

# Strategies for Homology-Based Identification of Eukaryotic Non-Coding RNA Genes

UNIVERSITÄT LEIPZIG  
Fakultät für Mathematik und Informatik  
Institut für Informatik

**DISSERTATION**

zur Erlangung des akademischen Grades

**DOCTOR RERUM NATURALIUM**  
**(Dr. rer. nat.)**

im Fachgebiet

**Informatik**

vorgelegt von

**Dipl.-Biol. Dipl.-Inf. Manuela Marz**

geboren am 6. Mai 1981 in Leipzig

Leipzig, den 6. Juli 2009



## Abstract

Homology search is one of the basic tasks in comparative genomics: any investigation of the evolution of a family of genes or other functional elements is necessarily based on the prior knowledge of a set of homologous sequences that are to be compared. The rapidly increasing collection of completely sequenced genomes serves as the primary source of such sequence data. In principle, this is a straightforward problem of approximate string matching. In practise, however, it turns out to be a complex problem for non-coding RNAs, as a consequence of the peculiarities of the selective forces that govern their evolution. Not only non-coding RNAs are short, severely limiting the information contained in their sequences, but they also evolve rapidly, with more constraints on their structure than on the underlying sequence. Sequence conservation is also distributed very unevenly along the molecule, so that short highly conserved blocks are separated by highly variable domains. This thesis collects a series of case studies of a variety of small RNA families with the overarching aim of a better understanding of general patterns in ncRNA evolution.

Already the comparably well-conserved group of spliceosomal snRNAs shows a surprising variability even within the limited phylogenetic range of metazoa. Several highly diverged snRNAs show that expansion domains have to be expected even in the most highly conserved RNA families. The comprehensive analysis of animal snRNAs shows that most metazoan phyla, with the notable exception of the nematodes, have two distinct spliceosomes and a full complement of 9 spliceosomal snRNAs. In contrast, spliced leader (SL) RNAs appear in several eukaryotic phyla. A detailed comparison of their structures exhibits more similarities among them than previously thought that favours the hypothesis that they derive from a common ancestor and have been lost independently in many clades over the competing theory of frequent independent innovations of SL *trans*-splicing. SmY RNAs are an enigmatic class of small pol-III transcripts that appear to be involved in *trans*-splicing in most nematodes. They are studied here for the first time from an evolutionary perspective. The U7 snRNA is involved in the processing of the 3' ends of histone genes. It is transcribed as the shortest known pol-II transcript, with virtually no conserved sequence beyond its Sm and histone binding sites. Nevertheless, a nearly complete inventory of U7 snRNA in deuterostomes is reported here.

While the snRNAs are still relative well-behaved, there are several RNA families that are even harder to deal with because they do not only evolve rapidly at sequence level but also exhibit dramatic variations in size and structure. For the U3

snoRNA, for instance, only a few protein-binding motifs and relative small core structure is ubiquitously conserved among eukaryotes. In fungi, the search for U3 homologs is further complicated by the presence of introns, usually a very rare phenomenon among ncRNA genes. Similar patterns of structural variation are observed for RNase MRP, where – in certain clades – large parts of the structure can be missing while short stems can be enlarged to complex domains comprising a substantial fraction of the entire molecule. The 7SK RNA and the telomerase RNA are probably the most extreme examples. The 7SK RNA was until recently considered specific to vertebrates because attempts to find homologs in invertebrates with sequence-based methods have remained unsuccessful. Here, we provide evidence that the 7SK RNA, and its protein partners, are present throughout the entire animal kingdom. The situation is a bit different for telomerase RNA: while it is common consensus that all organism with “normal” telomers also have telomerase RNA, no homolog in any invertebrate has yet been confirmed, although as a first step, promising candidates have been derived for the sea urchin.

Homology-based searches for ncRNAs require the combination of multiple search tools, ranging from simple `blast` and semi-global sequence alignments to descriptor-based surveys for characteristic structural motifs. Depending on the specifics of each RNA family different strategies have been employed, following the same general principle. First, candidate sets are extracted from genomic data, which are then filtered down to a set for which detailed manual analysis becomes feasible. The same methodology was also applied to annotate ncRNAs in two genomes: *Trichoplax adhaerens*, one of the most basal metazoans, and *Schistosoma mansoni*, a parasitic platyhelminth of high medical importance. In both cases the ncRNA inventory was expanded way beyond what little was known before.

In summary, the work presented has extended the knowledge base on non-coding RNA by hundreds of novel ncRNA gene sequences, which in turn were employed to refine and improve the consensus structure models of more than a dozen RNA families. Taken together, the data provide novel and unexpected insights into the structural variability of the ncRNAs and emphasize the importance of large-scale structural variations in ncRNA evolution.

# Acknowledgements

First of all I would like to thank my supervisor Peter F. Stadler. To thank you, Peter, would not fit on that small page, and this time it does not even fit in the appendix. Therefore I would like to refer to the Supplemental Material<sup>1</sup>.

I want to thank my collaborators Anne-Catherine, Astrid, Denise, Axel, Ivo, Julian, Josef, Magnus, Olivier, Ronny, whom I was always to spam with many questions.

One day I will learn how to fill in forms and how to administrate my own computer. But right now, I want to thank Petra and Jens for their superb work! Without you it would not have been possible to finish this work within that time!

This work was partly supported by the "Landesstipendium Sachsen", the DFG-funded *Graduierten Kolleg "Wissensrepräsentation"*, and the DFG funded *Bioinformatics Initiative* and SYNLET 043312.

At many points within the last years I got a lot of mental help from Sonja, Pe{dro,t{ra,er}}, as well as my room mates Sven, Jane, Dom, i-like-the-poker-gruber, who always (had to) go with me through all my ups and downs.

In the middle of these acknowledgements and being the centre of my life, I want to thank my family for always being there for me at any time. Thank you Micha, Thank you Ferdinand. Thank you Bruderherz. Danke Dir Mama.

Very big thanks to my go-students, which never let me forget natural stages from the roots of life, via development and to pro-view: Anne, Stefan, Sylvia, Anja, Maria, Sonja, Dieter, Jan, David, Fenchel, Anke, Christian, Theresa, Uta, Mandy, Bernd, Tommy, Maria, Georg, Tom, Konrad, Peter, Harald, Jochen. I want to thank Saijo Masatake Sensei teaching me basic shapes of life. Combining hobby and job is possible through people like Magnus, Josef, Georg and Ingo. Thanks for your special offer and contribution.

Thanks to Kerstin, Maria, Sonja, Anke and Monika for weekly nice catchy songs in my ears.

I want thank room number 326 (Lydia, Wint, Worschtel, Markus, Christian) and Dom for the table tennis meetings whenever I had the feeling my blood is not carrying enough oxygen to my brain. Thanks to some more people who helped me not getting out of control during this work: Dom for my right tennis-arm, Sven for my left climbing arm.

I want to thank Allan for the thaypho's.

What is a dissertation without competition? (Special thanks to you, Axel ;P)

Some "übelst gemehrte" acknowledgements to Konstantin, who inspired me whenever we met (mostly in a pub).

Being a mom means indispensable exchange with other parents' feelings. Thanks to Clara, Kristin, Bärbel and Mr. Drasdisch for helping me to find the way between work and family. Subsequently thanks to Jane and Steffi providing emergency Ferdi-sitting-numbers.

Last but not least special thanks to all people which enabled these acknowledgements :)

---

<sup>1</sup>[www.bioinf.uni-leipzig.de/~manja/thx/](http://www.bioinf.uni-leipzig.de/~manja/thx/)



# Contents

List of Figures . . . . .	iii
List of Tables . . . . .	vii
Abbreviations . . . . .	ix
<b>1 Introduction</b>	<b>1</b>
1.1 Evolution of ncRNAs . . . . .	2
1.2 Homology Based Search of ncRNAs . . . . .	4
1.3 Overview of this thesis . . . . .	7
1.4 List of Publication and Supplements . . . . .	9
<b>2 Tools Of The Trade</b>	<b>13</b>
2.1 General Homology Search . . . . .	13
2.1.1 Sequence Based Search . . . . .	13
2.1.2 Structure Based Search . . . . .	17
2.1.3 Pattern and Structure Based Search (by Hand) . . . . .	19
2.2 Specific ncRNA Search-Programs . . . . .	25
2.3 Verification of Predicted ncRNAs . . . . .	27
2.3.1 Multiple Alignments . . . . .	27
2.3.2 Promoter Analysis . . . . .	29
2.3.3 Synteny Information . . . . .	30
2.3.4 Target-Prediction . . . . .	30
2.4 Other Commonly Used Programs . . . . .	31
2.4.1 Multiple Candidates . . . . .	31
2.4.2 RNAz and Annotation Pipeline . . . . .	31
2.4.3 Phylogenetic Analysis . . . . .	32
2.4.4 Discarding Repeats . . . . .	32
2.4.5 Example for Homology Search of All Known ncRNAs . . . . .	33
<b>3 RNAs involved in mRNA processing</b>	<b>35</b>
3.1 Evolution of the Splicing Machinery . . . . .	36
3.2 <i>cis</i> -splicing with small nuclear RNAs . . . . .	39

---

3.2.1	Homology Search . . . . .	41
3.2.2	Specific Upstream Elements . . . . .	45
3.2.3	Clusters of snRNA genes . . . . .	46
3.2.4	Phylogenetic Analysis and Paralogs . . . . .	46
3.2.5	Secondary Structures . . . . .	49
3.2.6	Syntenic Conservation . . . . .	51
3.2.7	Pseudogenes . . . . .	52
3.2.8	Discussion . . . . .	54
3.3	<i>trans</i> -splicing with splice leader RNAs . . . . .	56
3.3.1	Re-evaluation of secondary structures . . . . .	58
3.3.2	Phylum specific alignments . . . . .	59
3.3.3	Ubiquitous Sequence Features . . . . .	61
3.3.4	Secondary Structure Analysis . . . . .	63
3.3.5	Discussion . . . . .	66
3.4	SmY RNAs . . . . .	69
3.4.1	Initial SmY sequences. . . . .	69
3.4.2	Homology searches and a representative seed alignment. . . . .	70
3.4.3	Phylogenetic diversity of SmY RNAs. . . . .	72
3.4.4	Discussion . . . . .	74
3.5	U7 RNA . . . . .	75
3.5.1	<i>Bona fide</i> U7 snRNA Sequences . . . . .	77
3.5.2	More Distant Homologs? . . . . .	81
3.5.3	Discussion . . . . .	83
3.6	Introns in Insects . . . . .	84
3.6.1	Computational identification of spliced RNAs in <i>Drosophila</i> genomes . . . . .	84
3.6.2	Novel spliced transcripts . . . . .	85
3.6.3	Novel spliced non-coding RNAs . . . . .	89
3.6.4	Novel mlncRNAs are mostly unstructured . . . . .	89
3.6.5	Experimental verification of predicted mlncRNAs . . . . .	89
3.6.6	Discussion . . . . .	91
<b>4</b>	<b>Highly divergent ncRNAs</b> . . . . .	<b>93</b>
4.1	U3 RNA . . . . .	94
4.1.1	Homology search . . . . .	94
4.1.2	Introns in U3 snoRNA genes . . . . .	98
4.1.3	Promoters of U3 snoRNAs . . . . .	98
4.1.4	Multiplicity of U3 snoRNA genes . . . . .	98
4.1.5	Concluding Remarks . . . . .	100
4.2	RNase MRP and RNase P . . . . .	101



---

4.2.1	Homology Based Search . . . . .	102
4.2.2	Secondary Structure . . . . .	103
4.2.3	Conclusion . . . . .	106
4.3	7SK RNA . . . . .	107
4.3.1	Phylogenetic Distribution of HEXIM . . . . .	108
4.3.2	Phylogenetic Distribution of LARP7 . . . . .	110
4.3.3	Revised Secondary Structure Model of 7SK RNA . . . . .	111
4.3.4	Homology Search for 7SK snRNAs . . . . .	113
4.3.5	Discussion . . . . .	116
4.4	Telomerase RNA . . . . .	120
4.4.1	Homology Based Search . . . . .	121
4.4.2	Pipeline for Prediction of Divergent Telomerase Candidates . . . . .	122
4.4.3	Results . . . . .	126
<b>5</b>	<b>NcRNA Screens</b> . . . . .	<b>129</b>
5.1	<i>Trichoplax adhaerens</i> . . . . .	130
5.1.1	Results . . . . .	131
5.1.2	Discussion . . . . .	138
5.2	<i>Schistosoma mansoni</i> . . . . .	141
5.2.1	Results & Discussion . . . . .	142
5.2.2	Conclusions . . . . .	154
<b>6</b>	<b>Conclusion</b> . . . . .	<b>155</b>
<b>A</b>	<b>Alternativ Alignment Listener</b> . . . . .	<b>163</b>
A.1	ComposAlign . . . . .	163
<b>B</b>	<b>Genomes and Accesionnumbers</b> . . . . .	<b>177</b>
B.1	Sources of Used RNA Sequences . . . . .	177
B.2	Sources of SL NcRNAs . . . . .	178
B.3	7SK sequences . . . . .	179
B.4	FTP Sites of Genome Assemblies . . . . .	179
<b>C</b>	<b>Secondary Structures of SL-RNAs</b> . . . . .	<b>187</b>
	<b>Literature</b> . . . . .	<b>192</b>



# List of Figures

1.1	Sketch of the post-ENCODE view of a mammalian transcriptome .	2
1.2	Origins of major ncRNA families. . . . .	3
1.3	From genes to functional elements in eukaryots. . . . .	6
2.1	Searching for ncRNAs. . . . .	14
2.2	Histogram of score distribution for U4atac, U17 and RNase MRP. Circles denote true homologous. . . . .	17
2.3	A simple example about the outcome of <code>rnabob</code> . . . . .	20
2.4	Rewriting <code>rnabob</code> descriptor files. . . . .	22
2.5	An efficient search tool for fragmented patterns in genomic sequences. . . . .	24
3.1	Distribution of intron classes . . . . .	37
3.2	Splicing types. . . . .	38
3.3	Phylogenetic network of Drosophilid U5 snRNAs. . . . .	47
3.4	Phylogenetic tree of insect U4 snRNAs. . . . .	48
3.5	Phylogenetic networks of teleost fish snRNAs. . . . .	49
3.6	U12 snRNA secondary structures of <i>Capitella capitata</i> and <i>Xenopus tropicalis</i> . . . . .	51
3.7	Secondary structure prediction of U1 snRNA . . . . .	52
3.8	Secondary structures of U11, U12, U6atac in <i>Acyrtosiphon pisum</i> , <i>Drosophila melanogaster</i> and <i>Homo sapiens</i> . . . . .	53
3.9	Number of blast hits versus cut-off <i>E</i> -value for 6 different genomes	55
3.10	Schematic drawing of a typical SL RNA . . . . .	56
3.11	Evolution of SL RNAs. . . . .	57
3.12	Neighbor-net analysis of SL1 and SL2 RNAs. . . . .	60
3.13	Rotifera and nematoda SL RNAs. . . . .	60
3.14	Consensus of Donor splice site and Sm binding site. . . . .	61
3.15	Common sequence features of SL-RNAs. . . . .	62
3.16	Two alternative secondary structures of stem I. . . . .	65
3.17	Sequence and structure similarities. . . . .	66

3.18	Three alternative scenarios for the evolution of SL RNA. . . . .	68
3.19	Consensus SmY RNA structure. . . . .	71
3.20	Phylogenetic distribution of 147 identified SmY RNA homologs. . . . .	73
3.21	The pathway of mammalian histone pre-mRNA biosynthesis [1]. . . . .	76
3.22	Clusters of U7snRNA genes in <i>Xenopus</i> and zebrafish. . . . .	77
3.23	Conserved elements in functional U7 snRNA gene. . . . .	79
3.24	Manually curated alignment of functional U7 snRNA sequence. . . . .	80
3.25	Comparison of U7 hairpin structures. . . . .	81
3.26	Best U7 candidates in lamprey, <i>Branchiostoma floridae</i> , and <i>Bombyx mori</i> . . . . .	83
3.27	Overview of the computational intron prediction procedure. . . . .	86
3.28	Evaluating characteristic intron evolution. . . . .	87
3.29	Examples of transcript-confirmed intron predictions. . . . .	88
3.30	Experimentally verified introns in mlncRNA transcripts. . . . .	90
3.31	MlncRNA69E2 is alternatively spliced. . . . .	91
4.1	Secondary structure model of U3 snoRNA for eukaryots. . . . .	97
4.2	Overview of U3 snoRNAs found in fungi and their introns. . . . .	99
4.3	Distribution of copy numbers of U3 snoRNA genes. . . . .	100
4.4	Schematic drawing of the consensus structures of P and MRP RNAs. . . . .	102
4.5	Secondary structure of RNase MRP containing P7. . . . .	105
4.6	Promoter Region of RNase MRP and RNase P. . . . .	107
4.7	Distribution of HEXIM1/2, LARP7, MePCE/BCDIN3, and 7SK RNA. . . . .	109
4.8	NeighborNet of all metazoan HEXIMs created with SplitsTree [2]. . . . .	110
4.9	Common structural elements of 7SK snRNAs. . . . .	112
4.10	Revised secondary structure model of 7SK RNA. . . . .	113
4.11	Predicted human 7SK RNA. . . . .	114
4.12	Consensus sequence of M5. . . . .	114
4.13	7SK-automaton. . . . .	115
4.14	The secondary structure of the <i>C. briggsae</i> 7SK RNA. . . . .	117
4.15	Accessibility of single nucleotides in human 7SK RNA. . . . .	118
4.16	Telomerase RNA structures of yeast and human. . . . .	121
4.17	Pseudoknot of Telomerase RNAs. . . . .	124
5.1	Trichoplax pre-rRNA cluster. . . . .	131
5.2	Secondary structures of major and minor snRNAs in <i>Trichoplax</i> . . . . .	133
5.3	Structural alignments of the <i>Trichoplax</i> RNase P RNA, SRP RNA, and U3 snoRNA. . . . .	134

