promote de novo methylation as well as maintenance of DNA methylation ([9]) in plants and animals ([89]). Several studies also give rise to the idea that histone methylation of specific loci might be guided by small RNAs. MicroRNAs target promoter regions of genes, whereas rasiRNAs shut down repeat-rich regions in the genome.

Components of small RNA biogenesis include Type III RNases that bind and cleave double stranded RNA (dsRNA), divided into three families. RNases of classes I and II are also found in small RNA pathways. Drosha is a class II enzyme that requires Pasha as cofactor. It cleaves pre-miRNAs from longer precursors, which are then further processed by Dicer, a Class III enzyme that has an N-terminal DExD/H-box helicase and PAZ (Piwi/Argonaute/Zwille) domains. It dices long dsRNA into 20nt-long duplexes with a typical 2nt overhang at the 3 end. The family of Argonaute proteins (AGOs) comprises a multitude of different members with various functions ([79]). AGOs consist of an N-terminal PAZ domain, also found in Dicer, and a C-terminal PIWI domain. The exact functions of the domains remain unresolved. However, the PIWI domain seems to bind to the 5 seed region of miRNAs, whereas the PAZ domain interacts with the 3-OH. Vertebrates have four AGOs (Ago 1-4, also known as eIFC1-4). Ago 2 is required for RNAi, whereas Ago1 acts in translational inhibition. Both interact with Dicer ([126]). Polymerases associated with small RNA biogenesis

include DNA polymerase II for miRNAs. Organisms with strong siRNA activity require another enzyme in order to multiply their response to parasitic RNAs. In plants, protozoans, and lower metazoans, RdRP (RNA dependent RNA polymerase) performs siRNA-primed synthesis of dsRNA, which is then cleaved by RISC and Dicer homologs.

tRNA-derived RNA fragments (tRFs) tRNA cleavage products have been associated with stress responses, development, alteration of tRNA structural stability, and other biological processes [105, 178, 179, 190, 76, 66, 19, 18, 25, 129]. However, it is only in the last year, and through whole transcriptome analysis, that the identification of new tRNA-derived small RNAs has increased [23, 98, 67, 52]. Deep sequencing of mixed HeLa cell extracts revealed that the most abundant tRNA-derived small RNAs are products of processing of Lysine, Valine, Glutamine and Arginine tRNAs. These tRNAs are almost exclusively processed from the 5' end, with cleavage by Dicer at the D-loop, resulting in small RNAs approximately 19 nt in length [23]. From the analysis of the global expression profile of small RNAs in human prostate cancer cell lines, three series of tRFs (tRNA-derived RNA fragments) have been discovered: tRF-5, tRF-3 and tRF-1. These sequences were reported as the second most abundant small RNAs (second only to miRNAs) and their names are derived from their precise alignment to the 5' and 3' ends of mature tRNA. The tRF-1 series is located downstream from the 3' end of the mature tRNA sequence. The tRF-5 series exhibits sizes ranging from 15 to 25 nt, tRF-3 ranges from 13 to 22 nt, and tRF-1 has a size distribution that does not correlate with the theoretical distribution expected if the cleavage of all pre-tRNAs gave rise to tRF-1 molecules [98]. A population of these small RNAs is actively produced in *Trypanosoma cruzi* [52], and their production was found to increase under conditions of nutritional stress. This population is preferentially restricted to specific isoacceptors and to the 5' halves of mature tRNAs. The importance of tRNA-derived small RNA to the global regulation of RNA silencing through differential Argonaute association suggests that small RNA-mediated gene regulation may be even more finely regulated than previously realized [67]. These sets correspond to small RNA produced from the 3' tRNA arm and from the region downstream from the 3' end of the mature tRNA. The precise start and end sites for these sets, at or near the tRNA ends, together with their nonrandom nature with respect to size and nucleotide composition at cleavage junctions, strongly suggests that these small RNAs are derived from tRNA cleavage in a specific manner [98, 67]. Recent findings of competition between mammalian RNAi-related gene silencing pathways shown that tRNA-derived small RNAs are involved in the global control of small RNA silencing through differential Argonaute association. tRF levels have minor effects on the abundance of miRNAs but more pronounced influence on the silencing activities of both miRNAs and siRNAs [67].

Origin of other functional molecules from tRNAs Short interspersed elements (SINEs) and long interspersed elements (LINEs) are transposable elements in eukaryotic genomes that mobilize through an RNA intermediate. Most eukaryotic SINEs are ancestrally related to tRNA genes, although the typical tRNA cloverleaf structure is not apparent for most SINE consensus RNAs. In all cases, SINEs harbor in their tRNA-related segment an internal promoter (composed of A and B boxes) recognized by RNA polymerase III. A general multistep model is available for the evolution of tRNA-related SINEs in eukaryotes [167].

Another tRNA-derived functional molecule is the rodent BC1 RNA, a small brain-specific non-messenger RNA. BC1 RNA is specifically transported into the dendrites of neuronal cells, where it is proposed to play a role in the regulation of translation near synapses [148]. A previous study demonstrated that the 5' domain of BC1 RNA was derived from Alanine tRNA. However, evidence indicates that changes accumulated during evolution have created an extended stem-loop that does not fold into the predicted canonical tRNA cloverleaf structure. BC1 RNA has been associated with fragile X syndrome (caused by the functional absence of the fragile X mental retardation protein, FMRP). However, the specific details of interactions between FMRP and BC1 RNA remain controversial, although BC1 RNA has been proposed to repress translation initiation at the level of 48S complex formation [148].

miR916 from *Chlamydomonas reinhardtii* Analysis of the transcriptome of the unicellular algae *C. reinhardtii* led to the discovery of the existence of miRNAs in unicellular organisms [123, 192]. These studies showed that *C. reinhardtii* contains putative evolutionary precursors of miRNAs and species of siRNAs resembling those in higher plants, indicating that complex RNA-silencing systems evolved before multicellularity and were a feature of primitive eukaryotic cells [123, 192].

In a pilot study, we previously identified a set of microRNAs derived from pre-miRNAs in which an imperfectly matched inverted repeat forms a partly double-stranded region, as observed in *Chlamydomonas*. This suggests that head-to-head or tail-to-tail arrangements of tRNAs might be an evolutionary source of small RNAs. See Tab. 1.1 and Fig. 1.3.

miRNA	tRNA begin	Bounds end	tRNA type	Anticodon	S.Intron	E. Intron	Bounds end	CS
02359	1	301	374	Ile	AAT	0	0	75.27
02359	2	84	11	Ile	AAT	0	0	75.27
mir916	1	69	142	Ile	AAT	0	0	72.81
mir916	2	864	791	Ile	AAT	0	0	72.81

Table 1.1: Identified tRNAs stretched on cre-miRNA 02359 and mir916, CS cover score from tRNAscan-SE.

1.3 Computational identification of tRNA candidates

RNA sequence analysis using covariance models Presumably because the function of noncoding RNAs is based on their three-dimensional secondary structure, RNA sequences appear to be adapted to the maintenance of a particular base-paired structure rather than the conservation of primary sequence [39]. In the preceding sections, we have mentioned how pivotal RNA secondary structure is to RNA function. Strong pairwise correlations in RNA sequence, usually manifested as Watson-Crick complementary pairing, correspond to an intermediate three dimensional RNA structure. Making use of these correlations, computational approaches and mathematical models aimed at the detection of RNA at the genome-wide scale have been recently developed. Since the inference of phylogenetic trees is based on multiple RNA sequence alignments, the demand for new automated RNA

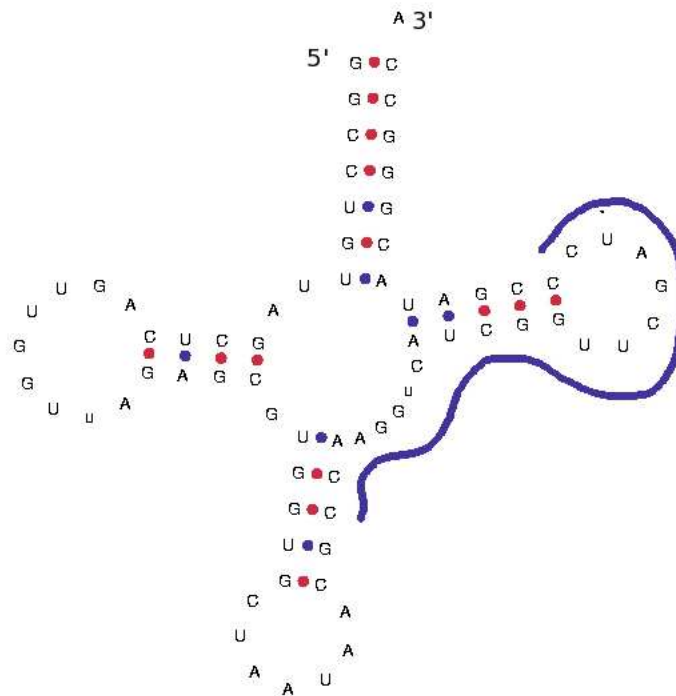


Figure 1.3: Stretch of the miR916 from *C. reinhardtii* over the Isoleucine tRNA.

detection methods, combining both structure prediction and multiple alignment approaches, has been increasing over the course of the last two decades. By 1994, Eddy & Durbin introduced a probabilistic model, called the “covariance model” (CM), for the detection of tRNAs. This model cleanly describes both the secondary structure and the primary consensus sequence of a tRNA. This new approach has helped to solve several RNA analysis problems, including consensus secondary structure prediction, multiple sequence alignment, and database similarity searching [39], and was later extended to the sequences of other families of structured RNAs [64, 53]. CMs are a generalization of hidden Markov models (HMMs), probabilistically rigorous models that were first described in a series of statistical papers by Leonard E. Baum and other authors in the second half of the 1960s and were first applied to speech recognition [143]. CM is a special case of profile stochastic context-free grammar. Tools like *Infernal* (“INFERENCE of RNA ALIGNMENT”) or *tRNAscan-SE* are sophisticated implementations of CMs developed for searching DNA sequence databases for RNA structure and sequence similarities. *Infernal* is used to search nucleic acid sequence databases for homologous RNAs, or to create new sequence- and structure-based multiple sequence alignments of RNA families [131]. *tRNAscan-SE* is used to detect tRNA genes and related sequences [111].

tRNA gene identification The *tRNAscan-SE* program [111] works in three phases.

Stage 1. The input sequence is analyzed using *tRNAscan* (an optimized version of *tRNAscan* 1.3 [46]) and the Pavesi algorithm. The latter is an implementation of the Pavesi

search algorithm [137] known as EufindtRNA. Results from both programs are then merged into one list of candidate tRNAs. The Pavesi algorithm is based on a modified version of the general weight matrix procedure; their algorithm relies on the recognition of two intragenic control regions, known as A and B boxes, a transcription termination signal, and the evaluation of the spacing between these elements [137].

- Stage 2. tRNAscan-SE is used to extract the candidate subsequences and pass these segments to covels, a covariance model search program [39]. Covels applies a tRNA covariance model (TRNA2.cm) that was made through the structural alignment of 1415 tRNAs from the 1993 Sprinzl database [164] with some editions [39]. To improve intron prediction, intron sequences were manually inserted into the Sprinzl alignment for 38 intron-containing tRNAs of known genomic sequence.
- Stage 3. tRNAscan-SE takes predicted tRNAs that have been confirmed with covels, logs odds scores of over 20.0 bits, trims the tRNA bounds to those predicted by covels, and runs the covariance model global structure alignment program coves [39] to get a secondary structure prediction. The tRNA isotype is predicted by identifying the anticodon within the coves secondary structure output. Introns are identified from this output as runs of five or more consecutive non-consensus nucleotides within the anticodon loop.
- Stage 4. tRNAscan-SE uses heuristics to try to distinguish pseudogenes from true tRNAs, primarily on the basis of a lack of tRNA-like secondary structure. A second tRNA covariance model (TRNA2ns.cm) was created from the same alignment, under the constraint that no secondary structure is conserved (this model is effectively just a sequence profile, or hidden Markov model). By subtracting a tRNA's similarity score (similarity to the primary structure-only model) from that using the complete tRNA model, a secondary structure-only score is obtained. In Bayesian terms, this difference can be viewed as the evidence for a complete tRNA model, as opposed to a sequence-only pseudogene model (lacking secondary structure). We observed that tRNAs with low scores for either component of the total score were often pseudogenes. Thus, tRNAs are marked as likely pseudogenes if they have either a score of <10 bits for the primary sequence component of the total score, or a score of <5bits for the secondary structure component of the total score.

The sensitivity of tRNAscan-SE relies on the identification of 99-100% of transfer RNA genes in a DNA genome while presenting less than one false positive per 15 gigabases. The selectivity is measured by its ability to avoid misidentifying non-tRNA sequences as true tRNAs. In human genomes, the program's false positive rate is zero except for the cases of tRNA-derived SINEs and tRNA pseudogenes [39].

Chapter 2

Genomic Organization of Eukaryotic tRNAs

2.1 Introduction

Understanding genome features is one of the major goals of computational genomics. Computational analysis is being increasingly used to decipher biological information from genome sequences and related data. The recently published literature points out the importance of studying the vast array of ways in which organisms' genomes are organized. For example, a comparison of the genomic organization of six major model organisms shows size expansion in parallel with the increase in complexity of the organism, e.g., the difference between the genome size of yeast and mammals is more than 300-fold, but the difference in overall gene number is only a modest 4- to 5-fold [92]. Many authors believe that studies of genomic organization will be the basis for future understanding of epigenetic mechanisms.

Though most efforts to study epigenetics have focused on the genome organization of protein-coding RNA, attention has lately been aimed at the organization of small non-coding regulatory RNAs. Although the most extensively studied among these are microRNAs, transfer RNAs (tRNAs) are among the most ancient genes. Mainly known as housekeeping RNAs, tRNAs have a pivotal function in protein translation. Progress in nucleic acid sequencing, especially in large-scale automated DNA sequencing of whole genomes, along with different algorithms that allow tRNA gene identification on a wide genome scale, now allow access to more detailed information regarding the number and the organization of tRNA genes at the genome level.

The diversity of tRNA genes in eukaryotes has been previously reported for 11 eukaryotic genomes [58]. In addition, a comparison of 50 genomes from the three domains of life (7 eukarya, 13 archaea, and 30 bacteria) [113] reveals domain-specific structural and functional features as well as a suggestive diversity of tRNA function. For example, the eukarya exhibit tRNA redundancy, with two or more proteins encoded by the same anticodon, in contrast to both archaeons and bacteria, in which the trend is evidently a low level of tRNA redundancy [113]. In spite of these advances in understanding tRNA diversity, little is known about the organization of tRNA in the genome.

The limited body of knowledge in this area includes surveys that point out the importance of tRNAs and genome organization. These studies include laboratory experimentation

that has shown how tRNA gene locations interfere with replication forks [35, 93, 5, 31, 82]. In addition, retrotransposable elements frequently target the vicinity of tRNA genes in order to avoid gene disruptions upon retrotransposition [20], indicating that tRNA genes are selected as chromosomal integration sites. Finally, this view is also supported by the dynamic of genomic rearrangements, losses and additions. Sites of gene gain and evolutionary breakpoints both tend to be associated with tRNA genes, as revealed by the comparison of an extinct ancestral yeast and *S.cerevisiae* [59]. Clearly, tRNA genome organization is an intriguing issue in RNA biology.

tRNA: relation of structure and aminoacylation

Experimental studies have shown that mispairs and tRNA helix irregularities are significantly more important for aminoacylation and translation than previously thought. The importance of tRNA helix irregularities shaped by G:A, C:A, and U:U mispairs has been documented through the use of site-directed mutagenesis studies [118, 13, 119]. Interestingly, tRNA sequence comparisons between genomic DNA and cDNA obtained from unprocessed primary transcripts has revealed nucleotide discrepancies in mitochondrial tRNAs from *Acanthamoeba castellanii*, the protist *Seculamonas ecuatoriense*, and plants [110, 115, 45, 100]. Those editing events correct mismatched C:A and U:U base pairs, which appear when folding the gene sequence into the standard cloverleaf structure. Then those introducing sequence changes and some specific studies of tRNA aminoacylation would be important to the identification and re-evaluation of more potential functional tRNA pseudogenes from Genomic sequences.

tRNA tandem array in Entamoeba

Clustering of tRNA genes is considered rare and tandem arrays of tRNA genes have been reported to date only in the amoeba *E. histolytica*. In the most related organism, the soil-living amoeba *Dictyostelium discoideum*, the tRNA genes are dispersed throughout the genome [40]. The unique organization of tRNA genes in *E. histolytica* [21] was discovered because 29% of all sequence reads were excluded from the raw *E. histolytica* HM-1:IMSS genome assembly [16] due to their repetitive nature. Analysis revealed that approximately two-thirds of these were derived from the ribosomal DNA episome, but that the rest were derived from tRNA arrays. This tRNA gene organization, in tandem arrays, makes up over 10% of the genome, with 25 distinct arrays having been described. They are composed of tandemly repeated units encoding between 1 and 5 tRNA acceptor types. It has also been reported that three of the arrays also encode 5S RNA and one encodes another RNA suspected to be a small nuclear RNA (snRNA) [21]. In addition to the high copy number, the tandemly arrayed organization of the tRNA genes is unprecedented and its origins are still unknown [21]. To detail this unique gene organization, studies of four other species of the genus *Entamoeba*, *E. dispar*, *E. moshkovskii*, *E.terrapinae*, and *E. invadens*, have been undertaken. This survey revealed that tRNA arrays appear to be a general feature of *Entamoeba*, but many questions regarding their origin and function remain [177].

Synteny

Synteny means “same thread” (or ribbon), a state of being together in location, as synchrony would be together in time. More strictly, in genetics, synteny refers to gene loci on the same chromosome, regardless of whether or not they are genetically linked by classic linkage analysis. Although the term was introduced in 1971 by John H. Renwick, of the London School of Hygiene and Tropical Medicine, at the 4th International Congress of Human Genetics in Paris, the term synteny nowadays is often used to refer to gene loci in different organisms located on a chromosomal region of common evolutionary ancestry [135]. For modern comparative genomics, it is a well-established inference that human and mouse species share around 200 homologous segments, i.e., chromosome chunks that contain a linear stretch of the same gene homologs in the two compared species. Based on the definition of synteny, the term “conserved synteny” corresponds to the (local) maintenance of gene content and order in certain chromosomal regions of related species [27, 99]. The increasing number of genome sequences and the improved analytical approaches being used today are clarifying angiosperm evolution and revealing patterns of differential gene loss after genome duplication and differential gene retention associated with evolution of some morphological complexity [175].

In the following section, we present a methodology for a comprehensive and exhaustive search of tRNA locations on a genome-wide scale. This is followed by an exploration of the genomic dynamics of tRNAs by searching for tRNA-bearing loci, especially addressing the issue of whether tRNAs can be found in syntenic locations. Finally, we will touch on the many questions arising from our results, particularly those about the sorts of forces that might give rise to the long-range patterns seen in these tRNA genomic distributions.

2.2 Methodology

Sequence data We retrieved 74 eukaryotic genome mainly from the following public resources: NCBI, Ensemble Genome Browser and Joint Genome Institute. For a detailed list of the individual genome assemblies we refer to [4]

tRNA detection Detection of tRNAs was performed by using tRNAscan-SE v.1.23 (April 2002) with default parameters, i.e., the TRNA2.cm covariance with strict filter parameter -32.1 was used to screen each genome for tRNAs and tRNA pseudogenes. All analyses were performed using both the set of all intact, putatively functional tRNAs identified by tRNAscan-SE and using all tDNA loci, i.e., the union of tRNA genes and tRNA pseudogenes.

The distinction of tRNA genes and pseudogenes necessarily relies of a set of heuristics implemented in tRNAscan-SE. These are well-founded in what is known about functional tRNA genes [74, 29, 152, 73, 60, 142]. Processing and recognition of specific tRNAs imposes stringent constraints on the sequence (and secondary structure) of tRNAs; several nucleotides of mature tRNAs need to be chemically modified in most species, imposing further constraints on the primary sequence. tRNAscan-SE’s consensus models implement these constraints with reasonable accuracy but by no means perfectly. In the absence of detailed experimental information on the expression and the functionality of a particular tDNA it

is of course impossible to distinguish between tRNA genes and tRNA pseudogenes with absolute certainty. For the statistical evaluation of genome-wide comparison reported here, however, the accuracy of tRNAscan-SE appears to be sufficient [111, 58, 172]. There are, however, several sources for errors, in particular in the presence of RNA editing e.g. in the mitochondrial tRNAs of many plants and the protist [110, 115, 45, 100]. Such organellar data are not considered in this contribution, however.

tRNA-geo pipeline The tRNA-geo pipeline is a Perl program that parses tRNAscan-SE output and produces summary information as well as overview graphics such as those shown in Figs. 2.1 and 2.7. First tDNA locations are sorted in consecutive order along each input sequence, distances are measures (see below for exact definitions), tDNA pairs and tDNA clusters are identified, summary statistics are computed. Graphics are produced using PSTricks macros and LaTeX.

tRNA-geo pipeline description

- **String definition.** Every tDNA is represented by a quadruple $P = (a, b, o, t)$, where $a < b$ are the start and end positions within each input sequence (chromosome, scaffold, or contig), and $o \in \{+, -\}$ is the orientation of the tRNAs. We say that two tDNAs are of the same type t if they belong the same isoacceptor family, i.e., if they code for the same aminoacid.
- **Distance definition:** The tRNA loci are ordered such that $P_i \prec P_j$ if and only if $a_i < a_j$. The distance between two consecutive loci P_i and P_{i+1} is defined as $\delta_i = a_{i+1} - b_i$.
- **Cluster definition:** A cluster C is a maximal sub-sequence of loci $C = (P_i, P_{i+1}, \dots, P_j)$ such that $\delta_k < 1000$ for $i \leq k < j$. The cut-off of 1000 was chosen because the overwhelming majority of consecutive tDNA pairs in the random control have larger distances while a large fraction of the tDNA pairs in the real data have smaller distances than this cut-off value, see Fig. 2.5 here.
- **Cluster class definition:** A cluster is called homogeneous if all its tDNAs are of the same type t_k ; otherwise, it is called heterogeneous. A sub-sequence consisting of two consecutive loci located within a cluster C is called a pair. The pair (P_i, P_{i+1}) is homogeneous if $t_i = t_{i+1}$ and heterogeneous otherwise. A pair has parallel orientation if $o_i = o_{i+1}$. For anti-parallel pairs, $o_i = -o_{i+1}$, we distinguish head-to-head $o_i = +$ and $o_{i+1} = -$ ($\leftarrow\rightarrow$), and tail-to-tail $o_i = -$ and $o_{i+1} = +$ ($\rightarrow\leftarrow$) orientations.

Simulations In order to investigate the statistical significance of the tDNA pairs we compare the genomic tDNA organization with randomized configurations. To this end, we remove the collection of tRNA genes and pseudogenes from the genome and re-insert them at positions chosen from a uniform distribution on the remaining sequence. Empirical p -values, defined as

$$p = \#\{i | y(i) \geq x\} / N, \quad (2.1)$$

where $y(i)$ is the number of clustered tRNAs in replicate i , x is the number of clustered tRNAs in the genome, are determined from $N = 50$ to $N = 1000$ random replicates. For large (insignificant) p -values, simulations were terminated at fewer replicates to save computer time.

Statistical tests Were performed using the R statistics environment [1]. In particular, Fisher’s exact test [48] with 2×2 contingency tables was used in order to test whether the observed proportion of homogenous and heterogeneous pairs depends strongly on whether tRNA pseudogenes are included in the analysis or not.

Synteny To analyze the synteny between species, we utilized two different pipelines depending on available genomic data and their interrelations in public data sources. The BioFuice [91] integration platform is used to analyze the synteny in eight different vertebrate species *Homo sapiens*, *Pan troglodytes*, *Pongo pygmaeus*, *Macaca mulatta*, *Mus musculus*, *Monodelphis domestica*, *Gallus gallus*, and *Xenopus tropicalis*. The analysis runs in several steps. Firstly, the Ensembl data source (version 53) is utilized to create the genomic mappings between the tRNAs and/or tRNA pseudogenes and at most five consecutive protein-coding flanking genes in both directions, up- and downstream. The number 5 was chosen pragmatically as a trade-off between the need to evaluate local information and the unavoidable incompleteness of genome annotations, whence homologs of many genes are missing in individual genomes. These genomic mappings are chromosome- and strand-specific, i.e., the resulting genes are located on the same chromosome and strand as the input tDNAs. Next, the resulting genes are associated to protein-coding genes of other mammalian species using the homologous data available in Ensembl Compara (version 53). These homology relationships between genes in different species are then filtered to focus on those genes flanking tRNAs. Finally, tDNAs of different mammals can be associated based on the genomic mappings to their flanking genes (gene-tRNA) and the homology relations between those (gene-gene).

We consider two alternatives for creating such tDNA relationships:

1. Two tDNAs are associated by the *single-sided linkage* relation if there is at least one homology relationship between their pre-selected flanking genes. Here we do not require that the homologous genes have the same relative orientation or relative location w.r.t. to the tDNAs.
2. Two tDNAs are associated by the *two-sided linkage* relation if there is at least one pair of homologous genes in both the up-stream and the down-stream region. Again, relative orientations are not taken into account. The tDNAs need to be located between the two homologous gene-pairs, however.

The Single-sided linkage relation turns out to be not very informative because many-to-many homology relations for large gene families and the relatively large regions used to define the synteny relation severely limit the sensitivity. We therefore limit a details discussion to the two-sided linkage relation.

For invertebrate genomes, synteny information was extracted directly from genome annotation using a custom-made pipeline based on Perl and awk scripts. For the nematodes

C. elegans, *C. briggsae*, *C. japonica*, *C. brenneri*, *C. remanei* we considered a region of 40.000nt up- and downstream of the tRNA loci. A pair of tDNAs was defined as syntenic if we could find in this range at least two orthologous proteins between them. The flanking proteins were taken from the genome annotation gff-files from Wormbase WS204. A list of orthologous proteins was computed using *OrthoMCL* [103] to determine if two proteins are ortholog. Tab. 2.4 summarizes the prevalence of tRNA synteny within the genus *Caenorhabditis*. The tDNAs in the genus *Drosophila* were analyzed in the same way. The flanking proteins were taken from Flybase (release FB2009_09). Since a sufficiently complete orthology annotation was not readily available, we used *ProteinOrtho* [96] for this purpose. The results are compiled in [4].

The fraction of syntenically conserved tDNAs was compared to the evolutionary distances for each pair of genomes in the three data sets described above. The evolutionary distance for the Vertebrates and Nematodes is gathered by the tree model underlying the UCSC 28-way alignments [121]. For the genus *Drosophila* the evolutionary distances are genomic mutation distances computed from 4-fold degenerated sites in all coding regions corrected for base composition as in [174].

Codon bias usage Codon bias usage was done by using the codon usage table from [128]. A total of 52988 codons corresponding to 217 CDS from *S. japonicum*, 8160 codons of 13 CDS from *S. mediterranea* and 161225 codons of 369 CDS from *S. mansoni* were analyzed.

RepeatMasker RepeatMasker [159] is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. Sequence comparisons in RepeatMasker are performed by so-called `cross match` program [63], an efficient implementation of the Smith-Waterman-Gotoh algorithm [160, 61]. The general purpose of this utility is to compare any two sets of DNA sequences. In our survey we have compared the set of detected tRNA genes against libraries of repeats [159].

2.3 Results and Discussions

For each of the 74 genomes included in our survey we collected summary statistics on the number of tRNA gene and tRNA pseudogenes as well as on their genomic clusters. To simplify the language, we will use the term “tDNA” to refer to both tRNA genes and tRNA pseudogenes, while “tRNA gene” will be reserved to loci with probably intact tRNA sequences. In practise, we use *tRNAscan-SE* to distinguish between tRNA genes and tRNA pseudogenes (see Methods for details).

We define two adjacent tRNA gene or tDNAs as “clustered” if their distance is less than 1000 nucleotides. This threshold is motivated by a statistical analysis of the distances between adjacent tDNA loci, which shows that at this distance we have to expect very few or no tDNA pairs in the genomes under investigation (see Methods for details). We then distinguish between *homogeneous clusters*, consisting of tDNA with the same isoacceptor family (i.e., coding for the same aminoacid), and *heterogeneous clusters*. Within clusters, we separately consider the three relative orientations $\rightarrow\rightarrow$, $\leftarrow\rightarrow$, and $\rightarrow\leftarrow$. Data have

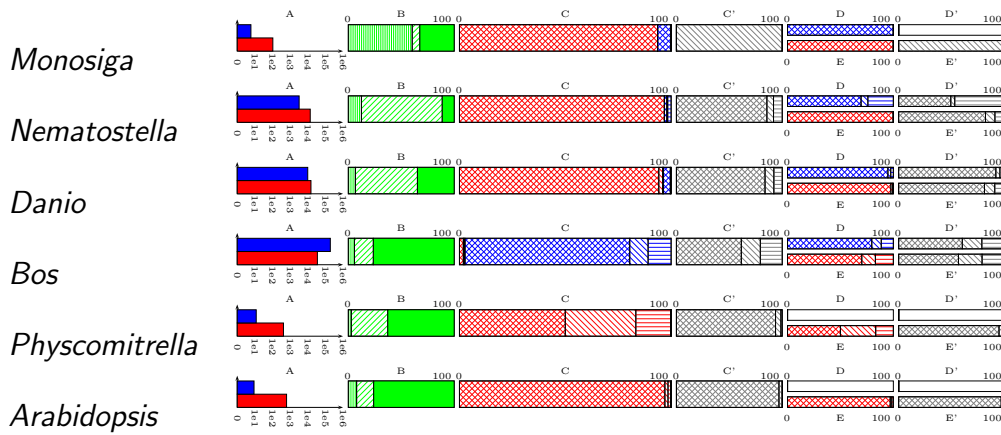


Figure 2.1: Summary of tRNA gene and tDNA statistics **A** Distribution of tRNA genes and tRNA pseudogenes

■ Natural logarithm of the total of tRNA pseudogenes ■ tRNAs. **B Fraction of tRNAs and tRNA pseudogenes in clusters:** ▨ homogeneous clusters ▧ heterogeneous clusters ■ not in clusters. **C Fraction of Homogeneous pairs:** ▨ →→, ▧ →←, ▨ ←→: tRNA pairs. ▨ →→, ▧ →←, ▨ ←→: tRNA pseudogenes pairs. **C' Fraction of Heterogeneous pairs:** ▨ →→, ▧ →←, ▨ ←→. In D and D' and E and E' holds the same rules as C and C' but the raw data is the result of filtering tRNAs or tRNAs pseudogenes respectively. D and D' fraction for pairs of tRNA pseudogenes and E and E' fraction for tRNAs.

been analyzed for putatively functional tRNA gene (as classified by tRNAscan-SE), and for all tDNAs. Fig. 2.1 shows a sample of a graphical representation of the survey results. The full figure comprising all 74 genomes is provided at Appendix A. Complete lists of tDNAs in gff format can be found at the website [4].

Despite an overall correlation with genome size, there does not seem to be a general trend in the number of tRNA genes. Although some mammals, for instance, exhibit tens or even hundreds of thousands of tDNA copy numbers, other mammalian genomes harbour only a few hundred copies. For instance, old world monkeys and great apes have about 616 ± 120 tDNAs, while the related bushbaby (*Otolemur garnetti*) exhibits 45225 tDNAs. The highest counts are reached for the cow and rat genomes with more than 100000 tDNAs. For the 12 sequenced *Drosophila* species, we find 320 ± 73 tDNAs. *Trichoplax adhaerans*, one of the most basal animals has no more than 50 tRNA genes, while the cnidaria *Nematostella vectensis* has more than 17000. Within teleosts, tDNAs range from about 700 in Tetraodontiformes to 20000 in zebrafish.

Variations by about an order of magnitude are also common in other major clades. *Naegleria gruberi*, for example has 924 tDNAs, while Kinetoplastids (*Leishmania* and *Trypanosoma* have only 91 and 65 copies).