---

5.4	Secondary structure of <i>Trichoplax adhaerens</i> RNase MRP RNA. . .	136
5.5	Comparison of the tRNA complement of <i>Schistosoma mansoni</i> and <i>Schmidtea mediterranea</i> . . . . .	143
5.6	Fragments of RNA Operons. . . . .	145
5.7	Diagram of rRNA subunits. . . . .	146
5.8	5s rRNA of schistosomes . . . . .	146
5.9	Secondary structures of the 9 snRNAs and the interaction com- plexes of U4/U6 and U4atac/U6atac, respectively. . . . .	147
5.10	SL RNA sequences and structure of schistosomes. . . . .	149
5.11	SRP RNA of schistosomes. . . . .	150
5.12	Multiple sequence alignments of the <i>pre</i> -miRNAs in <i>S. mansoni</i> . .	151
5.13	U7 RNA of <i>Schistosoma japonicum</i> . . . . .	153
A.1	Transposition probabilities between Markov states. . . . .	166
A.2	Data flow diagram of COMPOSALIGN. . . . .	168
A.3	12 motifs and assignment of instruments to the transposed motifs. .	171
A.4	Mapping of fly species to instruments. . . . .	172



# List of Tables

3.1	Three Major Splicing Mechanisms. . . . .	36
3.2	Approximate copy number of snRNA genes . . . . .	42
3.3	Paralog groups of major spliceosomal snRNAs within major animal clades. . . . .	49
3.4	MFE of alternative SL secondary structures. . . . .	64
3.5	Previously published SmY RNA sequences. . . . .	70
3.6	Trusted U7 snRNA sequences. . . . .	78
4.1	Summary of the homology-based survey of U3 snoRNA. . . . .	95
4.2	Structural overview of RNase MRP stems. . . . .	104
4.3	Expected number of the templates. . . . .	122
4.4	Telomerase Pseudoknot <b>rnabob</b> -pattern included in <b>TR-PK-finder</b> . . . . .	124
5.1	Proximal sequence element (PSE) of snRNAs in <i>Trichoplax</i> . . . . .	132
5.2	Small nucleolar RNAs in <i>Trichoplax</i> . . . . .	135
5.3	RNAz screens of <i>Trichoplax adhaerens</i> genome. . . . .	138
5.4	Non-coding RNA predictions from the sequenced genome of <i>S. mansoni</i> . . . . .	142
5.5	Copy number of snRNAs in <i>S. mansoni</i> . . . . .	147
6.1	Presence and absence of spliced leader RNAs and minor spliceosomal snRNAs. . . . .	157
A.1	Input example for COMPOSALIGN with three species. . . . .	169
C.1	Sequences, secondary structures, and folding energies of known SL RNAs. . . . .	187





# Abbreviations

A	Adenine
blast	Basic Local Alignment Search Tool
bp	Basepairs
C	Cytosine
CDK9	Cyclin-dependent kinase 9
CDS	Coding Sequence
CPU	Central Processing Unit
DIALIGN	DIagonal ALIGNment
DNA	Desoxyribonucleic acid
ENCODE	ENCyclopedia Of DNA Elements
erpin	Easy RNA Profile Identification
EST	Expressed Sequence Tags
FDR	False Discovery Rate
Fig	Figure
G	Guanine
Gyr	Gigayear
HDE	Histone Downstream Element
IUPAC	International Union of Pure and Applied Chemistry
JGI	Joint Genome Institute
LARP	La-related protein
LHB	Late Heavy Bombardment
LSU	Large Subunit
LUCA	Last Universal Common Ancestor
MFE	Minimum Free Energy
MEME	Multiple EM for Motif Elicitation
miRNA	Micro RNA
lncRNA	mRNA-like non-coding RNA
mRNA	Messenger RNA
My	Million years
NCBI	National Center for Biotechnology Information

ncRNA	Non-coding RNA
nt	Nucleotides
OCT	Octamer
ORF	Open Reading Frame
P-TEFb	Positive Transcription Elongation Factor b
Pol	Polymerase
PSE	Proximal Sequence Element
RNA	Ribonucleic acid
RNaseP	Ribonuclease P
RNP	Ribonucleoprotein
rRNA	Ribosomal RNA
SCFG	Stochastic Context-Free Grammars
SL	Spliced leader
SMN	Survival of Motor Neuron
snoRNA	Small nucleolar RNA
snRNA	Small nuclear RNA
snRNP	Small nuclear Ribonucleoproteins
SRP	Signal Recognition Particle
SSU	Small Subunit
SVM	Support Vector Machine
T	Thymine
Tab	Table
TR	Telomerase RNA
TERT	Telomerase Reverse Transcriptase
tmRNA	Transfer-messenger RNA
TR	Telomerase RNA
tRNA	Transfer RNA
U	Uracil
UCA	Unknown Common Ancestor
UTR	Untranslated region

# Chapter 1

## Introduction

Comparing all three domains of life (Bacteria, Archaea and Eukarya) by 16S rRNA [3, 4], tRNA [5, 6] and protein [7] analysis the assumption of *one* origin of life, the last universal common ancestor (LUCA) is commonly believed among scientists.

The literature virtually agrees on the existence of an “RNA-Protein World” stage preceding the divergence of Bacteria, Archaea and Eukarya [8–11]. Reasons are (1) the inflexible double strand DNA molecule can exist as information storage only; (2) synthesis of proteins is possible without DNA, but not without RNA; (3) Desoxyribonucleotides are synthesized from ribonucleotides; (4) DNA-replication starts with the synthesis of an RNA primer; (5) RNA is foldable in three dimensions acting as information storage, information transmitter (mRNA), with structural functions (ncRNAs, ribonucleotides; see sec. 1.1); (6) RNAs may have catalytic abilities (ribozymes, like RNase P, sec. 4.2).

Considering life without DNA, the importance of ncRNAs as the only information storage is expound. Therefore this thesis about the evolution of ncRNAs has a large impact. In the scenario of LUCA with a RNA genome only, the transition to DNA genomes independently occurred twice (Bacteria and Archaea+Eukarya) [12] or even thrice [13], possibly mediated by viral entities [8]. NcRNAs are independently of DNA of such an importance to a cell for regulation, that they are involved in all kinds of processes. Specific regulation of possible “transcription”-like [11] processes could have been simply lifted to the novel DNA genome, thereby using the same regulatory sequences and the same protein factors. In [8] it is hypothesized, that the transition of DNA genome was possible only from an ancestral state in which the protein-production was regulated at least predominantly at the level of translation. Transcription regulation must have been a later innovation.

In this thesis we are interested in the evolution of ncRNAs.

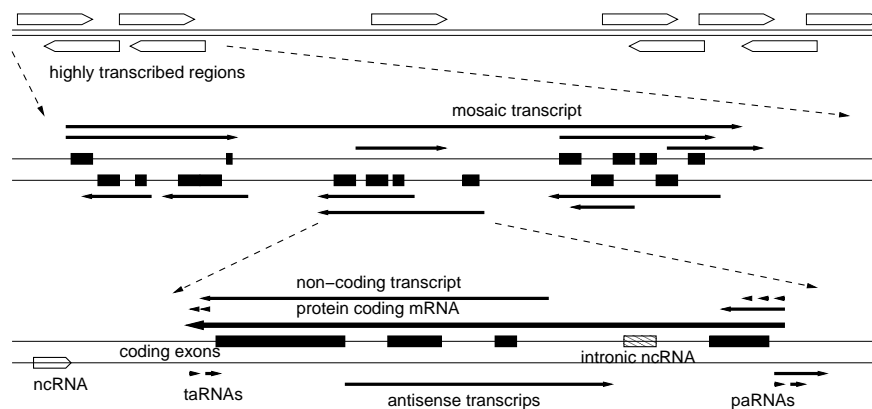


Figure 1.1: Sketch of the post-ENCODE view of a mammalian transcriptome (adapted from [17]). Highly transcribed regions consist of a complex mosaic of overlapping transcripts (arrows) in both reading-directions. These transcripts link together the locations of several protein coding genes (coding exon indicated by black rectangles). Conversely, multiple transcription products, many of which are non-coding, are processed from the same locus as a protein coding mRNA.

## 1.1 Evolution of ncRNAs

A decade ago, the genome was seen as a linear arrangement of separated individual genes which are predominantly protein-coding, with a small set of ancient non-coding “house-keeping” RNAs such as tRNA and rRNA dating all the way back to an RNA-World. However, in contrast to this simple view more recent studies reveal a much more complex genomic picture. The ENCODE Pilot Project [14], the mouse cDNA project FANTOM [15], and a series of other large scale transcriptome studies, e.g. [16], leave no doubt that the mammalian transcriptome is characterized by a complex mosaic of overlapping, bi-directional transcripts and a plethora of non-protein coding transcripts arising from the same locus, Fig. 1.1.

This newly discovered complexity is not unique to eukaryots. Similar high-throughput studies in invertebrate animals [18, 19] and plants [20] demonstrate the generality of the mammalian genome organization among higher eukaryots. Even the yeasts *Saccharomyces cerevisiae* and *Saccharomyces pombe*, whose genomes have been considered to be well understood, are surprising us with a much richer repertoire of transcripts than previously thought [21–24]. Even in bacteria, an unexpected complexity of regulatory RNAs was discovered in recent years [25].

Given the importance and ubiquity of non-coding RNAs (ncRNAs) and RNA-based mechanisms in all extant lifeforms, it is surprising that we still know relatively little about the evolutionary history of most RNA classes, although a series of systematic studies have greatly improved the understanding since the first attempt at a comprehensive review of this topic [26].

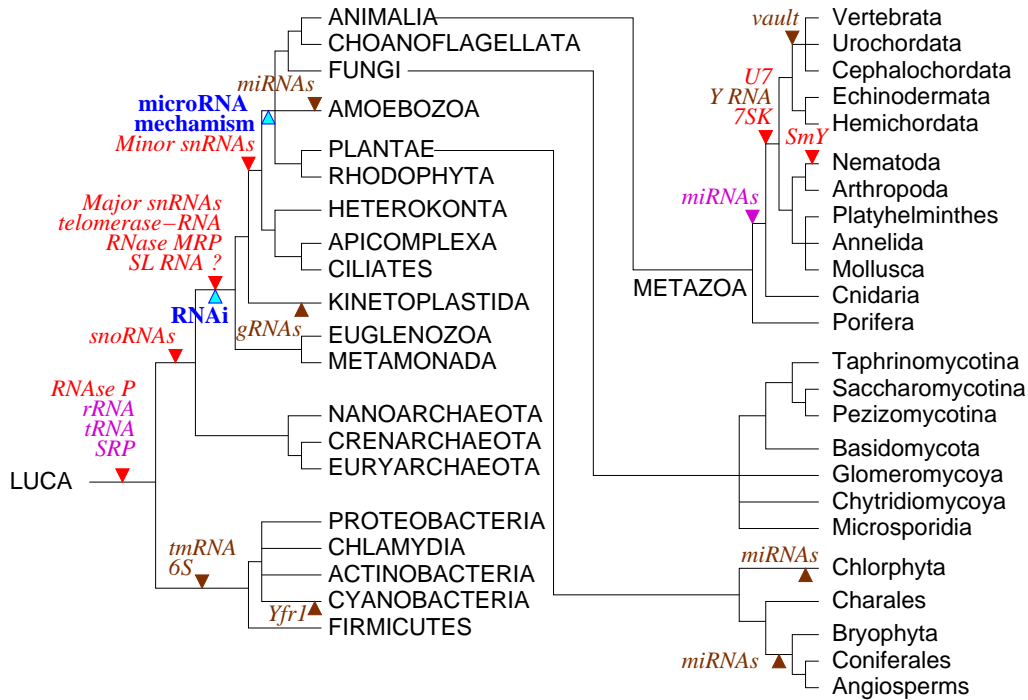


Figure 1.2: Origins of major ncRNA families. The origin of ncRNA families is marked leading to the last common ancestor of the known representative. The microRNA families of (a) eumetazoa (animals except sponges), (b) the slime mold *Dictyostelium*, (c) embryophyta (land plants), and (d) the green algae *Chlamydomonas* are non-homologous. In addition, the putative origin of the RNAi mechanism and the microRNA pathways is indicated. NcRNAs coloured red are examined in detail in this thesis, whereas ncRNAs coloured purple indicate minor examination and brown no examination in this thesis.

There are strong reasons to conclude that LUCA was preceded by simpler life forms that were based primarily on RNA. In this RNA-World scenario [27, 28], the translation of RNA into proteins, and the usage of DNA [29] as an information storage device are later innovations. The wide range of catalytic activities that can be realized by relatively small ribozymes [30, 31] as well as the usage of RNA catalysis at crucial points of the information metabolism of modern cells provides further support for the RNA-World hypothesis. Multiple ancient ncRNAs are involved in translation: the ribosome itself is an RNA machine [32], tRNAs perform a major part of the decoding on the messenger RNAs, and RNase P, another ribozyme, is involved in processing of primary tRNA transcripts. The signal recognition particle, another ribonucleoprotein (RNP), also interacts with the ribosome and organizes the transport of secretory proteins to their target locations.

On the other hand, most functional ncRNAs do not date back to the LUCA but are the result of later innovations. Some crucial “housekeeping” functions involve domain-specific ncRNAs. Eukaryotes, for instance have invented the splicing machinery involving several small spliceosomal RNAs (snRNAs), while bacteria use tmRNA to free stalled ribosomes and the 6S RNA as a common transcriptional regulator.

The innovation of ncRNAs is an ongoing process. In fact, most experimental surveys for ncRNAs report lineage-specific elements without detectable homologies in other species. An overview of evolutionarily older ncRNA families is compiled in Fig. 1.2 without claim to completeness. Many RNA classes, however, such as Y RNAs and vault RNAs, and most bacterial ncRNA families have not been studied in sufficient detail to date their origin with certainty. Some of them thus might have originated earlier than shown.

## 1.2 Homology Based Search of ncRNAs

In biology, homology refers to any similarity between characteristics that is due to their shared ancestry. Many scientists have argued for a long time about the definition of homology. “Any similarity between characteristics” does not only refer to DNA. It is a consensus consistency of organs, apparatus, corpus structures, physiological processes or even behaviours.

An established example for homology on a morphological layer are front extremities of humans and pectoral fins of dolphins. The similarity of homologous organs may be lost during evolution through further developments. Therefore three criteria for verifying homology exists: (1) With the *criterion of position* organs of different species can be recognized as homologs, due to an identical positioning. The long, acuminate canine of gorillas and the much smaller canine of humans can be localized between incisors and premolars. (2) Characteristics can be homologous by the *criterion of consistency*. The swim bladder of fish and the lung of humans are homologs, evolutionary intermediated stages are lungs of lungfish, amphibians and reptiles. (3) The third *criterion of specific quality* shows in case of many matches in constitution homology has to be assumed. Placoid scales of sharks and teeth of humans are similarly built-on: a cavity, dentine and enamel.

At the level of DNA “homology” is also used for genes, which in different species have similar or identical functions. Their sequence arose from a common ancestor. The *criterion of position* can be determined by adjacent genes. If two genes

of different organisms have highly similar DNA sequences, it is likely that they are homologous by the *criterion of consistency*. However, one has to be careful: Sequence similarity may also arise without common ancestry. Short sequences may be similar by chance and other sequences might be similar because both were selected to bind to a particular molecule. Such sequences are similar but not homologous. The *criterion of specific quality* plays a decisive role in terms of non-coding RNAs, since their constitution is not only based on nucleotides rather than their structure. An extreme form of this criterion will be presented and discussed in section 3.3.

We here distinguish between orthologs and paralogs, which are two fundamentally different types of homologous genes that evolved, respectively, by vertical descent from a single ancestral gene and by duplication [33]. Bioinformaticians use homology search for the prediction of so far unknown occurrence of genes in a certain organism based on vertical descent. For this purpose a known piece of DNA, RNA or proteins – in this thesis usually a non-coding RNA gene – of an organism is used as query. One of the nowadays more than 5300 sequenced genomes<sup>1</sup>, ESTs or other genomic sequences are used as database to search a homologous sequence of the query sequence. Genomes used in this thesis can be viewed in App. B.4. Known ncRNAs are available and obtained for this thesis from **Rfam**, **NCBI**, **NonCode** and various other sources as described in each section and in App. B.1.

At first glance this search seems to be straightforward. Although this is fundamental several difficulties cannot be dismissed. Organisms and therewith their genomics develop during evolution. Hence, it might be difficult to find homologous genes, if query and source organism are divergent. The *missing link method*, i.e. searching with the query in organisms phylogenetically located "between" the query and source organism, is a fashionable and well-founded evolutionary solution.

Homology search of non-coding RNAs (ncRNAs) is a very fundamental and complex challenge. This feature is justified in the function of ncRNAs. In the case of protein-coding genes (Fig. 1.3a) the transcribed RNA is processed and a certain combination of three nucleotides are translated in a specific amino acid. The resulting chain of amino acids acts within the cell, the function proteins are coded mainly in the sequence of nucleotides. NcRNAs are likewise transcribed and processed, Fig. 1.3b. However, they are not translated into proteins. Instead they have a specific structure and interact this way with other molecules (proteins, DNA or RNA). Consequently, homology search of ncRNAs is not only based on the sequence of nucleotides rather than a specific structure.

---

<sup>1</sup>26.04.2009, <http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome>

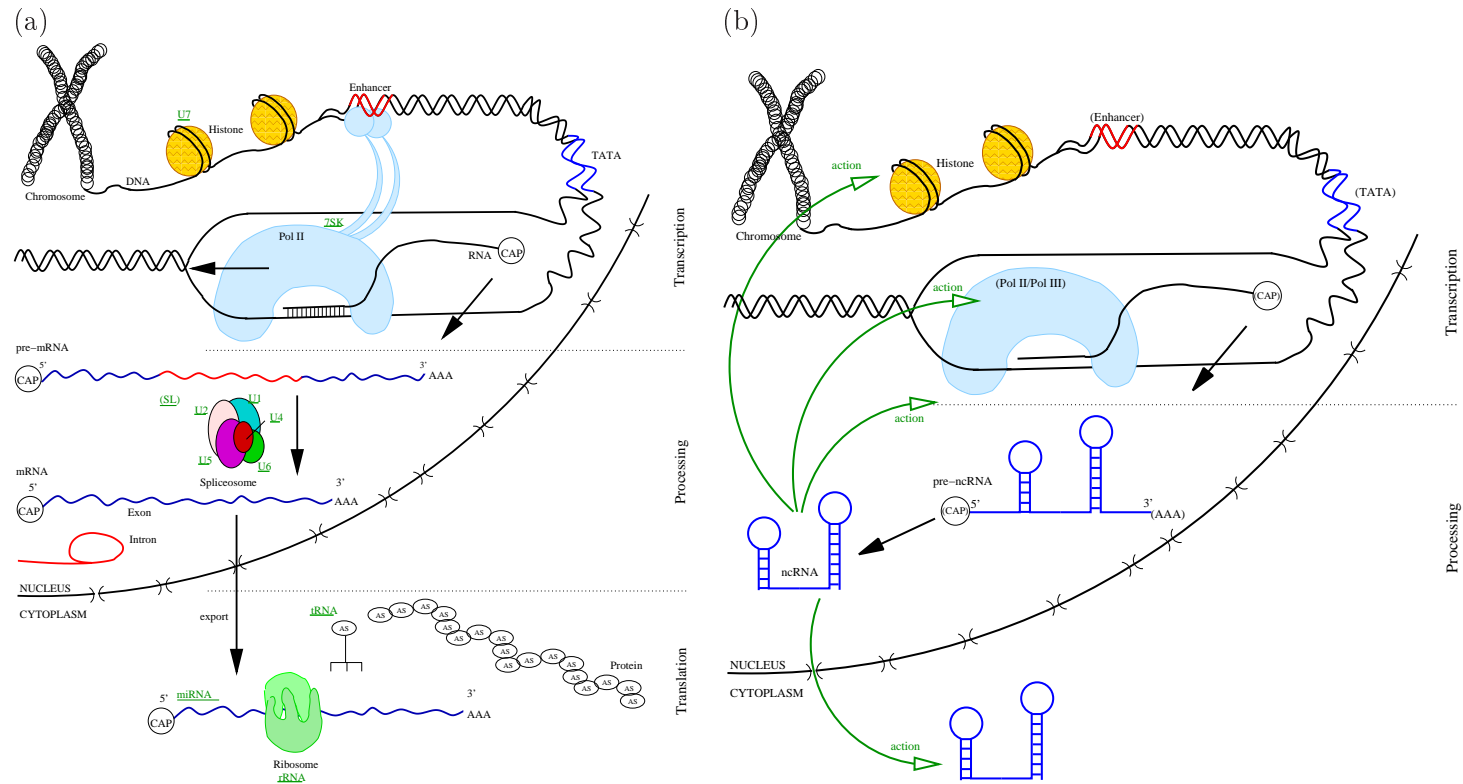


Figure 1.3: (a) From genes to proteins in eukaryotes. Transcription, RNA processing and translation. During Transcription polymerases transcribe DNA to mRNA. The pre-mRNA contains introns and exons and is protected against exonucleases by 5' cap and 3' polyadenylation. The processing step prepares RNA for leaving the cell nucleus. The spliceosome removes introns. In the cytoplasm the mRNA is translated to proteins by ribosomes and tRNA adapter molecules. For regulation of the synthesis of proteins non-coding RNAs are needed and in this picture underlined. (b) From genes to functional RNAs in eukaryotes. The step of transcription produces pre-mRNA with exons, introns, 5' cap and 3' polyadenylation as for the transcription for proteins. During processing pre-mRNA matures. Afterwards ncRNA acts in different parts of the cell as tRNA, rRNA, snRNA, snoRNA, miRNA, RNase P etc., indicated by arrows.



At present, we have an incomplete understanding of genes coding for proteins within the genome and even less of an understanding of genes not coding for proteins (non-coding genes). Experimental studies using a variety of different techniques, from tiling arrays [34–37] to cDNA sequencing [38–40], and unbiased mapping of transcription factor binding sites [41] agree that a substantial fraction of the genome is transcribed and that non-protein-coding RNAs (ncRNAs) are the dominating component of the transcriptome.

An as-yet unsatisfactorily resolved question is whether novel transcripts lacking protein-coding capacity (non-coding transcripts) have a biological function as such, or whether they rather represent “biological noise” (i.e., selectively neutral transcription) [42]. Analogous to the analysis of protein-coding genes, a combination of both experimental and computational techniques seems necessary to address this question.

### 1.3 Overview of this thesis

The next chapter introduces all programs used in this thesis for the prediction of a various number of non-coding RNAs. It provides a general overview and classification of these programs.

In chapter 3 an evolutionary overview of chemical reactions of splicing is given. Splicing is a main processing step for proteins and non-coding RNAs and an indispensable part of this thesis. We will see that the ability to remove parts of the transcripts or producing different products is finally a regulation step existing in different variations in eukaryots, bacteria and archaea.

Non-coding RNAs involved in processing are examined. For *cis*-splicing (Section 3.2) with the major spliceosome, ancient snRNAs U1, U2, U4, U5 and U6 are needed, whereas U11, U12, U4atac, U5 and U6atac act for the minor spliceosome. *trans*-splicing (Section 3.3) involves a splice leader (SL) as mini-exon. This splice reaction exists in just a few organisms spread wide over the phylogenetic tree of eukaryots. We flavour in this thesis for a possible origin of all these diverged SL-RNAs. Recently SL RNAs were observed together with another small class of non-coding RNAs: We show the phylogenetic distribution of SmY RNAs in nematodes (Section 3.4) and ask the question of a proposed interaction of SL2 RNA and SmY RNA. In Section 3.5 the histone-mRNA processing unit of U7RNP, namely the non-coding U7 snRNA, is examined extensively in an evolutionary context.

MRNA-like-ncRNAs (mlncRNAs) are non-coding RNA transcripts, which are processed just as normal mRNAs, but carry only very small ORFs or no ORFs at all.

Transcriptional control, [14, 43–45], tissue specific differential expression [46], alternative splicing and polyadenylation [47] of mlncRNAs does not seem to differ from those of protein coding polymerase II products. Some of them remain in the nucleus. In Section 3.6 we examine mlncRNA in a large scale for insects.

For the introduced ncRNAs involved in processing the discovery and therefore a fundamental overview of their evolution is possible to obtain. For some long, highly derived ncRNA (Chapter 4) a homology search within the 243 assembled eukaryotic genomes is not straightforward. In case of atypical snoRNA U3, Section 4.1, which is essential for processing of 18S rRNA transcripts into mature 18S rRNAs [48], at least within subgroups boxes C', B, C and D are present in a conserved way. Each box has 6 to 10 nt and by chance a box occurs 732.421 or 2.861 times, respectively in the human genome ( $3 \cdot 10^9$  nt). To make the problem more complex, for some species it is known that U3 contains introns [49, 50]. With the information of similar sequences between these boxes, information of a conserved secondary structure, a known distance range between conserved sequence motifs and finally specific written programs for this purpose, however, in most cases it is possible to identify the only U3 of an organism's genome, Section 4.1.

In the case of RNase MRP or RNase P (Section 4.2) the secondary structure can vary dramatically [51]. Only one stem (P10/12) may consist of 15nt (*Yarrowia lipolytica*) or 280nt (*P. anserina*). Although there exist no precisely calculating pseudoknot programs with information of the main functional and interacting part of the gene, it is possible to determine and find the only RNase MRP or RNase P sequence in most eukaryotic organisms. The most divergent ncRNAs consist of very less (7SK RNA, Section 4.3) or no (Telomerase, Section 4.4) sequence similarity at all. The homologous sequences share structural features and their functionality only. In these cases specific programs are developed to identify in very little cases their homologous genes. It is my great pleasure to present you the newly discovered 7SK RNA and Telomerase sequences in this thesis.

In the last chapter 5 we want to show different comparative approaches to predict functional RNA secondary structures and provide a detailed screen and comparison of genomes. The computational approach is based on the observation that structural constraints imply specific mutational patterns visible at the sequence level. Beside RNAz [52, 53], which considers structural conservation and stability of the putative structures in terms of predicted folding energies, we use a combination of introduced methods to examine all nowadays known ncRNAs [54].

Altogether I am proud to present even in the case of highly divergent non-coding RNAs, such as 7SK RNA, U3 snoRNA, RNase MRP, Telomerase, the evidence of proven functionality by wet-lab experiments of our collaborators.

## 1.4 List of Publication and Supplements

This thesis is based on the following research papers:

(Shared first authors are marked with an asterisk)

**Marz M**, Schön A, Rosenblad MA, Stadler PF, *RNase P and RNase MRP in Fungi*, in preparation

**Marz M**, Rosenblad MA, Schön A, Stadler PF, *RNase P and RNase MRP in Animals and Plants*, in preparation

**Marz M**, Donath A, Stadler PF, Bensaude O, *Evolution of 7SK RNA and its Protein Partners in Metazoa*, submitted

Ingalls T, Martius G, Hellmuth M, **Marz M**, Prohaska SJ *Converting DNA to Music: ComposAlign*, accepted for GCB 2009.

**Marz M**, Stadler PF, *Phylogentic range of U3 snoRNA in eukaryots*, submitted

Yusuf D, **Marz M**, Stadler PF, Hofacker IL, *Bcheck: a wrapper tool for RNaseP RNA gene prediction*, in preparation.

Copeland CS\*, **Marz M\***, Rose D\*, Hertel J\*, Brindley PJ, Santana CB, Kehr S, Attolini CSO, Stadler PF, *Non-coding RNA Annotation of the Schistosoma mansoni Genome*, submitted

**Marz M**, Vanzo N, Stadler PF, *Temperature-Dependent Structural Variability of RNAs: Spliced Leader RNAs and their Evolutionary History*, resubmitted

Hiller M, Findeiss S, Lein S, **Marz M**, Nickel C, Rose D, Schulz C, Backofen R, Prohaska SJ, Reuter G, Stadler PF, *Conserved introns reveal novel transcripts in Drosophila melanogaster*, Genome Res (2009), **19**, 1290–1300; DOI 10.1101/gr.090050.108

Hertel J, de Jong D, **Marz M**, Rose D, Tafer H, Tanzer A, Schierwater B, Stadler PF, *Non-Coding RNA Annotation of the Genome of Trichoplax adhaerens*, Nucleic Acids Res (2009), **37**, 1602–1615, DOI 10.1093/nar/gkn1084

Jones TA\*, Otto W\*, **Marz M\***, Eddy SR, and Stadler PF, *A Survey of Nematode SmY RNAs*, RNA Biology (2009), **6**, 5–8

**Marz M**, Kirsten T, Stadler PF, *Evolution of Spliceosomal snRNA Genes in Metazoan Animals*, J.Mol.Evol. (2008), **67**, 594–607, DOI 10.1007/s00239-008-9149-6

Donath A, Findeiß S, Hertel J, **Marz M**, Otto W, Schulz C, Stadler PF, Wirth

S, *Non-Coding RNAs*, Evolutionary Genomics, Caetano-Anolles, Gustavo, Wiley, 2008, in press

**Marz M**, Mosig A, Stadler BM, Stadler PF., *U7 snRNAs: a computational survey.*, J Mol Evol. (2008), **66**:107–115, DOI 0.1007/s00239-007-9052-6

Gruber AR\*, Koper-Emde D\*, **Marz M\***, Tafer H\*, Bernhart S, Obernosterer G, Mosig A, Hofacker IL, Stadler PF, Benecke BJ., *Invertebrate 7SK snRNAs.*, J Mol Evol. (2008), **66**, 107–115, DOI 10.1007/s00239-007-9052-6

ENCODE Project Consortium, *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.*, Nature (2007), **447**, 799–816, doi:10.1038/nature05874

Washietl S, Pedersen JS, Korbelt JO, Stocsits C, Gruber AR, Hackermüller J, Hertel J, **Lindemeyer M**, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigo R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF., *Structured RNAs in the ENCODE selected regions of the human genome.*, Genome Res. (2007), **17**, 852–864, DOI 10.1101/gr.5650707

For the following chapters supplemental material is available at:

# Chapter	Short Titel	Page	<a href="http://www.bioinf.uni-leipzig.de/">www.bioinf.uni-leipzig.de/</a> ***
3.2	snRNAs	39	Publications/SUPPLEMENTS/08-001
3.3	SL RNAs	56	Publications/SUPPLEMENTS/09-009
3.4	SmY	69	Publications/SUPPLEMENTS/09-001 Publications/SUPPLEMENTS/09-013
3.5	U7	75	Publications/SUPPLEMENTS/07-010
3.6	Fly Introns	84	Publications/SUPPLEMENTS/08-021
4.1	U3	94	Publications/SUPPLEMENTS/09-021
4.2	MRP/P	101	Publications/SUPPLEMENTS/09-023
4.3	7SK	107	Publications/SUPPLEMENTS/07-021 Publications/SUPPLEMENTS/08-008 Publications/SUPPLEMENTS/09-010
4.4	Telomerase	120	Publications/SUPPLEMENTS/09-022
5.1	<i>Trichoplax</i>	130	Publications/SUPPLEMENTS/08-024
5.2	<i>Schistosoma</i>	141	Publications/SUPPLEMENTS/08-014
A	<i>ComposAlign</i>	163	Publications/SUPPLEMENTS/09-017



## Chapter 2

# Tools Of The Trade

This chapter introduces all programs used in this thesis for the prediction of a various number of non-coding RNAs. A classification of programs and a general workflow can be obtained in Fig. 2.1. Beside a description of **RNAz** and a subsequently annotation pipeline for screening full genomes, in the last part of this chapter (Sec. 2.4) we provide information for other used programs, e.g. for Target Prediction Tools, Repeat Filter Tools, Tools for phylogenetic analysis or synteny information.

### 2.1 General Homology Search

Programs comparing primary sequences (**Blast**, **GotohScan**) are used for homology search of highly conserved non-coding RNA (tRNAs, rRNAs, snRNAs, SRP RNA) or between closely related organisms, Sec. 2.1.1. This thesis comprises mainly low conserved or divergent ncRNAs homologs. Depending on the degree of derived sequences, homology search might be performed by secondary structure based programs (Sec. 2.1.2) or programs using a pattern search methods and covariance models (Sec. 2.1.3).

#### 2.1.1 Sequence Based Search

##### **Blast**

The **Blast**-package [55] provides various programs for calculating high scoring local alignments between a query sequence and a target database. Query and target could consist of DNA or protein sequences. In this thesis the following

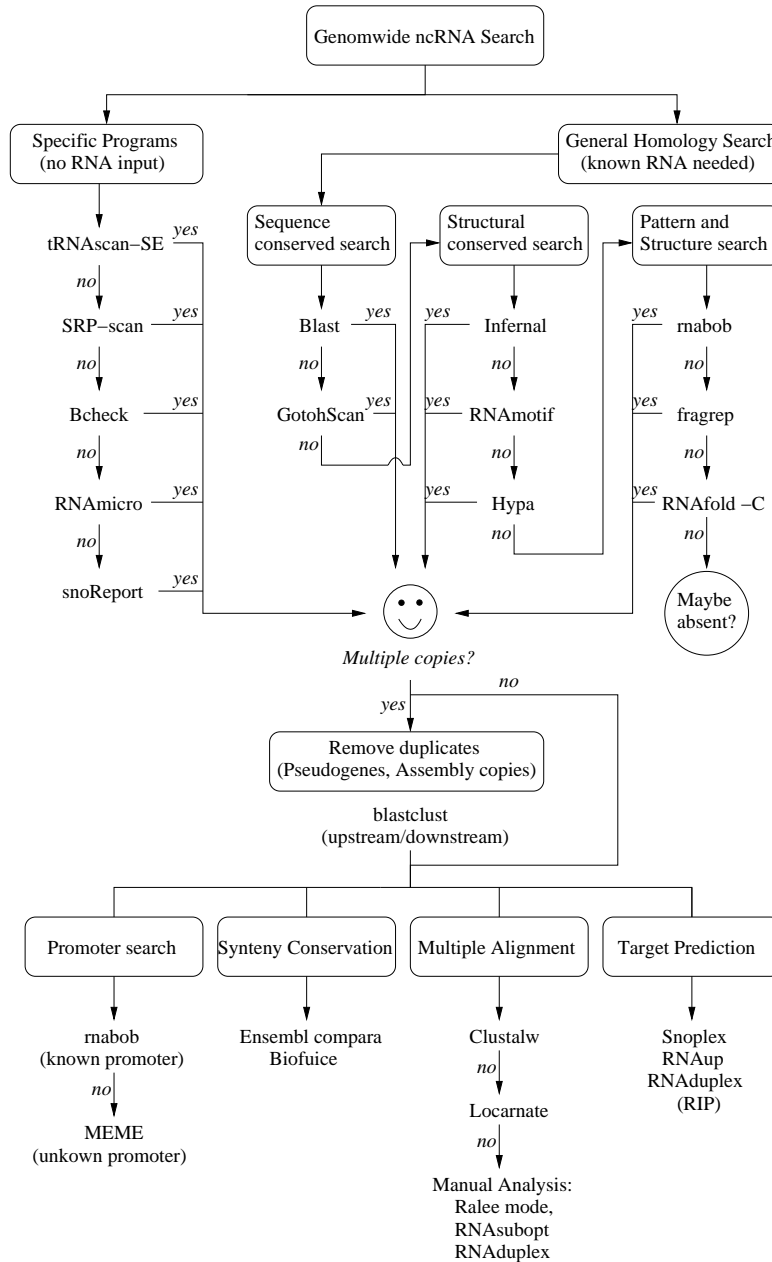


Figure 2.1: Searching for ncRNAs. Methods of homology search for ncRNAs can be divided in the following classes: (1) General Homology Search, which might be sequence based, structure based or a mixture of existing programs; (2) Specific ncRNA Search Programs were developed for large common groups of ncRNAs; (3) Verification of Predicted ncRNAs is an indispensable step for computationally verifying obtained sequences; (4) For further specific tasks programs for target prediction, phylogenetic analysis, synteny information, promoter verification and much more are available.



programs are used: **blastn** and **blastp** for comparison of nucleotide and protein sequences, respectively; **blastx** for comparison of six-frame translation products of nucleotide query and database; **tblastn** for comparison a protein query with a nucleotide database and **PSI-blast** for identifying distant relatives of proteins.