Surprisingly, the variation is very small in the “green lineage”. Spermatophyta show little variation with 706 ± 96 loci, the basal land plants *Physcomitrella patens* (432 tDNAs) and *Selaginella moellendorffii* (1290 tDNAs) and even the unicellular algae *Volvox carteri* (1051 tDNAs) and *Chlamydomonas reinhardtii* (336 tDNAs) have similar numbers.

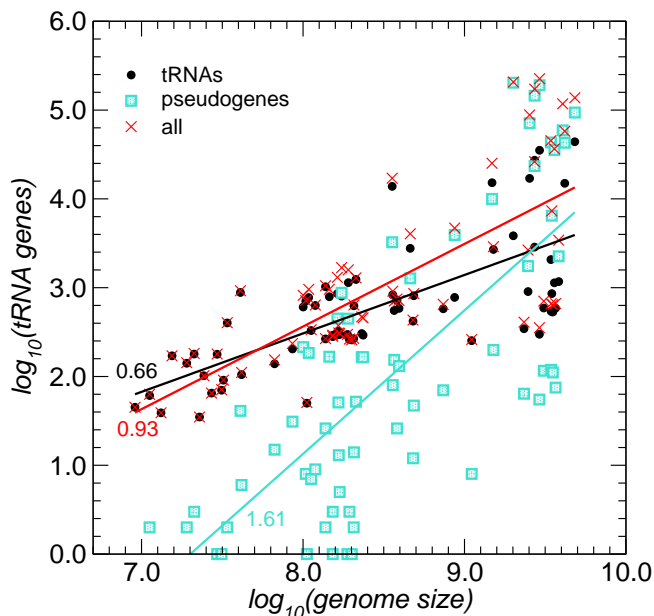


Figure 2.2: Correlation of the number of tDNAs with genome size. The slopes of the three regressions are significantly different: Intact tRNA gene (\bullet , $\alpha = 0.658 \pm 0.076$), tRNA pseudogenes (\square , $\alpha = 1.615 \pm 0.181$), total number of tDNAs (\times , $\alpha = 0.930 \pm 0.096$)

Despite the often large variation among even closely related lineages, we observe the expected correlation between the number of tDNAs with genome size, Fig. 2.2. The correlation is significant, with correlation coefficient $\rho \in (0.71 \dots 0.76)$, but subject to a high level of variation reflecting large differences in the evolutionary history of different lineages. While the total number of tDNAs scales approximately linearly with genome size, $\alpha = 0.93 \pm 0.10$, the growth in the number of intact, probably functional tRNA genes is much slower, consistent with $N^{2/3}$. The number of tRNA pseudogenes, on the other hand, grows faster than linearly, $\sim N^{1.61 \pm 0.18}$. The reasons for this difference in scaling remains unclear. One may speculate that selective forces maintain only a limited number of functional tDNA copies causing the sub-linear growth of intact tRNA genes with genome size, while the duplication/deletion mechanism acts towards a uniform coverage of the genome with a rate that is to a first approximation constant throughout eukaryotic genome, accounting for the linear growth of the total number of tDNAs.

Several selective forces could act on the tRNA genes and/or all tDNA loci to cap their number. The bias towards small deletions over insertions observed in [92] is one potential candidate that is independent of special properties of tRNAs. Variations in codon usage might provide another selection-based explanation for the variation of tDNA copy numbers. In eubacteria, a correlation between tRNA abundance, tRNA gene copy number, and codon usage is well established [146]. Whether codon bias causes tDNA copy number variation or *vice versa* remains topic of an intense debate, however. A mechanistic explanation describing the coevolution of codon usage with tRNA gene content is given in [69]. It remains unclear to what extent the correlation of tRNA copy numbers and codon usage carries over to eukaryotic genomes. A detailed investigation in *Schistosoma mansoni* and *Schistosoma japonicum* finds no correlation between tRNA gene numbers and codon usage, while a statistically significant but still very weak correlation is observed in *Schmidtea mediterranea* [26]. In *Nasonia*, the correlation of codon usage and the copy numbers of tRNA genes appears to be restricted to highly expressed genes. The strength of this correlation decreases with GC-content in plant genomes [125].

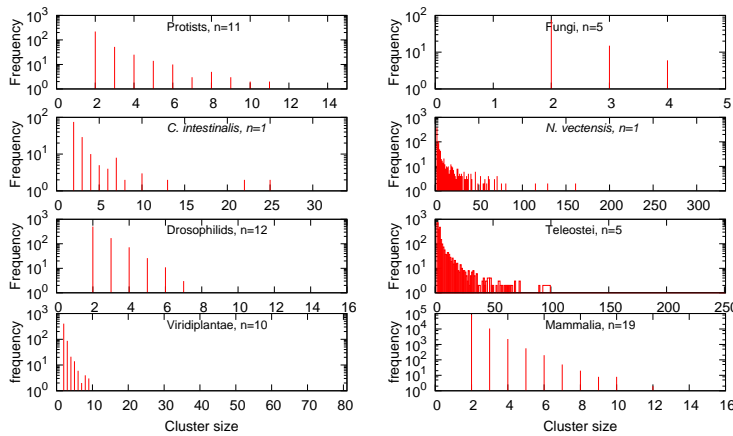


Figure 2.3: Distribution of tDNA clusters sizes for several lineages for which multiple sequenced genomes are available as well as some examples of individual genomes. Most tDNA clusters are small, and the frequency of long clusters rapidly decreases.

In any case, codon usage cannot be employed to explain the observed differences in tDNA copy numbers that span several orders of magnitude. These huge fluctuations, which are observed both within some lineages and between closely related lineages, argues against a mechanism that relies on selection on the tRNAs. Instead, the more than linear scaling of tRNA pseudogenes with genome size suggests a faster tDNA turnover in larger genomes — after all, pseudogenes and gene relics are steps in the evolutionary degradation of genes.

Survey on Nematode

Nevertheless, only little is reported about genome-wide distribution and organization of tRNA loci in most species. Nematodes are no exception, we investigated the organization of tDNAs in more detail including four species of different Nematode genus.

For the genus *Caenorhabditis*, the total number of predictions (including pseudogenes) is $1,126 \pm 314.39$. After removing pseudogenes, the numbers reduce to 833.4 ± 194.28 standard deviation, varying from 606 in *C. elegans* to 1,139 in *C. brenneri*. For other nematodes *B. malayi*, *P. pacificus*, *M. hapla* and *M. incognita* the total is 643 ± 611.43 and is reduced to 527.75 ± 520.61 after removing pseudogenes. Compared to *D. melanogaster* (299 tRNAs genes; 5 tRNA pseudogenes) these numbers are much higher.

Most tDNAs in *Caenorhabditis* are isolated genes, although $\sim 20\text{-}41\%$ of the total tDNA predictions are organized in clusters. For comparison, in other nematode species only $\sim 11\text{-}23\%$ tRNA genes occur clustered, 84% in the planarian *S. mediterranea*, and 66% in the insect *D. melanogaster*. An indirect measure of cluster structure is seen on C and C' in Fig. 2.4. It turned out, that 50% of pairs have the same orientation (direct duplication) or the opposite orientation (inverse duplications). The exception to this feature is shown for the species *P. pacificus* and *B. malayi* where more than 70% of pairs are in the same orientation.

tDNA clusters

In order to investigate the propensity for the formation of tDNA clusters, we consider the cumulative distribution of consecutive tDNA pairs as a function of their genomic distance.

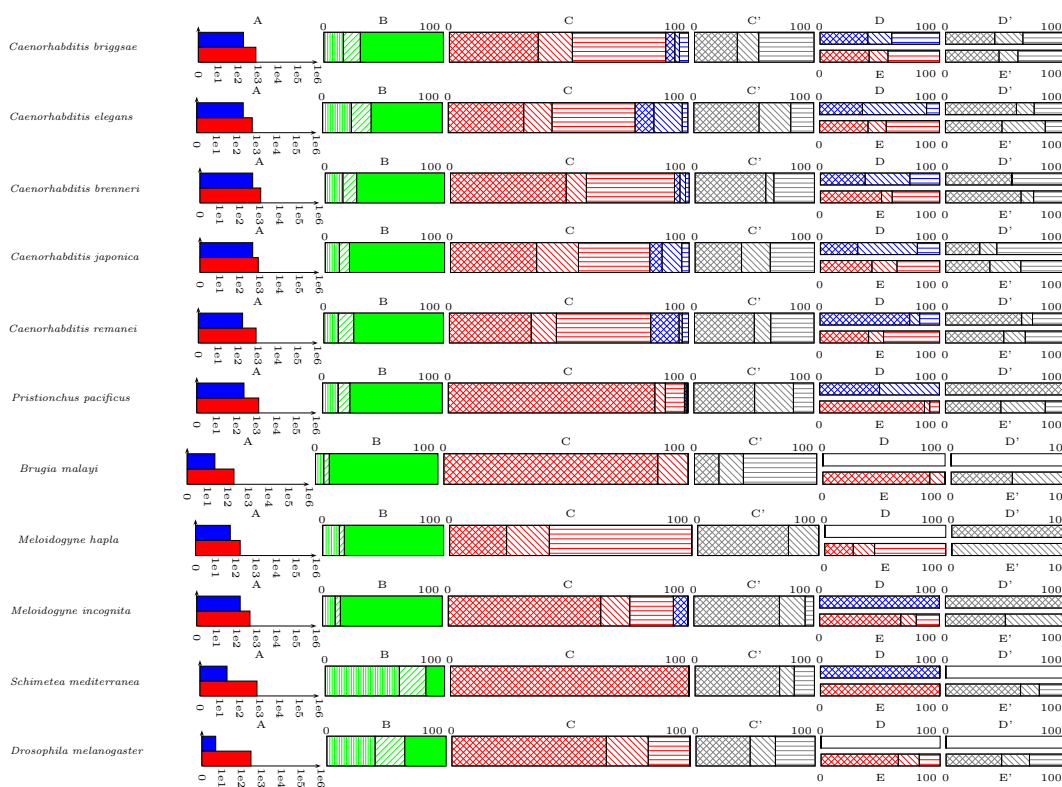


Figure 2.4: Summary of tRNA gene and tDNA statistics **A** Distribution of tRNA genes and tRNA pseudogenes

B Natural logarithm of the total of tRNA pseudogenes **C** tRNAs. **B** Fraction of tRNAs and tRNA pseudogenes in clusters: homogeneous clusters heterogeneous clusters not in clusters. **C** Fraction of Homogeneous pairs: $\rightarrow\rightarrow$, $\rightarrow\leftarrow$, $\leftarrow\rightarrow$: tRNA pairs. $\rightarrow\rightarrow$, $\rightarrow\leftarrow$, $\leftarrow\rightarrow$: tRNA pseudogenes pairs. **C'** Fraction of Heterogeneous pairs: $\rightarrow\rightarrow$, $\rightarrow\leftarrow$, $\leftarrow\rightarrow$. In D and D' and E and E' holds the same rules as C and C' but the raw data is the result of filtering tRNAs or tRNAs pseudogenes respectively. D and D' fraction for pairs of tRNA pseudogenes and E and E' fraction for tRNAs.

Based on a statistical evaluation of the distances between adjacent tDNAs (see Methods) and Fig. 2.5, we define two tDNAs to be clustered in the genome if they are located within 1000nt.

Not surprisingly, in species with small tDNA copy number, clusters typically are rare. In *Trichoplax adherens*, for instance, all tDNAs are isolated. There is no clear-cut relation between tDNA copy number and clustering, however. In *Nematostella vectensis* 89% of the tDNAs appear in clusters. In mammals, which have even larger tDNA copy numbers, less than a quarter of the tDNAs appear in clusters. Again, there do not appear to be any large-scale phylogenetic regularities. In teleost fishes, for example, the stickleback *Gasterosteus aculeatus* has 87% clustered tDNAs, in zebrafish this number reaches 65%. On the other hand, pufferfishes and medaka (*Oryzias latipes*) have predominantly isolated tDNAs. Similarly, large variation appears in other clades, see Figs. 2.1 and [4]. Higher primates

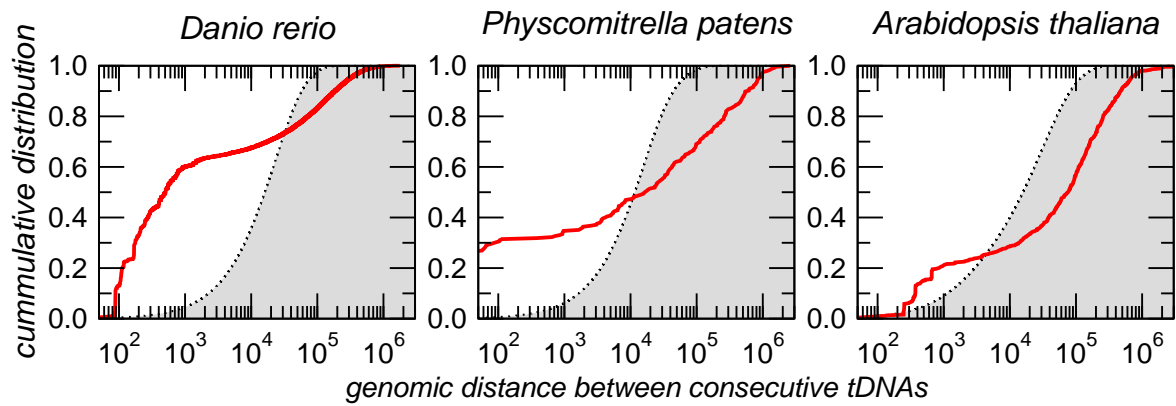


Figure 2.5: Cumulative distribution of tDNA pairs distances. Measured data are shown in red, random expectation from randomly placed tDNAs are shown as gray background. At a distance of 1000nt the vast majority of clusters cannot be explained by the random background.

have 17% to 36% of their tDNAs in clusters, with the exception of the bushbaby *Otolemur garnettii*, with only 5.6% of its 45225 tDNAs located in clusters. In plants there are also no clear regularities. The fraction of clustered tDNAs stays below 25% in Spermatophyta, while the chlorophyceae *Volvox carteri* and *Chlamydomonas reinhardtii* have 41% and 56% of their tDNAs localized in genomic clusters.

Most tDNA clusters are small, containing only a few co-localized tRNA genes. Typically, the frequency of larger clusters quickly decreases, at least approximately following an exponential distribution. This is particularly obvious in the case of mammals and drosophilids. In some cases, however, longer clusters are more abundant. Exceptionally large tDNA gene clusters, with fifty and more members, are observed for example in *Nematostella* and in the genomes of teleost fishes, Fig. 2.3.

The internal structure of tDNA clusters also differs widely between lineages. Fig. 2.1 and Appendix C, summarize the relative abundances of homogeneous and heterogeneous clusters, respectively. More precisely, we record the fraction of adjacent tDNA pairs coding for the same amino acid. While *Tetrahymena*, *Monosiga*, and the drosophilids exhibit mostly homogeneous pairs, we observe mostly heterogeneous pairs in kinetoplastids, *Nematostella*, clawed frog, and zebrafish, see Fig. 2.6 for an example.

In order to further investigate the structure of heterogeneous clusters we determined how often combinations two isoacceptor families appear in adjacent pairs. These data are conveniently represented in triangular matrices such as those in Fig. 2.7. Homogeneous clusters populate the main diagonal, whereas heterogeneous pairs are represented by off-diagonal entries. As for other features of the genomic tRNA distribution there are neither strong common patterns among all organisms investigated, nor are there systematic phylogenetic patterns. While *Monosiga*, for example, has almost exclusively homogeneous pairs, other species exhibit a wide variety of heterogeneous pairs. In *Danio*, for instance, K-N, K-S, and R-T are most frequent. In the cow genome, many clusters involve tRNA pseudogenes, which are much less prevalent in the other three examples. In the cow, C-C

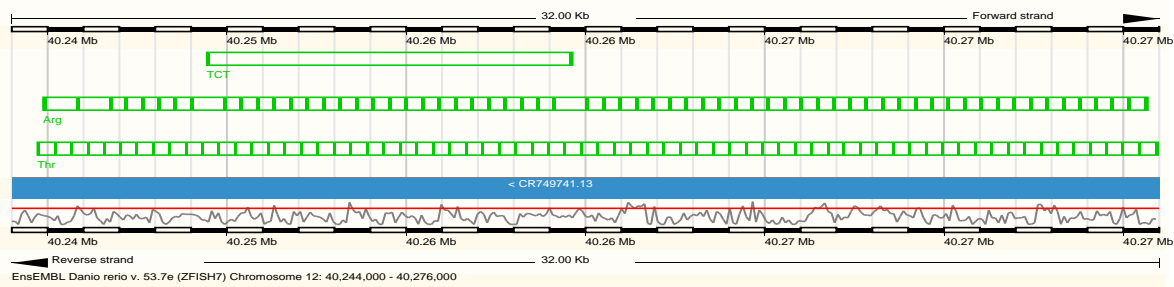


Figure 2.6: Example of heterogeneous tDNA cluster consisting of multiple copies of tRNA-Arg(TCT) and tRNA-Thr(AGT or TGT). Two tRNA pseudogenes with anti-codon TCT are interspersed.

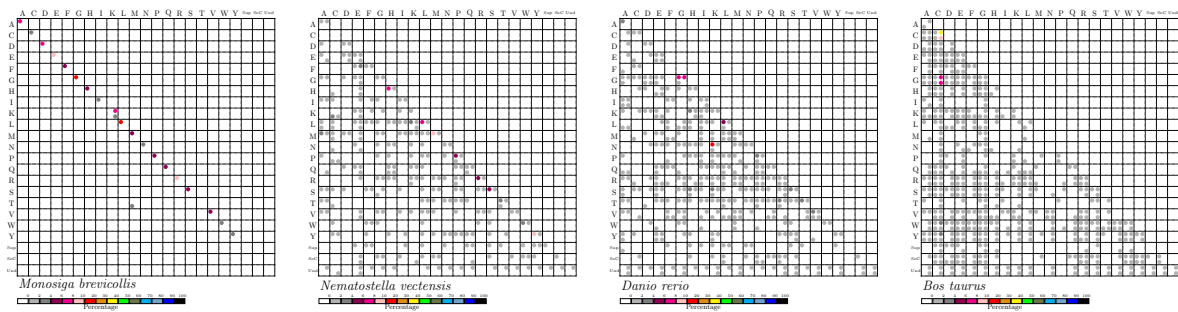


Figure 2.7: Relative abundance tRNA isoacceptor families located consecutively within tRNA clusters. Four data points are shown for each combination of amino acids: Top: pairs in the same reading direction; below: pairs in opposite reading direction. Left: pairs of presumably functional tRNA, right: pairs of tRNA pseudogenes. The last three rows and columns refer to putative Suppressor, SeC, and tRNA pseudogenes of undetermined isoacceptor class, resp.

pseudogenes account for more than 30% of the pairs.

A comprehensive collection of co-occurrence tables is provided in [4]. See in Appendix B an example for *Drosophila* genus. Not surprisingly, there is a general trend towards more complex co-occurrence matrices for species with larger numbers of tDNAs.

Most adjacent tDNA pairs in both homogeneous and heterogeneous clusters have parallel orientation. If the arrangements were random, we would expect that 50% of pairs are of this type. In many cases, e.g. *Arabidopsis*, *Selaginella*, *Xenopus*, or *Danio*, nearly all pairs are in parallel. Among the anti-parallel pairs, some species have a strong bias for either head-to-head (e.g. primates, and *Cryptococcus*) or tail-to-tail arrangements (*Oryza* and *Caenorhabditis*). Even within primates, the ratio of head-to-head and tail-to-tail pairs varies considerably.

In most species with very large tDNA copy numbers we can expect some tDNA clusters to appear by chance. We tested this by randomizing the tDNA locations (see Methods for details). The results for eutherian mammals are compiled in Tab. 2.1 and in the Appendix C, a full list of random pair configuration is given in [4]. In most genomes, there are significantly more tDNA pairs than expected, suggesting a mode of tDNA evolution of

favours the formation of local clusters. Local DNA duplications, also underlying the copy number variations within many populations (see e.g. [133, 138] and the references therein), are of course the prime suspects.

We observe significant under-representations of tDNA pairs only in a few species with very high tDNA counts: *Dasypus novemcinctus*, *Felis catus*, and *Loxodonta africana*. At present, we have no biological explanation for this observation. See Tab. 2.1

Table 2.1: Comparison of observed and expected number of tRNA pairs. The expectation values are computed by placing the tRNAs at uniformly random position in the genome. Empirical p values are computed from 50 to 1000 replicates.

Species	Observed	Expected	p-value
<i>B. taurus</i>	28452	22790	0
<i>C. familiaris</i>	4858	4271,27	0
<i>D. novemcinctus</i>	7918	11498,56	1
<i>M. domestica</i>	7402	914	0
<i>E. telfairi</i>	49	9.35	0
<i>E. caballus</i>	72	4.42	0
<i>F. catus</i>	8792	11816.7	1
<i>G. gorilla</i>	40	0.08	0
<i>H. sapiens</i>	97	0.27	0
<i>L. africana</i>	1645	3553.2	1
<i>M. mulata</i>	168	0.23	0
<i>M. murinus</i>	42	0.06	0
<i>M. musculus</i>	1001	425.5	0
<i>O. anatinus</i>	27015	25008	0
<i>O. lemur</i>	1364	1285	0
<i>P. troglodytes</i>	78	0.25	0
<i>P. pygmaeus</i>	83	0.28	0
<i>O. cuniculus</i>	118	37.15	0
<i>R. rattus</i>	28198	16148	0

Clusters of tDNAs have been implicated in interfering with the DNA replication forks [35]. The tDNA clusters might thus be instrumental in orchestrating the timing of DNA replication. On the other hand, replication fork pause sites are associated with genomic instability [93, 5, 31, 82] and hence may contribute to the rapid evolution of these tDNA clusters. Furthermore, retrotransposable elements tend to select tRNA genes as chromosomal integration sites [20], apparently in order to avoid gene disruptions upon retrotransposition. A recent comparison of yeast genomes associated genomic rearrangements, losses, and additions with tRNA genes [59]. Taken together, tDNA clusters thus appear as highly dynamic unstable genomic regions.

Synteny

Transfer RNAs have been reported to behave similar to repetitive elements as far as their genomic mobility is concerned. They appear to evolve via a rapid duplication-deletion mechanism that ensures that copies of tRNA genes within a genome are usually more similar to each other than tRNA gene of different species [187, 186]. In *E. coli*, for example, the

Teleosteo	Ho-pairs raw	Ho-pairs filtered	He-pairs raw	He-pairs filtered	P-value	ODDS
<i>D. rerio</i>	5124	5895	9543	7743	$1E+4$	0.71
<i>T. rubripes</i>	93	94	38	29	0.39	0.76
<i>G. aculeatus</i>	1987	2033	1157	1027	0.01	0.87
<i>O. latipes</i>	166	169	120	96	0.19	0.79
<i>T. nigroviridis</i>	66	66	51	40	0.41	0.79

Table 2.2: Fisher test results for Teleosteos species. Ho-pair and He-pair for Homogeneous and Heterogenous pair total counts. Raw without filter. P-value for independence test. ODDS of proportions.

rate of tRNA gene duplication/deletion events has been estimated to be about one event every 1.5 million years [187]. We are not aware of (semi)-quantitative estimates from eukaryotes. Our analysis is consistent with this mechanism (see below).

Since tRNA genes with the same anticodon are typically nearly identical, the only way to estimate rates of tRNA gene turnover is to determine, for each tRNA-bearing locus, whether tDNAs can be found in a syntenic locations in evolutionarily related species. We have determined such data here for eight selected species, including six mammals, namely the Catarrhini *Homo sapiens*, *Pan troglodytes*, *Pongo pygmaeus*, and *Macaca mulatta*, the rodent *Mus musculus*, and the Marsupialia *Monodelphis domestica*. The data set includes also more distant vertebrates *Gallus gallus* and *Xenopus tropicalis* to investigate whether there are tDNAs with very stable genomic locations.

Tab. 2.3 shows the results for the one- and two-side linkage analysis (see Section Methods). The number of related synteny regions based on the single-side linkage analysis is significantly higher than the region number created by the two-side linkage analysis. Since the latter analysis approach is more restrictive, the results between both analysis approaches also differ. While synteny regions in related species are mostly assigned by the single-side linkage analysis, the results of the two-side linkage analysis are more differentiated. Therefore, we discuss only the results of two-side linkage analysis in the following.

Within Catarrhini, tDNA locations are quite well conserved. For instance 80% (394/493) of human tDNA regions are conserved in the chimp, and there are still 63% (284/450) of the rhesus tDNA locations recovered in chimp. Somewhat surprisingly, there is also a large fraction of syntenic loci between mouse and opossum (80% [19,466/24,352] of the mouse loci and 76% [16,634 of 21,810] of the opossum loci). We suspect that the large fraction is confounded by the large overall number of tDNA loci and the rather larger intervals of five flanking genes used to define synteny, which taken together cover a substantial fraction of the genome. A second group of comparisons identified only a small number of syntenically conserved loci. Asymmetric results, which large retention in one direction is observed when the tDNA numbers are dramatically different. This concerns the comparisons between Catarrhini, on the one hand, and opossum and mouse on the other hand. Between frog and Catarrhini, finally, there is only a small number of syntenically conserved tDNAs.

We also analyzed the tDNA mobility in two invertebrate clades, drosophilids and nematode genus *Caenorhabditis*. Within these nematodes, we observe a rather high degree of syntenic conservation, ranging from 45% between *C. elegans* and *C. japonica* up to 84% for the most closely related pair *C. remanei* and *C. brenneri*. In general, conservation levels are consistent to the known phylogeny of the *Caenorhabditis* species [90]. For the genus

Table 2.3: Quantity structure of linkage analysis results in vertebrates: The upper right triangle quantifies the single-sided linkage results whereas the lower left triangle represents the number of two-sided linkage analysis results.

		Homo Sapiens	Pan Troglodytes	Pongo Pygmaeus	Macaca Mulatta	Mus Musculus	Monodelphis Domestica	Gallus Gallus	Xenopus Tropicalis						
Homo Sapiens	493		22,312	20,143	17,154	453,512	167,537	985	6341						
		444	494	438	488	438	387	442	22,206	442	20,398	366	155	383	332
		0.9	0.98	0.89	0.95	0.89	0.86	0.9	0.91	0.9	0.94	0.74	0.93	0.78	0.57
Pan Troglodytes	504	8,641		22,375	17,073	176,153	182,022	1,048	6,201						
		366	394	497	503	498	391	504	22,963	503	20,847	400	160	395	364
		0.73	0.8	0.99	0.98	0.99	0.87	1	0.94	1	0.96	0.79	0.96	0.78	0.62
Pongo Pygmaeus	512	8,673	7,556		16,838	179,360	183,128	1,033	6,585						
		349	368	330	375	494	390	512	22,716	512	20,797	411	158	450	348
		0.68	0.75	0.64	0.74	0.96	0.87	1	0.93	1	0.95	0.8	0.95	0.88	0.58
Macaca Mulatta	450	6,301	5,881	5,488		152,984	152,619	909	5,550						
		309	368	284	363	286	333	393	22,588	393	20,646	332	156	355	347
		0.69	0.75	0.63	0.72	0.64	0.65	0.87	0.93	0.87	0.95	0.74	0.94	0.79	0.58
Mus Musculus	24352	6,212	5,958	6,289	5,151		10,073,201	65,044	106,441						
		2,030	382	2,294	369	2,126	395	2,211	351	24,336	21,809	20,643	166	20,716	422
		0.08	0.77	0.09	0.73	0.09	0.77	0.09	0.78	1	1	0.85	1	0.85	0.72
Monodelphis Domestica	21810	3,395	3,766	3,677	4,017	190,815		67,383	106,233						
		1,750	318	2,071	363	1,846	35	2,123	353	16,634	19,466	19,290	166	19,306	416
		0.08	0.65	0.09	0.72	0.08	0.07	0.1	0.78	0.76	0.8	0.88	1	0.89	0.71
Gallus Gallus	166	44	46	43	42	1,398	1,560		569						
		38	38	38	39	35	38	32	38	132	1,169	130	1276	142	236
		0.23	0.08	0.23	0.08	0.21	0.07	0.19	0.08	0.8	0.05	0.78	0.06	0.86	0.4
Xenopus Tropicalis	586	77	74	79	59	912	802	16							
		24	72	25	69	23	65	26	53	126	708	115	630	16	12
		0.04	0.15	0.04	0.14	0.04	0.13	0.04	0.12	0.22	0.03	0.2	0.03	0.03	0.07

Each table entry is organized as follows

associations
dom # ran
cov(d) cov(r)

The top row lists the number of synteny associations; # dom and # ran are the sizes of domain and range, i.e., the numbers of tDNAs in the two species. Below the coverage, i.e., the fraction of syntenically conserved tDNAs in the two species, is indicated.

Table 2.4: Syntenic conservation of tDNAs: The table shows the fraction of tRNA loci between pairs of species. Every field contains the fraction of tDNAs of the species in the column, for which we could find a syntenic position in the row species.

	tDNA	<i>C. briggsae</i>	<i>C. remanei</i>	<i>C. brenneri</i>	<i>C. elegans</i>	<i>C. japonica</i>
<i>C. briggsae</i>	958	-	0.84 809	0.82 788	0.72 691	0.68 647
<i>C. remanei</i>	958	0.74 712	-	0.73 696	0.63 603	0.55 528
<i>C. brenneri</i>	1587	0.61 962	0.63 997	-	0.49 783	0.48 763
<i>C. elegans</i>	820	0.77 629	0.75 617	0.73 602	-	0.68 558
<i>C. japonica</i>	1307	0.46 607	0.48 633	0.49 634	0.45 589	-

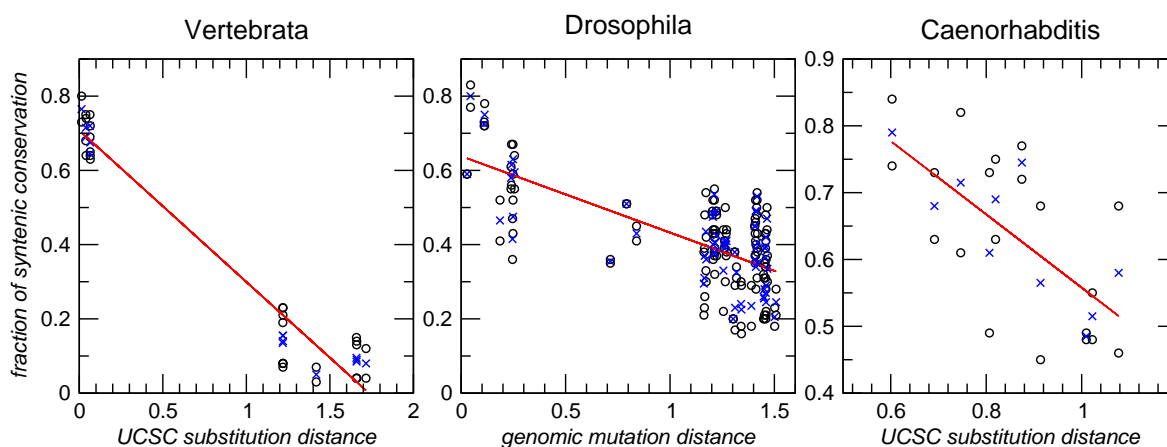


Figure 2.8: Correlation of syntenic conservation of tDNA loci with genomic distance. Estimates for each pairwise comparisons (\circ) and averages over the two comparisons for each pair of species (\times) are shown. For vertebrates and nematodes distances were extracted from trees provided through the UCSC browser, for Drosophilds, corrected mutation distances were used (see Methods for details). Because of the large number of tDNA loci *Mus musculus* and *Monodelphis domestica* were not used for the correlation.

Clade	ρ	slope
Vertebrates	-0.968	-0.41 \pm 0.02
Drosophila	-0.678	-0.21 \pm 0.02
Caenorhabditis	-0.638	-0.55 \pm 0.16

Drosophila with the twelve common representatives, on the other hand, there is much less syntenic conservation. The lowest value is 17% (*D. wilstoni* and *D. persimilis*). The best conserved tDNA arrangements are observed between the two closely related species *D. simulans* and *D. sechellia* with 78%. On average, the percentage of conservation is just around 50% or less. Full data are shown in Tab. 2.4 for nematodes and in [4] for Drosophila.