Before using **Blast**, an index file is created by **formatdb**, which avoids redundant copies of databases, therefore the efficiency is enhanced by faster searches in smaller databases. Afterwards, a search of small regions is performed, that are exactly identical in both sequences (seeds). The word size of **blastn** is by default 11 nucleotides and limited to 7 nucleotides. Both sides of the seeds are extended in order to obtain a good longer alignment.

For the local Smith-Waterman algorithm a matrix  $F$  indexed by  $i$  and  $j$  is constructed. Each index stands for a position in each sequences  $p$  and  $q$ .  $F_{i,j}$  is the score of the best alignment between the initial segment  $p_{1..i}$  and  $q_{1..j}$ .

$$F_{i,j} = \max \begin{cases} 0, \\ F_{i-1,j-1} + \sigma(p_i, q_j), \\ F_{i-1,j} - d, \\ F_{i,j-1} - d \end{cases} \quad (2.1)$$

$F_{i,j}$  is calculated recursively,  $F_{0,0} = 0$ ,  $d$  stands for gap costs and  $\sigma(p_i, q_j)$  for match/mismatch costs. Subsequently a traceback from the maximal entry of  $F$  is performed in order to obtain the best sequence.

There are two main ways to estimate the significance of the alignment scores: The Bayesian approach via model comparison and the classical approach by extreme value distribution. For details see Chapter 2 of [56].

The program **formatdb** indexes databases. Results are cut by **fastacmd**.

In the following a typical application of **Blast** in combination with other programs is described for snRNA search:

In a first automatic step we used a local installation of NCBI **blast** (v.2.2.10) with default parameters and  $E < 10^{-6}$  to find candidate sequences in closely related genomes. If successful, the results of this search were aligned to the query sequence using **ClustalW** (v.1.83), Sec. 2.3.1. After a manual inspection using **ClustalX**, the consensus sequence of the alignment was again used as a blast query with the same  $E$ -value cutoff.

If this automatic search was not successful, the best `blast` hit(s) were retrieved and aligned to a set of known snRNAs from related species. Candidate sequences were retained only when a visual inspection left no doubt that they were true homologs. This manual analysis step included a check whether the phylogenetic position of the candidate sequence in a neighborjoining tree was plausible, taking into account that the sequences are short and some parts of the alignments are of low quality.

In cases where no snRNA homologs were found as described above, we searched the genome again with a much less stringent cutoff of  $E < 0.1$  (or even larger in a few cases) and extracted all short hits together with 200nt flanking sequence. We used Sean Eddy’s `rnabob`, Sec. 2.1.3 with a manually constructed structure model to extract a structure-based match within the selected regions and attempted to align the candidate sequences manually to a structure-annotated alignment of snRNAs in the `Emacs` editor using the `ralee mode` mode [57], Sec. 2.3.1.

Finally, the resulting alignments of snRNAs were used to derive search patterns for `RNAmotif` [58] and `Erpin` [59], Sec. 2.1.2. To this end, the consensus structure of the alignment was computed using `RNAalifold` [60], Sec. 2.3.1 and converted into a form suitable as input for the two search programs.

Although some non-coding RNA homologs can not be identified by `Blast`, this way is the usually the first, fastest and mostly used approach for RNA detection.

### GotohScan

Since `Blast` fails to identify many of the ncRNAs that are reasonably expected to be present in certain genomes, e.g. homologs of U4atac, U3 snoRNA, and RNase MRP RNA of *Trichoplax adhaerens*, a full dynamic programming approach is used [61]. Instead of using a local (Smith-Waterman) implementation such as `ssearch` [62] or its partition function version [63], `GotohScan` suggests that a “semi-global” alignment approach is more natural for the homology search problems at hand, here the best match of the *complete* query sequence to the genomic DNA is sought. Due to relatively long insertion and deletions, the use of an affine gap cost model becomes necessary. This problem is solved by the following straight-forward modification of Gotoh’s dynamics programming algorithm [64].

Denote the query sequence by  $Q = q_1, q_2, \dots, q_m$  and the genomic “subject” sequence by  $P = p_1, p_2, \dots, p_n$ . Note that the problem is not symmetric since deletions of the ends of  $P$  do not incur costs, while deletions of the ends of  $Q$  are fully penalized. As usual, denote by  $S_{ij}$  the optimal alignment of the prefixes

$Q[1\dots i]$  and  $P[1\dots j]$ , respectively. The values of  $D_{ij}$  and  $F_{ij}$  are the optimal scores of alignments of  $Q[1\dots i]$  and  $P[1\dots j]$  with the constraint that the alignment is an insertion or a deletion, respectively. The recursions read

$$\begin{aligned} D_{ij} &= \max \{S_{i-1,j} + \gamma_o, D_{i-1,j} + \gamma_e, \} \\ F_{ij} &= \max \{S_{i,j-1} + \gamma_o, F_{i,j-1} + \gamma_e, \} \\ S_{ij} &= \max \{D_{ij}, F_{ij}, S_{i-1,j-1} + \sigma(p_i, q_j)\} \end{aligned} \quad (2.2)$$

$\gamma_o$  for open-gap penalty,  $\gamma_e$  gap-extend penalty,  $\sigma(a, b)$  for match/mismatch costs, with the initializations

$$\begin{aligned} S_{00} &= 0, \\ D_{0j} &= -\infty, \quad S_{0j} = F_{0,j} = \gamma_o + (j-1)\gamma_e, \\ F_{i0} &= -\infty, \quad S_{i0} = D_{i,0} = \gamma_o + (i-1)\gamma_e. \end{aligned}$$

In this full version, the algorithm requires  $\mathcal{O}(n \times m)$  time and memory, where  $n$  is the length of the genome and  $m$  is the length of the query sequence. For each endpoint, the alignment can be obtained by standard backtracing in  $\mathcal{O}(m^2)$  time and space.

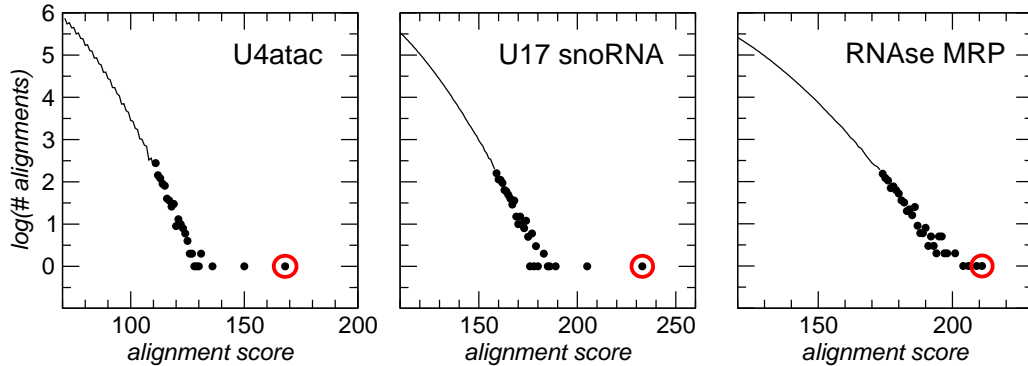


Figure 2.2: Histogram of score distribution for U4atac, U17 and RNase MRP. Circles denote true homologous.

The current C implementation of GotohScan stores a histogram of all the scores for each query sequence over all database sequences. Fig. 2.2 gives some example of score histograms.

### 2.1.2 Structure Based Search

Since sequence based search methods fail to identify many of the ncRNAs that are divergent but structurally conserved, depending on their function, I used programs using the information of the molecule folding.

Advantages and disadvantages of the following programs will be described in this section. **RNAmotif**, **Erpin** and **HyPa** were used for snRNA search, **Infernal** for SmY and SL RNA search. For longer diverse ncRNAs these methods turned out to be unsuitable. Known sequences/structures with small variations (point-mutations, varying loop-lengths) are easily detected. However, large variations (stem length of more than double sizes, included introns, etc.) were missed. This problem is a general problem for secondary structure search methods: It is totally unclear at *which* positions within the gene, larger variations occurred during evolution. Writing very variable motifs which simply allow larger evolutionary events everywhere yields a massive amount of candidates of which nearly all are false positives. This can be partly dynamically adopted in **RNAmotif** with the use of scoring functions, but again it is necessary to predict *where* variable regions are. And this is impossible.

### **Infernal**

**Infernal** [65] is a software package that allows to make consensus RNA secondary structure profiles, and use them to search nucleic acid sequence databases for homologous RNAs, or to create new structure-based multiple sequence alignments. To make a profile, one needs to have a multiple sequence alignment of an RNA sequence family, and the alignment must be annotated with a consensus RNA secondary structure. The program **cmbuild** takes an annotated multiple alignment as input, and outputs a profile. Then one uses that profile to search a sequence database for homologs, using the program **cmsearch**. One can also use the profile to align a set of unaligned sequences to the profile, producing a structural alignment, using the program **cmalign**. This allows to build hand-curated representative alignments of RNA sequence families, then use a profile to automatically align any number of sequences to that profile. This seed alignment/full alignment strategy combines the strength of stable, carefully human-curated alignments with the power of automated updating of complete alignments as sequence databases grow. This strategy is used to maintain the **Rfam** database of RNA multiple alignments and profiles. **Infernal** models are profile stochastic context-free grammars (profile SCFGs), which include sequence and RNA secondary structure consensus information. A large number of CPUs are needed to use it for serious work.

### **RNAmotif**

The **RNAmotif** program [58] searches a database for RNA sequences that match a "motif" describing secondary structure interactions. A match means that the

given sequence is capable of adopting the given secondary structure, but is not intended to be predictive. Matches can be ranked by applying scoring rules that may provide finer distinctions than just matching to a profile. `RNAmotif` program is an extension of earlier programs `rnamot` and `rnabob` [66–68], Sec. 2.1.3. The nearest-neighbour energies used in the scoring section are based on refs. [69, 70].

### Erpin

The program `Erpin` (Easy RNA Profile IdentificatioN) [59] is an RNA motif search program developed by Daniel Gautheret and André Lambert. Unlike most RNA pattern matching programs, `Erpin` does not require users to write complex descriptors before starting a search. Instead `Erpin` reads a sequence alignment and secondary structure, and automatically infers a statistical SSP (secondary structure profile). A novel Dynamic Programming algorithm then matches this SSP onto any target database, finding solutions and their associated scores. In the latest version `Erpin` computes E-values for matches.

### HyPa

The program `HyPa` [71] allows the user to search for hybrid patterns over an index constructed by the provided `mkaffix.sh` script. `HyPa` requires a query file containing the pattern descriptions in the provided language `HyPaL` and an index as input. The database, called `HyPaLib` (for Hybrid Pattern Library) [72], contains annotated structural elements characteristic for certain classes of structural and/or functional RNAs. These elements are described in `HyPaL` along with motifs consisting of sequence features and structural elements together with sequence similarity and thermodynamic constraints. Because of limitations in space and time this approach is practically not applicable.

### 2.1.3 Pattern and Structure Based Search (by Hand)

One of the main problem in bioinformatics is homology search of ncRNAs which are neither conserved in their sequence nor in their structure. For each of such divergent gene classes specific programs have to be invented. In this thesis some general approaches will be shown for RNase MRP, RNase P, U3, 7SK and Telomerase, Sec. 4. In this case handicrafts is needed: A combined search of `rnabob`, `Fragrep`, `RNAsubopt` and `RNAduplex` is used for this purpose.

`rnabob`

The program `rnabob` [68] has been utilised at various occasions during this thesis. This program is an extremely fast pattern searching program for RNA sequences, secondary and tertiary structures and even pseudoknots. However, this program lacks two essential features: (1) A non-marginal amount of results are not considered by `rnabob` and are therefore not part of the output; (2) `rnabob` is accident-sensitive in terms of structural variations such as point insertions/deletions. Both problems are discussed below with several methods to resolve these shortcomings.

**Lost `rnabob` Results** The program `rnabob` [68] is an implementation of Daniel Gautheret's `RNAmot` [66, 67] with a different underlying algorithm using a non-deterministic finite state machine with node rewriting rules.

Conserved sequence pattern and structural coherencies are specified in a descriptor file. A regular expression tree is built after starting `rnabob`, which is used to search in the sequence database (fasta, gcg, embl, genbank and other formats are possible).

During the search on 7SK RNA we discovered for our purpose an unintentional feature of `rnabob`. In Fig. 2.3 a simple example sketch this problem.

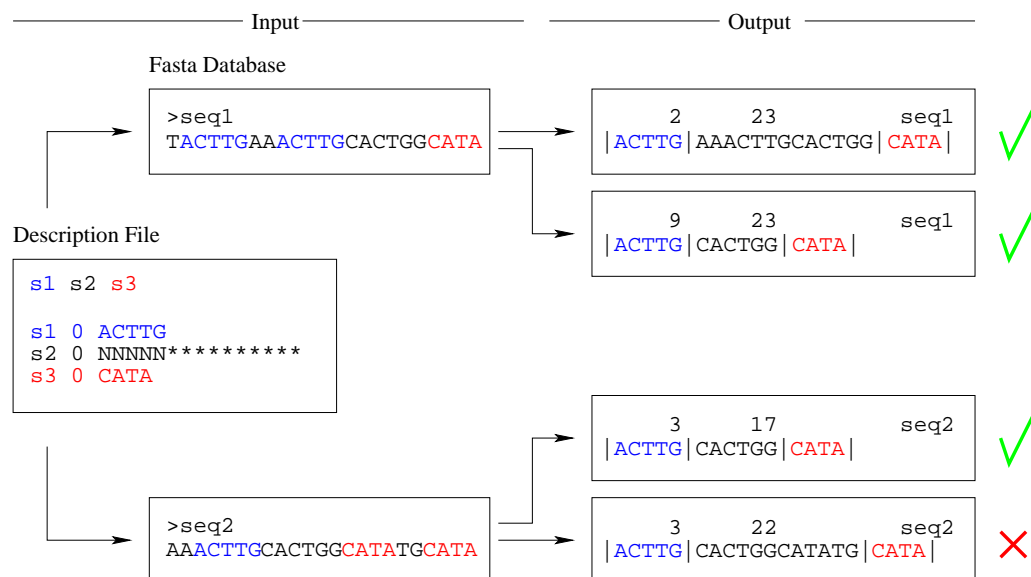


Figure 2.3: A simple example about the outcome of `rnabob`. The description file specifies a pattern of `ACTTG`, a 5-15nt spacer and `CATA`. Running the descriptor on `seq1`, `rnabob` dumps two results. Running the descriptor on `seq2` just one of two results is calculated.

A description file with a variable number of nucleotides between two specified patterns (sequence based as in Fig. 2.3 or structure based) may lack a subset of solutions. In order to obtain all possible matches a reimplementaion of `rnabob` seems to be indispensable.

hi Manja,

```
It's probably possible to hack rnabob to do that, but the way it's
coded, there's no option for doing that at the moment. It's focused
on finding *a* match rather than *all* matches, and for any given
start point i on the sequence, it finds the first match (if any) and
ignores other possible alignments of the pattern to subsequences
starting at i.
[...]
```

```
You might have a look at the code yourself (might be faster than
waiting on me!). Since rnabob is just based on a hacked regular
expression matcher (the same parent code as Perl's code is based on,
I believe), and regexp matchers usually allow you to output all
matches instead of the first one, this might be easy to hack.
```

Sean

After the correspondence with Sean Eddy, the author of `rnabob`, several programmers tried to rewrite the code, which turned out not to be an easy hack at all.

**Brute Force Solution** Obviously, replacing each asterisk `*` by `N`, yielding  $n$  descriptor files instead of one file with  $n$  asterisks, solves the problem for a simple pattern with *one* variable region, only.

However, in a more difficult example as for snoRNA U3 (Fig. 2.4A) the number of descriptor files may blow up. U3 RNA has six sequence conserved parts: Boxes  $A'$ ,  $A$ ,  $C'$ ,  $B$ ,  $C$  and  $D$ . The general structure is also conserved, however to keep it simple here we concentrate on boxes  $C'$ ,  $B$ ,  $C$  and  $D$ , which are highly conserved among fungi. Between  $C'$  and  $B$  18-27nt,  $B$  and  $C$  31-160nt,  $C$  and  $D$  53-120nt may exist. This would result in  $6 + 129 + 67$  asterisks and  $6 \cdot 129 \cdot 67 = 51.858$  files would be needed with this brute force solution. The runtime would be much too large and therefore a smarter approach is needed.

**A Smarter Approach** To avoid within a variable length of nucleotides two occurrences of a pattern  $p$  with the length  $n$ , the length of an asterisk sequence should be at most  $2n - 1$ . If the number of asterisks  $m$  is smaller then the length of

(A)

```

s1 s2 s3 s4 s5 s6 s7 s8

s1 0 GATGA
s2 0 NNNNNNNNNNNNNNNNNNN [6]
s3 1 AGA
s4 0 GTGA
s5 0 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN [129]
s6 0 GATGATCT
s7 0 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN [67]
s8 0 TCTGA

```

(B)

```

for ((i=1;i<=17;i++)) do for ((j=1;j<=14;j++)); do
out="s1 s2 s3 s4 s5 s6 s7 s8\n\n
    s1 0 GATGA\n
    s2 0 NNNNNNNNNNNNNNNNNNN [6] \n
    s3 1 AGA\n
    s4 0 GTGA\n
    s5 0 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN"
for ((a=1;a<$i;a++)); do out=$out"NNNNNNN"; done;
out=$out"[15]\n
    s6 0 GATGATCT\n
    s7 0 NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNN"
for ((a=1;a<$j;a++)); do out=$out"NNNNN"; done;
out=$out"[9]\n
    s8 0 TCTGA\n"
echo -e $out > rnabob.$i.$j; done; done

```

Figure 2.4: Rewriting `rnabob` descriptor files. (A) Original `rnabob` descriptor for fungi snoRNA U3. *GATGA* – C'-Box, variable 18-27nt, *AGAGTGA* – B-Box, whereas among the first three nucleotides one might be substituted, 31-160nt spacer, *GATGATCT*–C-Box, 53-120nt spacer, *TCTGA*–D-Box. (B) Shell-script producing a minimal number of description files ( $17 \cdot 14 = 238$ ) for yielding the full solution set.

the following reduced pattern<sup>1</sup>  $p_r$  ( $m < n_r$ ), then the original `rnabob`-description file can be used without modifications.

If  $m > n_r$ , then  $n_r$  asterisks are replaced by *N*s from one description file to the next, so that  $\lceil m/n_r \rceil$  description files are needed to yield all desired results. In the example of snoRNA U3 (Fig. 2.4) `s2` is not changed, because the following pattern *AGAGTGA* with the length of 7 is larger than the number of asterisks (6). The line for `s5` contains 129 asterisks, the following pattern has a length of 8, therefore 17 files are needed to describe this line: the first one contains just 9 asterisks, the following file contains 5 *N*'s and 9 asterisks and so on. The 67 asterisks of `s7` are

<sup>1</sup>Within a reduced pattern prefix and suffix are not equivalent. If the pattern  $p$  is *CAGTCCCAG*, the reduced pattern  $p_r$  is *CAGTCC*



divided into 14 files. Because of the combination of `s5` and `s7`  $17 \cdot 14 = 238$  files are needed to yield all possible results.

An equivalent procedure has to be used for secondary structure description parts following a sequence of unknown length.

**Accident-sensitivity of `rnabob`** RNase P RNA is highly divergent among all organisms, however some parts P8 or P9 are highly conserved among deuterostomes in terms of the stem-length and loop-sequence. The highly variable stem P10 follows directly downstream of P9. The first 6-7 nucleotides basepair with a region upstream of P7. This feature was implemented in `rnabob`, a part of this was:

```
h10 s17 h10'

h10 0:0 NNNNNN*:*NNNNNN
s17 0   NNNNN[150]
```

Throughout the work of this thesis it was generally straightforward to determine the start and end of P10 in deuterostomes, with the exception of *Petromyzon marinus* in which case it was not possible at all. The reason is that within the lamprey genome a single nucleotide within the stem was inserted:

```
callorhinchus  CGG.AAGC.[].GCTCCG
petromyzon     GTGCAGCC.[].GGCTCAC
#=GC SS_cons   <<<<.<<<<.[].>>>>>>>
```

It would be possible to include this into the `rnabob` description, but you have to specify at *which* position the evolutionary deletion/insertion happened. In this thesis long patterns/stems were divided into smaller patterns/stems. The different results of the descriptors were joint, sorted and filtered by distances. Results with  $n - 1$  subpatterns were still examined. This method sounds a bit circuitous, however results were complete and `rnabob` is still (compared to other structure searching programs) extremely fast.

### Fragrep

Another program utilized for divergent ncRNA search in genomes uses the property of short but well conserved patterns separated by poorly conserved regions [73]. The **Fragrep** tool implements an efficient algorithm for detecting the pattern fragments that occur in a given order. For each pattern fragment the mismatch

tolerance and bounds on the length of the intervening sequences can be specified separately. Compared to `RNAmotif`, `Fragrep` provides a statistically well-motivated ranking scheme, which relieves the user from defining an individual scoring scheme as in `RNAmotif`. On the other hand, `Fragrep` does not search for explicit secondary structure constraints [73].

As described in Fig. 2.5 `Fragrep` calculates a searching pattern similar to the input of `rnabob`. The  $k$  conserved sequence patterns  $C$  may be found in a genome sequence  $T$  with  $m_i$  mismatches and  $k-1$  unconserved sequences  $X$  with lower and upper bounds of length. `Fragrep` provides  $p$ - and  $E$ -value-like ranking schemes that are computed from a dinucleotide-based Markov model.

The time complexity of `Fragrep` is bound to  $O(kL)$ , whereas  $L := \max_i L_i$  and  $L_i$  denoting the number of occurrences of  $C_i$  in  $T$ .

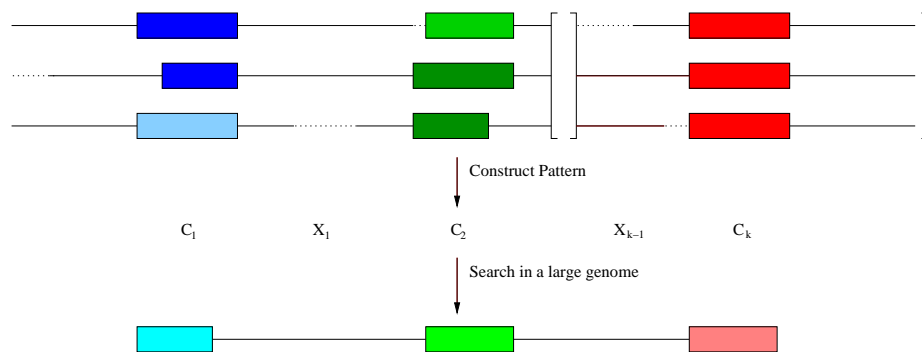


Figure 2.5: An efficient search tool for fragmented patterns in genomic sequences. `Fragrep` calculates for a given alignment a consensus pattern, whereas  $k$  conserved sequence fragments  $C_1, \dots, C_k$  are divided by  $X_1, \dots, X_{k-1}$  variable sequences. This pattern is searched in a large genomic context.

In this thesis `Fragrep` performed extensive searches of 7SK, U3 and Telomerase. Additionally, for U7 snRNA analysis we expanded the tool `aln2pattern`, the component of the `Fragrep` distribution that generates a collection of position weight matrices as search patterns with as “Sequence-Logo” style output derived from the `WebLogo` Postscript code [74].

Compared to `rnabob`, which is adequate for short conserved motifs in genomic contexts, `Fragrep` additionally is applicable for longer conserved parts (as described in the problem section of `rnabob`).

Searching for hairpins with `rnabob` and `Fragrep` I always suggest to pipe results to `RNAfold`, since the sequence `GTGTGTGTGTGTGTGTGT` would be matched by these programs with the query of a stem of length 6nt: `GTGTGT:GTGTGT`. However, `RNAfold` would reject this hairpin, in terms of a positive energy value.

## 2.2 Specific ncRNA Search-Programs

For some ncRNAs, such as tRNA, SRP, microRNAs, snoRNAs and RNase P, specific already existing programs were used for this thesis and discussed in the following section. Except the latter program, these programs are not the main subject of this thesis and used just for the last part (Chapter 5), since these programs and extensively homology of corresponding ncRNAs has been done previously.

### **tRNAscan-SE**

**tRNAscan-SE** identifies 99-100% of transfer RNA (tRNA) genes in DNA sequences while giving less than one false positive per 15 gigabases [75]. This program is extremely fast with  $\sim 30\,000$  bp/s.

We used **tRNAscan-SE** with default parameters to annotate putative tRNA genes in genome projects, e.g. for *Trichoplax adhaerens* (Section 5.1) and *Schistosoma mansoni* (Section 5.2). In the latter case the genome of the free-living platyhelminth *Schmidtea mediterranea* [76] was searched in order to obtain suitable data for comparison.

### **SRPscan**

A method for prediction of genes that encode the RNA component of the signal recognition particle (SRP) is developed by [77]. A heuristic search for the strongly conserved helix 8 motif of SRP RNA is combined with covariance models that are based on previously known SRP RNA sequences.

For annotation of ncRNAs in *Trichoplax*, platyhelminthes, cnidaria and nematods this program was used with default parameters, Chapter 5.

### **Bcheck**

We developed a RNase P specific gene finding tool, called **Bcheck** (in preparation) [78], which wraps **rnabob** pattern search and **Infernal** covariance validation. It has been developed and tested on both **Rfam** database and **GenBank** chromosome sequences of bacteria, archaea and recently eukaryots<sup>2</sup>. This program is developed with a decent speed and accuracy. **rnabob** descriptor models are built for short conserved regions I and S-domain.

---

<sup>2</sup><http://www.tbi.univie.ac.at/~dilmurat>

Hits are extended to both flanking sides. An `Infernal` covariance model of the whole RNase P gene filters by default with an E-value  $< 10^{-5}$ .

### RNAmicro

`RNAmicro` is a support vector machine (SVM) based approach that in conjunction with a non-stringent filter for consensus secondary structures, is capable of efficiently recognizing microRNA precursors in multiple sequence alignments [61]. For *Schistosoma* and *Trichoplax* we followed the general protocol as described in [61] to identify miRNA precursors, using all metazoan miRNAs listed in `miRBase` [79] [Release 11.0<sup>3</sup>]. The initial search was conducted by `Blast` with  $E < 0.01$  with the mature mature\* miRNAs as query sequences. The resulting candidates were then extended to the length of the precursor sequence of the search query and aligned to the precursors using `ClustalW` [80]. Secondary structures were predicted using `RNAfold` [81] for single sequences and `RNAalifold` [60] for alignments. Candidates that did not fold into miRNA-like hairpin structures were discarded. The remaining sequences were then examined by eye to see if the mature miRNA was well-positioned within the stem portion of each putative precursor sequence. For comparison, we used the final candidates to search the *S. japonicum* and *S. mediterranea* genomes to examine whether these sequences are conserved in Schistosomes and/or Platyhelminthes.

### snoReport

`snoReport` [82] is a combination of RNA secondary structure prediction and machine learning that is designed to recognize the two major classes of snoRNAs, box C/D and box H/ACA snoRNAs, among ncRNA candidate sequences. The `snoReport` approach deliberately avoids any usage of target information and instead uses the a pre-filter with SVM classifiers based on a small set of structural descriptors which are sufficient for a reliable identification of snoRNAs.

We compared all the known human and yeast snoRNAs that are annotated in the `snoRNAbase` [83] to the *S. mansoni* and *T. adhaerens* genome using `NCBI-blast`[55] and `Gotohscan` [61]. The search for novel snoRNA candidates was performed only on sequences that were not annotated as protein-coding or another ncRNA in the current *S. mansoni* assembly. The `SnoReport` program [82] was used to identify putative box C/D and box H/ACA snoRNAs on both strands. Only the best predictions, i.e., those that showed highly conserved boxes and

---

<sup>3</sup><http://microrna.sanger.ac.uk/sequences/>

canonical structural motifs, were kept for further analysis. The remaining candidates were further analysed for possible target interactions with ribosomal RNAs using `snoScan` [84] for box C/D and `RNAsnoop` [85] for box H/ACA snoRNA candidates, for details see section 2.3.4. In addition, the *S. mansoni* sequences were checked for conservation in *S. japonicum* and *S. mediterranea* using NCBI `Blast`. To estimate the number of false predictions we compared the candidate snoRNAs with common ncRNA databases, in particular `Rfam` [86] and `noncode` [87]. All sequences matching a non-snoRNA ncRNA were discarded.

## 2.3 Verification of Predicted ncRNAs

Computationally verification of predicted candidates is usually done by multiple alignments, promoter verification, conserved flanking regions (synteny) and through target prediction (valid mainly for snoRNAs). These methods will be described in following.

### 2.3.1 Multiple Alignments

Multiple alignments are usually obtained by generic programs as `ClustalW` (for nucleotides) and possibly by subsequent use for `RNAalifold` (for verification of secondary structure). A program combining structural and sequence feature is `Locarnate`, which is used mainly for long highly divergent ncRNAs, such as RNase MRP/P, 7SK or telomerase. However, sometimes (especially for the latter ncRNAs used in a wide genomic context) all programs do not align known short highly conserved boxes together. In such cases an alignment per hand is indispensable. In this thesis this is done by `Emacs ralee mode-mode` and intramolecular interactions are verified by also by hand with `RNA duplex` and `RNAsubopt`.

#### `ClustalW/ClustalX` and `RNAfold/RNAalifold`

`ClustalW` [88] accepts a wide range of input format, however in this thesis nucleotides and amino acids are used as input only. There are three main steps to produce a multiple alignment (1) Creation of pairwise alignments; (2) Construction of a phylogenetic tree, alternatively the user can specify such a tree; (3) Calculating a multiple alignment under the constraint of the given phylogenetic tree. This program is used mainly for predictions of all ncRNAs. Afterwards the alignment is viewed by `ClustalX` [88]. To prove in case of conserved ncRNAs a

common secondary structure, the output of `ClustalW` is calculated and the secondary structure is verified by `RNAalifold`. In cases of derived sequences new predictions may be folded by `RNAfold` with the `-C` option, allowing the user to specify a known structure and prove therefore the possibility of the molecule to fold into a known structure.

However for some divergent sequences this method was not suitable, therefore a combined sequence/structure alignment is needed.

### DIALIGN

While most multiple alignment methods are either purely global or purely local methods, `DIALIGN` is able to cope with a variety of different situations [89]. The program can find local similarities in a multiple sequence comparison even if these similarities involve only two sequences. These can be combined to one single multiple alignment and non-related regions between these regions are ignored. However, if sequences are globally related, `DIALIGN` will return a full global alignment.

Sequences for e.g. U7 RNA search or SRP RNA were aligned using `DIALIGN` to determine whether the characteristic up- and downstream elements were present.

### Locarnate

With `Locarnate`<sup>4</sup> [90] a novel approach for multiple alignments of RNAs is presented in the way that locality of RNA occurs as similarity of subsequences as well as similarity of only substructures. The approach extends `locARNA` by structural locality for computing all-against-all pairwise, structural local alignments. The final construction of the multiple alignments from the pairwise ones is delegated to `T-Coffee`. The paper systematically investigates structural locality in known RNA families. Benchmarking multiple alignment tools on structural local families shows the need for algorithmic support of this locality [90].

We used this method mainly for unknown structural parts as for RNase MRP and RNase P, stem P10-12 (Section 4.2); 5' part and stem loop IV of U3 snoRNA (Section 4.1) and 7SK RNA from stem M3 to M7 (Section 4.3).

---

<sup>4</sup><http://www.bioinf.uni-freiburg.de/Software/LocARNA/>

### Emacs ralee mode and the RNA Vienna Package

Within this thesis, structure annotated sequence alignments were manually modified in the Emacs text editor using the ralee mode mode [57] to improve local sequence-structure features based on secondary structure predictions for the individual sequences obtained from RNAfold, RNAsubopt, RNAduplex, RNAup or RNAcifold. [81]. The option `-C` allows to include known base pairing features from experiments or homologous genes. This way candidates for U3 snoRNA, RNase MRP, RNase P, 7SK and Telomerases were filtered extensively by these programs.

### ComposAlign

Beside visualised interpretation of alignments through ClustalX or ralee mode, a sonificated method of alignment interpretation is described in Appendix A.

## 2.3.2 Promoter Analysis

There are many genomic parts beyond the “gene” (in terms of the part which is transcribed), which are inevitable for transcription and its functionality. Although there are many regulatory units, some of them located many kilobases upstream/downstream of the gene, in most cases in this thesis we just examined 100nt upstream of transcription initiation. Within this region we can find various elements: GC-box (-90 nt, Pol II), CAAT-box (-70 nt, Pol II), proximal sequence elements (PSE, -60 nt, Pol III), Octamer motif (-54 nt), conserved -35 region, TATA box (-10 nt, Pol II and III) and  $\kappa$ B (-10 nt, Pol II). These elements were searched and identified with r nabob or MEME.

### r nabob

For this method a certain motif already has to be known. This can be modeled by r nabob (for a program description see section 2.1.3) and with a possible mutation rate searched in the upstream region of our candidates. This program is chosen for filtering many hundreds of candidates, e.g. 7SK (Section 4.3), Telomerase-template part (Section 4.4), or RNase MRP, RNase P (Section 4.2) or U3 snoRNA (Sec. 4.1).