The sequence conservation of syntenically conserved tRNAs is consistent with the duplication/deletion mechanisms. In [4] shows a neighbor-joining tree of the tRNA-Ala sequences of nematodes, which includes also a few additional species that are not part of the genome-wide survey. We find that syntenically conserved tRNAs genes are typically conserved with an identical sequence across species, even though some tRNAs with the same anticodon located elsewhere in the genome show small sequence variations.

The fraction of syntenically conserved tDNAs correlates with the divergence of the genomes at sequence level, Fig. 2.8. The correlation is significant even though the data is rather noisy, a fact that can be explained at least in part by the unavoidable artifacts resulting from our approach. Utilizing annotation data directly to determine local synteny is problematic, for instance, near members of very large recently duplicated gene families. In principle, syntenic conservation could be inferred more accurately from genome-wide alignments. Since tDNAs are treated like repetitive elements in the currently available

pipelines, this strategy cannot be employed in practice. Nevertheless, the method provides at least a crude estimate of the tDNA turnover rate, indicating the tDNAs are relocated at time-scales only 2-5 times slower than background mutation rate, i.e., at an evolutionary distance of 1 mutation per site, 20% to 60% of the tDNAs have been deleted or relocated in one lineage.

These values should be regarded as upper bounds of syntenic conservation, i.e., tDNA turnover is probably even faster. For example, the identity of the tDNA (i.e., its anticodon) was not used in the analysis. Despite of the high mobility of tDNAs there are some ancient conserved loci. We further investigated two of the 77 syntenic loci conserved between *Xenopus* and *Human* in which tDNAs with the same anticodon were retained. Manual inspection of the flanking protein coding genes confirmed synteny. Neither locus is syntenically conserved in stickleback, lamprey or lancet, however.

Codon bias usage in Platyhelminth

The variation of tRNA gene numbers might also be explained by the codon usage-tRNA optimization that has been fiercely debated. However, it is not clear whether codon usage drives tRNA evolution or *vice versa* and the codon bias specie-specific is still widely debated. A survey in bacterial genomes [69] has established a theory that describes the coevolution of codon usage with tRNA gene content however still remains to be evaluated whether this theory could explain the variations observed across Eukaryotic genomes. In this order of ideas we have used for comparison three free-living platyhelminth to search for codon usage bias.

Candidate tDNAs were predicted with tRNAscan-SE in the genomes of *Schistosoma mansoni*, *Schistosoma japonicum* and *Schmidtea mediterranea*. After removal of transposable element sequences (see below), tRNAscan predicted a total of 713 tRNAs for *S. mansoni* and 739 for *S. mediterranea*, while 154 tRNAs were found in the *S. japonicum* sequences. These included tRNAs encoding the standard 20 amino acids of the traditional genetic code, selenocysteine encoding tRNAs (tRNA^{sec}) [156] and possible suppressor tRNAs [8] in all three genomes. The tRNA^{sec} from schistosomes has been characterized, and is similar in both size and structure to tRNA^{sec} from other eukaryotes [78].

The tRNA complements of the three platyhelminth genomes are compared in detail in Fig. 2.9. The amino acids are represented in approximately equal numbers in *S. mansoni* and *Schmidtea*. Nevertheless, there are several notable deviations. *S. mansoni* contains many more leucine (86 vs. 46) and histidine (27 vs. 8) tRNAs, while serine (51 vs. 94), cysteine (21 vs. 44), methionine (21 vs. 44), and isoleucine (17 vs. 42) are underrepresented. In addition, there are several substantial differences in codon usage. In most cases, *S. mansoni* has a more diverse repertoire of tRNAs: tRNA-Asn-ATT, tRNA-Arg-CGC, tRNA-His-ATG, tRNA-Ile-GAT, tRNA-Pro-GGG, tRNA-Tyr-ATA, tRNA-Val-GAC are missing in *Schmidtea*. Only tRNA-Ser-ACT is present in *Schmidtea* but absent in *Schistosoma*. The tRNA complement of *S. japonicum*, on the other hand, differs strongly from its two relatives. Not only is the number of tRNAs decreased by more than a factor of four, *S. japonicum* also prefers anticodons that are absent or rare in its relatives, such as tRNA-Ala-GGC, tRNA-Cys-ACA, and Lys-CTT. On the other hand, no tRNA-Trp was found. Since the UGG codon is present in many open reading frames we interpret this a problem with the incompleteness of the

genome assembly rather than a genuine gene loss. The reduction in the number of tRNAs is also evident by comparing the number of tRNAs with introns: 27 in *S. mansoni* versus 5 in *S. japonicum*.

It has been shown recently that changes in codon usage, even while coding the same protein sequences, can severely attenuate the virulence of viral pathogens [24] by “de-optimizing” translational efficiency. This observation leads us to speculate that the greater diversity of the tRNA repertoire could be related to the selection pressures of the parasitic life-style of *S. mansoni*. The effect is not straight forward, however, because there is no significant correlation of tRNA copy numbers with the overall codon usage in both *S. mansoni* and *S. japonicum*, Fig. 2.9C. In contrast, a weak but statistically significant correlation can be observed in *Schmidtea mediterranea*. It would be interesting, therefore, to investigate in detail whether there are differences in codon usage of proteins that are highly expressed in different stages of *S. mansoni*’s life cycle, and whether the relative expression levels of tRNAs are under stage-specific regulation.

The most striking result of the tRNAscan-SE analysis was the initial finding of 1,135 glutamine tRNAs (Gln-tRNAs) in *S. mansoni* in contrast to the 8 Gln-tRNAs in *S. japonicum* and 65 in *S. mediterranea*. Nearly all of these (1,098 in *S. mansoni*) were tRNA-Gln-TTG. In addition, an extreme number of 1,824 tRNA-pseudogenes in *S. mansoni* (vs. 951 in *S. japonicum* and 19 in *S. mediterranea*) was predicted. Of these, 1,270 were also homologous to tRNA-Gln-TTG. These two groups of tRNA-Gln-TTG-derived genes (those predicted to be pseudogenes and those predicted to be functional tRNAs) totaled 2,368. These high numbers suggest a tRNA-derived mobile genetic element. We therefore ran the 2,368 *S. mansoni* tRNA-Gln-TTG genes through the RepeatMasker program [159]. Almost all of them (2,342) were classified as SINE elements. Further BLAST analysis revealed that these elements are similar to members of the Sm- α family of *S. mansoni* SINE elements [161]. Removal of these SINE-like elements yielded a total of 63 predicted glutamine-encoding tRNAs in *S. mansoni*. About 650 of 951 pseudogenes in *S. japonicum* derived from tRNA-Pro-CGG.

Homology-based analysis yielded similar, though somewhat less sensitive, results to those of tRNAscan-SE. For instance, a BLAST search in *S. mansoni* with Rfam’s tRNA consensus yielded 617 predicted tRNAs compared to the 663 predictions made by tRNAscan.

2.4 Conclusions

We have developed a pipeline based on tRNAscan-SE [111] to extract and analyze the locations of tRNA genes and pseudogenes of eukaryotic genomes. In our analysis, we focus not only on the number of tRNA genes, but also on their relative genomic locations, and in particular on the formation of tDNA clusters. Surprisingly, we found no distinctive clade-specific features or large scale trends, with the exception of the rather straightforward observation that the larger metazoan genomes typically tend to harbour large numbers of tDNAs.

In some species, large clusters of tDNAs occur. This effect has first been reported in *Entamoeba histolytica*. The origin of this gene organization in the genus *Entamoeba* clearly predates the common ancestor of the species investigated to date. Their function of the

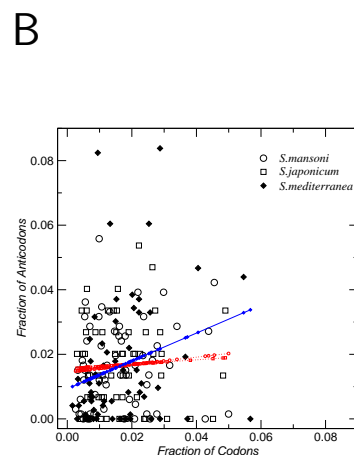
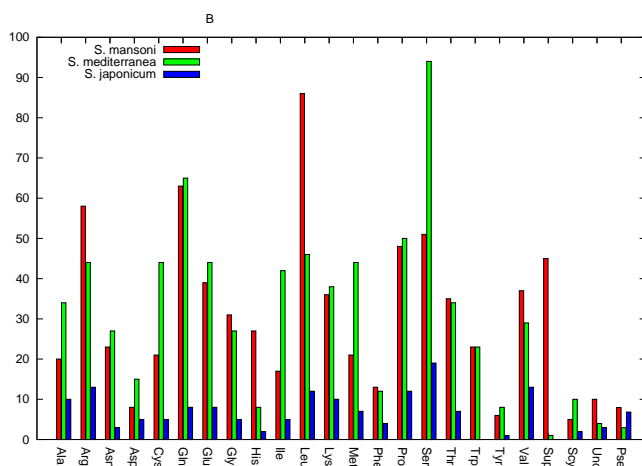
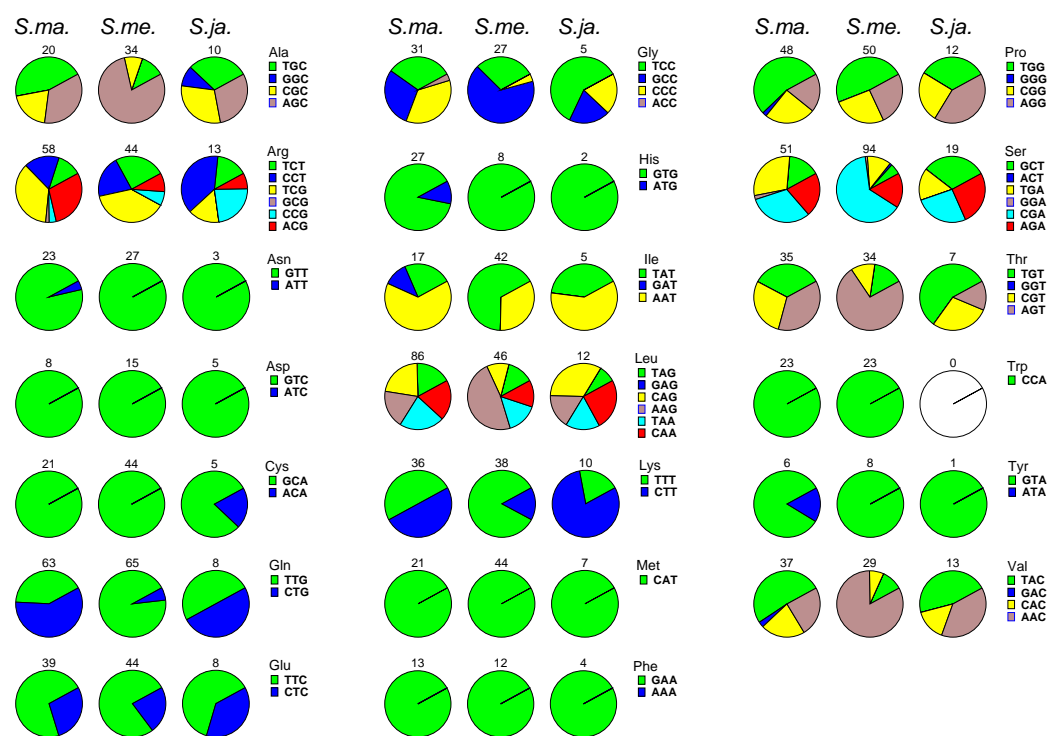


Figure 2.9: Comparison of the tRNA complement of *Schistosoma mansoni*, *Schistosoma japonicum*, and *Schmidtea mediterranea*. **A**: Comparison of anti-codon distributions for the 20 amino acids. Numbers below each pie-chart are the total number of tRNA genes coding the corresponding amino acid. Left columns: *S. mansoni*; middle columns: *S. mediterranea*; right columns: *S. japonicum*. **B**: Number of tRNAs encoding a particular amino acid. red: *S. mansoni*, blue: *S. japonicum*, green: *S. mediterranea*. Abbreviations: Sup: putative suppressor tRNAs (CTA, TTA); Scys: Selenocysteine tRNAs (TCA); Pseu: predicted pseudogenes; Und: tRNA predictions with uncertain anticodon; likely these are also tRNA pseudogenes. The Gln-tRNA derived repeat family (see text) is not included in these data. **C**: Comparison of codon usage and anti-codon abundance. No significant correlation is observed for the two schistosomes. For *S. mediterranea* there is a weak, but statistically significant, positive correlation: $t \approx 2.0$.

array-like structure remains unclear [177]. We report here that this phenomenon is not restricted to a particular clade of protists but rather appears independently in many times throughout eukaryotes.

In most eukaryotes, tRNAs are multi-copy genes with little or no distinction between paralogs so that orthology is hard to establish, in particular in the presence of tRNA gene clusters. As a consequence, the evolution of genomic tRNA arrangements is non-trivial to study over larger time-scales. Upper bounds on syntenic conservation can be estimated, however, by considering small sets of flanking protein coding genes for which homology information can be retrieved from existing annotation. We found that tRNAs change their genomic location at time-scales comparable to mutation rates: syntenic conservation fades at roughly the same evolutionary distances as sequence conservation in unconstrained regions.

The absence of large numbers of partially degraded tRNA copies in many of the investigated genomes provides a hint at the mechanisms of tRNA mobility: At least in part the relocation events appear to be linked to chromosomal rearrangements rather than mere duplication-deletion of the tRNA genes themselves. The latter mechanism, which appears to be prevalent e.g. in mitochondrial genomes [144], certainly also plays a role, since tRNA pseudogenes are readily observed in many species, as do tRNA retrogenes [183]. A link between tRNA loci, and in particular tRNA clusters, and chromosomal instability has been pointed out repeatedly in the literature, showing that tRNA genes can interfere with the replication forks [35, 93, 5, 31, 82]. The data collected here provide a basis to investigate this connection more systematically in the future.

Overall, the tRNA complement of Eukaryotes is highly dynamic part of the genomes whose organization evolves rapidly and in a highly lineage specific manner — a behavior that is in striking contrast to the extreme conservation of sequence and function of the tRNAs themselves.

Chapter 3

non-coding RNA identification from transcriptome data

3.1 Introduction

In recent years, several classes of small RNAs with a length of about 20 nt have been discovered. The most prominent of these are microRNAs (miRNAs), Piwi-interacting RNAs (piRNAs), and variants of endogenous small interfering RNAs (siRNAs) [122, 176]. In addition, small RNAs have been found to be associated with mRNA transcription start and stop sites [86, 169, 171]. Several studies have reported that well-known ncRNA loci are also processed to give rise to small RNAs. MicroRNA precursor hairpins, for instance, are frequently processed to produce additional “off-set RNAs” (moRNAs) that appear to function like mature miRNAs. These moRNAs were discovered in *Ciona intestinalis* [157], where they form an abundant class of processing products. At much lower expression levels, they can also be found in the human transcriptome [94]. Specific cleavage and processing of tRNAs has been observed in the fungus *Aspergillus fumigatus* [19] and later also found in human short read sequencing data [87]. Small nucleolar RNAs (snoRNAs) are also widely used as a source for specific miRNA-like short RNAs [44, 151, 170]. The same holds true for vault RNAs [162, 139].

The ENCODE project, focusing on high resolution on the analysis of 1% has also shown that RNA transcripts serve as a high source of new regulatory ncRNAs [14]. This focus by the ENCODE project should lead to future efforts to conduct studies to identify new ncRNAs in a variety of organisms and to elucidate their regulatory functions in the cell. Increasing attention directed to miRNAs as regulatory factors shaping cellular and organismal life, along with the discovery of many new miRNAs from deep sequencing data, has spurred the study of complete small ncRNA transcriptomes. These complete transcriptome studies have dated from those of kinetoplast mitochondria of *Leishmania tarentolae* and *Caenorhabditis elegans*, from which genomic features and new ncRNAs were reported as stem-bulge RNAs (sbRNAs) and snRNA-like RNAs (snlRNAs). These initial reports found that the majority of the *C. elegans* ncRNAs showed developmentally variable expression [30]. The same holds for *Aspergillus fumigatus*; an experimental screening of ncRNAs found that the majority were expressed under various growth conditions or during specific developmental stages [19]. In recent years, studies of this kind have extended to the investigation of small ncRNAs

associated with ribonucleoprotein particles (RNPs) from two different cellular systems and organisms: brain and HeLa cells from mouse and human, respectively [145]. Several current studies have focused on the detailed analysis of tRNA derived from transcriptome data. This topic will be addressed in the next chapter.

RNA quality control in Eukaryotes: source from which comes the transcriptome

Eukaryotic cells have numerous RNA quality controls that are important for shaping their transcriptomes [36]. The complexity of these systems implies a diversity of mechanisms involving nuclear and cytoplasmic RNAs that safeguard cells from abnormal mRNA function [81]. One example is messenger RNA surveillance systems, which monitor proper translation termination [6]. Another includes the many mechanisms that have evolved to degrade aberrant and nonfunctional RNAs. For instance, the exosome is a highly organized and regulated macromolecular machine that has only a few enzymatically active components and is present in both the nucleus and cytoplasm. The exosome continuously works to ensure adequate quantities and quality of RNAs by facilitating normal RNA processing and turnover, as well as by participating in more complex RNA quality-control mechanisms [112]. Thus, mRNA surveillance mechanisms play important roles both in depleting aberrant transcripts from cells and in maintaining the proper level of normal transcripts [6]. Ribosomes play a central role in some mRNA surveillance systems, but, for the production of the vast majority of ncRNAs in eukaryotes, ribosomes are not involved in quality control. One example of nuclear RNA surveillance for ncRNA is the tRNA surveillance pathway in *Saccharomyces cerevisiae*. This pathway utilizes polyadenylation to degrade hypomethylated tRNA_iMet through the function of a new poly(A) polymerase, Trf4p, along with the nuclear exosome [85]. The TRAMP (Trf4/Air2/Mtr4p Polyadenylation) pathway has recently also been shown to be associated with the degradation of rRNA and small nuclear/small nucleolar RNAs as well as the regulation of transcription from unannotated and/or silenced regions of the genome [17]. The topic of tRNA-associated degradation will be addressed further in the next chapter.

Mechanisms for stable RNA degradation under starvation conditions have been reported for bacteria [32] and some consequences of RNA quality control suggest that “normal” RNAs are subjected to degradation by quality control mechanisms. One key aspect of these pathways is that they could be seen as kinetic competitions between the normal rate of reaction in the life of an RNA and the quality-control event targeting the RNA for degradation. In short, the RNA processing creates a diverse pool of transcripts and only those that survive quality control accumulate to substantial levels [36].

The next section will summarize our efforts to define new features for use in the characterization and classification of three main classes of ncRNA (microRNA, snoRNA and tRNA) from a brain-tissue transcriptome library, based on deep-sequencing methods.

3.2 Methodology

Non-coding RNA identification and classification methods from transcriptome data Since next-generation sequencing technology is the current choice to determine content and class of ncRNA in the cell, several new methods have been developed that

contribute to the identification of ncRNAs. The first involves the mapping of transcriptome data to a reference genome. These tools include `segemehl` [72] a recently developed method based on a variant of enhanced suffix arrays that efficiently deals with both mismatches and insertions/deletions (indels). The authors of this method have introduced a matching model for short reads that can, in addition to mismatches, also cope with indels. `Segemehl` also addresses different types of error models encountered in transcriptomics. For example, it can handle the problem of leading and trailing contamination caused by primers and poly-A tails or length-dependent increases in error rates. In these contexts, it thus simplifies the tedious and error-prone trimming step. For efficient searches, this method utilizes index structures in the form of enhanced suffix arrays [72]. Another method, `MicroRsaizerS`, is a read mapping tool based on the rapid alignment of small RNA reads [43]. However, tools designed to detect or even to classify ncRNA are rare. The miRNA detection and analysis tool `miRanalyzer` [65] allows detection of all known miRNAs, finds all perfect matches against other libraries of transcribed sequences, and predicts new miRNAs. However, a comprehensive method for the classification of a complete set of known ncRNAs is not yet available. One tool that is available uses profiles of short sequence reads [84] to identify ncRNAs using two features derived from profiling data.

The dataset analyzed here was produced according to standard small RNA transcriptome sequencing protocols in the context of other projects and will be published in that context. In brief, total RNA was isolated from the frozen prefrontal cortex tissue using the TRIzol (Invitrogen, USA) protocol with no modifications. Low molecular weight RNA was isolated, ligated to the adapters, amplified, and sequenced following the Small RNA Preparation Protocol (Illumina, USA) with no modifications. All small RNAs, 17-28nt long, were mapped to the human genome (NCBI36.50 Release of July 2008) using `segemehl` [72], a method based on a variant of enhanced suffix arrays that efficiently deals with both mismatches as well as insertions and deletions. We required small RNAs to map with an accuracy of at least 80% thus only the best hit was selected. Reads mapping multiple times to the genome with an equivalent accuracy were discarded. After filtering the effective accuracy was $> 97\%$. Subsequently, all hits were sorted by their genomic position. Two reads were assigned to the same putative ncRNA locus, i.e. cluster, if separated by less than 100nt or 39nt. Clusters consisting of less than 10 reads were discarded because of their low information content.

Cluster definition The mapped reads were then sorted by genomic position. Two reads were assigned to the same putative ncRNA locus if they are separated by less than 39nt based on the following cluster definition:

Definition 1 : A string S is a read, i.e. is a sequence that in the data library comprises a number of times the read was sequenced. Every S has a length l , $17 \leq l \leq 28$, where $l = 17$ ($l = 28$) represent the minimum *min* (maximum *max*) string length. Since every S has a starting point a and an ending point b , then S is represented by (a, b) and $l = b - a + 1$.

Definition 2 : Given a sorted set of strings $\mathbf{C} = \{S_n\}_n$, ($S_n = (a_n, b_n)$) we define the distance δ_{n+1} between two consecutive strings S_n, S_{n+1} as $\delta_{n+1} = a_{n+1} - b_n$.

Definition 3 : A cluster is a sorted set of strings $\mathbf{C}=\{S_n\}_n$ (with at least 10 elements S_n) such that the distance δ_{n+1} satisfies $\delta_{n+1} \leq 39$ for each $S_n \in \mathbf{C}$. This bound 39 is obtained by adding 11 (the maximum difference between the length of two strings) and 28 the *max* length of a string. See Fig. 3.1 for a visual representation of the bound definition. For defining cluster including flanking regions of 100 nt the bound was increased to 100.

Blockbuster Once ncRNA loci were defined, we faced the problem of dividing consecutive reads into blocks to detect specific expression pattern. Note that this task is different from the segmentation of e.g. tiling array profiles [77] since we cannot *a priori* restrict ourselves to non-overlapping blocks. Due to biological variability and sequencing inaccuracies, the read arrangement does not always show exact block boundaries. We have developed the `blockbuster` tool that automatically recognizes blocks of reads. In the first step, a mapped read u with start and end positions a_u and b_u is replaced by a Gaussian density ρ_u with mean $\mu_u = (b_u + a_u)/2$ and variance σ_u^2 . We set $\sigma_u = s|(b_u - a_u)/2|$, where s is a parameter that is used to tune the resolution. For each locus, these gaussian densities are added up separately for the two reading directions. The resulting curves f^+ and f^- that exhibit pronounced but smooth peaks centered at blocks of reads with nearly identical midpoints, Fig. 3.10, middle panel. Now we use a greedy procedure to extract the reads that belong to the same block:

1. Determine the location \hat{x} of the highest peak.
Set $B = \emptyset$ and $\delta = 0$.
2. Include in the block B all reads u such that
 $\hat{x} \in [\mu_u - (\sigma_u + \delta), \mu_u + (\sigma_u + \delta)]$.
Set δ to the standard deviation of the $\mu_u, u \in B$ and repeat step (ii) until not further reads are included in u
3. Compute $f_B = \sum_{u \in B} \rho_u$, output B , remove the reads in B , and set $f \rightarrow f - f_B$.

This procedure iteratively extracts blocks in an order that intuitively corresponds to their importance, Fig. 3.10. Since the area under a peak equals the number of reads in the block the height of the peaks provides a meaningful trade-off between the coherence of a block and its expression level. We therefore suggest to use the height of the peak to define the stop condition for `blockbuster`. Here, we used $s = 0.5$, a value that requires blocks to be well separated to be recognized as distinct.

We remark that block-detection could alternatively be performed using Gaussian deconvolution approaches, which are commonly used e.g. in chromatography [182] and many areas of spectroscopy. For the present application, the additional computational efforts do not seem justified, however. Furthermore, we still would need a heuristic to associate individual reads to peaks.

Classification and description based on Random Forests approach The Random Forest approach is a machine learning approach developed by Leo Breiman [15] and Adele Cutler. Random Forests [®] is the original implementation of the algorithm. The Random Forest approach has also been implemented in WEKA [3]. Random Forests, a meta-learner comprised of many individual trees, was designed to operate quickly over large datasets and, more importantly, to be diverse, by using random samples to build each tree in the forest [136]. In short, Random Forests is a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution as the other trees in the forest. Generalization errors depend on how large the forests become, the strength of the individual trees in the forest, and the correlation between them [15]. Random Forests grows many classification trees. To classify a new object from an input vector, the input vector is added to each of the trees in the forest. Each tree presents a classification, and we say that the tree "votes" for that class. The forest chooses the classification having the most votes (out of all the trees in the forest). Each tree grows as follows: if N is the number of cases in the training set, sample N cases at random - but with replacement, from the original data. This sample will be the training set for growing the tree. Then, if there are M input variables, a number $m \ll M$ is specified such that at each node, m variables are selected at random out of the M and the best split on these m is used to split the node. The value of m is held constant during the forest growing, then each tree is grown to the largest extent possible. There is no pruning [2].

Two data objects generated by random forests are important for the classification, oob (out-of-bag) and proximities. When the training set for the current tree is drawn by sampling with replacement, about one-third of the cases are left out of the sample. This oob data is used to get a running unbiased estimate of the classification error as trees are added to the forest [15, 2]. It is also used to get estimates of variable importance. After each tree is built, all of the data are run down the tree, and proximities are computed for each pair of cases. If two cases occupy the same terminal node, their proximity is increased by one. At the end of the run, the proximities are normalized by dividing by the number of trees. Proximities are used in replacing missing data, locating outliers, and producing illuminating low-dimensional views of the data. Prototypes are a way of getting a picture of how the variables relate to the classification. The final output of prototypes showed at the screen, for continuous variables are the result of standardize by subtracting the 5th percentile and dividing by the difference between the 95th and 5th percentiles. In words of Breiman [2] "For the j th class, the model find the case that has the largest number of class j cases among its k nearest neighbors, determined using the proximities. Among these k cases the median, 25th percentile, and 75th percentile for each variable is predicted. The medians are the prototype for class j and the quartiles give an estimate of its stability. For the second prototype, we repeat the procedure but only consider cases that are not among the original k , and so on".

In order to investigate whether the short reads patterns carry information on the particular ncRNA class from which they originate, we selected three distinct ncRNA classes tRNAs ($n = 87$), miRNAs ($n = 218$) and snoRNAs ($n = 129$) and set up a machine learning approach based on the WEKA [188] implementation of the Random Forest learning scheme [188, 15] with the number trees set to 100.

To define input data we have defined descriptors that represent features from the transcriptome profile. Based on a visual inspection of the mapped reads, ten features were selected to train the random forest model:

- Number of blocks within a cluster (*blocks*),
- Length of a cluster (*length*).
- Number of nucleotides covered by at least two blocks (*nt overlap*),
- Number of overlapping blocks (*block overlap*).
- Maximum and minimum block height in a cluster (*max and min block height*).
- Mean block height in a cluster (*mean block height*).
- Maximum and minimum distance between consecutive blocks (*max and min distance*).
- Mean distance between consecutive blocks (*mean distance*).

Two different training sets were built by randomly sampled of the original data set. The Random Forest was training with training sets of size 250 and 150.

3.3 Results

Cluster detection By cluster definition, two consecutive reads were assigned to the same putative cluster if they are separated by less than 39nt. See in methods cluster definition. This bound was defined based on the maximum *max* and minimum *min* read length. See Fig. 3.1 for schematic representations. In order to further investigate the hidden structure of the brain-tissue transcriptome library we determined how often clusters appear into some specific range of height (number of reads), distances of adjacent clusters and Cluster length.

For our purpose, we have explored the robustness of the mapping of *segemehl* [72]. this tool uses a variant of enhanced suffix arrays that efficiently deals with both mismatches as well as insertions and deletions. Thus it is expected that the structure of the mapped library varies when the number of allowed operations increased, when allowing more deletions and/or mismatches and/or insertions. Therefore we explore the library structure for K0, K1 and K4 *segemehl* operations i.e. 0 (for not operations), 1 (at least one operation) and 4 (four operations).

We observe that the number of clusters increased as the *segemehl* operations increase $K0 < K1 < K4$. See Tab. 3.1 and Fig. 3.2. The more abundant height is 2 to 10 reads corresponding to $\approx 90\%$ of clusters. Only the 6% of the clusters gather reads that can be comparable with some intensity signal of experimental RNA expression and one small fraction ($\approx 0.25\%$) of clusters comprise more that 50 reads. Then about ≈ 7240 clusters might correspond to annotated loci and to new loci candidates.

We also counted how often two consecutive clusters are located to some thresholds of distance, see Tab. 3.2 and Fig. 3.2. Some clusters are overlapping in opposite sense

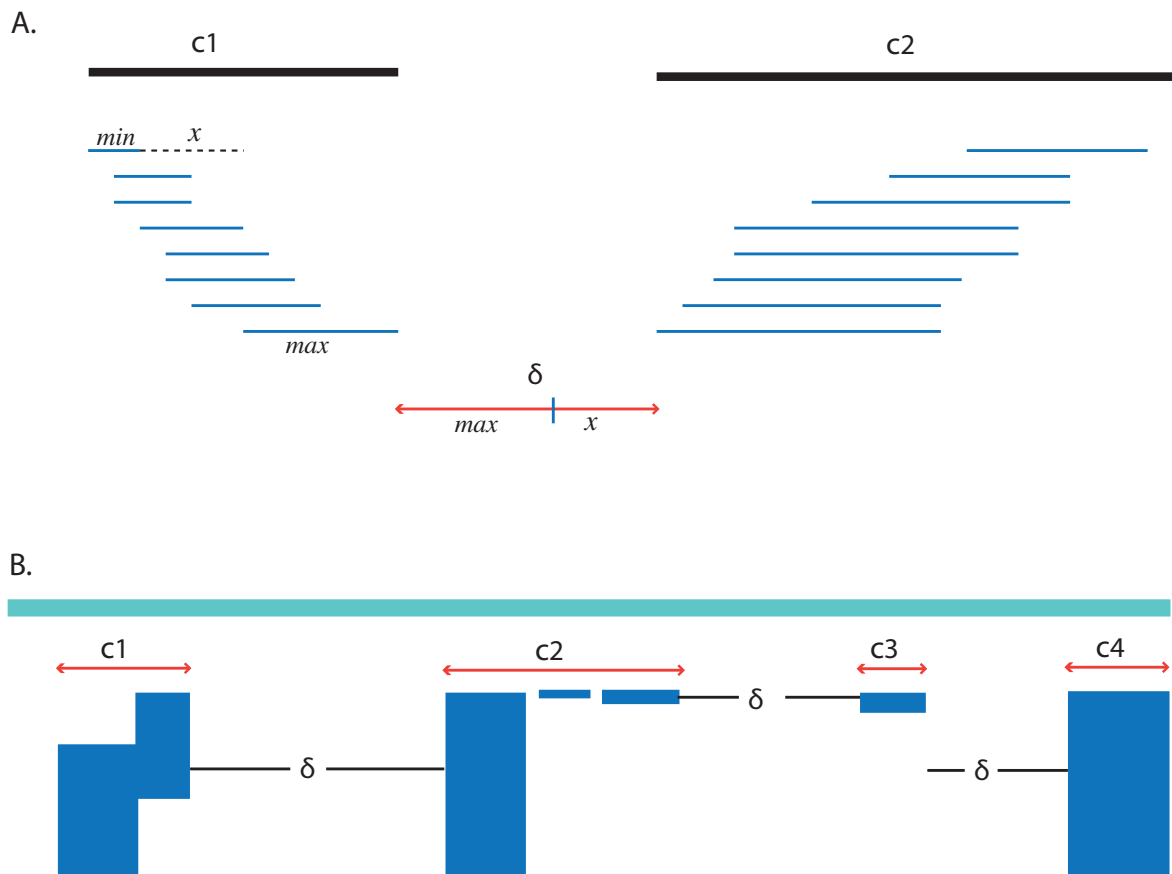


Figure 3.1: A. Bound definition δ . This bound 39 is obtained by adding 11 (x the maximum difference between the length of two strings) and 28 the max length of a string. B. Different patterns of blocks. δ is the bound determining whether a read belongs to a cluster or not. C1, C2 and C3 are different cluster patterns.