## MEME

**MEME** (Multiple EM for Motif Elicitation) is one of the most widely used tools for searching novel motifs in sets of biological sequences. Applications include the discovery of new transcription factor binding sites and protein domains. **MEME** works by searching for repeated, ungapped sequence patterns that occur in the DNA (or protein) sequences provided by the user [91, 92]. This program is mainly used for all non-coding RNAs. In case of snRNAs we discovered with **MEME** (v.3.5.0) motifs upstream of the sequences for analysis of regulators and other possible dependencies. They were manually compared with previously published sequence elements. We visually compared the **MEME** patterns with the upstream elements in related species from the following literature sources: [93] (general motifs), [94–97] (human), [98, 99] (chicken), [100] (insects), [101] (*Bombyx mori*), [102] (*Strongylocentrotus purpuratus*), [103] (*Caenorhabditis elegans*).

### 2.3.3 Synteny Information

In order to assess whether ncRNA genes are mobile in the genome, we determined their flanking protein-coding genes. We used the **ensembl compara** annotation [104] to retrieve homologous proteins in other genomes and compared whether these homologs also have adjacent ncRNAs. This method was mainly used for snRNAs. For consistency, this analysis was performed based on **ensembl** (release 46) [105] using the data integration platform **BioFuice** [106]. More precisely, for each human snRNA  $G$  we examined the relation of the left homologous  $L_H(G)$  and right homologous  $R_H(G)$  of flanking protein coding genes  $L(G)$  and  $R(G)$  on both sides of  $G$ . We only considered annotations in  $L_H(G)$  and  $R_H(G)$ , respectively, if the sequence distance between  $G_H$  and  $L_H(G)$  and  $R_H(G)$  was not more than twice (five times for mammals) the distance between  $G$  and  $L(G)$  and  $R(G)$ .

### 2.3.4 Target-Prediction

RNA-protein interaction is a mysterious unsolved problem. For RNA-RNA interaction several program are developed, however most of them do not mirror *in vivo* interactions. It is not known how minimum free energy is influenced, especially if these interaction are short sequences containing bulges. Programs like **RNAduplex** [107] or **RIP** [108] are developed for a closer estimation of RNA-RNA interaction.

For snoRNA target prediction a reliable program has recently been written. The targets of the novel box H/ACA snoRNA candidate are computed using the novel



run-time efficient `snoplex` program [85]. This tool implements a dynamic programming algorithm to compute the binding energy of the snoRNA sequence to its target together with the energy of the snoRNA structure itself. In order to assess putative binding sites, `snoplex` furthermore considers the initial energy of the snoRNA structure, the energy that is necessary to open the target site and the duplex energy which is also depended on the surrounding snoRNA structure. Given a snoRNA sequence, `snoplex` scans the target RNA sequence and returns the set of thermodynamically most stable interaction structures.

## 2.4 Other Commonly Used Programs

### 2.4.1 Multiple Candidates

Multiple copies within an organism might be founded by unfinished assemblies. Contigs might contain single genomic locations multiple times. Therefore `blastclust` is used to filter out identical sequences. In cases of e.g. snRNAs, each gene is supposed to be present in a larger copy number. Some of them might even be pseudogenes. A combination of `blastclust` and `MEME` (for functional promoter analysis, see above) is used to estimate the correct number of functional genes.

`blastclust` is a program within the standalone `Blast` package used to cluster either protein or nucleotide sequences. The program begins with pairwise matches and places a sequence in a cluster if the sequence matches at least one sequence already in the cluster. In the case of nucleotide sequences, the `megablast` algorithm is used<sup>5</sup>.

### 2.4.2 RNAz and Annotation Pipeline

An important sub-class, which includes the housekeeping ncRNAs, has evolutionarily conserved secondary structures. These ncRNAs can be identified by methods such as `RNAz` [52] and `EvoFold` [109] that search for regions with an excess of mutations that maintain the secondary structure.

We used `multiz` [110] or `NcDNAalign` [111] to produce an alignment of the reference genome and closely related genomes (e.g. *Trichoplax* as reference genome, *Nematostella*, and *Hydra*). Only the blocks that contained the reference genome and at least one of the two cnidarian species were used for further analysis.

---

<sup>5</sup><http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html>

These sets of input alignments were passed to the **RNAz** [52] pipeline and processed in the same way: Alignments longer than 120nt are cut into 120 slices in 40nt steps. In a series of filtering steps sequences were removed from the individual alignments or alignment slices if they were (a) shorter than 50nt, or (b) contained more than 25% gap characters or (c) had a base composition outside the definition range of **RNAz**. All preprocessing steps were performed using the script `rnazWindows.pl` of the current release of the **RNAz** package. Overlapping slices with a positive ncRNA classification probability of  $p > 0.5$  were combined using `rnazCluster.pl` to a single annotation element, which we refer to as *locus*. In order to estimate the false discovery rate (FDR) of the screen we repeated the entire procedure with shuffled input alignments using `rnazRandomizeAln.pl`.

**RNAz** [52] has been proved to yield results in wide variety of species, from screens of the human genome compared against (mostly) mammalia [112, 113], teleost fishes [114], urochordates [115], nematodes [116], flies [117], yeasts [118], and *Plasmodium* [119]. In brief, **RNAz** is a machine learning tool that determines for a slice of aligned genomic DNA whether it encodes a structured RNA depending on measures of thermodynamics stability and evolutionary conservation [52].

Here **RNAz** was used as for yielding introns of mlncRNAs in flies (Section 3.6), ncRNA candidates in *Trichoplax*, *Schistosoma* and nematods (Chapter 5).

### 2.4.3 Phylogenetic Analysis

Phylogenetic Analysis were used in different parts of this thesis. The exemplary use of **SplitsTree** is shown here at snRNAs. snRNA are short sequences and in addition there are several highly variable regions. We uses split decomposition [120] and the neighbour net [121] algorithm (as implemented as part of the **SplitsTree4** package [2]) to construct phylogenetic networks rather than phylogenetic trees. The advantage of these method is that they are very conservative and that the reconstructed networks provide and easy-to-grasp representation of the considerable noise in the sequence data.

### 2.4.4 Discarding Repeats

**RepeatMasker** [122] screens DNA sequences for interspersed repeats and low complexity DNA sequences. The output of the program is a detailed annotation of the repeats that are present in the query sequence as well as a modified version of the query sequence in which all the annotated repeats have been masked. Sequence

comparisons in `RepeatMasker` are performed by the program `cross_match`, an efficient implementation of the Smith-Waterman-Gotoh algorithm.

For *D. melanogaster* we downloaded the `RepeatMasker` annotation from the UCSC genome browser excluding simple repeats and low complexity regions. We discarded introns overlapping a repeat with at least 10%.

#### 2.4.5 Example for Homology Search of All Known ncRNAs

We employed the following five steps for homology search in *S. mansoni*:

(a) Candidate sequences for ribosomal RNAs, spliceosomal RNAs, the spliced leader and the SRP RNA, we performed `blast` searches with  $E < 10^{-3}$  using the known ncRNA genes from the NCBI and `Rfam` databases. For the snRNA set, see [123]. For 7SL RNA we used **X04249**, for 5S and 5.8S rRNAs we used the complete set of `Rfam` entries, for the SSU and LSU rRNAs, we used **Z11976** and **NR\_003287**, respectively. The spliced-leader SL RNAs were searched using SL-RNA entries from `Rfam` and the sequences reported in [124]. For more diverged genes such as minor snRNAs, RNase MRP, 7SK, and RNase P, we used `GotohScan` [61], an implementation of a full dynamic programming alignment with affine gap costs. In cases where no good candidates were found we also employed descriptor-based search tools such as `rnabob`<sup>6</sup>.

(b) In a second step, known and predicted sequences were aligned using `ClustalW` [80] and visualized with `ClustalX` [125]. To identify functional secondary structure, `RNAfold`, `RNAalifold`, and `RNAcofold` [126] were used. Combined primary and secondary structures were visualized using `stockholm-format` alignment files in the `emacs` editor utilizing `ralee` mode [57]. Alignments are provided in the Supplemental Material.

(c) Putatively functional sequences were distinguished from likely pseudogenes by analysis of flanking genomic sequence. To this end, the flanking sequences of snRNA and SL RNA copies were extracted and analyzed for conserved sequence elements using `meme` [91]. Only snRNAs with plausible promoter regions were reported.

(d) Additional consistency checks were employed for individual RNA families, including phylogenetic analysis by neighbor-joining [127] to check that candidate sequences fall at phylogenetically reasonable positions relative to previously known homologs. For RNase MRP RNA candidates, `RNAduplex`<sup>7</sup> was used to find the pseudoknot structure. In order to confirm that the SL RNA candidate was indeed

---

<sup>6</sup> <http://selab.janelia.org/software.html>

<sup>7</sup> <http://www.tbi.univie.ac.at/RNA/RNAduplex.html>

*trans*-spliced to mRNA transcripts, we searched the *FAPESP Genoma Schistosoma mansoni* website <http://bioinfo.iq.usp.br/schisto/> for ESTs including fragments of the predicted SL RNA. We found 52 ESTs with **blast**  $E < 0.001$  spanning the predicted region of the SL RNA (nt 8-38), indicating that this RNA did indeed function as a spliced leader.

(e) Accepted candidate sequences were used as **blast** queries against the *S. mansoni* genome to determine their copy number in the genome assembly.

## Chapter 3

# RNAs involved in mRNA processing

Already during the transcription of a gene within eukaryotic cells a procedure of preparing pre-mRNA for translation starts outside of the nucleus containing several steps. A modified guanosine is linked through a 5,5-triphosphat bond to the 5'-end of the pre-mRNA. This 5'-cap is involved in binding to the ribosome for translation and furthermore protects mRNA against 5' exonucleases. After transcription termination the 3' end of the transcript is usually immediately polyadenylated. About 30-200 adenines facilitate the export of mRNA from the nucleus and protect the transcripts against degradation. Beside RNA-Editing, splicing is a more basic and indispensable step of processing. Introns are removed by major or minor spliceosome from pre-mRNA and exons are joint together. The evolution of the splicing machinery and the corresponding non-coding RNAs involved in *cis*-splicing are examined in detail in an evolutionary context in section 3.1 and 3.2, respectively.

Another protein-dependent spliceosome acts in some eukaryotic phyla in the form of *trans*-splicing. For this processing step a leader sequence derived from a small non-coding RNA, containing a hypermodified cap, is transferred to a 5' polycistronic transcript. These SL-RNAs present in single phyla wide-spread over of the phylogenetic tree of eukaryots. In section 3.3 various possible secondary structures are calculated and a possible common origin of SL RNAs is presented. Recently, for nematodes new investigated non-coding SmY RNA was proposed to interact SL2 RNA. A complete overview of existing SmY RNAs via homology search is performed in section 3.4. Additionally, we discuss our finding of SmY-SL2 RNA-RNA interaction.

Table 3.1: Splicing Mechanisms. Three major mechanisms, (A), (B), and (C) can be distinguished [130]. Group I [131] and group II [132] (which include the group III introns) are self-splicing. However, Group II introns also share several characteristic traits, including the lariat intermediate, with spliceosomal introns and might share a common origin. The splicing of eukaryotic tRNAs and all archaeal introns uses specific splicing endonucleases, reviewed in [133]. The spliceosomal machinery does not distinguish between protein coding mRNAs and mRNA-like ncRNAs.

Domain	(A)	(B)		(C)
	group I	group II	spliceosomal	endonuclease
Bacteria	+	+	–	–
Archaea	–	–	–	tRNA, rRNA, mRNA
Eukaryota	+	+	“mRNA”	tRNA

The interaction of U7 RNP with the histone downstream element (HDE) replaces the polyadenylating step and is therefore crucial for the correct processing of histone 3' elements. An extensively homology search of the only involved non-coding RNA U7 is performed in section 3.5.

Many eukaryotic transcripts consists of mRNA-like non-coding RNAs (mlncRNAs). These capped and polyadenylated non-coding RNAs are additionally often spliced. In section 3.6 we show a comprehensive genome-wide comparative genomics approach searching for short conserved introns in order of identifying conserved transcripts with a high specificity.

### 3.1 Evolution of the Splicing Machinery

In eukaryots, introns of protein-coding mRNA and mRNA-like ncRNAs are spliced out of the primary transcript by the spliceosome, a large RNP complex which consists of up to 200 proteins and five small ncRNAs [128]. Mounting evidence suggests that these snRNAs (Section 3.2) exert crucial catalytic functions in the splicing process [129]. Spliceosomal splicing is one of four distinct mechanisms, see Tab. 3.1 and Fig. 3.1 for details.

The spliceosomal machinery itself may be present in three distinct variants in eukaryotic cells, Fig 3.2. The dominant form is the *major spliceosome* which contains the snRNAs U1, U2, U4, U5 and U6 and removes in *cis* introns delimited by the canonical donor-acceptor pair GT-AT (as well as some AT-AC, GC-AG and some other underrepresented introns). A recent report on the expression of a U5 snRNA candidate in *Giardia* [134], a protozoan with few introns, suggests that

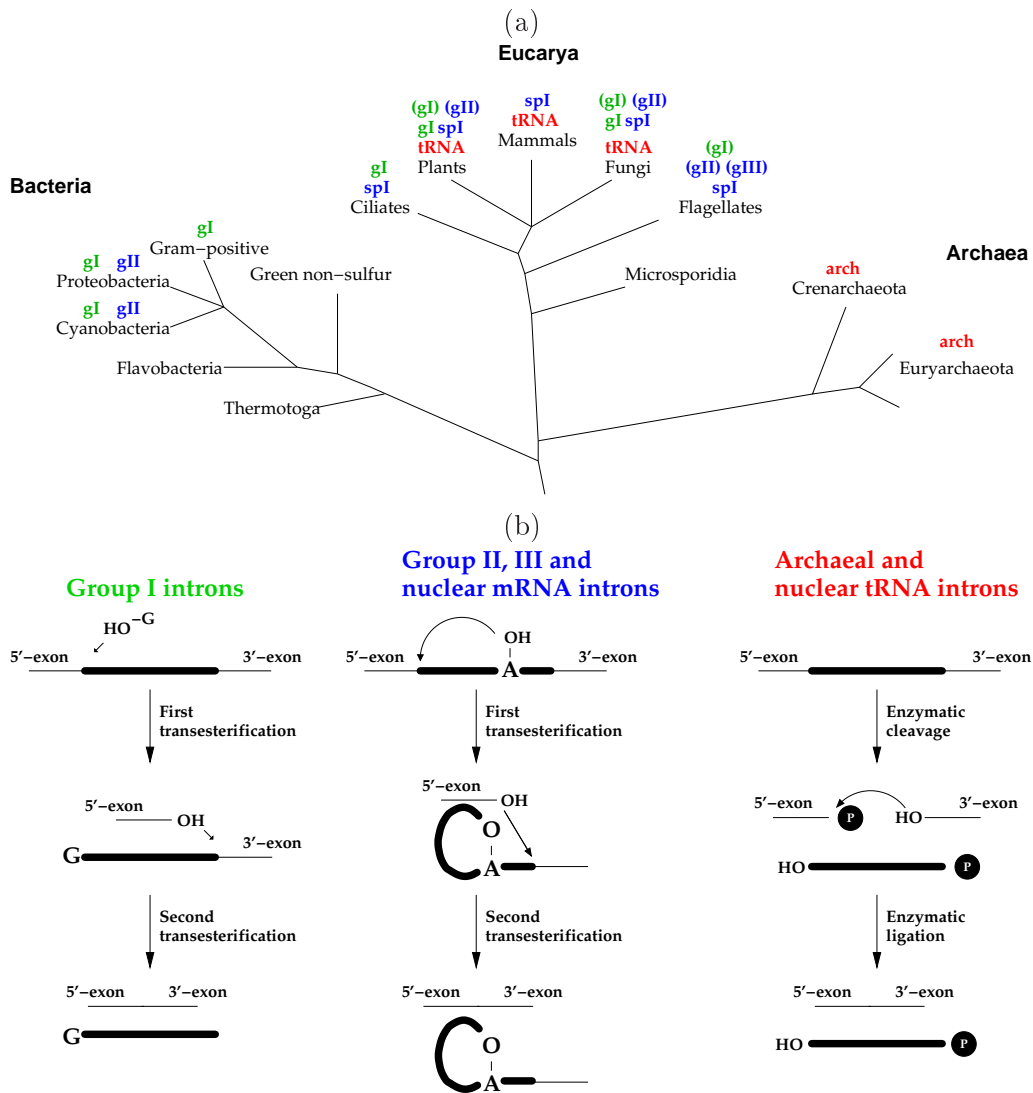


Figure 3.1: (a) Phylogenetic tree showing the known distribution of the different classes of introns that are colour-coded according to their splicing mechanism as shown in (b): arch – archaeal (red), tRNA – nuclear tRNA (red); gI – group I (green); gII – group II (blue), gIII – group III (blue); spI – nuclear mRNA (blue) (also called spliceosome introns). Mitochondrial and chloroplast introns are given in brackets. (b) The three mechanisms of introns removal. Group I introns (green) are removed by the two transesterification reactions that are illustrated. The subsequent circularization of some group I introns is not shown. Group II, group III and nuclear mRNA introns (blue) are also excised by two consecutive transesterifications, that are outlined, to produce ligated exons and an intron lariat. Archaeal and nuclear tRNA introns (red) are excised by a splicing endoribonuclease that generates 5'-OH and 2',3'-cyclic phosphates and then the exons are ligated. Figure taken from [130] and modified. No distances are given in the tree.