(negative distance) $\approx 4-5\%$ or are adjacent (0 distance) or are colocated in a distance over 100nt. Cluster overlapping in negative sense might correspond to loci that probably regulate on *trans*. About 13% of the cluster are located on a boundary less than 100 nt. The more frequently observed length was from 16 to 49 nt, followed from 50 to 99nt. The longest cluster corresponds to a repetitive motive region.

Since we required small RNAs to map with an accuracy of at least 80% , only the best hit was selected. Reads mapping multiple times to the genome with an equivalent accuracy were discarded. After filtering the effective accuracy was increased to 97%. Finally, to assign clusters or co-located clusters to a locus we increased the bound to 100 nt. Clusters consisting of less than 10 reads were discarded. Once clusters were annotated and linked, 434 of 852 clusters were found within regions of annotated miRNA, tRNA and snoRNA loci. See Tab. 3.4

Operations	Cluster size: read numbers		
	2 to 10	11 to 49	Over 50
K0	93.7	6.05	0.25
K1	93.54	6.1	0.35
K4	92.71	6.78	0.51

Table 3.1: Fractions of clusters and size of the cluster

Operations	Cluster distance in nucleotides					
	<0	0	1 to 49	50 to 99	100 to 199	Over 200
K0	5.25	0.17	6.17	7.92	9.51	70.98
K1	5.23	0.17	6.12	7.85	9.48	71.16
K4	4.03	0.14	4.98	6.56	8.56	75.73

Table 3.2: Fractions of Clusters and their distance locations.

Operations	Cluster length in nucleotides				
	16 to 49	50 to 99	100 to 199	200 to 299	Over 300
K0	87.04	11.59	1.31	0.06	0.004
K1	86.88	11.7	1.36	0.07	0.005
K4	88.77	10.04	1.12	0.06	0.01

Table 3.3: Fraction of Clusters length.

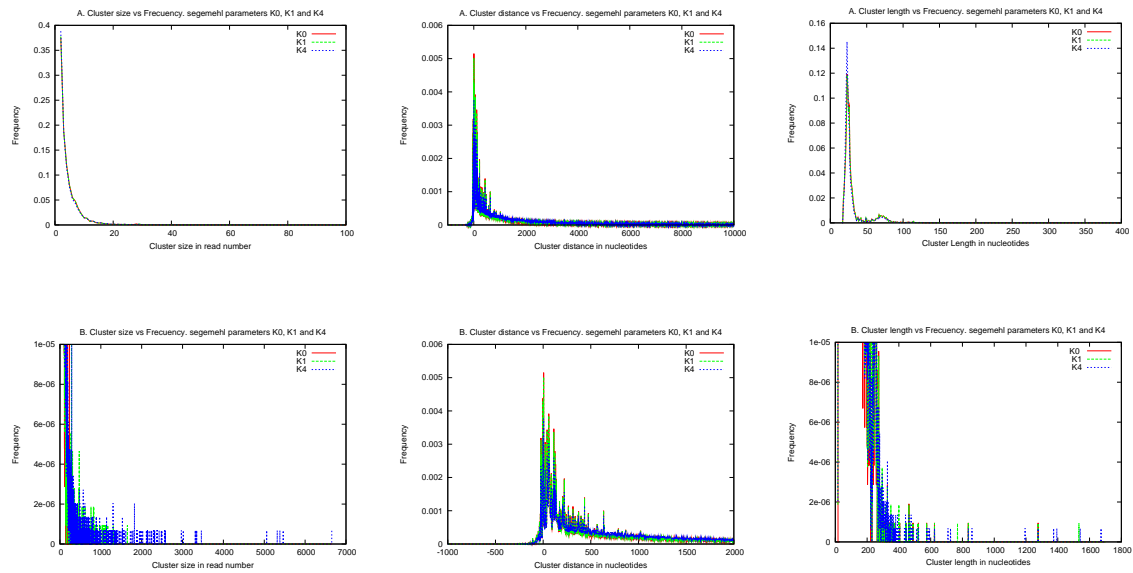


Figure 3.2: Distribution of frequency of clusters by size, distance and length. In B zoom of A on a specific range

ncRNA characterization from the Brain library We have recently developed the tool `blockbuster` [94] to simplify the task of identify pattern of block expression in genome-wide analyzes. The program merges mapped HTS reads into *blocks* based on their location in the reference genome (Fig. 3.3a). After the assembly of blocks, specific block patterns for several ncRNA classes can be observed. For example, miRNAs typically show 2 blocks corresponding to the miR and miR* positions (Fig. 3.3b). A similar processing can be observed for snoRNAs (Fig. 3.3c). On the other hand, tRNAs show more complex block patterns with several overlapping blocks (Fig. 3.3d).

Here, we used a width parameter of $s = 0.5$, a value that requires blocks to be well separated to be recognized as distinct. We required a cluster to have at least 2 blocks. In the following we refer to the number of reads comprised in a block as the *block height*. Using the cluster definition we identified 852 clusters across the whole human genome. By using `blockbuster` 2,538 individual blocks and 85,459 unique reads were identified. 434 clusters were found within annotated ncRNA loci [miRBase v12 (727 entries), tRNAscan-SE (588 entries) and snoRNAbase v3 (451 entries)], see Tab. 3.4.

We then computed secondary structures (using `RNAfold` [71]) to assess the relationship of reads and structure. For each read, the base pairing probabilities were calculated for the sequences composed of the read itself and 50nt of flanking region both up- and downstream. These data were also collected separately for reads found within annotated miRNA, tRNA, and snoRNA loci, respectively.

Little is known, however, about the mechanisms of these processing steps and their regulation. Here, we show that the production of short RNAs is correlated with RNA secondary structure and therefore exhibits features that are characteristic for individual ncRNA classes. The specific patterns of mapped HTS reads thus may be suitable to identify and classify the ncRNAs from which they are processed. We explore here to what extent such an approach is feasible in practice. The 5'-ends of reads arising from known snoRNAs preferentially map just upstream of the C- and ACA-boxes. This indicates the correlation of mapping patterns with processing steps and thus with structural properties of snoRNAs (Fig. 3.5). Based on earlier findings that miRNA-like products are derived from snoRNAs [170] and the observation that miRNA transcripts tend to have higher blocks (Tab. 3.4), the two peaks shown in the Figure 3.5 (left) probably represent small RNAs produced from the 5'- and 3'-hairpins of the HACA (see also Fig. 3.3c). CD-snoRNAs show, in contrast to the HACA-snoRNAs, only a single prominent peak at the 5'-end (Fig. 3.5, middle). An increased number of 5'-ends of HTS reads is also observed just upstream of loops of tRNAs (Fig. 3.5 (right)).

RNA class	source	loci found	blocks/cluster (mean)	reads/cluster (median)
microRNAs	miRBase v12	218	2.42 ± 1.04	4535.33
tRNAs	tRNAscan SE	87	3.22 ± 1.92	183.95
snoRNAs	snoRNAbase v3	129	2.60 ± 1.66	127.5

Table 3.4: In total 434 of 852 clusters were found within regions of annotated miRNA, tRNA and snoRNA loci. While the average number of blocks is similar for all three ncRNA classes, the number of reads differs significantly among the classes.

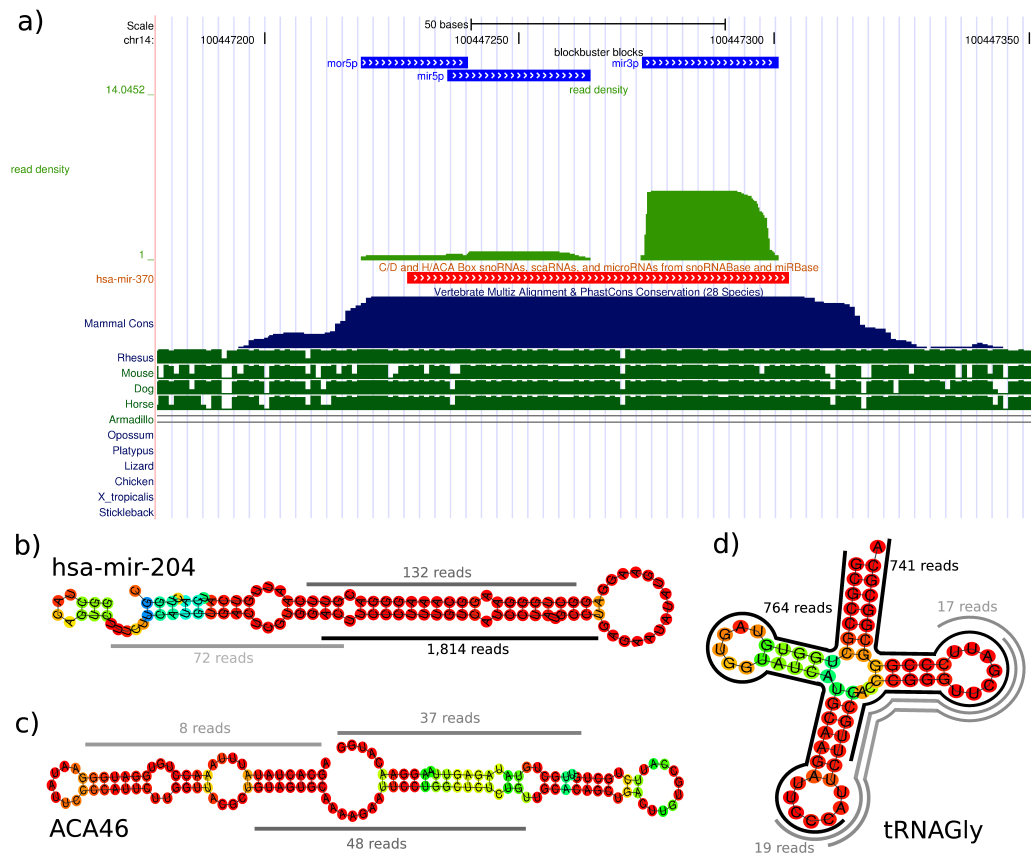


Figure 3.3: Non-coding RNAs exhibit specific block patterns. (a) Distribution of short reads at the hsa-mir-370 locus. There are three clearly distinct blocks of reads: they correspond to moR (5'-end), miR* (center) and miR (3'-end) transcripts. The conservation pattern is shown below. (b) The class of miRNAs often shows a block pattern of two or three separated blocks. (c) snoRNAs tend to have miRNA-like mature and star blocks at their 5' and 3' hairpins with minor overlaps, while a series of overlapping blocks is striking for the tRNA class (d).

The pairing probabilities of bases covered by HTS reads are significantly increased (Fig. 3.4b). Just upstream the 5'-end of these reads, the median base pairing probability increases sharply and reaches a level of > 0.9 . At the 3'-end the base pairing probability drops again. However, median base pairing probabilities of bases covered by the center of reads drop down to 70%. Although this effect is boosted by reads found within miRNA loci, it can also be observed unambiguously for reads within snoRNA and tRNA loci (Fig. 3.4a).

The observation that blocks reflect structural properties of ncRNAs was exploited to train a random forest classifier to automatically detect miRNAs, tRNAs and snoRNAs. After visual inspection of block patterns for some representatives of these classes, ten features were selected. Their evaluation reveals significant statistical differences among the chosen ncRNA classes (Fig. 3.6). As expected, the number of reads mapped to miRNA loci (minimum and maximum block height) clearly distinguishes miRNAs from other ncRNA classes. In contrast to tRNAs and snoRNAs the maximum block distance of miRNAs shows

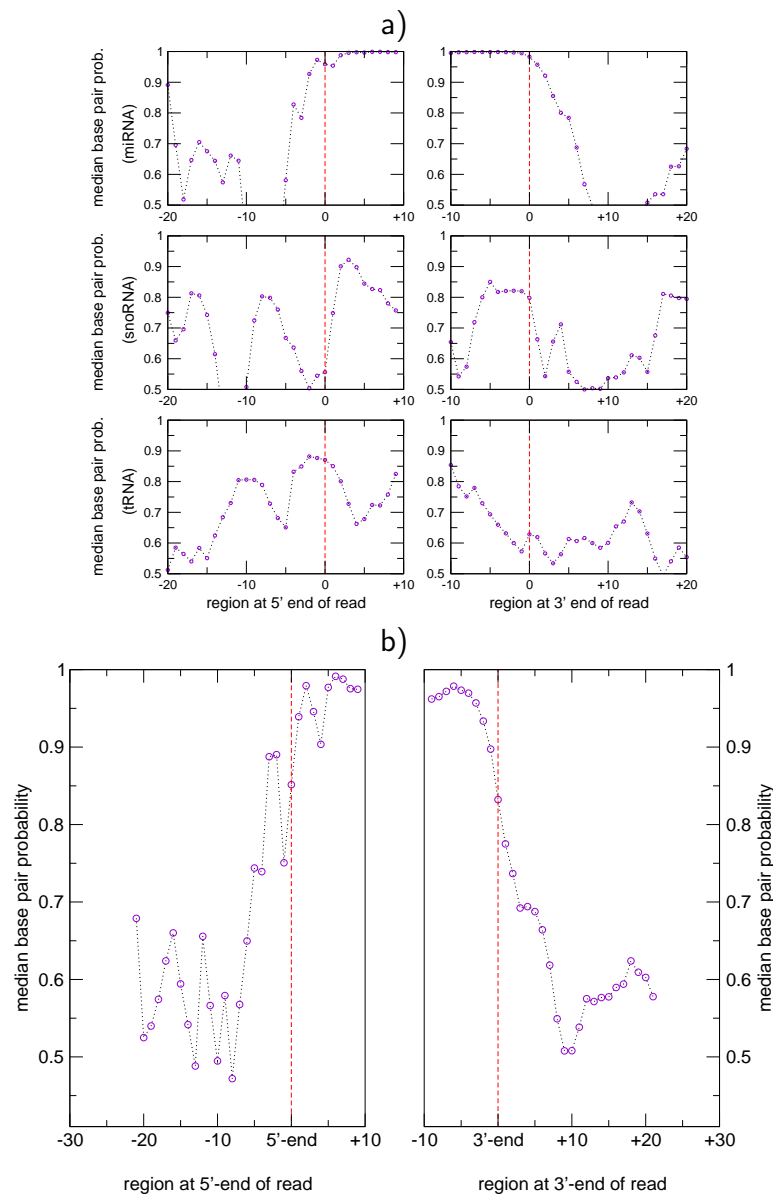


Figure 3.4: Base pairing probabilities increase at the 5'-end and decrease at the 3'-end of reads mapped to ncRNA loci. (a) The 3'- and 5'-ends are indicated by dashed lines. The median base pairing probability increases sharply at the 5'-ends (upper left) and drops again at the 3'-ends of reads mapped to miRNA loci (upper right). A similar – but attenuated – effect is observed for snoRNAs (middle panel) and tRNAs (lower panel). (b) The median base pairing probabilities at 5'- (left panel) and 3'- ends (right panel) for all reads within the 852 clusters. The 5'- and 3'-ends are indicated by dashed vertical lines.

a very narrow distribution around 40nt, reflecting the distance between miR and miR* transcripts. Furthermore, the class of tRNAs frequently shows more block overlaps than snoRNAs and miRNAs. The distance of blocks is an important feature for snoRNAs: the maximum block as well as the minimum block distance is higher compared to both tRNAs

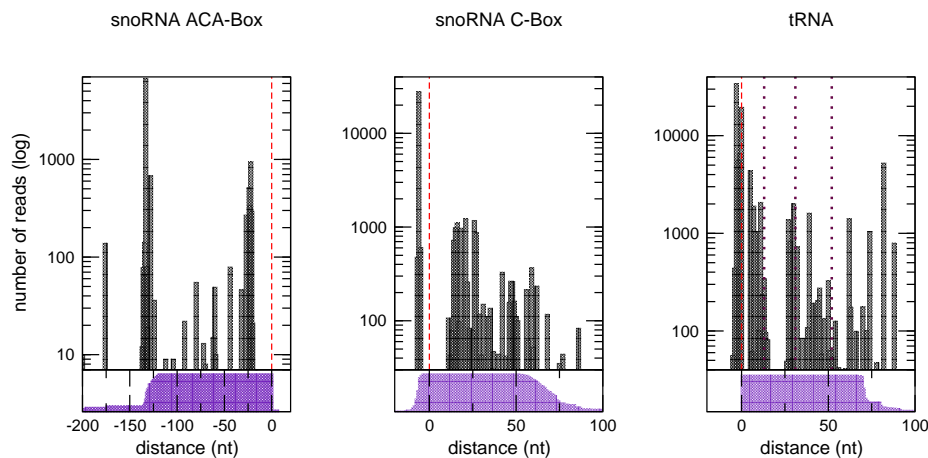


Figure 3.5: HTS data reflects structural properties of ncRNAs. Upper panels show the number of 5'-ends of mapped HTS reads (bars) relative to aligned the 5'-ends (dashed vertical lines) of 27 ACA boxes (left), 81 CD boxes (middle) and 87 tRNAs (right). The area in the lower panel represents the number of boxes and tRNAs present at the distance relative to their aligned start sites. In accordance with Taft et al. [170] a sudden and sharp increase of 5'-ends is seen just upstream of the snoRNAs' ACA and C boxes, resp., indicating that read blocks reflect structural properties of snoRNAs. Similarly, the number of 5'-ends increases just upstream of the tRNA and the relative start sites of its three loop regions (dotted lines). Downstream the start sites there is a sudden drop in the number of reads.

and miRNAs.

The random forest model was repeatedly trained with randomly chosen annotated loci and different training set sizes in order to determine predictive values (PPV) and recall rates. For the training sets comprising 150 clusters the random forest model shows a positive predictive value > 0.7 for all three ncRNA classes. The recall rate for miRNAs is well above 80%. However, with a rate of ≈ 0.55 the recall of snoRNAs and tRNAs is relatively poor (Tab. 3.5). For larger training sets containing 250 clusters, the positive predictive value (PPV) is > 0.8 for all classes. For miRNAs the classification achieves recall rates and PPVs of > 0.9 . Likewise, the recall rates for snoRNAs and tRNAs rise to 0.7-level. In summary, for both training set sizes and all classes the random forest model achieves PPVs and recall rates of ≈ 0.8 .

We applied the classifier to unannotated ncRNA loci. A list of miRNA, snoRNA, and tRNA candidates predicted is available from the supplementary page [95]). This resource includes the original reads, their mapping accuracy and their mapping location in machine-readable formats. Furthermore, the page provides links to the UCSC genome browser to visualize the block patterns. For microRNAs and snoRNAs, we also indicate whether the candidates are supported by independent ncRNA prediction tools.

The 29 miRNA predictions contained 3 miRNAs (hsa-mir1978, hsa-mir-2110, hsa-mir-1974) which have already been annotated in the most recent miRBase release (v.14), as well as a novel member of the mir-548 family, and another locus is the human ortholog of the bovine mir-2355. In addition, we found two clusters antisense to annotated miRNA

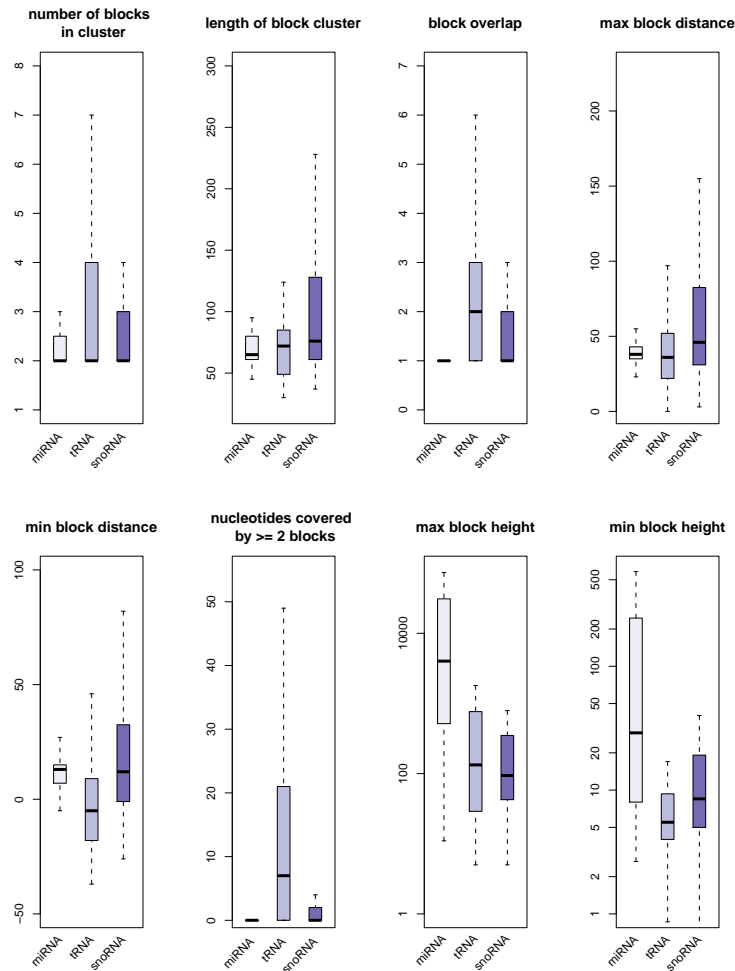


Figure 3.6: Box plots for 8 different features selected to train the random forest classifier. The number of reads mapped to miRNA loci alone (max block height and min block height) effectively distinguish miRNAs from other ncRNAs. Likewise, the distribution of block distances seems to be a specific feature for miRNAs. Compared to other regions, tRNA loci frequently show block overlaps of two or more blocks. The minimum block distance shows a median overlap of ≈ 5 nt for blocks in within tRNA loci. SnoRNAs typically have longer block distances than the other classes.

loci (hsa-mir-219-2 and hsa-mir-625). Such antisense transcripts at known miRNA loci have been reported also in several previous publications [55, 163, 12, 181], lending further credibility to these predictions.

For the tRNAs and snoRNAs we expect a rather large false positive rate. The 78 tRNA predictions are indeed contaminated by rRNA fragments, but also contain interesting loci, such as sequence on Chr.10 that is identical with the mitochondrial tRNA-Ser. SnoReport [68], a specific predictor for HACA snoRNAs based on sequence and secondary features, recognizes 44 (20%) of our 223 snoRNAs predictions.

To help to understand and use the various options implemented in RandomForest model,

	#loci	PPV		recall	
		mean	sdev	mean	sdev
Training size 250					
all	852	0.889	0.015	0.799	0.015
miRNA	227	0.932	0.020	0.918	0.023
tRNA	287	0.860	0.040	0.683	0.046
snoRNA	143	0.819	0.032	0.694	0.060
other	195				
Training size 150					
all	852	0.827	0.020	0.698	0.027
miRNA	236	0.900	0.027	0.847	0.041
tRNA	348	0.755	0.044	0.580	0.062
snoRNA	115	0.733	0.057	0.525	0.071
other	153				

Table 3.5: Positive predictive values (PPV) and recall rates for training sets of size 150 and 250. For each set size means, medians and standard deviations are calculated from 20 randomly sampled training sets.

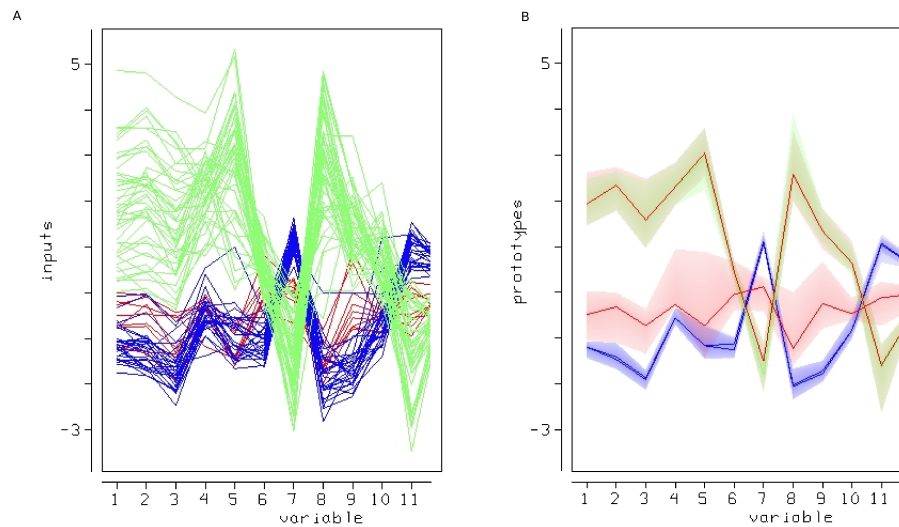


Figure 3.7: A. Parallel coordinate displays are used to represent input variables on the Random Forest approach. Green for microRNA, blue for snoRNA and red for tRNAs. B. Prototypes showing a view of how the variable relate to the classification by each class.

we use an example to visualize proximities and some other steps of the RadomForest as it was originally implemented by Breiman [15, 2]. In this example the model was repeatedly trained with the complete set of the 434 annotated loci. The outcome is similar to the classification when the model was trained with a size of 150. Since the classes tRNAs and snoRNAs showed a recall mean around 0,525 and 0.580 respectively we intended to explore further information about how they are computed. In our example, Fig. 3.8 darker

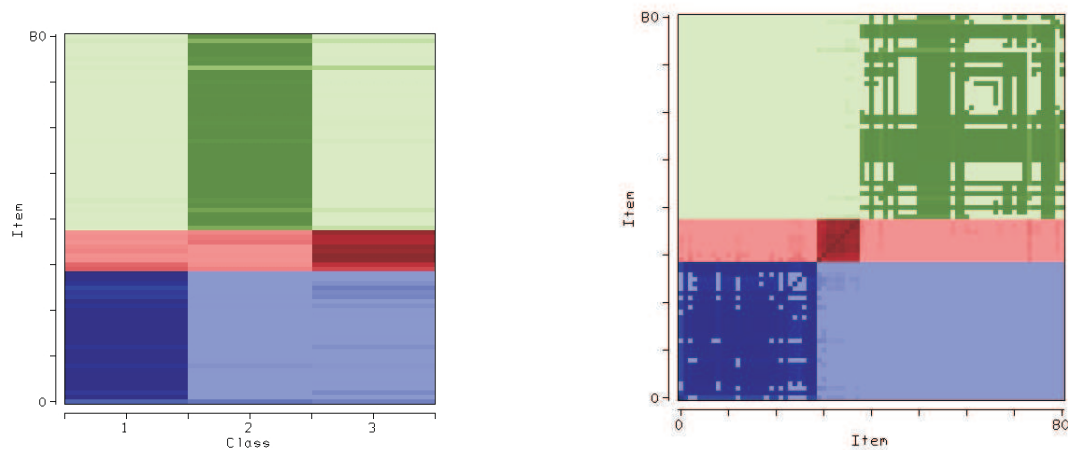


Figure 3.8: Heatmaps used to represent the votes for each class and proximity matrices. The cells are colored according to the true class. In green microRNA class, red for tRNA class and blue for snoRNA. Plot at the left correspond to votes to classify ncRNAs. A darker cell indicates that a higher percentage of trees voted for that case, when it was out-of-bag. Plot at the right shows a heatmap of the proximity matrix. The cells are colored according to the true class. Darker cells indicate higher proximities.

cell indicates that a higher percentage of trees voted for that case for each class. See for example that snoRNAs and tRNAs receive votes for misclassified cases and the proximity matrix shows that true classes are composed of higher proximities indicated by darker cells. Prototypes are a way of getting a picture of how the variables relate to the classification. See the prototypes for our test depicted in Fig. 3.7.

Identification of other structured ncRNAs Short RNAs are processed from virtually all structured ncRNAs. Complex read patterns are observed, for instance, for the 7SL (SRP) RNA and the U2 snRNA. Y RNAs, which have a panhandle-like secondary structure produce short reads mostly from their 5' and 3' ends, see Fig. 3.9.

In a recent study, [158] found that in the tunicate *Ciona intestinalis*, half of the identified miRNA loci encode up to four distinct, stable small RNAs. These additional RNAs, termed *miRNA-offset RNAs* (moRs), are generated from sequences immediately adjacent to mature miR and miR* loci. Like mature microRNAs, they are about 20nt long, developmentally regulated, and appear to be produced by RNase III-like processing from the pre-microRNA hairpin. This observation prompted us to specifically search for analogous pattern in human small RNA sequencing libraries. In the brain libraries we found 78 annotated microRNA loci that exhibit blocks of reads at positions characteristic for moRNAs. See the locus *mir-125b-1* with expression on mor sequences Fig. 3.10.

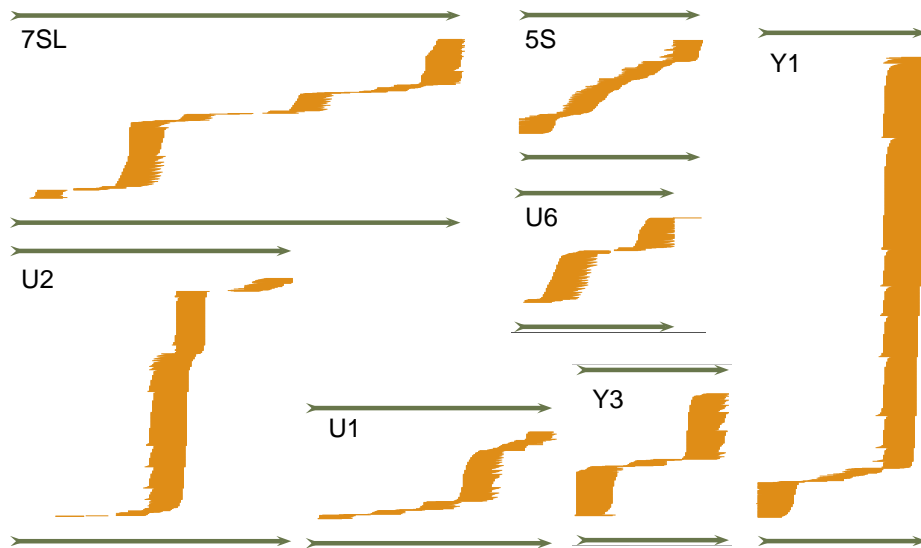


Figure 3.9: Short reads are produced from a wide variety of structured ncRNAs. Green arrows indicate the ncRNA gene and its reading direction, individual short reads are shown as orange lines. The same scale is used for all examples.

3.4 Discussion

In extension of previous work establishing that various ncRNA families produce short processing products of defined length [87, 157, 170], we show here that these short RNAs are generated from highly specific loci. The dominating majority of reads from short RNAs originates from base paired regions, suggesting that these RNAs are, like miRNAs, produced by Dicer or other specific RNAases. For example, specific cleavage products have recently been reported for tRNAs [180]. In this work we show that the block patterns are characteristic for three different ncRNA classes and thus suitable to recognize additional members of these classes. For instance, the random forest trained with loci annotated in the mirBase v12 predicted five additional miRNAs reported in the mirBase release 14 as well as two “antisense microRNA”.

Comparison with other ncRNA classification approach from transcriptome data

Since the identification of non-coding RNAs using small RNA libraries is a new issue in computational science we first aimed to classify three already known and well annotated ncRNAs. After preprocessing and data reduction steps, a set of different variables, i.e. features, was chosen to describe the short read data in a compact fashion.

To reduce the amount of data but simultaneously keeping as much information as possible, consecutive hits were condensed to blocks using an automatized approach. The block definition relies on the representation of read hit locations as gaussian distributions. An iterative statistical procedure is used to merge neighboring gaussians. All hits merged via the gaussian approach are represented by a single block. The block height represents the number of reads that have been consolidated in a block.

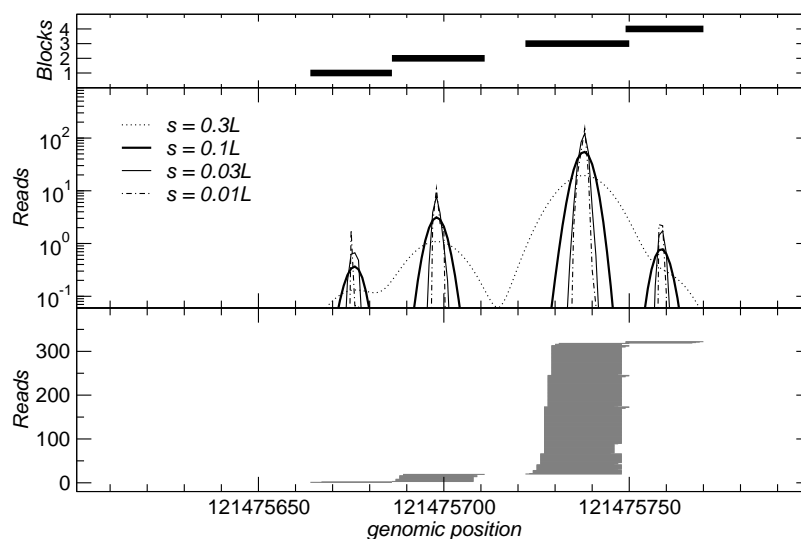


Figure 3.10: Decomposition of the cluster of reads at the *mir-125b-1* locus [97] on chr.11 (bottom panel) The blockbuster algorithm replaces each read by Gaussian profile centered at the midpoint of the read. The middle panel shows the superposition $f(i)$ of these profiles for four different width of the Gaussian, here chosen to be fraction of the read lengths L . Clusters (top) panels are identified as sets of reads whose midpoints are located is close to the peaks of $f(i)$. Clusters 2 and 3 correspond to the *miR-125b-1* and *miR-125b-1**.

After each addition of gaussians, the standard deviation of the resulting gaussian "decides" if further gaussians have to be included. Thus, the algorithm automatically adjusts the blocks to hit pattern. The user defined parameter s tunes the resolution of the merger [94].

A different approach by [84] is based on the assembly of contiguously overlapping consecutive tags into tag-contigs (TC). Thus, only a single distance parameter definition allows a static and genome-wide scaling of the blocks.

The blocks are the basis and starting point for the classification approach. Compared to [84], the number of features is significantly larger.

The afore mentioned features **block-height** and **block-length** are somewhat similar to **tag-depth** (maximum number of overlaying sequence reads per each base) and **tag-length** (length of the tag-contig) in [84]. However, the approach here uses additional features such as number of nucleotides covered by at least two blocks, number of overlapping blocks, mean distance between blocks etc.

In terms of accuracy, the presented approach using a Random Forest classifier [15] seems to be more robust since positive predictive values and recall rates can be calculated for the classification

Compared to other machine learning techniques, the Random Forest classification exhibits a good accuracy. It automatically selects a set of variables in an essentially unbiased

way. It is furthermore efficient to run on large data bases [2]. In contrast, [84] presents a classification of ncRNAs based on visual comparison of scatter plots for only two features.

3.5 Conclusions

The block patterns for the evaluated ncRNAs show some interesting characteristics. Although miRNA loci accumulate far more reads than tRNAs and snoRNA loci, the reads are extremely unevenly distributed across the blocks. For tRNAs we observe series of overlapping blocks that are specific enough to separate this class from other classes with high positive predictive values.

However, the successful prediction of miRNAs heavily depends on the height of the blocks, i.e. the number of reads that map to a potential locus. In comparison tRNAs and snoRNAs show significantly lower positive predictive values and recall rates. A relatively large training set is required to achieve PPV's $> 80\%$. Obviously, the selection of appropriate features is crucial for the success of the presented approach. Hence, the random forest classifier is not sufficient as it stands and the identification of other characteristic features is subject to further research. The integration of secondary structure information of cluster regions is likely to enhance the prediction quality.

Beyond the classification by means of soft computing methods, this survey shows that HTS block patterns bear the potential to greatly improve and simplify ncRNA annotation. Given the striking relationship of HTS reads and secondary structure for some ncRNA classes, block patterns may also be used in the future to directly infer secondary structure properties of non-coding RNAs from transcriptome sequencing data. In this context, although not shown here, block patterns may also help to identify new classes of RNAs directly from transcriptome sequencing data.

Chapter 4

Computational analysis of tRNA-derived small RNAs

4.1 Introduction

Whole transcriptome analysis has greatly facilitated the identification of new small regulatory RNAs. Recent filtering of deep sequencing data has revealed the existence of abundant small RNAs derived from tRNAs [23, 98, 67, 52]. However, the history of tRNA-derived products date from its biogenesis *per se* [140] to multiple mechanism of degradation stress related and general RNA quality control in eukaryotes [36]. The first small product is produced through pre-tRNA transcript processing, during which 5' leader and 3'trailer sequences must be removed and/or intervening sequences must be spliced (for review see [189, 129, 75]). Products of tRNA cleavage have been associated with responses to stress [105, 178, 179, 190] or phosphate deficiency in *Arabidopsis* [76]; the regulation of development in bacteria *Streptomyces coelicolor* [66]; and conidiation in *Aspergillus fumigatus* [19]. These responses involve alterations to the mechanisms that assure tRNA structural stability and the universality of tRNA modifications, which are normally extensively and extremely stable [18, 25, 129]. Some tRNAs lacking specific modifications are subject to degradation pathways [141] and to rapid tRNA decay. These degradation pathways do not leave intermediate tRNA products, as has been reported for the degradation of an endogenous tRNA species at a rate typical of mRNA decay. This demonstrates a critical role for nonessential modifications conferring increased tRNA stability and cell survival [184, 7]. However, aside from eukaryotic quality control, much tRNA cleavage is regulated according to tRNA type and tissue expression. Newly discovered small RNAs exhibit unique characteristics suggesting independent pathways of tRNA processing, and are unlikely to be the result of non-specific degradation [88]. Furthermore, tRNA has been shown to produce a Dicer-dependent small RNA [10, 23, 88] and recently discovered human tRNA-derived small RNAs have been shown to interact with the RNAi machinery, mainly with Argonaute and Dicer [67]. All of these data favor the idea that tRNA cleavage is unlikely to be the consequence of non-specific degradation but instead is probably a controlled process, the biological significance of which remains to be elucidated.

Less is known about the patterns of tRNA products from different species. Nowadays, there are various approaches to the identification and classification of small RNA in tran-

scriptome data [95]. Although numerous tools are available for the detection of microRNAs [50, 43, 65], no tools are available for the exploration of the repertoire of cleaved tRNA products. Thus, in order to analyze and classify patterns of small RNA derived from tRNA families, we have developed a new approach based on the classification of tRNA-short-read-block patterns found in small RNA libraries from *Bos taurus*, *Caenorhabditis elegans*, *Canis familiaris*, *Drosophila melanogaster*, *Gallus gallus*, *Homo sapiens*, *Macaca mulatta*, *Mus musculus* and *Rattus norvegicus*. First, RNA libraries are mapped to the reference genome using `segemehl` [72], a method based on a variant of enhanced suffix arrays, and `blockbuster`, a tool for the detection of blocks associated with specific expression patterns [94], until every mapped tRNA is associated with a unique common system. Once the tRNA products have been associated with a vector of block patterns, a distance measure is defined to determine a comprehensive classification system for block patterns of tRNA families across the referenced genomes.

4.2 Methodology

Small RNA Libraries The following high-throughput genomic libraries has been used in this survey whose repository is at Gene Expression Omnibus (GEO), NCBI.

- *B. taurus* (cow): small RNA libraries prepared from cell line derived from the adult bovine kidney under normal conditions and upon infection of the cell line with the bovine herpesvirus, GSE15450 [57].
- *C. familiaris* (dog): small RNAs were sequenced from domestic dog lymphocytes, GSE10825 [50].
- *C. elegans* (worm): small RNAs were sequenced from mixed-stage, GSE5990 [62].
- *D. melanogaster* (fly): mixture of tissues from GSE9389.
- *G. gallus* (chicken): embryon tissue, GSE10686 [56].
- *M. musculus* (mouse): mouse embryonic stem cells, GSE12521 [10].
- *H. sapiens* (human) and *M. mulata* (monkey): brain libraries was produced according to standard small RNA transcriptome protocols. See [94].
- *R. rattus* (rat): rat hepatocytes, <http://web.bioinformatics.cicbiogune.es/> [65].

Reference genome *B. taurus* NCBI Build 4, *C. familiaris* NCBI Build 2, *C. elegans* WS204, *D. melanogaster* Flybase, dm3, *G. gallus* NCBI Build 2,1, *H. sapiens* NCBI36.50, *M. mulata* NCBI Build 1, *M. musculus* NCBI Build 37, *R. rattus* NCBI Build 4.

Mapping and tRNA loci detection The first step to define tRNA specific reads was the identification of tRNA loci from a collection of mapped HTS reads. Reads were mapped to every reference genome using `segemehl` ([72]). Our study makes use of the near-perfect sensitivity and specificity of `tRNAscan-SE`, which reliably determines the complete tRNA complement of eukaryotic genomes. After the mapping, the subset of tRNA loci that registered expression signal were selected to further steps. We required small RNAs to map with an accuracy of at least 80% and normalized the read occurrence so that the reads mapping to multiple tRNA loci get equally weight distributed across all their loci.

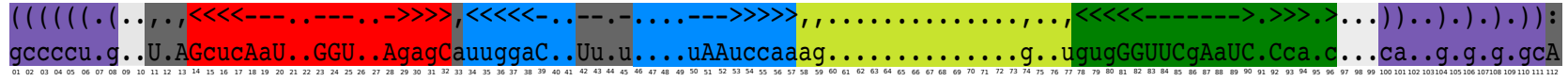
Common coordinate system for tRNAs For making a reliably comparison of tRNA block patterns across species we have defined a common coordinate system to normalize the mapping positions of the reads. Since our main major was to identify block pattern associated to the processed tRNAs and to the canonical tRNA secondary structure then *intronic*, *Sup* and *SeC* sequences were excluded. A total of 2171 sequences were used to built the common system. This common system was set to compare the block patterns of different tRNAs by a combination of structure and sequence alignments as is implemented in `Infernal 1.0` [37]. This approach allowed us to have scored both primary sequence and tRNA secondary structure conservation in conjunction with the covariance model corresponding to tRNA RFAM family RFAM database 9.1. In Fig. 4.2 the common system output is depicted. tRNA structural elements are highly differentiated by colors in the common system see Fig. 4.2A. The total length of the common system is 112 nt. We have also depicted our common system into the canonical enumeration proposed by Sprinzl [164].

Block patterns Individual reads were combined into blocks using `blockbuster` [94] as was described in chapter 3. We used a width parameter of $s = 0.2$, a value that requires blocks to be well separated to be recognized as distinct. A `Per1` pipeline was devised to map the original locations of reads and blocks onto the common set of coordinates for tRNAs. The final step of the pipeline produces tRNA-superfamily expression profiles by plotting new common locations. The plots are the results of an automatized combination of `PSTricks` macros and `LaTeX`. Profile plots of the tRNA superfamilies for the nine reference genomes are available from http://www.bioinf.uni-leipzig.de/~clara/Transcriptome0_2/

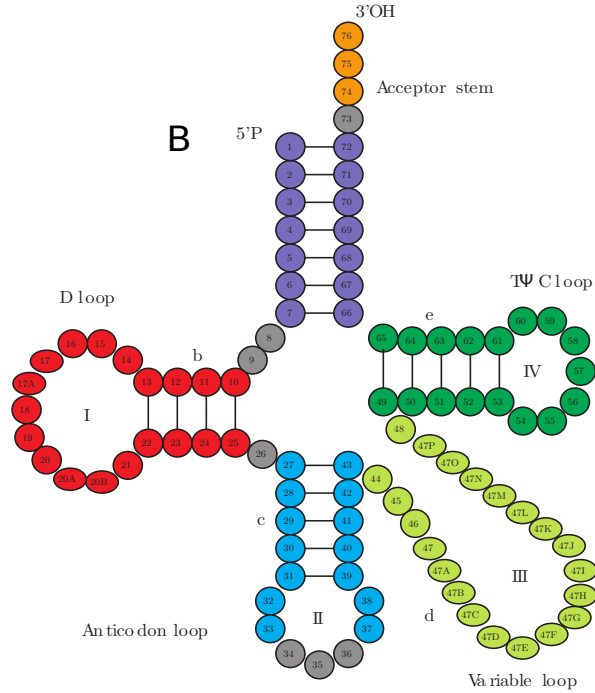
tRNA superfamilies To define the superfamilies, we clustered tRNA sequences (amino acids-specific) by using `BLASTClust`. To merge only highly similar tRNAs to families we used a quite stringent set of parameters (97% sequence similarity and a coverage of 100%).

Once superfamilies were defined, the expression of individual locus was merged to have a complete profile of expression by tRNA superfamily. We then used `blockbuster` to re-calculate tRNAsuperfamily-derived block patterns.

Superfamilies patterns characterization In order to quantify processing patterns similarity and to establish conservation across the species we have defined a set of block pattern descriptors to represent each tRNA superfamily similarities by using a Hierarchical cluster analysis [28]. The idea is to show which of a set of tRNA superfamilies are more similar to one another and to group these similar tRNA superfamilies in the same limb of a tree.



A



B



C

Figure 4.1: A. Common system B. Representation of common system using the canonical enumeration by Sprinzl [164]. C. Tridimensional representation of the tRNA structure.

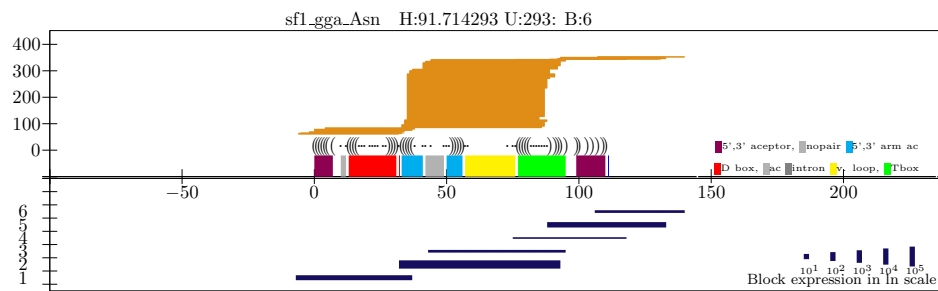


Figure 4.2: Block expression visualization of Asparagine tRNAsuperfamily1 for Chicken. In orange reads, in blue their corresponding blocks calculated by `blockbuster`. The common system output scheme separates reads and blocks.

Groups of tRNA superfamilies that are distinctly different are placed in other limbs. Then each of the tRNA superfamilies can be thought of a sitting in a m -dimensional space, defined by m descriptors characterizing a block pattern:

- Total expression by tRNA superfamily.
- Number of blocks within a tRNAsuperfamily.
- Maximum and minimum of expression fraction.
- Average of block expression.
- Maximum and minimum distance between consecutive blocks.
- Average of distance between consecutive blocks.
- Average of non-overlapping consecutive blocks.
- Average of overlapping consecutive blocks.
- Number of nucleotides overlapping by consecutive blocks.

Then we defined similarity on the basis of the Euclidean distance between two tRNA superfamilies in this m -dimensional space. The quantitative dissimilarity structure is stored in a matrix. Then initially each tRNA superfamily is assigned to its own cluster, and then a clustering algorithm proceeds iteratively, at each stage joining the two more similar clusters, continuing until there is just a single cluster.

We have also performed a test for assessing the uncertainty in hierarchical cluster analysis. For each cluster in hierarchical clustering, p-values are calculated via multiscale bootstrap resampling. We have used AU (Approximately Unbiased) p-value which was computed by multiscale bootstrap resampling (resampling size = 1000). AU p-values are a better approximation to unbiased p-value than BP value computed by normal bootstrap resampling [168]. Values on the edges of the clustering are AU p-values (%). Clusters with AU larger than 95% are highlighted by red, which are strongly supported by data. Hierarchical cluster analysis and p-values calculations were performed using the packages `hclust` and `pvclust` from R statistics environment [1].

4.3 Results

Mapped tRNA loci From the highest tRNA counts reached for the cow and rat genomes (with more than 150000 tRNAs and tRNA pseudogenes) just only its 0,25% and 0.01% are detected to show expression signal. Despite the large variation among all the species, for worm, fly, chicken, human and monkey more than its 50% of tRNAs shown expression signal. In Tab. 4.1 the corresponding counts and fraction for every species is summarized. Since our main major was to identify block pattern associated to the processed tRNAs and to the canonical tRNA secondary structure then mapped tRNA discarded correspond to mapped intronic, Sup and SeC tRNA loci that were filter out to define the common coordinate system.

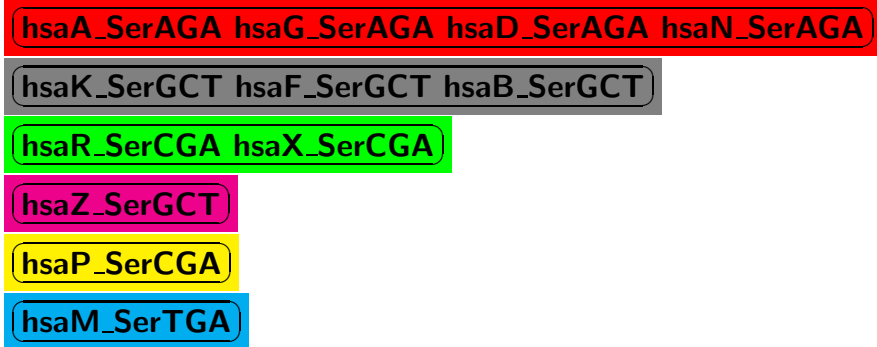
species	Cow	Worm	Dog	Fly	Chicken	Human	Mouse	Monkey	Rat
tRNAscan-SE	207156	820	87435	304	242	622	26200	520	172474
% Fraction of Total mapped tRNAs	0.25	47.2	0.05	46.38	59.92	55.14	1.82	78.65	0.01
Mapped tRNA non-discarded	430	347	36	139	138	309	403	354	15
Mapped tRNA discarded	85	40	5	2	7	34	75	55	8
tRNAsuperfamilies	129	60	22	33	51	98	143	132	8

Table 4.1: Counts of tRNAscan-SE predictions and fraction of mapped tRNAs for each species. Mapped tRNA discarded correspond to mapped intronic, Sup and SeC tRNA loci that were filter out to define the common coordinate system.

tRNA superfamilies To define tRNAsuperfamilies, we clustered tRNA sequences by each corresponding amino acid family. To merge only highly similar tRNAs to families we used BLASTClust with a quite stringent set of parameters (97% sequence similarity and a coverage of 100%). See an schematic representation of the output Fig. 4.3. Not surprisingly, in species with more mapped tRNAs the stringent non-redundant tRNA sequence set is bigger with some outliers as is observed in worm. See Tab. 4.1.

Thus far we have been able to proceed without difficulty to merge mapped HTS reads of individual mapped tRNAs into a super-read repository by tRNAsuperfamily. Further filtering steps allowed us to re-calculate with `blockbuster` block patterns of these supersets.

The new common system is also represented into the canonical numbering scheme proposed by Sprinzl [164], see Fig. 4.2B. The four base-paired stems: acceptor (in purple), D stem (in red), anticodon (in blue), T ψ C (in dark green) are depicted. Colors in gray represent non-paired regions and anticodon triplet. The variable loop is drawn in light green. In orange the 3'CCA end is also shown, although it is not part of the common system definition due that 3'CCA ends are added in the pre-tRNA processing steps. Further representation of this secondary structure into three-dimensional structure is depicted in Fig. 4.2C. Each structural element from the cloverleaf maps into its three-dimensional representation. The acceptor stem and T ψ C-arm stack together to form a continuous alpha-helix, while the D-arm and anticodon arm stack to form another continuous helix. Two RNA double helices cross by 90° to form a characteristic L-shaped tertiary structure [129].



```

hsaA_SerAGA hsaG_SerAGA hsaD_SerAGA hsaN_SerAGA
hsaK_SerGCT hsaF_SerGCT hsaB_SerGCT
hsaR_SerCGA hsaX_SerCGA
hsaZ_SerGCT
hsaP_SerCGA
hsaM_SerTGA

```

Figure 4.3: BLASTClust output example. The output consists of a file, one cluster to a line, of sequence identifiers separated by spaces. The clusters are sorted from the largest cluster to the smallest.

Patterns comparison The common coordinate system defined across the mapped set of tRNAs and the profiling visualization of tRNAsuperfamily expression allowed us to identify at first glance some interesting patterns resembling features observed in recent publications [23, 98, 67], see Fig. 4.4. Different sorts of tRNA-derived small RNAs seem to be classical processing products of pre-tRNA transcripts, i.e. products of the removal of 5' leader and 3' trailer extensions, see as example, evidence of 3' trailer products in Arginine tRNAsuperfamilies from human and chicken in Fig. 4.4. It is also appreciable to observe opposite examples of block patterns, ladder-type matching for Arginine tRNAsuperfamilies and semi-paired block patterns in Asparagine and Lysine tRNAsuperfamilies in human.

The analysis of the distribution of block expression for either the *tRNA general model* or individual *tRNA amino acid family model* also show similar patterns resembling the latest publications on tRNA-derived small RNAs [23, 98, 67, 52].

After merging individual tRNAsuperfamilies blocks into a big and unique tRNA repository of block expression *tRNA general model* or into *tRNA amino acid family model*, it is observed expression signal over tRNA regions previously reported as source of tRNA-derived fragments. At *tRNA general model expression* (see Fig. 4.5.) the number of 5'-ends increases just upstream of the tRNA and the relative start sites of its three hairpin regions (red, blue and green). Downstream of the the relative end sites of the loops there is a sudden drop in the number of reads. However the highly enrichment at 5' ends upstream of the tRNA has not previously reported. Surprisingly an even distribution at 3'-ends just downstream of the tRNA is not observed albeit there is low expression signal. Interestingly it is also observed an increased number of 5'-ends matching to the acceptor tRNA region either 5' arm or 3' arm. These regions have been also identified as regions from which small tRNA fragments might be derived.

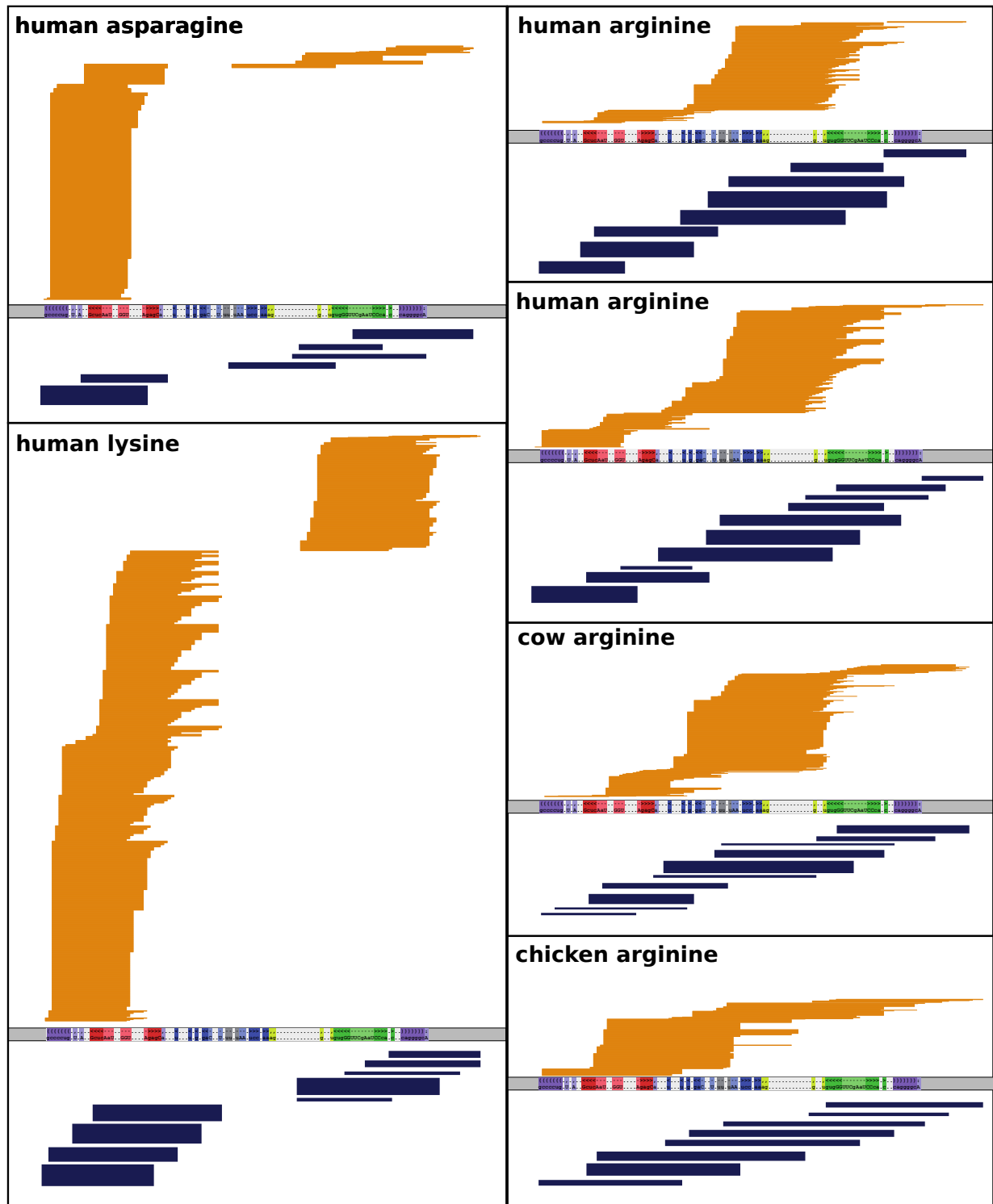


Figure 4.4: Different processing patterns for some tRNA superfamilies. In Orange reads and in Blue block detection by blockbuster.

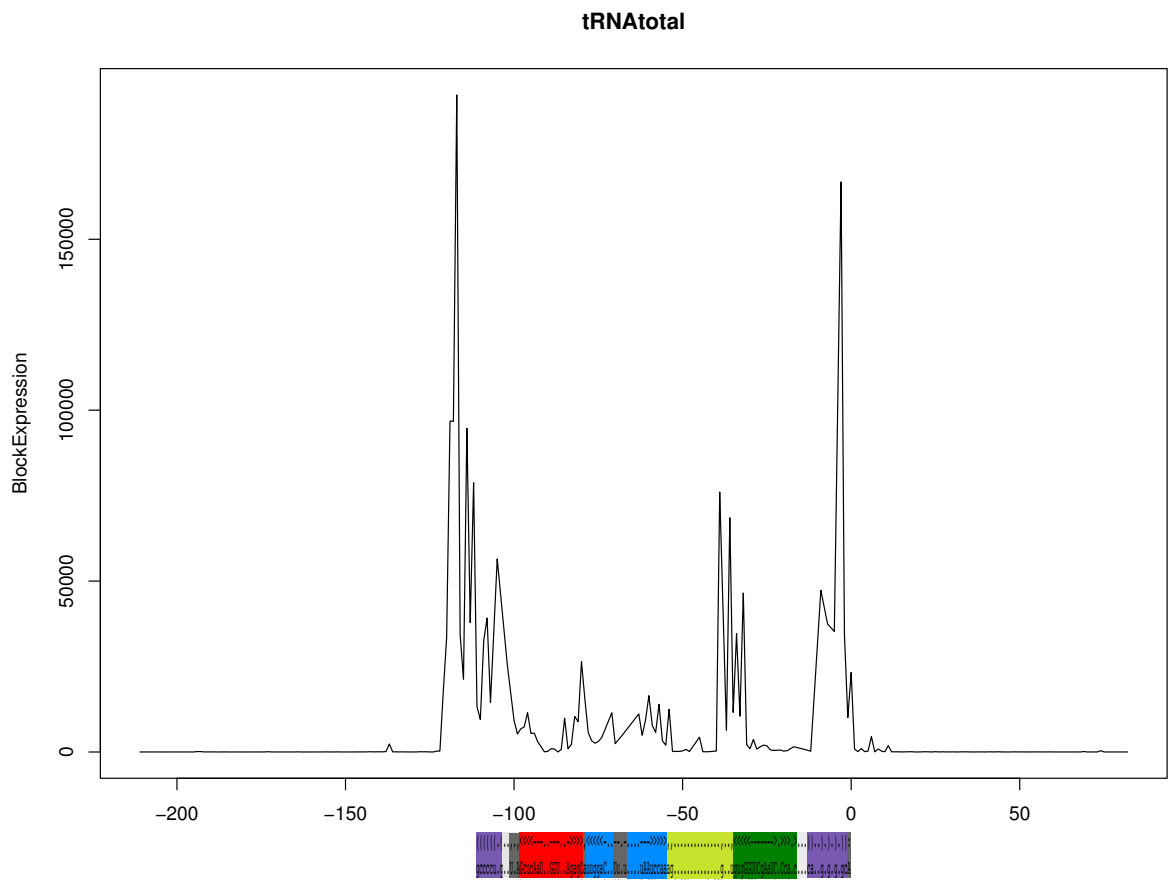


Figure 4.5: Block expression distribution for all the merged blocks. tRNA general model expression. The number of 5-ends increases just upstream of the tRNA and the relative start sites of its three hairpins regions (red, blue and green). Downstream of the the relative end sites of the loops there is a sudden drop in the number of reads. It is also observed an increased number of 5-ends matching to an acceptor tRNA region either 5' arm or 3' arm.

Distribution analysis by *tRNA amino acid family model* provides detailed visualization of patterns by specific family. Comparison of distribution of number of 5'ends among all the tRNA amino acid family let us to identify quite differentiable pattern types. See Fig. 4.6. For Alanine, Methionine, Glycine, Leucine, Phenylalanine and Proline families, the signal expression increases at the 5' end. These tRNA families exhibit an increased 5-ends upstream of the tRNA 5' end too. Over the 3' arm region there is an increased number of 5' ends for Serine, Threonine, and Cysteine families as well as increased lower expression downstream of the 3'end. For the remainder families the expression signal increases either 5' ends or 3' ends. Non well defined patterns are shown for Aspartic acid, Phenylalanine and Histidine families. Histidine family is the only example showing an expression enrichment at both regions outside of the 5' and 3' ends suggesting a processing type of an extended hairpin that might be shaped by extensions of both ends of the tRNA acceptor stem. Surprisingly there was not signal detection for any locus of Tyrosine tRNAs.