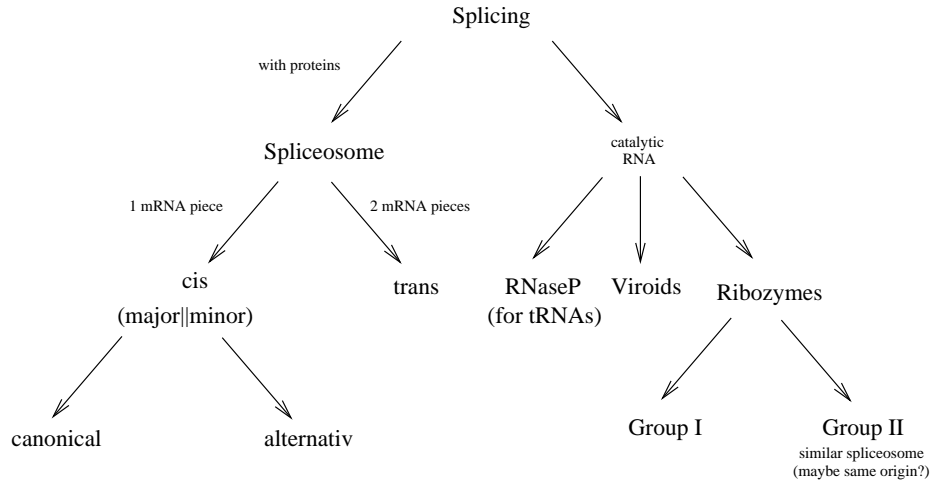


Figure 3.2: Splicing types. Splicing can be divided in splicing with and without proteins. In this thesis the non-coding RNAs interacting with protein components of the spliceosome is detailed described (Section 3.2,3.3). The event of removing part of the sequence (splicing) dates probably back until LUCA. Only eukaryots splice with proteins.

the spliceosome and its snRNA date back to the eukaryote ancestor. In general, snRNAs are subject to concerted evolution if they are present in multiple copies. Nevertheless, there is evidence for differential regulation of paralogous snRNA genes in several lineages [94, 123, 135], Sec. 3.2.

About 1 in 10 000 protein coding genes is spliced by the *minor spliceosome* [136] which is composed of the snRNAs U11, U12, U4atac, U5 and U6atac and acts in *cis* on AT-AC (rarely GT-AG) [137] introns. The snRNAs U11, U12, U4atac, and U6atac take on the roles of U1, U2, U4, and U6. Whereas, both U6 and U6atac are polymerase-III transcripts, all other spliceosomal snRNAs are transcribed by polymerase-II. Interestingly, the minor spliceosome can also act outside the nucleus and has a function in the control of cell proliferation [138]. Functional and structural differences between the two types of spliceosomes are reviewed in [139]. The snRNAs themselves are not only part of the spliceosomes but are also involved in transcriptional regulation [140].

The minor spliceosome is present in most eukaryotic lineages and traces back to an origin early in eukaryotic evolution [141–143]. Although it appears to have been lost in many lineages, most metazoa have a minor spliceosome, with the notable exception of nematodes such as *Caenorhabditis elegans* [136] and certain cnidaria [123, 144], Sec. 3.2. Nowadays, it is discussed if the minor spliceosome is completely absent or highly divergent among these organisms. Within fungi, minor spliceosomes have been reported only for zygomycota and some chytridiomycota.



Minor spliceosomes are also reported in oomycetes (Heterokonta) and streptophyta [144]. Whereas, Euglenozoa and Alveolata do not seem to have minor spliceosomes.

The third type of splicing is *spliced-leader-trans-splicing*. Here a “miniexon” derived from the non-coding spliced-leader RNA (SL RNA) is attached to the 5’ end of each protein-coding exon [145–147], Sec. 3.3. The corresponding spliceosomal complex contains the snRNAs U2, U4, U5, and U6, as well as an SL RNA [147].

The evolutionary origin of SL-*trans*-splicing was recently unclear. It has been described in tunicates, nematodes, platyhelminthes, cnidarians, euglenida, kinetoplastids [147], rotifera [146] and dinoflagellates [148]. Due to the rapid evolution and the small size of SL RNAs it is hard to determine whether examples from different phyla are true homologs or not. Thus, two competing hypotheses were previously discussed in the literature: (i) ancient *trans*-splicing and SL RNAs have been lost in multiple lineages and (ii) the mechanism has evolved independently as a variant of spliceosomal *cis*-splicing in multiple lineages. Recently, the second hypothesis is more and more rejected (see Sec. 3.3).

In nematodes polycistronic pre-mRNAs are *trans*-spliced into two or even more [149] distinct SL RNAs which provide the 5’ acceptor site for the first (SL1) and all subsequent (SL2) mRNA sequences. This leads to the formation of discrete monocistronic mRNAs that start with either the SL1 or the SL2 sequences [150]. In some organisms many (in case of *Trichinella spiralis* at least 15) highly polymorphic noncanonical splice leaders are known [149]. The individual spliced leaders vary in both size and primary sequence, showing a much higher degree of diversity that was previously thought.

### 3.2 *cis*-splicing with small nuclear RNAs

In most eukaryote lineages, introns are spliced out of protein-coding mRNAs by the spliceosome, a huge RNP complex consisting of about 200 proteins and five small non-coding RNAs [128]. These snRNAs exert crucial catalytic functions in the process [129, 151, 152] in three distinct splicing machineries. The *major spliceosome*, containing the snRNAs U1, U2, U4, U5 and U6, is the dominant form in metazoans, plants, and fungi, and removes introns with GT-AG (as well as rarely AT-AC and GC-AG) boundaries. Another class of “non-canonical” introns with AT-AC (and rarely GT-AG [137]) boundaries is excised by the *minor spliceosome* [136], which contains the snRNAs U11, U12, U4atac, U5, and U6atac. Just as the major spliceosome, the minor spliceosome is present across most eukaryotic lineages and traces back to an origin very early in the eukaryote evolution [141–144].

Recently it was found that the minor spliceosome can also act outside the nucleus and controls cell proliferation [138]. Functional and structural differences of two spliceosomes are reviewed in [139]. The third type of splicing the *SL-trans-splicing*, in which a “miniexon” derived from the non-coding spliced-leader RNA (SL) is attached to each protein-coding exon. The corresponding spliceosomal complex requires the snRNAs U2, U4, U5, and U6, as well as an SL RNA [147]. Due to the high sequence variation of the short SL RNAs, and the patchy phylogenetic distribution of SL-trans-splicing, the evolutionary origin(s) of this mechanism, which is active at least in chordates, nematodes, cnidarians, euglenozoa, and kinetoplastids, is still unclear.

Previous studies on the evolutionary origin of the spliceosomes have been performed predominantly based on homology of the most important spliceosomal proteins. Thus relatively little detail is known on the evolution of the snRNA sequences themselves beyond the homology of nine families of snRNAs across all eukaryotes studies so far [141–143, 153–155]. This may come as a surprise since it has been known for more than a decade that at least all of the snRNAs of the major spliceosome appear in multiple copies and that these paralogs are differentially regulated in at least some species, see e.g. [99, 156–159]. Very recently, however, some of these variants have been studied in more details, see e.g. [101, 135, 160–163] and the references therein. The only systematic study that we are aware of is the recent comprehensive analysis of 11 insect genomes [100] which reported that phylogenetic gene trees of insect snRNAs do not provide clear support for discernible paralog groups of U1 and/or U5 snRNAs that would correspond to the variants with tissue-specific expression patterns. Instead, the analysis supports a concerted mode of evolution and/or extreme purifying selection, a scenario previously described for snRNA evolution [164–166].

In this contribution we extend the detailed analysis of the nine spliceosomal snRNAs to metazoan animals. In particular in mammals, the analysis is complicated by a high copy number of snRNAs of the major spliceosome and an associated large number of pseudogenes [167]. We focus here on four questions: (1) Is there evidence for discernible paralog groups of snRNAs in some clades? A dominating mode of concerted evolution does not necessarily prevent this, as demonstrated by the existence of two highly diverged copies of both LSU and SSU rRNA in Chaetognatha [168, 169], which is probably associated with a duplication of the entire rDNA cluster. (2) Are there clades with deviant snRNA structures? The prime example for a highly divergent snRNA is the U11 in a subset of the insects [154]. (3) Are there interpretable trends in the copy number of snRNAs across metazoa? (4) How mobile are snRNA genes relative to the “background” of protein coding genes? In other words, to what extent are some or all of the snRNA

genes off-springs of a locus that remains stably linked to its context over large time-scales.

Over all, the published experimental evidence on metazoan snRNAs is very unevenly distributed. For example, a large and phylogenetically diverse set of U2 snRNA sequences is reported in [170], while most other snRNAs have been reported for a few model organisms only. A recent experimental screen for snRNAs in *Takifugu rubripes* [171] resulted in copies of eight snRNAs families. U4atac was missing, but a plausible candidate can easily be found by **blast**. Only a few sequences of minor spliceosomal snRNAs have been reported so far, mostly in a few model mammals [95] and in Drosophilids [100, 154].

### 3.2.1 Homology Search

Tab. 3.2 summarizes the results of the sequence homology search detailed in the Methods section, Sec. 2.1.1. Only sequences that passed all filtering steps and structure checks are reported as “homologs” in the following. We found that, with few exceptions, **blast**-based homology search strategies are in general sufficient to find homologs of all nine spliceosomal snRNAs in most metazoan genomes. The procedure is hard to automatize, however, since in many cases the initial **blast** hits have poor *E*-values, while a multiple sequence alignment then leaves little doubt that a true homolog has been found. This is in particular true for searches bridging large evolutionary distances, in particular when the search extends beyond bilateria.

With very few exceptions we found multiple copies of all five major spliceosomal RNAs that exhibited the typical snRNA-like promoter elements and were hence mostly likely functional copies of the genes. The snRNA copy numbers varied substantially between different clades. The genus *Caenorhabditis*, for example, was set apart from other nematodes by a two to threefold increase in the number of major spliceosomal snRNAs. In contrast, the snRNAs of the minor spliceosome were in most cases single-copy genes.

Many genomes, most notably mammalian genomes, contained a sizeable number of major snRNA pseudogenes. Table 3.2 therefore lists only candidates that have plausible snRNA-like promoter structure, that fit the secondary structures of snRNAs in related species, and that exhibit strong sequence similarity in the unpaired regions of the molecule. These are rather restrictive criteria. In the electronic supplement, we therefore provide a corresponding table that is based only on sequence homology.

Table 3.2: Approximate copy number of snRNA genes.

We list here only those sequences that (1) are consistent with the secondary structures of related snRNAs, (2) show substantial sequence conservation in the unpaired regions of these structures, and (3) have recognizable promoter motifs. In some cases none of the candidates satisfies all these criteria. Entries of the form  $S_0$  and  $P_0$  indicate that there is homologous sequence which however lacks structural similarity or recognizable promoter elements. The quality of the genome assembly is marked by the following symbols:  $\triangle$  - Traces,  $\square$  - Contigs,  $\diamond$  - Scaffolds,  $\spadesuit$  - Chromosomes.

Coverage	Species	U1	U2	U4	U5	U6	U11	U12	U4atac	U6atac
$\diamond$	<i>M. brevicollis</i>	0	0	0-1	0-2	1	0	0	0	0
$\triangle$	<i>Reniera sp</i>	2	0-1	2	3	2	1	1	0	3
$\diamond$	<i>Trichoplax adhaerens</i>	1	1	1	1	2	1	1	1	1
$\diamond$	<i>N. vectensis</i>	2	2	4	5	3	3	3	1	2
$\triangle$	7.45-8.33X <i>H. magnipapillata</i>	4	2	5	7	4	1	1	0	2
$\triangle$	0.05X <i>A. millepora</i>	0	2	0	2	2	0	0	0	0
$\triangle$	0.047X <i>A. palmata</i>	1	0	0	0	1	0	0	0	0
$\diamond$	<i>S. mansoni</i>	3	3	1	2	9	1	1	1	1
$\square$	<i>S. mediteranea</i>	2	$P_0$	3	2	2	0	0	0	0
$\triangle$	13.03X <i>L. gigantea</i>	3	8	11	2	7	2	1	0	2
$\triangle$	0.05X <i>B. glabrata</i>	$S_0$	2	0	1	$S_0$	0	0	0	0
$\triangle$	0.54X <i>P. lobata</i>	1	1	1	0	0	0	0	0	0
$\triangle$	0.012X <i>E. scolopes</i>	$SP_0$	1	0	0	0	0	0	0	0
$\triangle$	4.48X <i>A. californica</i>	4	2	4	10	8	1	1	0	1
$\diamond$	<i>C. capitata</i>	5	2	1	4	2	1	1	1	1
$\diamond$	<i>H. robusta</i>	6	8	4	7	4	0	1	1	1
$\triangle$	0.23X <i>H. bacteriophora</i>	2	2	0	2	1	0	0	0	0
$\triangle$	11.33X <i>B. malayi</i>	3	3	1	1	2	1	0	0	0
$\triangle$	12.15X <i>T. spiralis</i>	1	5	2	3	1	1	0	0	0
$\triangle$	11.24X <i>P. pacificus</i>	2	2	4	4	7	1	0	0	0
$\square$	<i>C. brenneri</i>	19	19	10	19	25	0	0	0	0
$\square$	<i>C. remanei</i>	14	11	5	13	15	0	0	0	0
$\triangle$	10.18X <i>C. japonica</i>	16	15	4	14	7	0	0	0	0
$\spadesuit$	<i>C. elegans</i>	10	17	4	9	15	0	0	0	0
$\spadesuit$	<i>C. briggsae</i>	9	10	4	10	22	0	0	0	0
$\triangle$	3.29X <i>D. pulex</i>	5	6	4	9	8	1	1	$PS_0$	1
$\triangle$	11.81X <i>P. humanus</i>	3	4	1	2	1	1	1	0	1
$\square$	<i>N. vitripennis</i>	7	4	3	5	5	1	2	1	2
$\triangle$	2.58X <i>I. scapularis</i>	4	4	3	4	3	0	1	0	1
$\triangle$	1.6X <i>A. pisum</i>	2	3	0	2	3	1	1	0	1
$\diamond$	<i>A. mellifera</i>	5	3	2	3	3	1	1	1	1
$\diamond$	<i>B. mori</i>	5	6	3	5	4	1	1	1	2
$\triangle$	0.75X <i>T. castaneum</i>	5	5	2	6	3	1	1	0	1
$\spadesuit$	<i>A. gambiae</i>	7	7	2	5	2	2	1	1	1
$\spadesuit$	<i>D. melanogaster</i>	5	6	3	7	3	1	1	1	1
$\spadesuit$	<i>D. ananassae</i>	9	8	2	4	2	1	1	1	1
$\spadesuit$	<i>D. erecta</i>	8	9	3	7	4	1	1	1	1
$\spadesuit$	<i>D. grimshawi</i>	7	6	3	7	3	1	1	1	2
$\spadesuit$	<i>D. mojavensis</i>	6	8	3	6	3	1	1	1	1
$\spadesuit$	<i>D. persimilis</i>	7	7	3	7	3	1	1	1	1
$\spadesuit$	<i>D. pseudoobscura</i>	7	7	3	6	3	1	1	1	1
$\spadesuit$	<i>D. sechellia</i>	7	6	3	7	3	1	1	1	1
$\spadesuit$	<i>D. simulans</i>	8	6	3	8	3	1	1	0	1
$\spadesuit$	<i>D. virilis</i>	6	8	3	6	2	1	1	2	1
$\spadesuit$	<i>D. willistoni</i>	8	9	3	8	$P_0$	1	1	1	0
$\spadesuit$	<i>D. yakuba</i>	8	7	3	8	3	1	1	1	1

Coverage	Species	U1	U2	U4	U5	U6	U11	U12	U4atac	U6atac
◇	<i>S. purpuratus</i>	5	7	9	8	3	2	3	1	1
△ 3.77X	<i>S. kowalevski</i>	7	4	4	5	4	1	2	0	3
◇	<i>C. savignyi</i>	3	2	3	7	2	1	1	1	1
◇	<i>C. instestinalis</i>	1	1	3	5	2	1	1	1	1
△ 7.8X	<i>O. dioica</i>	1	6	2	7	4	0	0	0	0
◇	<i>B. floridae</i>	8	3	5	9	4	1	1	0	1
△ 6.19X	<i>P. marinus</i>	6	5	8	9	5	1	2	$PS_0$	3
♠	<i>D. rerio</i>	5	4	4	7	3	1	1	1	1
♠	<i>O. latipes</i>	4	2	2	4	4	1	1	1	1
♠	<i>G. aculeatus</i>	6	2	4	7	3	1	1	1	1
◇	<i>F. rubripes</i>	5	5	3	6	4	1	1	1	1
♠	<i>T. nigroviridis</i>	4	5	3	5	2	1	1	0	1
◇	<i>X. tropicalis</i>	5	1	3	2	5	1	1	1	2
♠	<i>G. gallus</i>	1	1	1	2	4	1	1	1	1
△ 8.34X	<i>T. guttata</i>	2	5	2	3	2	1	1	0	1
△ 8.24X	<i>A. carolinensis</i>	14	6	2	6	5	1	2	1	1
♠	<i>O. anatinus</i>	5	2	2	4	6	1	1	1	1
♠	<i>M. domestica</i>	7	4	2	5	6	1	$PS_0$	1	1
♠	<i>M. musculus</i>	7	5	1	6	7	1	2	1	2
♠	<i>R. norvegicus</i>	4	10	1	4	5	4	1	1	1
♠	<i>C. familiaris</i>	6	5	2	4	5	1	1	1	1
♠	<i>B. taurus</i>	7	8	2	5	6	2	1	1	1
♠	<i>P. tropicalis</i>	7	2	2	7	8	1	1	3	1
♠	<i>H. sapiens</i>	8	3	2	5	7	1	1	3	1

It is surprisingly difficult to compare the present snRNA survey with previous reports on vertebrate snRNAs. The main reason for discrepancies in the count of snRNAs is that distinguishing functional snRNAs from pseudogenes is still an unsolved problem. In this contribution, we use a very stringent criterion by insisting on a recognizable promoter structure. In some cases, however, it is known that snRNAs have internal promoters only [172]. These cases constitute false negatives in Tab. 3.2. On the other hand, much of the published literature considers sequence similarity to the known functional genes as the only criterion, thus most likely leading to the inclusion of a substantial fraction of pseudogenes. For instance, ref. [173] counts 16 U1, 6 U2 and 44 U6 snRNAs in the human genome (compared to our 8, 3, and 7, resp.), while [94] report 5-9 U6 snRNA genes, consistent with our list. Similarly, only a fraction of the major spliceosomal snRNAs reported for the chicken genome in [174] pass our promoter analysis.