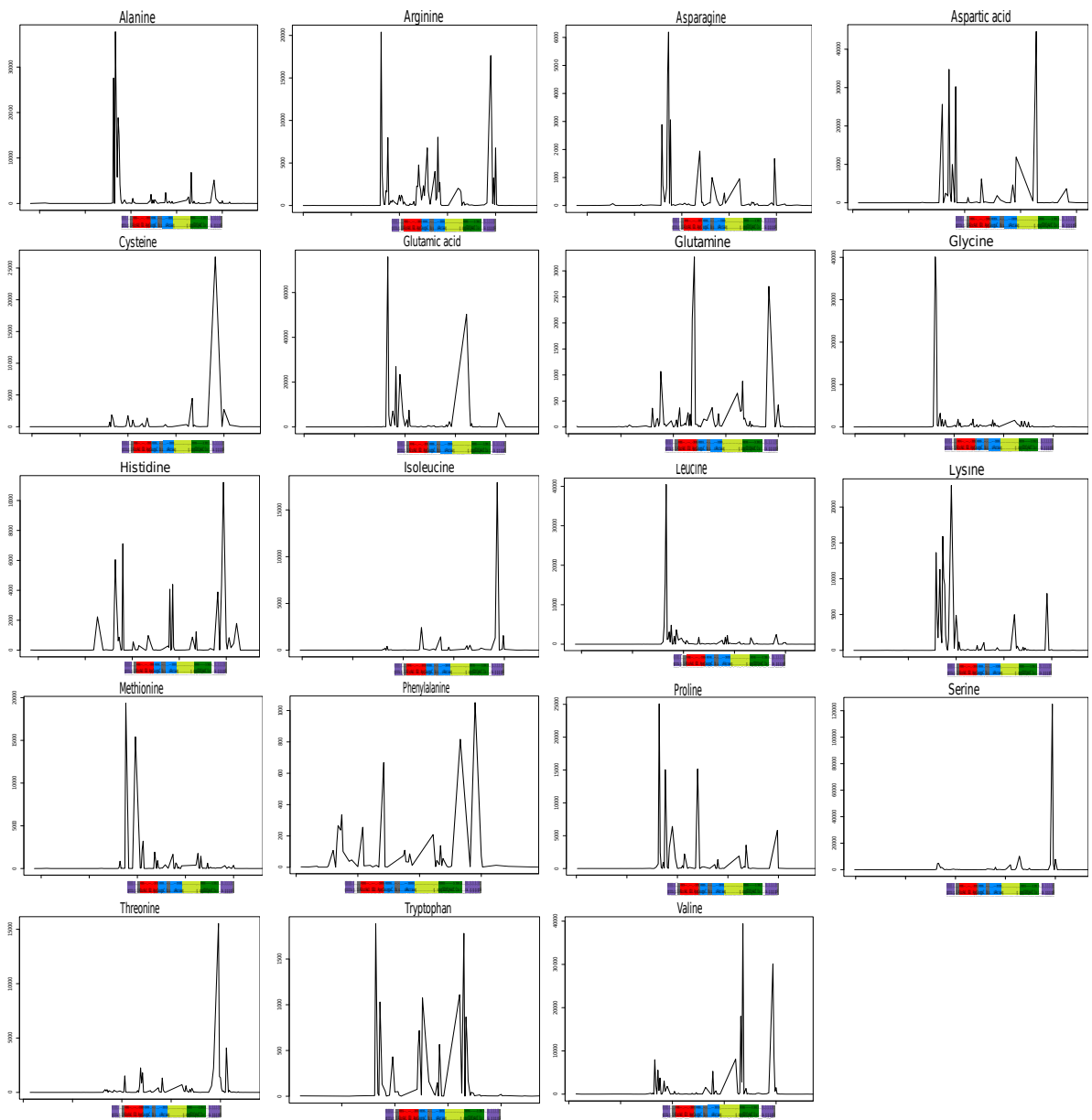


Figure 4.6: Block expression distribution for merged blocks. tRNA model expression by tRNA amino acid isoacceptor family. Not signal was detected for any locus of Tyrosine tRNAs. Y-axis for block expression in terms of number of 5'ends of the reads by position in X-axis. The system is anchored to the tRNA 3'end.

tRNA superfamily similarities In order to quantify processing patterns similarity and to establish conservation across the species we have defined a set of block pattern descriptors to represent each tRNA superfamily similarities by using a Hierarchical cluster analysis [28]. The idea is to show which of a set of tRNA superfamilies are more similar to one another and to group these similar tRNA superfamilies in the same limb of a tree. For all the species the test for assessing the uncertainty in hierarchical cluster analysis using p-values shows

p-values over the 80 %. Clusters with AU larger than 95% are highlighted by rectangles, which are strongly supported by data. Different main groups corresponds to tRNAs that have specific processing patterns. Some just only gather tRNA superfamilies with two or three blocks of expression. Some others gathering more intriguingly block expression patterns. See Figs. 4.7, 4.8 and 4.9 for some classification examples.

4.4 Discussion

In a first step, reads that correspond to an specific expression value were mapped to its corresponding reference genomes by using `segemehl` [72], a method based on a variant of enhanced suffix arrays and matching statistics that efficiently deals with mismatches, insertions and deletions. We required small RNAs to map with an accuracy of at least 80% and normalized the read occurrence so that the reads mapping to multiple tRNA loci get equally weight distributed across all their loci. The program `tRNAscan-SE` was used to identify tRNA loci. Only reads, which overlapped with predicted tRNA loci were used for all the previous analysis.

We have shown that read blocks reflect secondary structure properties of different non-coding RNAs [95]. Specially for tRNAs we showed that the number of reads increases just upstream of the 5-ends of the tRNA and the relative start sites of its three loop regions (See the Fig. 3.5) however, downstream the start sites there is a sudden drop in the number of reads, implying that a double stranded sequence is needed for the tRNA processing.

To compare the block patterns of different tRNAs, we defined a common system to normalize the mapping positions of the reads. The common system was defined by a combination of a structure and a sequence alignment of all the tRNA predictions (from `tRNAscan-SE`) using `Infernal 1.0`.

Different schematic visualizations were employed in this survey to identify most of the latest tRNA-derived fragments. Although tRNA fragments are generally considered to be random degradation products [23], here it was shown that enrichment at some specific tRNA regions presumably indicated that tRNAs can produce stable small RNAs. In first place, the significant number of the sequences that are derived from precise processing at the 5' or 3' ends of mature or precursor tRNAs were detected. Previously, it has been reported by [98] three series of tRFs (tRNA-derived RNA fragments), their names are derived from their precise alignment to the 5' and 3' ends of mature tRNAs: tRF-5 (over the 5' end), tRF-3 (over the 3'end), and tRF-1 series (3' trailer sequences). Those series were also detected in our survey. The 3' trailer sequence of Serine tRNAs that are trimmed during tRNA maturation, albeit low expression, was also detected in our survey, indicating that our results are in agreement with [98]. However, intermediate block of patterns associate to regions within the tRNA structure but offset of the regions covered in tRF-5 and tRF-3 fragments are observed. Those ladder-type patterns suggest the reads are the result of random degradation of tRNAs. However, rest to experimentally validate whether other regions are prone to produce functional small RNAs as for instance, it was reported in the analysis of the transcriptome of the unicellular algae *C. reinhardtii* [123, 192]. These studies showed that *C. reinhardtii* contains putative evolutionary precursors of miRNAs and species of siRNAs resembling those in higher plants, indicating that complex RNA-silencing systems

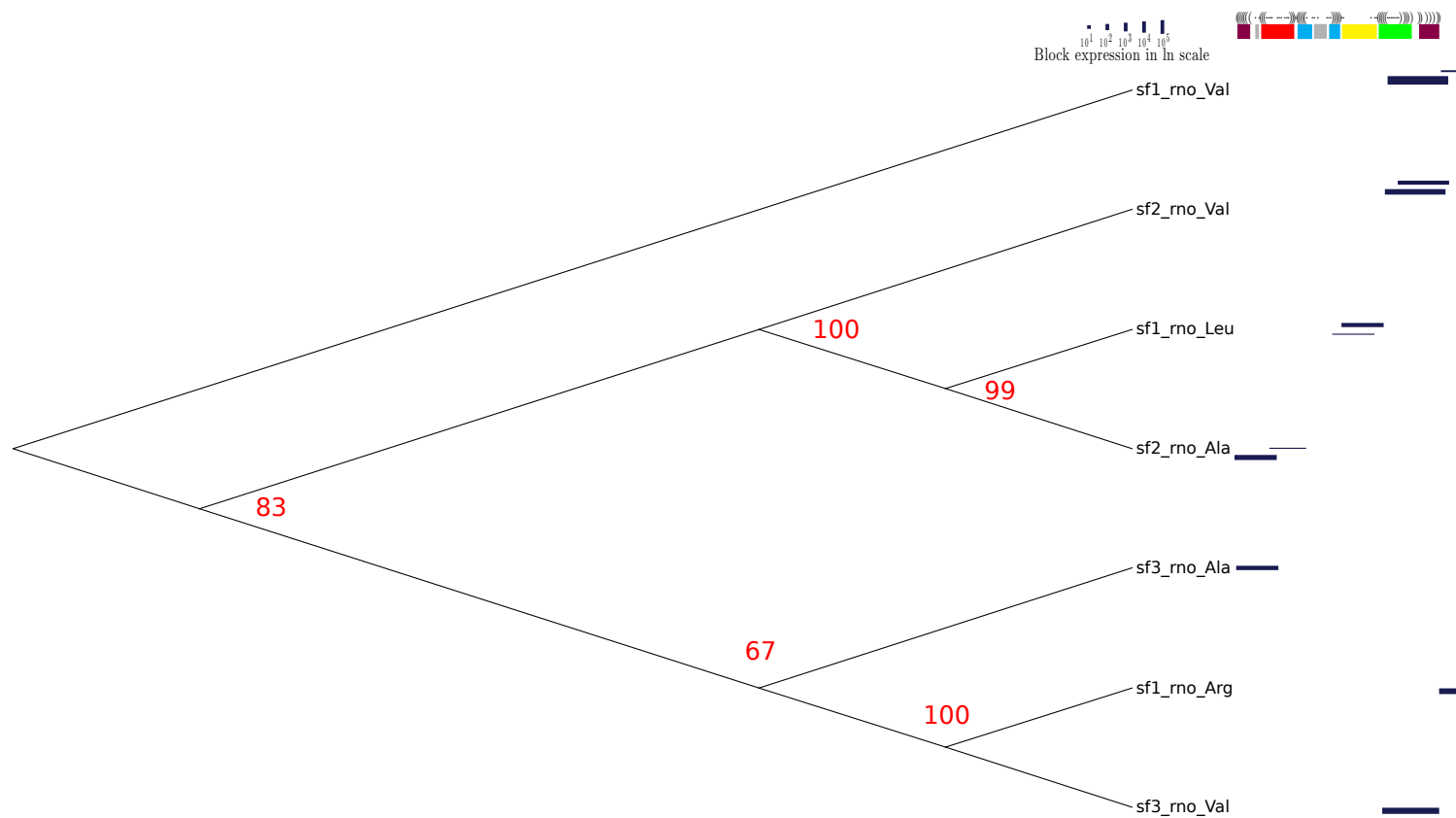


Figure 4.7: Hierarchical cluster analysis with p -values for block expression patterns in rat. Values on the edges of the clustering are AU p -values. (Approximately Unbiased) p -value which was computed by multiscale bootstrap re-sampling (re-sampling size = 1000). Clusters with AU larger than 90% are highlighted by red, which are strongly supported by data. Hierarchical cluster analysis and p -values calculations were performed using the packages `hclust` and `pvclust` of R statistics environment [1].

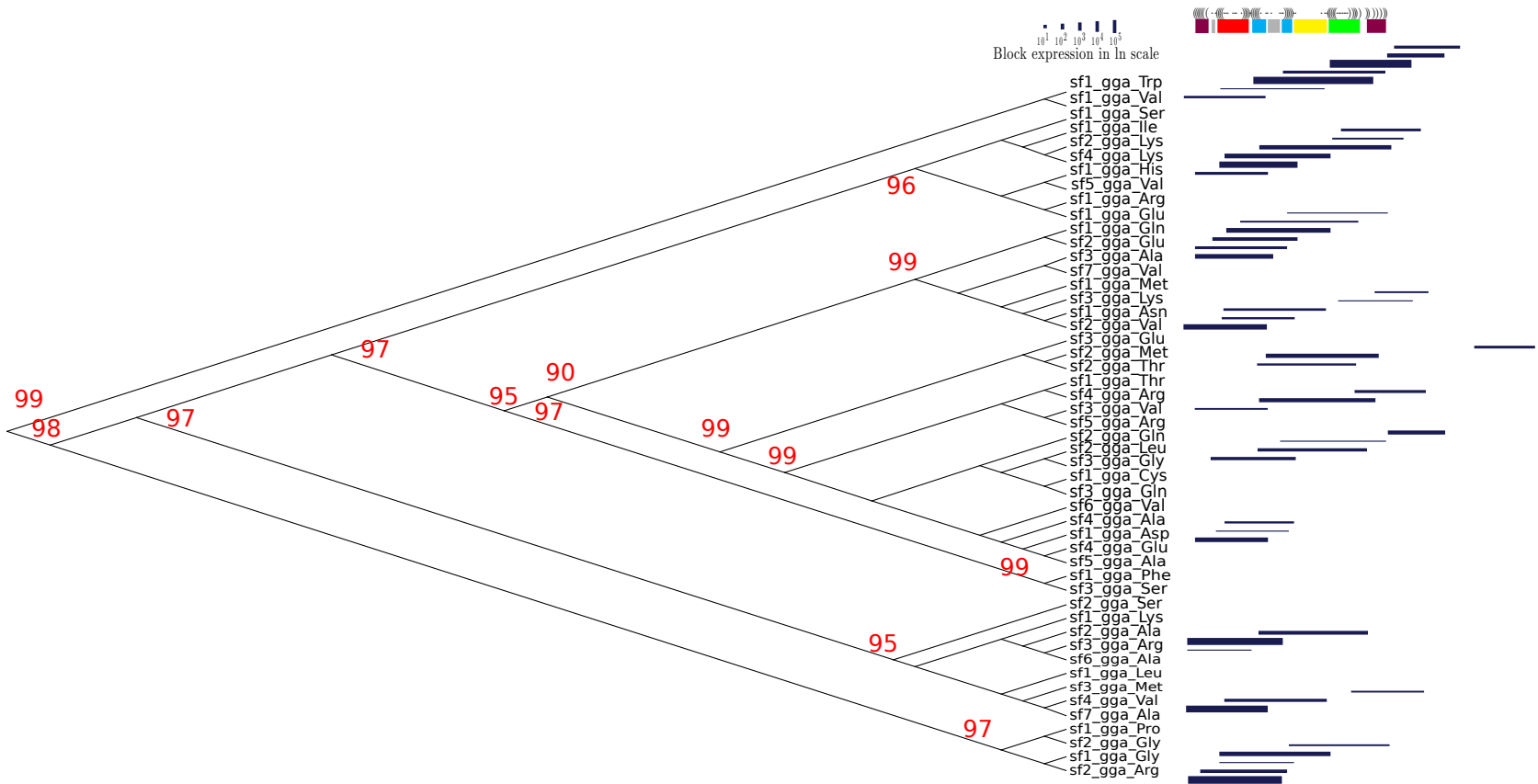


Figure 4.8: Hierarchical cluster analysis with p-values for block expression patterns in chicken. Values on the edges of the clustering are AU p-values. (Approximately Unbiased) p-value which was computed by multiscale bootstrap re-sampling (re-sampling size = 1000). Clusters with AU larger than 90% are highlighted by red, which are strongly supported by data. Hierarchical cluster analysis and p-values calculations were performed using the packages hclust and pvclust of R statistics environment [1].

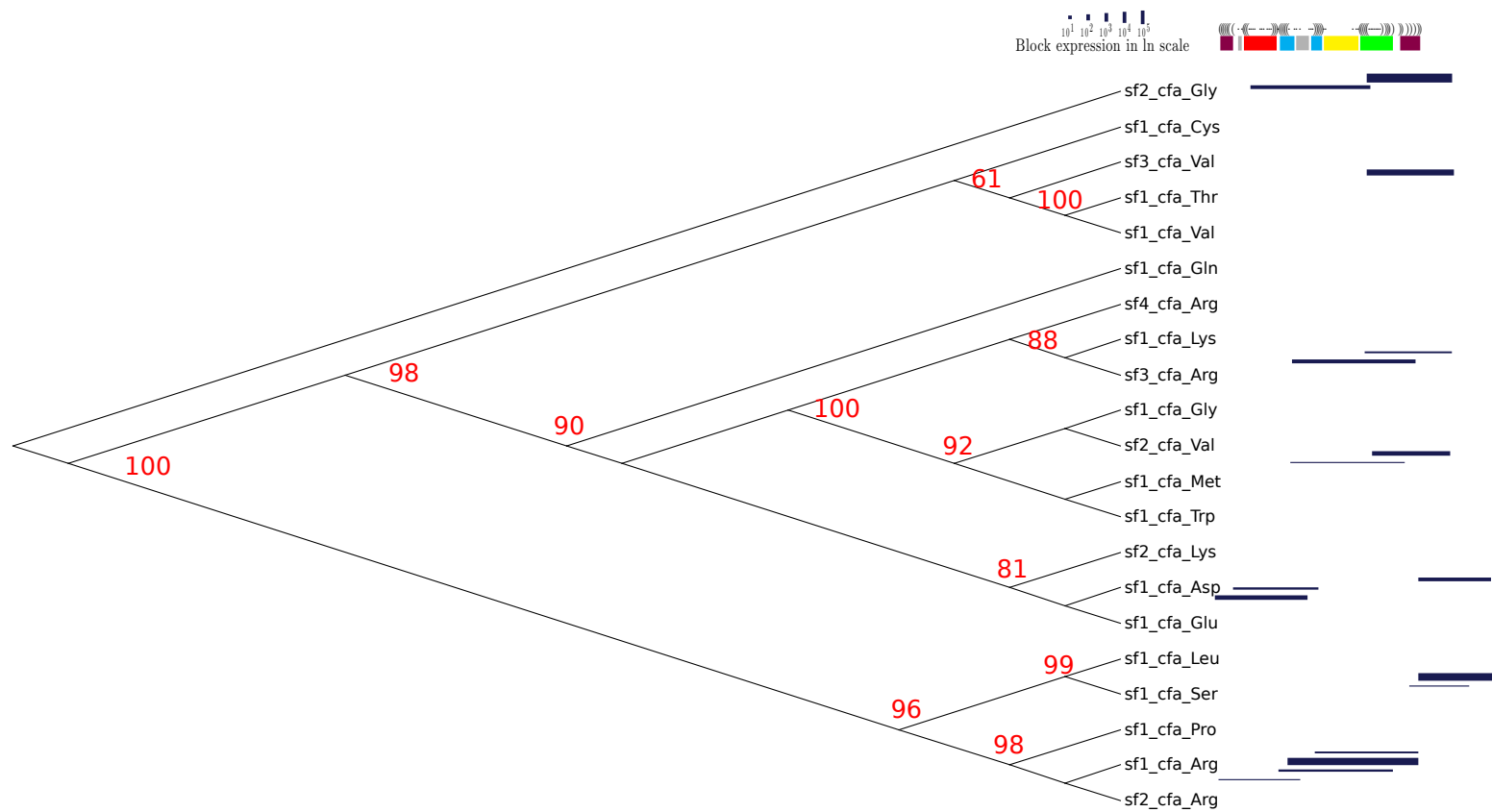


Figure 4.9: Hierarchical cluster analysis with p-values for block expression patterns in dog. Values on the edges of the clustering are AU p-values. (Approximately Unbiased) p-value which was computed by multiscale bootstrap re-sampling (re-sampling size = 1000). Clusters with AU larger than 90% are highlighted by red, which are strongly supported by data. Hierarchical cluster analysis and p-values calculations were performed using the packages hclust and pvclust of R statistics environment [1].

evolved before multicellularity and were a feature of primitive eukaryotic cells [123, 192]. In a pilot study, we previously identified a set of microRNAs derived from pre-miRNAs in which an imperfectly matched tRNA inverted repeat forms a partly double-stranded region, as observed in *Chlamydomonas*. The location of the processed region correspond to a partial end of the T ψ C loop, the variable loop and the 3' end arm of the anticodon stem. A detailed location is shown in Tab. 1.1 and Fig. 1.3.

However, there is also some interesting patterns showed here, suggesting that many tRNA-derived reads do not correspond to random degradation products as has been suggested by [98, 65]. Not all the tRNA families are processed to produce small RNAs evenly distributed. The fact that some specific loci are able to produce small RNA in an specific fashion suggest to us that previous tRNA-derived small RNAs are more diversified that previously though.

We has been also able to identify new sets of tRNA isoacceptors able to produce tRF1 besides Serine tRNA families. For example, Cysteine or Threonine tRNA families. These type of tRF-1 has been associated to cell proliferation in mammals as a different pathway of pre-tRNA transcript processing by the tRNA 3' endonuclease ELAC2 in the cytoplasm [98] or cytosolic RNase Z [67]. Some remarkable is that function of 3' trailers sequences has been previously reported in Bacteria [107] thus it is expected that other tRNA loci might be prone to produce functional tRF-1. A population of small RNAs is actively produced in *Trypanosoma cruzi* [52], and their production was found to increase under conditions of nutritional stress. This population is preferentially restricted to specific isoacceptors and to the 5' halves of mature tRNAs. We also have identified intermediate products of cleavage tRNA as previously reported for stress conditions or/and some other biological pathway [105, 178, 179, 190, 76, 66, 19, 10]. These products have been reported as stable products opposite to the rapid degradation of preexisting Valine tRNA reported by [7]. However still rest to identify many pathways related to both tRNA processing and degradation functions as was mentioned above.

4.5 Conclusions

After mapping short read data to tRNAs, merging consecutive reads to blocks, normalizing these blocks based on a common system for tRNAs, we were able to show that our block patterns are characteristic for some tRNA families within one species and even conserved between species. Furthermore we present an classification approach based on block patterns descriptors, in order to build a tree that shows which tRNA families are similar.

In agreement with previous classification system [95] our survey shows that HTS block patterns bear the potential to greatly improve the identification of tRNA-derived small RNAs from whole transcriptome data. From the analysis of the global expression profile three series of tRFs (tRNA-derived RNA fragments) have been detected: tRF-5, tRF-3 and tRF-1 in agreement with previous results. These small RNAs were reported as the second most abundant small RNAs (second only to miRNAs). The importance of tRNA-derived small RNA to the global regulation of RNA silencing through differential Argonaute association suggests that small RNA-mediated gene regulation may be even more finely regulated than previously realized [67].

Chapter 5

Conclusions and outlook

5.1 Conclusions

In this thesis are discussed new findings of three novel aspects of tRNA biology: genome organization, preliminary transcriptome data analysis, and the classification of a novel class of tRNA-derived small RNAs from transcriptome data.

New aspect about the tRNA genomic organization were widely discussed. In our analysis, we focus not only on the number of tRNA genes, but also on their relative genomic locations, and in particular on the formation of tDNA clusters. Surprisingly, we found no distinctive clade-specific features or large scale trends, with the exception of the rather straightforward observation that the larger metazoan genomes typically tend to harbour large numbers of tDNAs. In some species, large clusters of tDNAs occur and we discussed here that this phenomenon is not restricted to a particular clade of protists but rather appears independently in many times throughout eukaryotes.

In most eukaryotes, tRNAs are multi-copy genes with little or no distinction between paralogs so that orthology is hard to establish, in particular in the presence of tRNA gene clusters. As a consequence, the evolution of genomic tRNA arrangements is non-trivial to study over larger time-scales. Upper bounds on syntenic conservation can be estimated, however, by considering small sets of flanking protein coding genes for which homology information can be retrieved from existing annotation. We found that tRNAs change their genomic location at time-scales comparable to mutation rates: syntenic conservation fades at roughly the same evolutionary distances as sequence conservation in unconstrained regions.

The absence of large numbers of partially degraded tRNA copies in many of the investigated genomes provides a hint at the mechanisms of tRNA mobility. The data collected here provide a basis to investigate the connection of tRNA gene arrangements and genome organization in general.

Recent development of deep sequencing technologies allows to identify small regulatory RNAs. Here, we focus on the identification of small ncRNAs. Our transcriptome survey results also describe new insights of ncRNAs. The block patterns for the evaluated ncRNAs show some interesting characteristics. Although miRNA loci accumulate far more reads than tRNAs and snoRNA loci, the reads are extremely unevenly distributed across the blocks. For tRNAs we observe series of overlapping blocks that are specific enough to separate this

class from other classes with high positive predictive values.

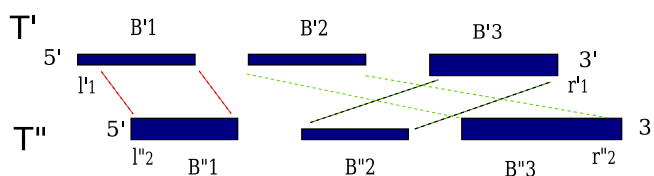
Beyond the classification by means of soft computing methods, this survey shows that HTS block patterns bear the potential to greatly improve and simplify ncRNA annotation. Given the striking relationship of HTS reads and secondary structure for some ncRNA classes, block patterns may also be used in the future to directly infer secondary structure properties of non-coding RNAs from transcriptome sequencing data. In this context, although not shown here, block patterns may also help to identify new classes of RNAs directly from transcriptome sequencing data.

The analysis of tRNA-derived small RNAs also presents results of tRNA processing patterns not previously described in a comparative way, neither for tRNA families or comparison across species. Here, we focused on the identification of small RNA fragments derived from tRNAs. After mapping transcriptome sequencing data to reference genomes we searched for specific short read patterns reflecting tRNA processing. In this context, a common tRNA coordinate system based on conservation and secondary information has been devised. That allows a vector representation of processing products and thus a comparison of different tRNAs. We report patterns of tRNA processing that seem to be conserved across species. The analysis suggests that every tRNA has a specific pattern and thus undergoes a characteristic maturation. It remains to be clarified how these tRNA shreds are processed and if they have any functional implications.

5.2 Outlook

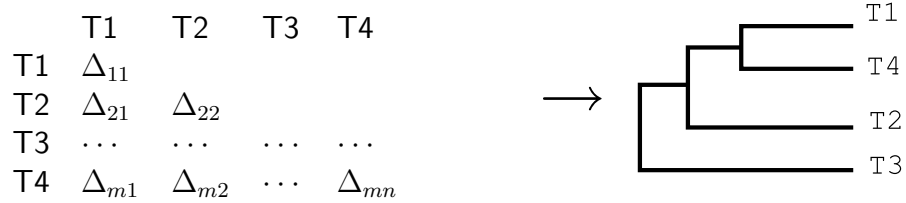
tRNA genomic organization The genomic organization of tRNAs shows complex lineage-specific patterns characterized by and extensive variability that is in striking contrast to the extreme levels of sequence-conservation of the tRNA genes themselves. Our comprehensive analysis of Eukaryotic tRNA distributions provides a basis for further studies into the interplay of tRNA gene arrangements and genome organization in general.

tRNA-derived small RNA fragments To compare the block patterns of tRNA-derived small RNAs, we plan to define a greedy alignment model. We aim at to use δ as a distance function between blocks $\delta(B', B'') = |b' - b''| + |e' - e''| + q|\log h' - \log h''|$ where b is the start position of the common tRNA coordinate system, e the end position, h the fraction of reads in a particular block and q a scaling factor for expression of each block. However using the definition above, blocks may be cross aligned. We then greedily exclude a block or keep the block when we move from the left to right.



To extract a matching between the blocks of the two loci T by simply greedily choosing the pair with minimal $\delta(B', B'')$, removing B' and B'' from the list and repeating the

procedure. Sum up the contributions for the individual pairs of blocks. For any left-overs just use a fixed extra penalty.

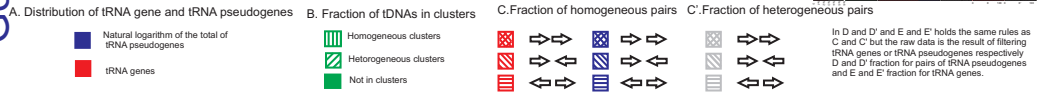
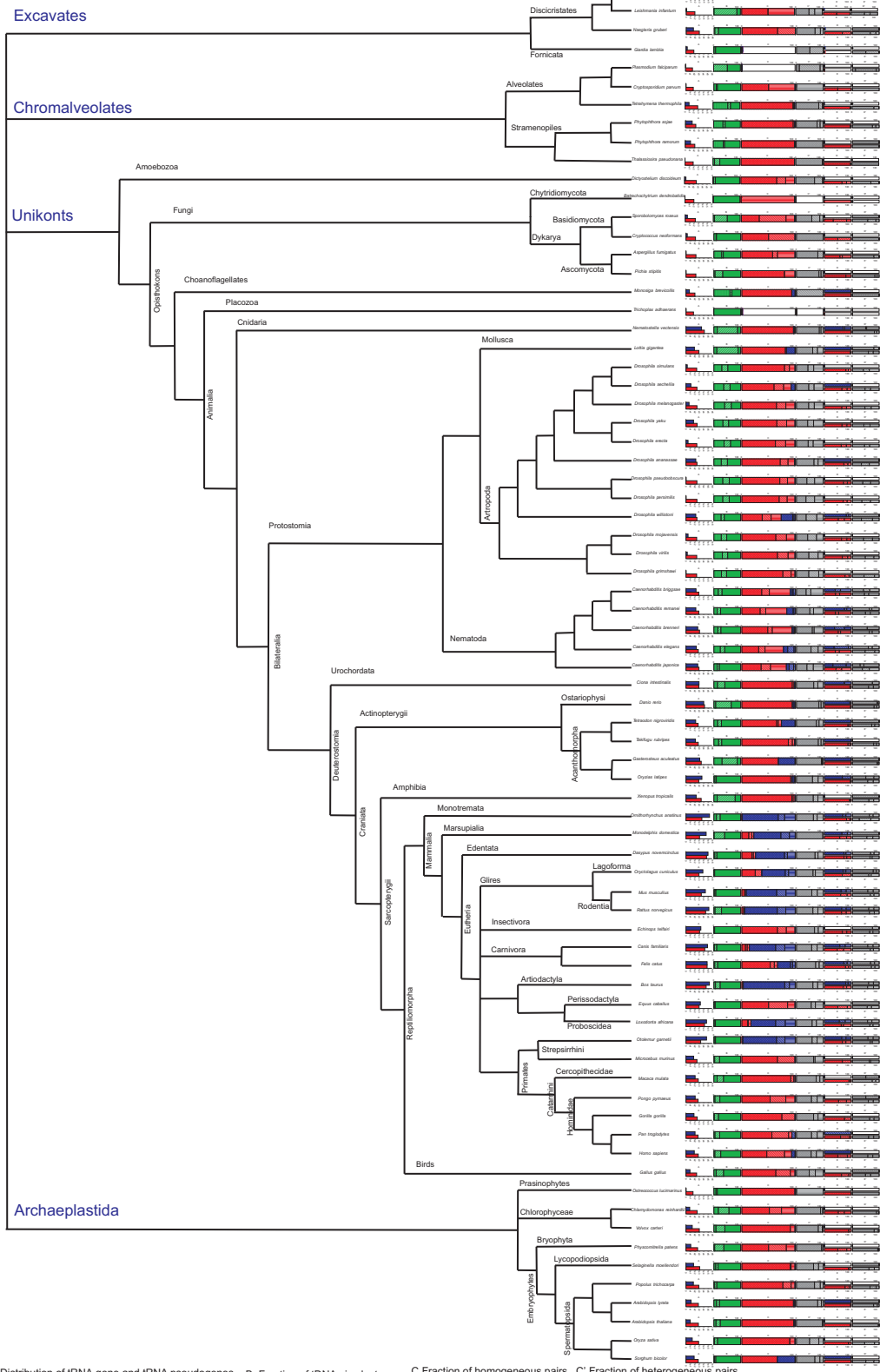


Since the steps described above give us a distance measure $\Delta(T', T'')$ between any two tRNAs T' and T'' now we plan to use this matrix distance to cluster the tRNA-short-read-block patterns for each species, and also for the entire dataset of a few different species.

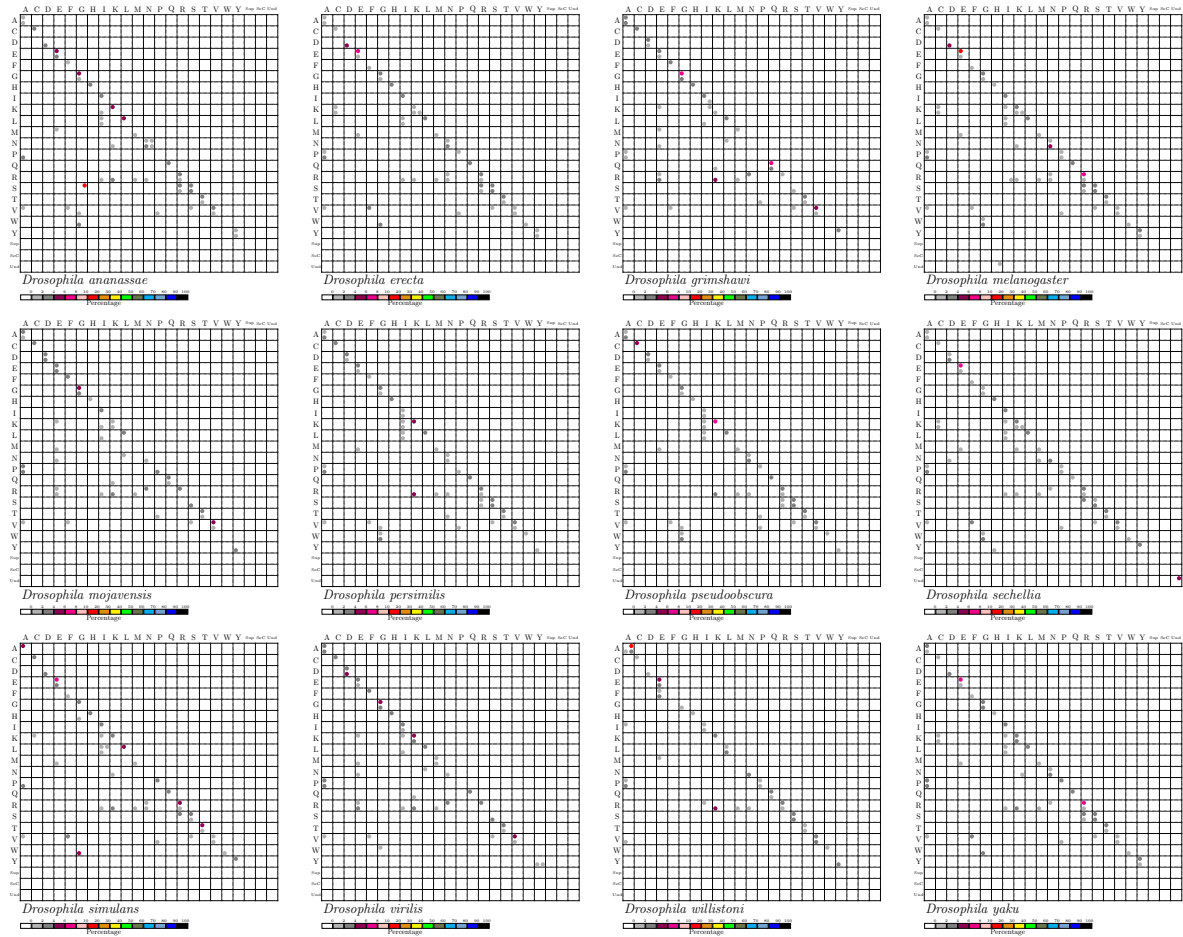
In the future we want to present a way of how these block patterns can be aligned, in order to build a tree that shows insight which tRNA superfamilies are similar under this criteria.

Appendix A

Genomic distribution of tRNA genes and tRNA pseudogenes detected by tRNAscan-SE



Apendix B



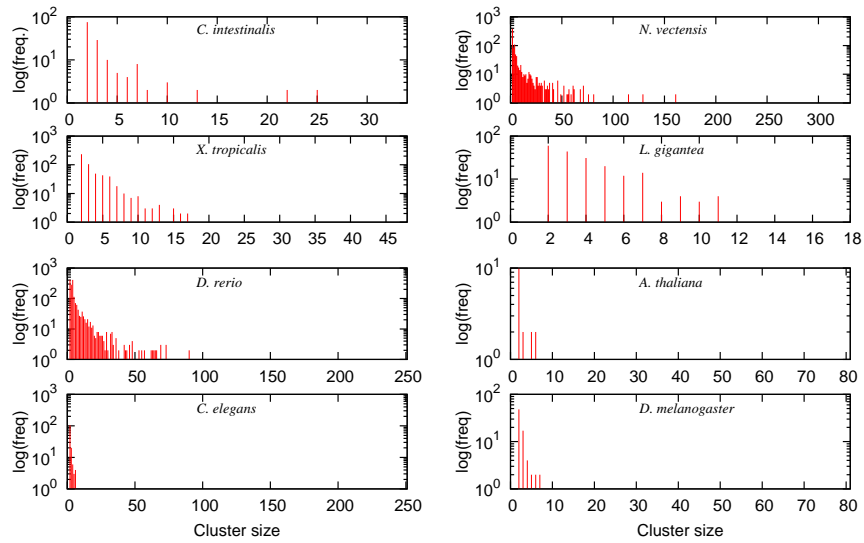
Relative abundance tRNA isoacceptor families located consecutively within tRNA clusters in *Drosophila* genus. Four data points are shown for each combinator of amino acids: Top: pairs in the same reading direction; below: pairs in opposite reading direction. Left: pairs of presumably functional tRNA, right: pairs of tRNA pseudogenes. The last three rows and columns refer to putative Suppressor, SeC, and tRNA pseudogenes of undetermined isoacceptor class, resp.

Appendix C

Species	Counts and fraction of Cluster and cluster structure																		
	Total	Pse.	tRNA	GS	O.	R.	P	Ho++	Ho+-	Ho-	He++	He+-	He-	RHo++	RHo+-	RHo-	RHe++	RHe+-	R He-
Teleosts																			
Danio rerio	25194	9986	15208	1.48×10 ⁹	14667	679.69	0	4994	105	25	7994	795	754	26.17	12.77	12.77	313.72	157.12	157.13
Tetraodon nigroviridis	707	154	553	3.67×10 ⁸	117	1.95	0	60	2	4	38	5	8	0.04	0.02	0.02	0.94	0.47	0.47
Takifugu rubripes	716	131	585	3.93×10 ⁸	131	1.96	0	82	7	4	31	4	3	0.05	0.02	0.02	0.93	0.46	0.46
Gasterosteus aculeatus	4046	1274	2772	4.62×10 ⁸	3144	57.18	0	1982	4	1	1078	39	40	3.05	1.32	1.32	25.58	12.97	12.95
Oryzias latipes	4695	3919	776	8.69×10 ⁸	286	35.41	0	158	8	0	72	25	23	4.38	2.19	2.19	13.32	6.67	6.67
Mammalians																			
Ornithorhynchus anatinus	206915	203073	3842	2.01×10 ⁹	27015	25008.22	0	8227	1484	2396	9446	2090	3372	4951.19	2475.69	2475.35	7552.34	3777.16	3776.48
Monodelphis domestica	36542	35408	1134	3.61×10 ⁹	7402	914.11	0	99	34	23	7169	36	41	121.01	60.58	60.49	336.05	167.92	168.05
Dasylops novemcinctus	137797	93845	43952	4.82×10 ⁹	7918	11498.56	1	1741	192	448	3849	607	1081	2447.66	642.51	643.47	5089.8	1337.31	1337.8
Oryctolagus cuniculus	7324	6466	858	3.47×10 ⁹	118	37.15	0	11	3	2	63	9	30	5.8	2.9	2.9	12.77	6.39	6.39
Mus musculus	26264	23401	2863	2.72×10 ⁹	1001	425.51	0	250	78	75	343	122	133	77.96	38.99	38.97	134.86	67.37	67.35
Rattus norvegicus	172474	145265	27209	2.72×10 ⁹	28198	16148.13	0	3512	929	1022	13927	4303	4505	1689.01	844.74	844.69	6385.04	3191.61	3193.04
Echinops telfairi	3426	2255	1171	3.83×10 ⁹	49	9.35	0	4	1	1	29	6	8	0.38	0.19	0.19	4.3	2.15	2.15
Canis familiaris	88179	71169	17010	2.53×10 ⁹	4858	4271.27	0	426	157	178	2332	829	936	339.71	169.5	169.75	1796.13	898.29	897.88
Felis catus	117583	59100	58483	4.06×10 ⁹	8792	11816.7	1	1604	246	234	4808	886	1014	1435.56	717.16	717.98	4473.48	2237.22	2235.3
Bos taurus	225600	190329	35271	2.92×10 ⁹	28452	22790.52	0	6151	697	876	12716	3721	4291	2125.85	1063.02	1063.46	9269.66	4634.56	4633.98
Equus caballus	2656	1752	904	2.47×10 ⁹	72	4.42	0	7	5	2	32	9	17	0.4	0.2	0.2	1.81	0.9	0.9
Loxodonta africana	57804	42827	14977	4.18×10 ⁹	1645	3553.2	1	204	70	85	751	207	328	387.44	193.65	193.96	1389.3	694.93	693.93
Otolemur garnettii	45225	43155	2070	3.43×10 ⁹	1364	1285.4	0	531	133	162	314	101	123	394.91	197.63	197.6	247.71	123.82	123.73
Microcebus murinus	354	55	299	2.91×10 ⁹	42	0.06	0	4	2	0	23	2	11	0	0	0	0.03	0.01	0.01
Macaca mulata	706	116	590	3.10×10 ⁹	168	0.23	0	67	4	2	70	13	12	0	0	0	0.11	0.06	0.06
Pongo pygmaeus	659	119	540	3.44×10 ⁹	83	0.28	0	9	5	3	37	14	15	0.01	0	0	0.13	0.07	0.07
Gorilla gorilla	409	64	345	2.34×10 ⁹	40	0.08	0	3	1	0	23	6	7	0	0	0	0.04	0.02	0.02
Pan troglodytes	643	111	532	3.52×10 ⁹	78	0.25	0	9	6	1	35	13	14	0.01	0	0	0.12	0.06	0.06
Homo sapiens	663	75	588	3.67×10 ⁹	97	0.27	0	8	5	2	45	16	21	0.01	0	0	0.13	0.07	0.06

Table 5.1: Summary of counts and fractions of Clusters in Teleosts and Mammalian species. Pseu: tRNA pseudogenes, GS: genome size, O: Observed pairs, R: pairs of the random simulation, P: p-value, Homogeneous configurations and He: Heteorgeneous configurations, ++: →→, +-: →←, and -: ←→

Appendix D



Distribution of tRNA clusters sizes for several species for which multiple sequenced genomes are available as well as some examples of individual genomes. Most tRNA clusters are small, and the frequency of long clusters rapidly decreases.

List of Figures

1.1	tRNA secondary structure	4
1.2	Summary	5
1.3	Stretch of the miR916 from <i>C. reinhardtii</i>	13
2.1	Summary of tRNA gene and tDNA statistics	21
2.2	Correlation of the number of tDNAs with genome size.	22
2.3	Distribution of tDNA clusters sizes for several lineages	23
2.4	Summary of tRNA gene and tDNA statistics	24
2.5	Cumulative distribution of tDNA pairs distances	25
2.6	Example of heterogeneous tDNA cluster	26
2.7	Relative abundance tRNA isoacceptor families located consecutively	26
2.8	Correlation of syntenic conservation of tDNA loci with genomic distance.	30
2.9	Comparison of the tRNA complement of Platyhelminth	33
3.1	Cluster detection	41
3.2	Distribution of frequency of clusters by size, distance and length.	42
3.3	Non-coding RNAs exhibit specific block patterns	44
3.4	Base pairing probabilities of reads mapped to ncRNA loci	45
3.5	HTS data reflects structural properties of ncRNAs	46
3.6	Box plots for 8 different features selected to train the random forest classifier	47
3.7	Parallel coordinate displays on the Random Forest approach	48
3.8	Heatmaps used to represent the votes for each class and proximity matrices	49
3.9	Short reads are produced from a wide variety of structured ncRNAs.	50
3.10	Decomposition of the cluster of reads at the <i>mir-125b-1</i>	51
4.1	Common system definition for tRNAs	56
4.2	Block expression visualization of Asparagine tRNAsuperfamily1 for Chicken	57
4.3	BLASTClust output example	59
4.4	Different processing patterns for some tRNAsuperfamilies	60
4.5	Block expression distribution for all the merged blocks	61
4.6	Block expression distribution for merged blocks	62
4.7	Hierarchical cluster analysis with p -values for block patterns (rat)	64
4.8	Hierarchical cluster analysis with p -values for block patterns (chicken)	65
4.9	Hierarchical cluster analysis with p -values for block patterns (dog)	66

List of Tables

1.1	Identified tRNAs	12
2.1	Comparison of observed and expected number of tRNA pairs	27
2.2	Fisher test results for Teleosteos species	28
2.3	Quantity structure of linkage analysis results in vertebrates	29
2.4	Syntenic conservation of tDNAs	29
3.1	Fractions of clusters and size of the cluster	42
3.2	Fractions of Clusters and their distance locations	42
3.3	Fraction of Clusters length	42
3.4	In total 434 of 852 clusters were found within regions of annotated ncRNA loci	43
3.5	Positive predictive values (PPV) and recall rates for training sets	48
4.1	Counts of tRNAscan-SE predictions and fraction of mapped tRNAs	58
5.1	Summary of counts and fractions of Clusters in Teleosts and Mammalian	73

Bibliography

- [1] The R project for statistical computing. <http://www.r-project.org/>.
- [2] Random forests. <http://www.stat.berkeley.edu/~breiman/RandomForests/>.
- [3] Random forests weka. Weka 3.5, University of Waikato, <http://www.cs.waikato.ac.nz/ml/weka/>.
- [4] Supplementary data in machine-readable form, 2009. <http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-050>.
- [5] A. Admire, L. Shanks, N. Danz, M. Wang, U. Weier, W. Stevens, E. Hunt, and T. Weinert. Cycles of chromosome instability are associated with a fragile site and are increased by defects in DNA replication and checkpoint controls in yeast. *Genes & Dev.*, 20:159–173, 2006.
- [6] N. Akimitsu. Messenger RNA surveillance systems monitoring proper translation termination. *J Biochem*, 143(1):1–8, 2008.
- [7] A. Alexandrov, I. Chernyakov, W. Gu, S. L. Hiley, T. R. Hughes, E. J. Grayhack, and E. M. Phizicky. Rapid tRNA decay can result from lack of nonessential modifications. *Mol. Cell*, 21:87–96, Jan 2006.
- [8] A. Ambrogelly, S. Palioura, and D. Söll. Natural expansion of the genetic code. *Nat Chem Biol*, 3:29–35, 2007.
- [9] Mette M. F. van der Winden J. Matzke A. J. Aufsatz, W. and M. Matzke. RNA-directed DNA methylation in arabidopsis. *Proc Natl Acad Sci U S A*, 99(NIL):1649916506, 2002.
- [10] J. Babiarz, J. Graham, Y. Wang, D. Bartel, and R. Blelloch. Mouse es cells express endogenous shRNAs, siRNAs, and other microprocessor-independent, dicer-dependent small rnas. *Gen Develop*, 22:2773–2785, 2008.
- [11] A. Barski, S Cuddapah, K. Cui, T-Y. Roh, D.E Schones, Z. Wang, G. Wei, Chepelev, and K. Keji Zhao. High-resolution profiling of histone methylations in the human genome. *Cell*, 129(4):823–37, 2007.
- [12] W. Bender. MicroRNAs in the Drosophila bithorax complex. *Genes Dev.*, 22:14–19, Jan 2008.

- [13] P.J. Beuning, F. Yang, P. Schimmel, and K. Musier-Forsyth. Specific atomic groups and RNA helix geometry in acceptor stem recognition by a tRNA synthetase. *Proc Natl Acad Sci U S A*, 94(19):10150–4, 1997.
- [14] E. Birney, J. A. Stamatoyannopoulos, A. Dutta, R. Guigo, T. R. Gingeras, E. H. Margulies, Z. Weng, M. Snyder, E. T. Dermitzakis, R. E. Thurman, M. S. Kuehn, and et. al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447:799–816, Jun 2007.
- [15] L. Breiman. *Machine Learning: Random forests*. Springer Netherlands, 2001.
- [16] L. Brendan, I. Anderson, R. Davies, U. Alsmark, and et al. The genome of the protist parasite *Entamoeba histolytica*. *Nature*, 433(7028):865–8, 2005.
- [17] K. Callahan and J. Butler. TRAMP complex enhances RNA degradation by the nuclear exosome component Rrp6. *J Biol Chem*, 285(6):3540–7, 2010.
- [18] I. Chernyakov, J. Whipple, L. Kotelawala, E. Grayhack, and Eric M. Phizicky. Degradation of several hypomodified mature tRNA species in *Saccharomyces cerevisiae* is mediated by Me exonuclease Rat1 and Xrn1. *GENES DEVELOP.*, 22:1369–1380, 2008.
- [19] J. Christoph, M. Rederstorff, J. Hertel, P. F. Stadler, I. L. Hofacker, M. Schrettl, H. Haas, and A. Hüttenhofer. Small ncRNA transcriptome analysis from *Aspergillus fumigatus* suggests a novel mechanism for regulation of protein-synthesis. *Nucleic Acids Res.*, 36:2677–2689, 2008.
- [20] T. Chung, O. Siol, T. Dingermann, and T. Winckler. Protein interactions involved in tRNA gene-specific integration of *Dictyostelium discoideum* non-long terminal repeat retrotransposon TRE5-A. *Mol Cell Biol*, 27(24):8492–501, 2007.
- [21] C. G. Clark, I. K. Ali, M. Zaki, B. J. Loftus, and N. Hall. Unique organisation of tRNA genes in *Entamoeba histolytica*. *Mol. Biochem. Parasitol.*, 146:24–29, 2006.
- [22] N. Cloonan, A. Forrest, G. Kolle, B. Gardiner, G. Faulkner, M. Brown, D. Taylor, A. Steptoe, S. Wani, G. Bethel, A. Robertson, A. Perkins, S. Bruce, C. Lee, S. Ranade, H. Peckham, J. Manning, K. McKernan, and S. Grimmond. Stem cell transcriptome profiling via massive-scale mRNA sequencing. *Nat Methods*, 5(7):613–9, 2008.
- [23] C. Cole, A. Sobala, C. Lu, S. R. Thatcher, A. Bowman, J. W. Brown, P. J. Green, G. J. Barton, and G. Hutvagner. Filtering of deep sequencing data reveals the existence of abundant Dicer-dependent small RNAs derived from tRNAs. *RNA*, 15:2147–2160, 2009.
- [24] J. R. Coleman, D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, and S. Mueller. Virus attenuation by genome-scale changes in codon pair bias. *Science*, 320:1784–1787, 2008.

- [25] L. A. Copela, G. Chakshumathi, R. L. Sherrer, and S. L. Wolin. The La protein functions redundantly with tRNA modification enzymes to ensure tRNA structural stability. *RNA*, 12:644–654, Apr 2006.
- [26] C. Copeland, M. Marz, D. Rose, J. Hertel, P. Brindley, C. Bermudez Santana, S. Kehr, Stephan-Otto C., and P.F. Stadler. Non-coding rna annotation of the *Schistosoma mansoni* genome. *BMC Genomics*, 10:464, 2009.
- [27] N. Copeland, N. Jenkins, and S. O'Brien. Genomics. Mmu 16—comparative genomic highlights. *Science*, 296(5573):1617–8, 2002.
- [28] M Crawley. *The R Book*. John Wiley & Sons, 2007.
- [29] C. de Duve. Transfer RNA: the second genetic code. *Nature*, 333(12):117–118, 1988.
- [30] W. Deng, X. Zhu, G. Skogerb, Y. Zhao, Z. Fu, Y. Wang, H. He, L. Cai, H. Sun, C. Liu, B. Li, B. Bai, J. Wang, D. Jia, S. Sun, H. He, Y. Cui, Y. Wang, D. Bu, and R. Chen. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: genomic features, biogenesis, and expression. *Genome Res.*, 16:20–29, Jan 2006.
- [31] A. M. Deshpande and C.. Newlon. DNA replication fork pause sites dependent on transcription. *Science*, 272:1030–1033, 1996.
- [32] M. Deutscher. Degradation of stable RNA in bacteria. *J Biol. Chem.*, 278:45041–45044, 2003.
- [33] M. Di Giulio. The origin of the tRNA molecule: implications for the origin of protein synthesis. *J. Theor. Biol.*, 226:89–93, 2004.
- [34] M. Di Giulio. Formal proof that the split genes of tRNAs of *Nanoarchaeum equitans* are an ancestral character. *J Mol Evol.*, 69:505–511, 2009.
- [35] S. Di Rienzi, D. Collingwood, M.K. Raghuraman, and B. Brewer. Fragile genomic sites are associated with origins of replication. *Genome Biol Evol*, 2009(NIL):350–63, 2009.
- [36] M. Doma and R. Parker. RNA quality control in Eukaryotes. *Cell*, 131:644–654, Nov 2007.
- [37] S. R. Eddy. E. P. Nawrocki, D. L. Kolbe. Infernal 1.0: Inference of RNA alignments. *Bioinformatics*, 25:1335–1337, 2009.
- [38] J.R. Ecker and R. W. Davis. Inhibition of gene expression in plant cells by expression of antisense RNA. *Proc Natl Acad Sci U S A*, 83(15):5372–6, 1986.
- [39] S. R. E Eddy and R. Durbin. RNA sequence analysis using covariance models. *Nucl. Acids Res*, 22:2079–2088, 1994.
- [40] L. Eichinger, J.A. Pachebat, G. Glockner, and et al. (97 co authors). The genome of the social amoeba *Dictyostelium discoideum*. *Nature*, 435:43–57, 2005.

- [41] M. Eigen and R. Winkler-Oswatitsch. Transfer-RNA, an early gene? *Naturwissenschaften*, 68:282–292, 1981.
- [42] Manfred Eigen, Björn F. Lindemann, M. Tietze, Ruthild Winkler-Oswatitsch, Andreas W. M. Dress, and Arndt von Haeseler. How old is the genetic code? Statistical geometry of tRNA provides an answer. *Science*, 244:673–679, 1989.
- [43] A. Emde, M. Grunert, D. Weese, K. Reinert, and S. Sperling. MicroRazerS: rapid alignment of small RNA reads. *Bioinformatics*, 26(1):123–4, 2010.
- [44] C. Ender, A. Krek, M. R. Friedlander, M. Beitzinger, L. Weinmann, W. Chen, S. Pfeffer, N. Rajewsky, and G. Meister. A human snoRNA with microRNA-like functions. *Mol. Cell*, 32:519–528, Nov 2008.
- [45] J. Fey, J.H. Weil, K. Tomita, A. Cosset, A. Dietrich, I. Small, and L. Marechal-Drouard. Role of editing in plant mitochondrial transfer RNAs. *Gene*, 286(1):21–4, 2002.
- [46] A. Fichant and C. Burks. Identifying potential tRNA genes in genomic DNA sequences. *J. Mol. Bio.*, 220:659–671, 1991.
- [47] Bhattacharyya S. N. Filipowicz, W. and N. Sonenberg. Mechanisms of posttranscriptional regulation by microRNAs: are the answers in sight? *Nat Rev Genet*, 9(NIL):102–104, 2008.
- [48] R. A. Fisher. On the interpretation of χ^2 from contingency tables, and the calculation of p . *J. Royal Stat. Soc.*, 85:87–94, 1922.
- [49] F. E. Frenkel, M. Chaley, E. V. Korotkov, and K. G. Skryabin. Evolution of tRNA-like sequences and genome variability. *Gene*, 335:57–71, 2004.
- [50] M. Friedlander, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knospel, and N. Rajewsky. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotechnol*, 26(4):407–15, 2008.
- [51] K. Fujishima, J. Sugahara, M. Tomita, and A. Kanai. Sequence evidence in the archaeal genomes that tRNAs emerged through the combination of ancestral genes as 5' and 3' tRNA halves. *PLoS One*, 3:e1622, 2008.
- [52] M. R. Garcia-Silva, M. Frugier, J. P. Tosar, A. Correa-Dominguez, L. Ronalte-Alves, A. Parodi-Talice, C. Rovira, C. Robello, S. Goldenberg, and A. Cayota. A population of tRNA-derived small RNAs is actively produced in *Trypanosoma cruzi* and recruited to specific cytoplasmic granules. *Mol Biochem Parasitol*, Feb 2010.
- [53] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman. Rfam: updates to the RNA families database. *Nucleic Acids Res.*, 37:D136–140, Jan 2009.
- [54] R. F. Gesteland and J. F. Atkins, editors. *The RNA World*. Cold Spring Harbor Laboratory Press, Plainview, NY, 1993.

- [55] E. A. Glazov, K. Kongsuwan, W. Assavalapsakul, P. F. Horwood, N. Mitter, and T. J. Mahony. Repertoire of bovine miRNA and miRNA-like small regulatory RNAs expressed upon viral infection. *PLoS ONE*, 4:e6349, 2009.
- [56] Evgeny A Glazov, Pauline A Cottee, Wesley C Barris, Robert J Moore, Brian P Dalrymple, and Mark L Tizard. A microRNA catalog of the developing chicken embryo identified by a deep sequencing approach. *Genome Res*, 18(6):957–64, 2008.
- [57] Evgeny A Glazov, Kritaya Kongsuwan, Wanchai Assavalapsakul, Paul F Horwood, Neena Mitter, and Timothy J Mahony. Repertoire of bovine miRNA and miRNA-like small regulatory RNAs expressed upon viral infection. *PLoS One*, 4(7):e6349, 2009.
- [58] J. M. Goodenbour and T. Pan. Diversity of tRNA genes in eukaryotes. *Nucl. Acids Res.*, 34:6137–6146, 2006.
- [59] J. Gordon, K. Byrne, and K. Wolfe. Additions, losses, and rearrangements on the evolutionary route from a reconstructed ancestor to the modern *Saccharomyces cerevisiae* genome. *PLoS Genet*, 5(5):e1000485, 2009.
- [60] S. Goto-Ito, T. Ito, M. Kuratani, Y. Bessho, and S. Yokoyama. Tertiary structure checkpoint at anticodon loop modification in tRNA functional maturation. *Nat Struct Mol Biol*, 16(10):1109–15, 2009.
- [61] O Gotoh. An improved algorithm for matching biological sequences. *J. Mol. Biol*, 162, 1982.
- [62] R Graham, J. Calvin, C Player, M Axtell, W Lee, C Nusbaum, H. Ge, and D. Bartel. Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in *C. elegans*. *Cell*, 127(6):1193–207, 2006.
- [63] P Green. Crossmatch. <http://www.phrap.org/phredphrap/general.html>.
- [64] S. Griffiths-Jones, A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. Rfam: an RNA family database. *Nucleic Acids Res.*, 31:439–441, Jan 2003.
- [65] M. Hackenberg, M. Sturm, D. Langenberger, J.M. Falcon-Perez, and A. Aransay. miRanalyzer: a microRNA detection and analysis tool for next-generation sequencing experiments. *Nucleic Acids Res*, 37(Web Server issue):W68–76, 2009.
- [66] H. J. Haiser, F. V. Karginov, G. J. Hannon, and M. A. Elliot. Developmentally regulated cleavage of tRNAs in the bacterium *Streptomyces coelicolor*. *Nucleic Acids Res.*, 36:732–741, Feb 2008.
- [67] D. Haussecker, Y. Huang, A. Lau, P. Parameswaran, A. Fire, and M. Kay. Human tRNA-derived small RNAs in the global regulation of RNA silencing. *RNA*, (1):1–5, 2010.
- [68] J. Hertel, I. L. Hofacker, and P. F. Stadler. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24:158–164, Jan 2008.

- [69] P. Higgs and W. Ran. Coevolution of codon usage and tRNA genes leads to alternative stable states of biased codon usage. *Mol Biol Evol*, 25(11):2279–91, 2008.
- [70] P. G. Higgs, D. Jameson, H. Jow, and M. Rattray. The evolution of tRNA-leu genes in animal mitochondrial genomes. *J. Mol. Evol.*, pages 435–445, 2003.
- [71] I. L. Hofacker and P. F. Stadler. Memory efficient folding algorithms for circular RNA secondary structures. *Bioinformatics*, 22:1172–1176, May 2006.
- [72] S. Hoffmann, C. Otto, S. Kurtz, C. Sharma, P. Khaitovich, P. F. Stadler, and J. Hackermueller. Fast mapping of short sequences with mismatches, insertions and deletions using index structures. *PLoS Comp. Biol.*, 5(9):e1000502., 2009.
- [73] M.J Hohn, H-S. Park, P. O'Donoghue, M. Schnitzbauer, and D. Söll. Emergence of the universal genetic code imprinted in an RNA record. *Proc Natl Acad Sci U S A*, 103(48):18095–100, 2006.
- [74] P. Hou, Y-M. Schimmel. A simple structural feature is a major determinant of the identity of a transfer RNA. *Nature*, 333(12):140–145, 1988.
- [75] J. Hsieh and C. A. Fierke. Conformational change in the *Bacillus subtilis* RNase P holoenzyme–pre-tRNA complex enhances substrate affinity and limits cleavage rate. *RNA*, 15:1565–1577, Aug 2009.
- [76] L-C. Hsieh, S.I Lin, A. Chun-Chieh, J. Chen, W-Y. Yi Lin, C. Tseng, W. Li, and T. Chiou. Uncovering small RNA-mediated responses to phosphate deficiency in *Arabidopsis* by deep sequencing. *Plant Physiol*, 151(4):2120–32, 2009.
- [77] Wolfgang Huber, Joern Toedling, and Lars M. Steinmetz. Transcript mapping with high-density oligonucleotide tiling arrays. *Bioinformatics*, 22:1963–1970, 2006.
- [78] N. Hubert, R. Walczak, C. Sturchler, E. Myslinski, C. Schuster, E. Westhof, P. Carbon, and A. Krol. RNAs mediating cotranslational insertion of selenocysteine in eukaryotic selenoproteins. *Biochimie*, 78:590–596, 1996.
- [79] G. Hutvagner and M. J. Simard. Argonaute proteins: key players in RNA silencing. *Nat Rev Mol Cell Biol*, 9(NIL):22–32, 2008.
- [80] A. Iacoangeli, T. S. Rozhdestvensky, N. Dolzhanskaya, B. Tournier, J. Schutt, J. Brosius, R. B. Denman, E. W. Khandjian, S. Kindler, and H. Tiedge. On BC1 RNA and the fragile X mental retardation protein. *Proc. Natl. Acad. Sci. USA*, 105:734–739, 2008.
- [81] O. Isken and L. Maquat. Quality control of eukaryotic mRNA: safeguarding cells from abnormal mRNA function. *Genes & devp.*, 21:1833–1856, 2007.
- [82] A S Ivessa, B A Lenzmeier, J B Bessler, L K Goudsouzian, S L Schnakenberg, and V A Zakian. The *saccharomyces cerevisiae* helicase Rrm3p facilitates replication past nonhistone protein-DNA complexes. *Mol. Cell*, 12:1525–1536, 2003.

- [83] D. S. Johnson, A. Mortazavi, R. Myers, and B Wold. Genome-wide mapping of in vivo protein-DNA interactions. *Science*, 316(5830):1497–502, 2007.
- [84] C. Jung, M. Hansen, I. Makunin, D. Korbie, and J.S. Mattick. Identification of novel non-coding RNAs using profiles of short sequence reads from next generation sequencing data. *BMC Genomics*, 11:77, 2010.
- [85] S. Kadaba, X. Wang, and J. Anderson. Nuclear RNA surveillance in *Saccharomyces cerevisiae*: Trf4p-dependent polyadenylation of nascent hypomethylated tRNA and an aberrant form of 5S rRNA. *RNA*, 12(3):508–21, 2006.
- [86] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermuller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammana, and T. R. Gingeras. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488, Jun 2007.
- [87] H. Kawaji, M. Nakamura, Y. Takahashi, A. Sandelin, S. Katayama, S. Fukuda, C. O. Daub, C. Kai, J. Kawai, J. Yasuda, P. Carninci, and Y. Hayashizaki. Hidden layers of human small RNAs. *BMC Genomics*, 9:157, 2008.
- [88] H. Kawaji, M. Nakamura, Y. Takahashi¹, A. Sandelin, S. Katayama, S. email, Fukuda, C. Daub, C. Kai, J. Jun Kawai, J. Yasuda, P. Carninci, and Y Hayashizaki. Hidden layers of human small RNAs. *BMC genomics*, (9):157, 2008.
- [89] H. Kawasaki and K. Taira. Induction of dna methylation and gene silencing by short interfering rnas in human cells. *Nature*, 431(NIL):211–7, 2004.
- [90] K. Kiontke and D. H. A. Fitch. The phylogenetic relationships of *Caenorhabditis* and other rhabditids. In The *C. elegans* Research Communit, editor, *Wormbook*. Wormbook, 2005. doi/10.1895/wormbook.1.11.1, <http://www.wormbook.org>.
- [91] T. Kirsten and E. Rahm. BioFuice: Mapping-based data intergation in bioinformatics. In Ulf Leser, Felix Naumann, and Barbara Eckmann, editors, *Data Integration in the Life Sciences*, volume 4075 of *Lect. Notes Comp. Sci.*, pages 124–135, 2006. Third International Workshop, DILS 2006.
- [92] C.H. Kuo and H. Ochman. Deletional Bias across the Three Domains of Life. *Genome Biol Evol*, 2009(NIL):145–52, 2009.
- [93] K. Labib and B. Hodgson. Replication fork barriers: pausing for a break or stalling for time? *Embo reports*, 8(4):8492–501, 2007.
- [94] D. Langenberger, C. Bermudez-Santana, J. Hertel, S. Hoffmann, P. Khaitovich, and P. F. Stadler. Evidence for human microRNA-offset RNAs in small RNA sequencing data. *Bioinformatics*, 25:2298–2301, Sep 2009.
- [95] D. Langenberger, C. I. Bermudez-Santana, P. F. Stadler, and S. Hoffmann. Identification and classification of small rnas in transcriptome sequence data. *Pac Symp Biocomput*, pages 80–87, 2010.

- [96] M. Lechner, L. Steiner, and S. J. Prohaska. Proteinortho — orthology detection tool, 2009. <http://www.bioinf.uni-leipzig.de/~marcus/software/proteinortho/>.
- [97] Y S Lee, H K Kim, S Chung, K S Kim, and A J Dutta. Depletion of human micro-RNA miR-125b reveals that it is critical for the proliferation of differentiated cells but not for the down-regulation of putative targets during differentiation. *Biol Chem.*, 280:16635–16641, 2005.
- [98] Y S Lee, Y Shibata, A Malhotra, and A Dutta. A novel class of small RNAs: tRNA-derived RNA fragments (tRFs). *Genes Dev.*, 23:2639–2649, 2009.
- [99] J. Lehmann, P.F Stadler, and S. Prohaska. SynBlast: assisting the analysis of conserved synteny information. *BMC Bioinformatics*, 9(NIL):351, 2008.
- [100] J. Leigh and F. Lang. Mitochondrial 3' tRNA editing in the jakobid *Seculamonas ecuadoriensis*: a novel mechanism and implications for tRNA processing. *RNA*, 10(4):615–21, 2004.
- [101] T. Lengauer. *Bioinformatics: from Genomes to Therapies*. WILEY-VCH Verlag GmbH & Co., Weinheim, 2007.
- [102] H. Li, J. Ruan, and R. Durbin. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, 18:1851–1858, Nov 2008.
- [103] L. Li, Christian J. Stoeckert Jr, and David S. Roos Roos. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, 13:2178–2189, 2003.
- [104] R. Li, Y. Li, K. Kristiansen, and J. Wang. SOAP: short oligonucleotide alignment program. *Bioinformatics*, 24:713–714, Mar 2008.
- [105] Y. Li, J. Luo, H. Zhou, J. Liao, L. Ma, Y. Chen, and L. Qu. Stress-induced tRNA-derived RNAs: a novel class of small RNAs in the primitive eukaryote *Giardia lamblia*. *Nucleic Acids Res*, 36:6048–6055, 2008.
- [106] Y. Li, J. Luo, H. Zhou, J. Liao, Li Ma, Y. Chen, and L. Qu. Stress-induced tRNA-derived RNAs: a novel class of small RNAs in the primitive eukaryote *Giardia lamblia*. *Nucleic Acids Res*, 36:6048–6055, 2008.
- [107] Z. Li, X. Gong, V. H. Joshi, and M. Li. Co-evolution of tRNA 3' trailer sequences with 3' processing enzymes in bacteria. *RNA*, 11:567–577, May 2005.
- [108] H. Lin, Z. Zhang, M. Q. Zhang, B. Ma, and M. Li. ZOOM! Zillions of oligos mapped. *Bioinformatics*, 24:2431–2437, Nov 2008.
- [109] R. Lister, R. O'Malley, J. Tonti-Filippini, B. Gregory, C. Berry, A. Millar, and J. Ecker. Highly integrated single-base resolution maps of the epigenome in *Arabidopsis*. *Cell*, 133(3):523–36, 2008.
- [110] K.M Lonergan and M.V. Gray. Editing of transfer RNAs in *Acanthamoeba castellanii* mitochondria. *Science*, 259(5096):812–6, 1993.

- [111] T. M. Lowe and S. R. Eddy. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.*, 25:955–964, 1997.
- [112] S. Lykke-Andersen, D. Brodersen, and T.H. Jensen. Origins and activities of the eukaryotic exosome. *J Cell Sci*, 122(Pt 10):1487–94, 2009.
- [113] C. Marck and H. Grosjean. tRNomics: analysis of tRNA genes from 50 genomes of Eukarya, Archaea, and Bacteria reveals anticodon-sparing strategies and domain-specific features. *RNA*, 8:1189–1232, 2002.
- [114] C Marck, R Kachouri-Lafond, I Lafontaine, E Westhof, B Dujon, and H. Grosjean. The RNA polymerase III-dependent family of genes in hemiascomycetes: comparative RNomics, decoding strategies, transcription and evolutionary implications. *Nucleic Acids Res.*, 34:1816–1835, 2006.
- [115] L. Marechal-Drouard, R. Kumar, C. Remacle, and I. Small. RNA editing of larch mitochondrial tRNA(His) precursors is a prerequisite for processing. *Nucleic Acids Res*, 24(16):3229–34, 1996.
- [116] M. Margulies, M. Egholm, W. Altman, S. Attiya, J. Bader, L. Bemben, J. Berka, M. Braverman, Y. Chen, Z. Chen, S. Dewell, Lei Du, J. Fierro, X. Gomes, B. Godwin, W. He, S. Helgesen, C.H. Ho, G. Irzyk, S. Jando, M. Alenquer, T. Jarvie, K. Jirage, J-B. Kim, J. Knight, J.R. Lanza, J. Leamon, S. Lefkowitz, M. Lei, J. Li, K. Lohman, H. Lu, V. Makhijani, K. McDade, M. McKenna, E. Myers, E. Nickerson, J. Nobile, R. Plant, B. Puc, M. Ronan, G. Roth, G. Sarkis, J. Simons, J. Simpson, M. Srinivasan, K. Tartaro, A. Tomasz, K. Vogt, G. Volkmer, S. Wang, Y. Wang, M. Weiner, P. Yu, R. Begley, and J. Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, 437(7057):376–80, 2005.
- [117] J. Marioni, C. Mason, S. Mane, M. Stephens, and Y. Gilad. RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Res*, 18(9):1509–17, 2008.
- [118] Chen Y-M, Foss K, Schneider J, McClain, W. Association of transfer RNA acceptor identity with a helical irregularity. *Science*, 241:1681–1684, 1988.
- [119] W.C. McClain. Surprising contribution to aminoacylation and translation of non-WatsonCrick pairs in tRNA. *PNAS*, 103:4570–4575, 2006.
- [120] R J McFarlane and S K. Whitehall. tRNA genes in eukaryotic genome organization and reorganization. *Cell Cycle*, 8:3102–3106, 2009.
- [121] W. Miller, K. Rosenbloom, R. C. Hardison, M. Hou, J. Taylor, B. Raney, R. Burhans, D. C. King, R. Baertsch, D. Blankenberg, S. L. Kosakovsky, A. Nekrutenko, B. Gardine, R. S. Harris, S. Tyekucheva, M. Diekhans, T. H. Pringle, W. J. Murphy, A. Lesk, G. M. Weinstock, K. Lindblad-Toh, R. A. Gibbs, E. S. Lander, A. Siepel, D. Haussler, and W. J. Kent. 28-way vertebrate alignment and conservation track in the UCSC genome browser. *Genome Res.*, 17:1797–1808, 2007.

- [122] D. Moazed. Small RNAs in transcriptional gene silencing and genome defence. *Nature*, 457:413–420, Jan 2009.
- [123] A Molnár, F Schwach, D J Studholme, E C Thuenemann, and D C Baulcombe. miRNAs control gene expression in the single-cell alga *Chlamydomonas reinhardtii*. *Nature*, 447:1126–1129, 2007.
- [124] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Barbara Wold. Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nat Methods*, 5(7):621–8, 2008.
- [125] P Mukhopadhyay, S Basak, and T C Ghosh. Nature of selective constraints on synonymous codon usage of rice differs in GC-poor and GC-rich genes. *Gene*, 400:71–81, 2007.
- [126] D. Murphy, B. Dancis, and J. R. Brown. The evolution of core proteins involved in microRNA biogenesis. *BMC Evol Biol*, 8(NIL):92, 2008.
- [127] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320(5881):1344–9, 2008.
- [128] Y. Nakamura. Codon usage database. NCBI-GenBank, Flat File Release 160.0 [June 15 2007], <http://www.kazusa.or.jp/codon/>.
- [129] K. Nakanishi and O. Nureki. Recent progress of structural biology of tRNA processing and modification. *Mol. Cells*, 19:157–166, Apr 2005.
- [130] C. Napolii, C. Lemieux, and R. Jorgensen. Introduction of a Chimeric Chalcone Synthase Gene into Petunia Results in Reversible Co-Suppression of Homologous Genes in trans. *Plant Cell*, 2(4):279–289, 1990.
- [131] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25:1335–1337, May 2009.
- [132] H. Nishihara and Okada N. Smit A. F. Functional noncoding sequences derived from SINEs in the mammalian genome. *Genome Res*, 16:864–874, 2006.
- [133] M Nozawa, Y Kawahara, and M. Nei. Genomic drift and copy number variation of sensory receptor genes in humans. *Proc Natl Acad Sci U S A.*, 104:20421–20426, 2007.
- [134] J. O’Rourke and M. Swanson. Mechanisms of RNA-mediated disease. *J Biol Chem*, 284(12):7419–23, 2009.
- [135] E. Passarge, B. Horsthemke, and R. A. Farber. Incorrect use of the term synteny. *Nat Genet*, 23(4):387, 1999.
- [136] N. Pater. Enhancing Random Forest Implementation in Weka. *Machine learning conference paper for ECE591Q*, Nov:1–10, 2005.

- [137] A. Pavesi, F. Conterio, A. Bolchi, G. Dieci, , and S. Ottonello. Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucl. Acids Res*, 22:1247–1256, 1994.
- [138] G. H. Perry, F. Yang, T. Marques-Bonet, C. Murphy, T. Fitzgerald, A. S. Lee, C. Hyland, A. C. Stone, M. E. Hurles, C. Tyler-Smith, E. E. Eichler, N. P. Carter, C. Lee, and R. Redon. Copy number variation and evolution in humans and chimpanzees. *Genome Res.*, 18:1698–1710, 2008.
- [139] H. Persson, A. Kvist, J. Vallon-Christersson, P. Medstrand, A. Borg, and C. Rovira. The non-coding RNA of the multidrug resistance-linked vault particle encodes multiple regulatory small RNAs. *Nat. Cell Biol.*, 11:1268–1271, Oct 2009.
- [140] E. M. Phizicky. Have tRNA, will travel. *Proc. Natl. Acad. Sci. U.S.A.*, 102:11127–11128, Aug 2005.
- [141] E. M. Phizicky and J. D. Alfonzo. Do all modifications benefit all tRNAs? *FEBS Lett.*, 584:265–271, Jan 2010.
- [142] J. Putz, R. Giege, and C. Florentz. Diversity and similarity in the tRNA world: overall view and case study on malaria-related tRNAs. *FEBS Lett*, 584(2):350–8, 2010.
- [143] L. Rabiner. A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE*, 77:257286, 1989.
- [144] T. A. Rawlings, T. M. Collins, and R. Bieler. Changing identities: tRNA duplication and remolding within animal mitochondrial genomes. *Proc. Acad. Natl. USA*, 100:15700–15705, 2003.
- [145] M. Rederstorff, S. H. Bernhart, A. Tanzer, M. Zywicki, K. Perfler, M. Lukasser, I. L. Hofacker, and A. Huttenhofer. RNPomics: Defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles. *Nucleic Acids Res*, Feb 2010.
- [146] E. Rocha. Codon usage bias from tRNA's point of view: Redundancy, specialization, and efficient decoding for translation optimization. *Genomes Res.*, 14:2279–2286, 2004.
- [147] S. Rodin, Ohno. S., and A. Rodin. Transfer RNAs with complementary anticodons: could they reflect early evolution of discriminative genetic code adaptors? *Proc. Natl. Acad. Sci. USA*, 90:4723–4727, 1993.
- [148] T. S. Rozhdestvensky, A. M. Kopylov, J. Brosius, and A. Hüttenhofer. Neuronal BC1 RNA structure: evolutionary conversion of a tRNA(Ala) domain into an extended stem-loop structure. *RNA*, 7:722–730, 2001.
- [149] F. Sanger, G.M. Air, B. Barrell, N. L. Brown, A. R. Coulson, J.C. Fiddes, C. A. Hutchison, P. M. Slocombe, and M. Smith. Nucleotide sequence of bacteriophage phi-x174 DNA. *Nature*, 265:687–695, 1977.

- [150] F. Sanger, S. Nicklen, and A.R. Coulson. DNA sequencing with chain-terminating inhibitors. *PNAS*, 74:5463 – 5467, 1977.
- [151] A. A. Saraiya and C. C. Wang. snoRNA, a novel precursor of microRNA in *Giardia lamblia*. *PLoS Pathog.*, 4:e1000224, Nov 2008.
- [152] P. Schimmel, R. Giege, D. Moras, and S. Yokoyama. An operational RNA code for amino acids and possible relationship to genetic code. *Proc Natl Acad Sci U S A*, 90(19):8763–8, 1993.
- [153] S.C. Schuster. Next-generation sequencing transforms today's biology. *Nature methods*, 5(1):16–18, 2008.
- [154] H. Shaheen, R. Horetsky, S. Kimball, A. Murthi, L. Jefferson, and A. Hopper. Retrograde nuclear accumulation of cytoplasmic tRNA in rat hepatoma cells in response to amino acid deprivation. *Proc Natl Acad Sci USA*, 104(21):8845–8850, 2007.
- [155] J. Shendure, G. Porreca, N. Reppas, X. Lin, J.P. McCutcheon, A. Rosenbaum, M. Wang, K. Zhang, R. Mitra, and G. Church. Accurate multiplex polony sequencing of an evolved bacterial genome. *Science*, 309(5741):1728–32, 2005.
- [156] K Sheppard, P M Akochy, and D. Söll. Assays for transfer RNA-dependent amino acid biosynthesis. *Methods*, 44:139–145, 2008.
- [157] W. Shi, D. Hendrix, M. Levine, and B. Haley. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat. Struct. Mol. Biol.*, 16:183–189, Feb 2009.
- [158] W. Shi, D. Hendrix, M. Levine, and B. Haley. A distinct class of small RNAs arises from pre-miRNA-proximal regions in a simple chordate. *Nat Struct Mol Biol.*, 16:183–189, 2009.
- [159] A. F. A. Smit, R. Hubley, and P. Green. RepeatMasker. version, open-3.2.5 [RMLib: 20080611], <http://www.repeatmasker.org/>.
- [160] M.S. Smith.T.F., Waterman and W.M. Fitch. Comparative biosequence metrics. *J. Mol. Biol*, 18, 1981.
- [161] L D Spotila, H Hirai, D M Rekosh, and P T Lo Verde. A retroposon-like short repetitive DNA element in the genome of the human blood fluke, *Schistosoma mansoni*. *Chromosoma*, 97:421–428, 1989.
- [162] P. F. Stadler, J. J. Chen, J. Hackermuller, S. Hoffmann, F. Horn, P. Khaitovich, A. K. Kretzschmar, A. Mosig, S. J. Prohaska, X. Qi, K. Schutt, and K. Ullmann. Evolution of vault RNAs. *Mol. Biol. Evol.*, 26:1975–1991, Sep 2009.
- [163] A. Stark, N. Bushati, C. H. Jan, P. Kheradpour, E. Hodges, J. Brennecke, D. P. Bartel, S. M. Cohen, and M. Kellis. A single Hox locus in *Drosophila* produces functional microRNAs from opposite DNA strands. *Genes Dev.*, 22:8–13, Jan 2008.

- [164] S. Steinberg, A. Misch, and M. Sprinzl. Compilation of trna sequences and sequences of trna genes. *Nucl. Acids Res*, 21:3011–3015, 1993.
- [165] R. Strausberg and S. Levy. Promoting transcriptome diversity. *Genome Res*, 17(7):965–8, 2007.
- [166] J Sugahara, K Kikuta, K Fujishima, N Yachie, M Tomita, and A. Kanai. Comprehensive analysis of archaeal tRNA genes reveals rapid increase of tRNA introns in the order thermoproteales. *Mol Biol Evol.*, 25:2709–2716, 2008.
- [167] F. J. Sun, S. Fleurdépine, C. Bousquet-Antonelli, G. Caetano-Anollés, and J. M. Deragon. Common evolutionary trends for SINE RNA structures. *Trends Genet.*, 23:26–33, 2007.
- [168] R. Suzuki and H. Shimodaira. pvclust: An R package for hierarchical clustering with p-values. <http://www.is.titech.ac.jp/~shimo/prog/pvclust/>.
- [169] R. J. Taft, E. A. Glazov, N. Cloonan, C. Simons, S. Stephen, G. J. Faulkner, T. Lassmann, A. R. Forrest, S. M. Grimmond, K. Schroder, K. Irvine, T. Arakawa, M. Nakamura, A. Kubosaki, K. Hayashida, C. Kawazu, M. Murata, H. Nishiyori, S. Fukuda, J. Kawai, C. O. Daub, D. A. Hume, H. Suzuki, V. Orlando, P. Carninci, Y. Hayashizaki, and J. S. Mattick. Tiny RNAs associated with transcription start sites in animals. *Nat. Genet.*, 41:572–578, May 2009.
- [170] R. J. Taft, E. A. Glazov, T. Lassmann, Y. Hayashizaki, P. Carninci, and J. S. Mattick. Small RNAs derived from snoRNAs. *RNA*, 15:1233–1240, Jul 2009.
- [171] R. J. Taft, C. D. Kaplan, C. Simons, and J. S. Mattick. Evolution, biogenesis and function of promoter-associated RNAs. *Cell Cycle*, 8:2332–2338, Aug 2009.
- [172] A. Takashi, I. Toshimichi, O. Yasuo, U. Hiroshi, K. Makoto, K. Shigehiko, Y. Yuko, M. Akira, and I. Hachiro. tRNADB-CE: tRNA gene database curated manually by experts. *Nucleic Acids Res.*, 37:D163–D168, 2009.
- [173] P. B. Talbert and S. Henikoff. Chromatin-based transcriptional punctuation. *Genes Dev.*, 23:1037–1041, 2009.
- [174] K Tamura, S Subramanian, and S Kumar. Temporal patterns of fruit fly (*Drosophila*) evolution revealed by mutation clocks. *Mol Biol Evol*, 21(1):36–44, Jan 2004.
- [175] H. Tang, J. Bowers, X. Wang, R. Ming, M. Alam, and A. Paterson. Synteny and collinearity in plant genomes. *Science*, 320(5875):486–8, 2008.
- [176] A. Tanzer, M. Riester, J. Hertel, C.I Bermudez-Santana, J. Gorodkin, I. Hofacker, and P.F. Stadler. *Evolutionary Genomics*. John Wiley & Sons, 2010.
- [177] B. Tawari, I. K. Ali, C. Scott, M. A. Quail, M. Berriman, N. Hall, and C. G. Clark. Patterns of evolution in the unique tRNA gene arrays of the genus *Entamoeba*. *Mol. Biol. Evol.*, 25:187–198, 2008.

- [178] D. M. Thompson, C. Lu, P. J. Green, and R. Parker. tRNA cleavage is a conserved response to oxidative stress in eukaryotes. *RNA*, 14:2095–2103, Oct 2008.
- [179] D. M. Thompson and R. Parker. The RNase Rny1p cleaves tRNAs and promotes cell death during oxidative stress in *Saccharomyces cerevisiae*. *J. Cell Biol.*, 185:43–50, Apr 2009.
- [180] D.M. Thompson and R. Parker. Stressing Out over tRNA Cleavage. *Cell*, 138:215–219, 2009.
- [181] D. M. Tyler, K. Okamura, W. J. Chung, J. W. Hagen, E. Berezikov, G. J. Hannon, and E. C. Lai. Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci. *Genes Dev.*, 22:26–36, Jan 2008.
- [182] G. Vivó-Truyols, J. R. Torres-Lapasió, A. M. van Nederkassel, Y. Vander Heyden, and D. L. Massart. Automatic program for peak detection and deconvolution of multi-overlapped chromatographic signals: Part II: Peak model and deconvolution algorithms. *J. Chromatography A*, 1096:146–155, 2005.
- [183] J. N. Volff and J. Brosius. Modern genomes with retro-look: retrotransposed elements, retroposition and the origin of new genes. *Genome Dyn.*, 3:175–190, 2007.
- [184] X. Wang, H. Jia, E. Jankowsky, and J. T. Anderson. Degradation of hypomodified tRNA(iMet) in vivo involves RNA-dependent ATPase activity of the DExH helicase Mtr4p. *RNA*, 14:107–116, Jan 2008.
- [185] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet*, 10(1):57–63, 2009.
- [186] F. Wei, S. Li, and H. R Ma. Computer simulation of tRNA evolution. *J. Phys. A: Math. Theor.*, 42:345101, 2009.
- [187] M. Withers, L. Wernisch, and M. dos Reis. Archaeology and evolution of transfer RNA genes in the *Escherichia coli* genome. *RNA*, 12:933–942, 2006.
- [188] I. Witten and E. Frank. *Data Mining: practical machine learning tools and techniques*. Morgan Kaufmann publications, 2005.
- [189] S. Wolin and A.G. Matera. The trials and travels of tRNA. *Gene. Deve.*, 13:1–10, 1999.
- [190] S. Yamasaki, P. Ivanov, G-F. Hu, and A. Anderson. Angiogenin cleaves tRNA and promotes stress-induced translational repression. *J Cell Biol*, 185(1):35–42, 2009.
- [191] Q. Zhao, O. Caballero, S. Levy, B. Stevenson, C. Iseli, S. de Souza, P. Galante, D. Busam, M. Leversha, K. Chadalavada, Y. Rogers, J. Venter, A. Simpson, and R. Strausberg. Transcriptome-guided characterization of genomic rearrangements in a breast cancer cell line. *Proc Natl Acad Sci U S A*, 106(6):1886–91, 2009.

- [192] T. Zhao, G. Li, S. Mi, S. Li, G. Hannon, X. Wang, and Y. Qi. A complex system of small RNAs in the unicellular green alga *Chlamydomonas reinhardtii*. *Genes Dev*, 21(10):1190–203, 2007.

Curriculum Vitae

Personal Information

Name	Clara Isabel Bermudez Santana
Address	Ritterstr. 12, 04109 Leipzig
Email	clara@bioinf.uni-leipzig.de
Date and place of birth	11.06.1973, Chiquinquira, Colombia
Nationality	Colombian
Marital status	Married
06/2000-02/2001	Maternity leave

High school Education

02/1984-12/1989	Liceo Femenino de Cundinamarca, Bogota, Colombia
-----------------	--

University Studies

02/1990-08/1997	Biology, National University of Colombia, Bogota, Colombia Diplom: Biology Diplom work: A model to characterize the secondary structure of tRNA isoacceptors of Escherichia coli by using graph theory". August, 1, 1997.
02/2001-04/2004	Magister Scientiae, National University of Colombia, Bogota, Colombia Diplom : Magister Scientiae Magister work : A model to establish structure-biological function correlations based on a tRNA characterization model, April 14, 2004.
since 10/2006	Ph.D. scholarship holder from DAAD-Alecol program. Leipzig University, Bioinformatics group.

Job experience

- | | |
|-----------------|--|
| Since 01/2006 | Ternured Assistant Professor. |
| 01/2005-12/2005 | Assistant Professor.
Department of Biology, National University of Colombia, Bogota. |
| 06/2004-12/2004 | Adjunct scientist: Evolutive Biology and Molecular Biology Group and the Theoretical Chemistry Group, Faculty of Science, Universidad Nacional de Colombia. Research project : " <i>Molecular descriptors to model tRNAs based on quantum and graph theoretical approaches</i> ". Grant: Colciencias. |
| 01/2002-06/2004 | Adjunct scientist: Evolutive Biology and Molecular Biology Group and the Theoretical Chemistry Group, Faculty of Science, Universidad Nacional de Colombia. Research project: " <i>tRNA structure model</i> " |
| 01/1999-12/2001 | Adjunct scientist: Evolutive Biology and Molecular Biology Group and the Theoretical Chemistry Group, Faculty of Science, Universidad Nacional de Colombia. Research project: <i>Characterization of tRNAs of different organisms using a theoretical model of weighted graphs</i> . Grant: Banco de la Republica of Colombia and DIB. |

Publications of Thesis

- **Bermudez-Santana, C.**, Stephan-Otto, C., Kirsten, T., Engelhardt, J. Prohaska, S., Steigle, S., Stadler, P.F. (2010). Genomic Organization of Eukaryotic tRNAs. BMC Genomics. In press.
- Tanzer, T., Riester, M., Hertel, J., **Bermudez-Santana, C.**, Gorodkin, J. Hofacker, I. Stadler. P.F. *Evolutionary Genomics and Systems Biology: Chapter: Evolutionary genomics of microRNAs and their relatives*. ISBN: 978-0-470-19514-7, March 2010, Wiley-Blackwell
- Marz, M., Tafer, H., Hertel, J., Bartschat, S., Kehr, S., Rose, D., Otto, W., Donath, A., Tanzer, A., **Bermudez-Santana, C.**, Gruber, A., Juhling, F., Engelhardt, J., Busch, A., Stadler, P.F., Dieterich, C. (2010). Comparative Analysis of Non-Coding RNAs in Nematodes Submitted.
- Langenberger D, **Bermudez-Santana, C.**, Stadler, P. F. Hoffmann, S. (2010). Identification and classification of small RNAs in transcriptome sequence data. Pacific Symposium on Biocomputing 15:81-87.
- Langenberger D, **Bermudez-Santana, C.**, Hertel J, Hoffmann S, Khaitovitch P, Stadler PF: (2009). Evidence for Human microRNA-Offset RNAs in Small RNA Sequencing Data", Bioinformatics, doi:10.1093/bioinformatics/btp419. 25(18):2298-2301.
- Copeland, C., Marz, M., Rose, D., Hertel, J., Brindley, P., **Bermudez-Santana, C.**, Kehr, S., Stephan-Otto, C., Stadler, P. F. (2009). Homology-Based Annotation of Non-coding RNAs in the Genomes of *Schistosoma mansoni* and *Schistosoma japonicum*. BMC Genomics, 10:464:1-13.

Publications: Other contributions

- Galindo, J.F. **Bermudez, C.**, Daza. E.E. tRNA Chemical Structure from a Quantum and Graph Theoretical Perspective, 2006. J. Theor. Biol. Volume 240: 4, 574-582.
- Galindo, J.F., **Bermudez, C.**, Daza E.E. *A classification of central nucleotides induced by the influence of neighboring nucleotides in triplets.* Journal of Molecular Structure. 2006, J. of Mol. Struct. THEOCHEM. Volume 769:1-3, 103-109.
- **Bermudez, C.**, Daza, E.E., Andrade. E. *Characterization and Comparison of Escherichia coli transfer RNAs by graph theory based on secondary structure.* Journal of Theoretical Biology, 1999, Vol. 197: 193–205

Talks

- Convention 23th TBI Winter-seminar- Computational Mathematics and Theoretical Biology in Bled, Slovenia, 14/02/2010-21/02/2010. Talk: “ *Searching tRNA processing patterns in transcriptome sequencing data.*”
- Herbstseminar Bioinformatik 2009 Vysoka Lipa (Decin), Czech Republic. 21/10/2009-25/10/2009. Talk: “ *tRNA pair arrangements*” .
- Herbstseminar Bioinformatik 2007. Studeny, Czech Republic. 27/10/2007-31/10/2007. Talk: “ *Patterns of mir916 in Chlamydomonas reinhardtii*” .
- Convention 23th TBI Winter-seminar- Computational Mathematics and Theoretical Biology in Bled, Slovenia, 17/02/2008-24/02/2008. . Talk :“ *Configurations of tRNA pair arrangements*” .
- Convention 22th TBI Winter-seminar- Computational Mathematics and Theoretical Biology in Bled, Slovenia, 18/02/2007-24/02/2007. Talk: “ *Doctorate Project*” .

Posters

- 2010 At the 14th International Conference on Research in Computational Molecular Biology (RECOMB 2010), August. Lisboa, Portugal. Searching tRNA processing patterns in transcriptome sequencing data. **Bermudez-Santana, C.**, Langenberger, D., Hoffmann, S., Stadler, P.F.
- 2009 At German Conference of Bioinformatic 2009, Halle, Germany, Sept. 28-30. tRNA Clusters. **Bermudez-Santana, C.**, Stephan-Otto, C., Kirsten, T., Prohaska, S., Steigle, S., Engelhardt, J., Stadler, P.F.
- 2009 At German Conference of Bioinformatic 2009, Halle, Germany, Sept. 28-30. Classification and Identification of Non-coding RNAs using High Throughput Sequencing Data, Langenberger, D., Hoffmann, S., **Bermudez-Santana, C.**, Stadler, P.F.
- 2006 Seminar of advanced studies in molecular design”, SEADIN. January 16 to 20, La Habana, Cuba. A structure–activity relation of the tRNA^{Ala} acceptor stem based on the charge distribution. Marin, R.M, C.I. **Bermudez, C.**, Daza, E.E.

- 2005 XXXI congress of theoretician chemists of latin expression, 2005. Islas Margaritas, Venezuela. Structure–activity correlations of the tRNA^{Ala} acceptor stem by a quantum chemical characterization. Marin, R., **Bermudez, C.**, Daza, E.E.
- 2005 XXXI congress of theoretician chemists of latin expression, QUITEL, 2005. Islas Margaritas, Venezuela. Anticodons classification by an isoprotonic set. Galindo, J.F., **Bermudez, C.**, Daza, E.E.
- 2005 XXXI congress of theoretician chemists of latin expression, QUITEL, 2005. Islas Margaritas, Venezuela. tRNA Structure from a Graph and Quantum Theoretical Perspective. Galindo, J.F., **Bermudez, C.**, Daza, E.E..
- 1997 7th International conference on mathematical Chemistry and 3rd Girona seminar molecular Similarity University of Girona, Girona (Spain), May 26-31 1997. Graph Modelling of tRNA isoacceptor molecules **Bermudez, C.**, Andrade, E. , Daza, E.E.

Awards

- | | |
|------|--|
| 2006 | Awarded with a scholarship at Colombian DAAD-Alecol program contest. |
| 2004 | Magister Scientiae work: with Honors |
| 1999 | 2nd prize: Best Undergraduate works. National University of Colombia |
| 1997 | Diplom work in Biology: with Honors. |

Other courses

- | | |
|-------------|---|
| 02/99-07/99 | Set Theory and Topology.
Mathematics Department, National University of Colombia. |
| 11/97-12/97 | Local training seminar on multivaried statistical methods with applications in human and social sciences.
Programme de recherche et d'enseignement en statistique applique, Laboratoire de methodologie du traitemet des Donnes, Bruxelles-Belgique, Universite Libre de Bruxelles, Department of Mathematics, Faculty of Science, Universidad Nacional de Colombia, |
| 04/96-05/96 | Mathematics for graduate students.
Department of Mathematics, Faculty of Science, Universidad Nacional de Colombia. Colombia. |

Communication skills

- | | |
|-------------|--|
| 10/06-03/07 | German Language Course. InterdaF at Herder Institute, Leipzig University. |
| 1999 | Certificate of achievement in English language. Colombian-American Center, Bogota, Colombia. |

Languages

- Perl, LaTeX, PSTricks and HTML.
Fundaments in ANSI C, PHP, Postscript and MySQL.
- Spanish, English and German (Fundaments).

Selbständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder Dienstleistungen als solche gekennzeichnet.

Clara Isabel Bermudez Santana

Leipzig, 13. September 2010