For Drosophilids, on the other hand, our analysis is almost identical to the results of [100, Tab.1] and the data reported in [101]. Furthermore, we come close the results of a comparative genomics screen for non-coding RNAs in *C. elegans* [116], which reported 12 U1, 19 U2, 5 U4, 13 U5, and 23 U6, i.e., only a few more candidates than our present purely homology-based approach. A comparative screen of the two *Ciona* species for evolutionary conserved structured RNAs [115] missed a small number of snRNA genes that we identified as most likely functional ones.

In a few species we failed to identify individual major spliceosomal snRNAs (e.g. *A. pisum* U4, *H. bacteriophora* U4, and *S. mediterranea* U2). Minor spliceosomal

snRNAs are more often missing. In those cases where only some of the major or minor snRNAs remain undetected, the missing family member most likely escaped our detection procedure for one of several reasons:

(1) in the case of unassembled incomplete genomes for which only shotgun reads were searched, the snRNA may be located in the not yet sequenced fraction of the genome or it might not be completely contained within at least one single shotgun read.

(2) The snRNA in question may be highly derived in sequence. (For instance, the U11 snRNA in Drosophilids [154] cannot be found by a simple `blast` search starting from non-insect sequences. It can be found however, by the combination of very un-specific blast and subsequent structure search as described in Sec. 3.2.1.)

(3) In some cases we list a “0” in Tab. 3.2 even though there is recognizable sequence homology in the genome. In these cases we were not able to identify the snRNA-like promoter elements and/or the secondary structure did not fit the expectations. These cases are marked in the table.

(4) It is conceivable that some species had lost a particular snRNA and replaced it by corresponding snRNA from the other spliceosome. The observation that U4 may function in both the major and minor spliceosomes [175] shows that such a replacement mechanism might indeed be evolutionarily feasible.

In our data set, we most frequently were unable to find a U4atac homolog. We cannot know, of course, whether we missed these cases due to poor sequence conservation or due to loss of the gene. For instance, we did not recover a plausible U4atac candidate for the hemichordate *Saccoglossus kowalevski* despite the fact that the U4atac sequence of the sea urchin *Strongylocentrotus purpuratus* was easily retrieved.

Surprisingly, we found neither a canonical U6 nor a canonical U6atac in *Drosophila willistoni*. A highly derived U6 homolog has no recognizable snRNA-like promoter structure and exhibits substantial deviations from the consensus structure, see section 3.2.5. Interestingly, it is aligned to the functional U6 RNAs of the other 11 Drosophilids in the genome-wide “12-Fly” Pecan alignment<sup>1</sup>, which respects syntenic conservation. This strongly suggests that *D. willistoni* has indeed a highly derived U6 snRNA. According to known annotation the sequence is not located in an intron. The absence of external promoter elements has also been observed for one of the human U6 snRNAs [172], hence the prediction is not at all implausible. Similarly, the U4atac candidate from *Daphnia pulex* deviates substantially from other arthropod sequences. It is possible that in some or all of these cases the snRNA is present in the genome but is not contained in the currently available

---

<sup>1</sup><http://www.sanger.ac.uk/Users/td2/pecan-CAF1>

genomic sequence data. This is most likely the case for the missing minor spliceosomal snRNAs of *Ixodes scapularis*, *Pediculus humanus*, or *Drosophila willistoni*.

In some cases, however, we failed to identify all four minor spliceosomal snRNAs. Consistent with previous work [136] we found no convincing homologs of the minor spliceosomal snRNAs U11, U12, U4atac, or U6atac in any of the nematode genomes, suggesting that the minor spliceosome was lost early in the nematode lineage. Nevertheless, we find some `blast` hits for minor spliceosomal snRNAs in some nematode genomes.

Our analysis furthermore suggests the possible loss of the minor spliceosome in *Oikopleura dioica*, while a complete complement of minor spliceosomal snRNAs was found in the genus *Ciona*. It is unclear, however, whether this is an artifact due to limitations of available shotgun traces.

Our survey provides evidence that most metazoan clades for which genomic sequences are available have retained the minor spliceosome. For many groups, such as Annelida or Cnidaria, we are not aware of earlier references to the existence of minor spliceosome.

### 3.2.2 Specific Upstream Elements

The classical snRNA-specific PSE and TATA elements that have been described in detail for several vertebrates [93, 94] are highly conserved. This appears to be an exception rather than the rule, however: the snRNA upstream elements are highly diverse across metazoa. Our analysis agrees with the recent observation that in Drosophilids there is a rapid turnover in the upstream sequences. Even though the PSE is fairly well-conserved within Drosophilids, it already differs substantially between the major insect groups [100]. Similarly, within the nematodes conservation of upstream elements is limited to the genus level. In general, the PSE of U11, U12 and U4atac is much less conserved than their counterpart in major spliceosomal snRNA genes. For the purpose of this study, the relatively well-conserved elements were used to discriminate functional snRNAs from likely pseudogenes. We concentrated on PSE and TATA elements for this purpose because other snRNA-associated upstream elements, such as SPH, OCT, CAAT-box, GC-box, -35-element and *Inr* are even less well conserved:

A GC-box was identified in *Caenorhabditis* at a non-canonical position (about -68nt). These elements are different for each single snRNA class: U1 GGACGG (44/52 sites), U2 TGGCCG (38/60 sites) and for U5 CGGCCG (39/46 sites). However, also among a single snRNA this element varies a lot: insects have a U1

GC-box GCGCTG at about -75nt (15/39 sites). About half of the U6 sequences of basal deuterostomes show the CAAT-box motif TGCCAAGAA at the known position of -70nt. Interestingly, we found related motifs in the upstream region of Drosophilids U11 (GACCAATAT, -33nt) and other insects U5 snRNA (TTCCAATCA, -28nt). The Octamer motif (OCT, ATTTGCAC) was found in 6 of 7 sequences of basal deuterostomes at the known position of -54nt upstream of U6atac. However, in 12 of 14 Drosophilids sequences, the closely related motif ATTTGCTT was found at position -33nt. About 35nt upstream of U11 and U12 snRNAs of teleosts we found the motif GTGACA and TGCACA, respectively. The *Inr* element of U1 snRNA was found in each species. For teleost fishes and Drosophilids we found a complete set of this element for all snRNAs. However, the element show substantial sequence variations both between different genes in the same species and between homologous genes in different species. We refer to the electronic supplement for further details and lists of identified sequence elements.

### 3.2.3 Clusters of snRNA genes

In Mammalia, we observe linkage of tandem copies of U2 snRNAs, see also [176, 177], while there are no clusters of distinct snRNAs. In *Drosophila*, there are surprisingly constant patterns of snRNA clusters: (a) U2-U5 clusters are observed 4-6 times per genome, (b) there are one or two U1-U2 clusters, and (c) 3-9 tandem copies of snRNAs. Two species deviated therefrom. In *D. ananassae*, we find no U2-U5 cluster, but instead 7 U1-U2, one U4-U5 cluster and 4 other tandem copies, while the *D. willistoni* lacks the U4-U5 cluster but contains 10 U2-U5 pairs and 6 tandem copies. Teleost fishes also have a common pattern: there are one or two U1-U2 pairs and 2-6 tandem copies. In general, however, snRNA do not appear in clusters throughout metazoan genomes.

In several species, linkage of snRNAs with 5S rRNA has been observed [164, 165, 178–181]. We found only one further example of this type: in *Daphnia pulex* 5S and U5 snRNA are separated by only 308bp.

### 3.2.4 Phylogenetic Analysis and Paralogs

Like ribosomal RNAs, spliceosomal RNAs are subject to *concerted evolution* [182–184], i.e., one observes that paralogous sequences in the same species are more similar than orthologous sequences of different species. Multiple molecular mechanisms may account for this phenomenon: gene conversion, repeated unequal crossover, and gene amplification (frequent duplications and losses within family), see [165]





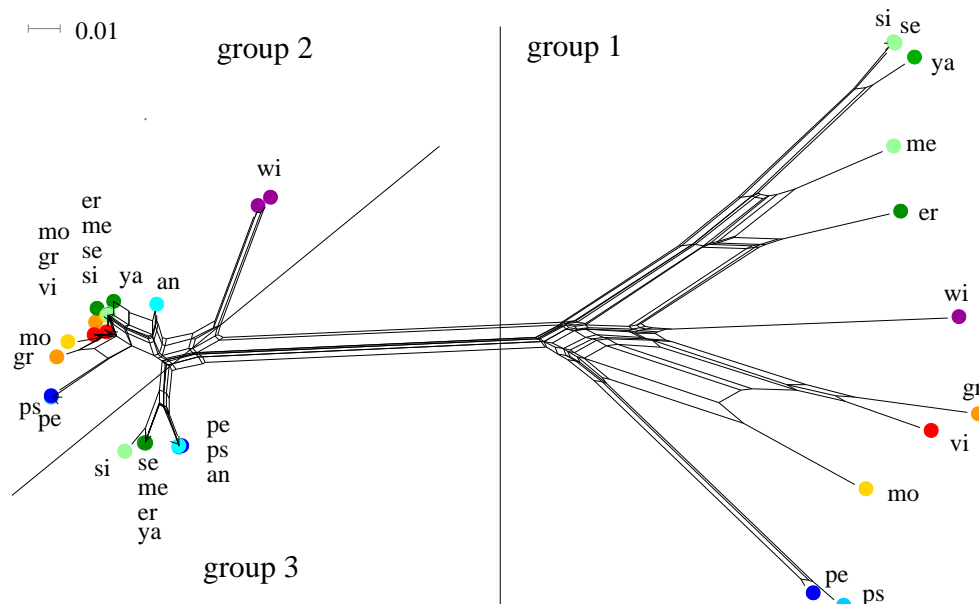


Figure 3.4: Phylogenetic tree of insect U4 snRNAs. In this case we can distinguish three paralog groups within the Drosophilids. me – *D. melanogaster*, er – *D. erecta*, si – *D. simulans*, se – *D. sechellia*, ya – *D. yakuba*, wi – *D. willistoni*, gr – *D. grimshawi*, mo – *D. mojavensis*, vi – *D. virilis*, pe – *D. persimilis*, ps – *D. pseudoobscura*, an – *D. ananassae*.

Fig. 3.3, for example shows that the U5 variants described by [135] do not form clear paralog groups beyond the closest relatives of *Drosophila melanogaster*. On the other hand, there is some evidence for distinguishable paralogs outside the melanogaster subgroup. The situation is much clearer for the Drosophilid U4 snRNAs, where three paralog groups can be distinguished, see Fig. 3.4. One group is well separated from the other two and internally rather diverse. The other two groups are very clearly distinguishable for the melanogaster and obscura group (see [186]). For *D. virilis*, *D. mojavensis*, *D. grimshawi*, and *D. willistoni* we have two nearly identical copies instead of two different groups of genes.

Table 3.3 summarizes the presence of recognizable paralog groups within major animal groups. Within the genus *Caenorhabditis* we find evidence for the formation of U5 paralog groups in *C. remanei*, *C. brenneri*, and *C. briggsae* to the exclusion of *C. elegans* and *C. japonica*. Evidence for paralog groups of U1 snRNA in Drosophilids remains ambiguous due to the small sequence differences.

In teleost fishes, we find clearly recognizable paralog groups for U2, U4, and U5 snRNAs. Surprisingly, the medaka *Oryzias latipes* has only a single group of closely related sequences, despite the fact that for U4, the split of the paralogs appear to predate the last common ancestor of zebrafish and fugu, Fig. 3.5.

Table 3.3: Paralog groups of major spliceosomal snRNAs recognizable within major animal clades. The symbol ● denotes clearly distinguishable paralog groups and refers to the supplemental material for details, ? indicates ambiguous cases, = means that all paralogous genes have identical sequences.

Clade	U1	U2	U4	U5	U6
Annelids	–	–	–	–	=
Nematods	–	–	–	–	=
Caenorhabditis	–	–	–	●	=
Insects	–	–	–	–	=
Drosophilids	?	–	●(Fig.3.4)	●Chen:05	=
Teleosts	–	●(Fig.3.5a)	●(Fig.3.5b)	●(Fig.3.5c)	–
Tetrapoda	–	–	–	–	–
Mammalia	–	–	–	●	–

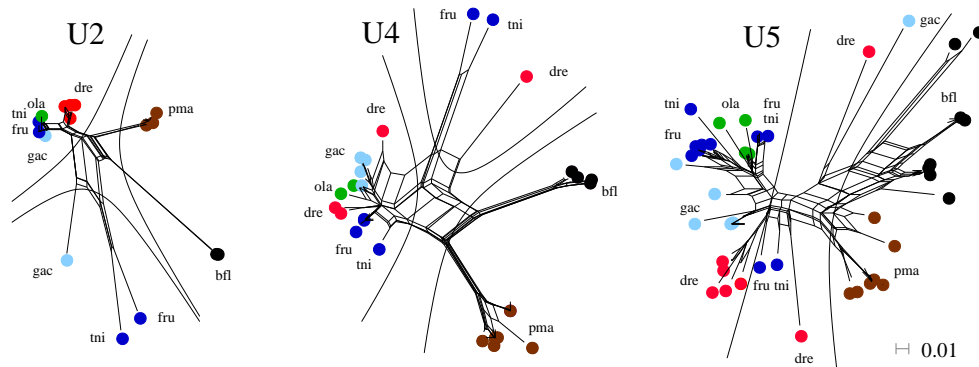


Figure 3.5: Phylogenetic networks of teleost fish snRNAs. Species abbreviations: fru – *Fugu rubripes*, tni – *Tetraodon nigrovirdis*, gac – *Gasterosteus aculeatus*, ola – *Oryzias latipes*, dre – *Danio rerio*, pma – *Petromyzon marinus*, bfl – *Branchiostoma floridae*.

Neither the two rounds of genome duplications at the root of the vertebrates nor the teleost-specific genome duplication has led to recognizable paralog groups of snRNAs. In particular, minor snRNA genes are single-copy genes in teleosts.

### 3.2.5 Secondary Structures

The spliceosomal snRNAs have evolutionarily well-conserved secondary structures [153]. These structures received substantial interest in the past, as exemplified by the following non-exhaustive list of references covering a diverse set of animal species: *Homo sapiens* U1 [187], U2 [188], U4 [189], U5 [158, 190], U6 [188], U11 [95, 143, 191], U12 [95, 143, 191] and U4atac [192], *Rattus norvegicus* U1 [189], U4 [189], U5 [189], *Gallus gallus* U4 [189], U5 [190], *Xenopus laevis* U1 [193], U2 [194],

*Caenorhabditis elegans* U1, U2, U5, U4/U6 [103], *Drosophila melanogaster* U1 [187, 195], U2 [195], U4 [195], U5 [195], U4atac/U6atac, U6atac/U12 [196], *Bombyx mori* U1 [197], U2 [198], *Asselus aquaticus* U1 [199], *Ascaris lumbricoides* U1, U2, U5, U4/U6 [200]. Large changes in snRNA structures over evolutionary time were recently reported for hemiascomycetous yeasts [201]. The comprehensive survey of snRNA sequences throughout metazoa set the stage for a comparably detailed analysis of metazoan snRNA structures. In order to assess structural variations, we constructed structure annotated sequence alignments of all snRNA families. The complete set of alignments and consensus structure models is provided (in Stockholm format) as part of the electronic supplement.

In general we find that snRNA sequences vary more in paired regions than in the loops. The sequence variations almost exclusively comprises compensatory mutations that leave the secondary structures intact. As an example, Fig. 3.6 shows the structures of the U12 snRNA of *Xenopus tropicalis* and *Capitella capitata*. The sequences have few paired nucleotides in common.

Structural variations are typically limited. In Fig. 3.7 we use the U1 snRNAs as a typical example for the evolutionary variation of snRNAs across the metazoa. Overall the structures are extremely well conserved with small variations in the length of the individual stems. With several notable exceptions this is true for all metazoan snRNAs [123].

As reported previously [135], the second stem of U5 snRNA shows some variations. More interestingly, the minor spliceosomal snRNAs tend to be derived in insects. This has been reported previously in particular for U11 in Drosophilids [100, 154]. We found substantial structural variations also for drosophilid U12 snRNAs: there are massive insertions in and after Stem III, while Stem I and II show mispairings. Furthermore, Stem II of U6atac is completely deleted in all examined insects. Details are compiled in the electronic supplement.

Most surprisingly, *Acyrtosiphon pisum* exhibits highly derived structures for all four minor spliceosomal snRNAs, see Fig. 3.8.

The U2 snRNA of *Schmidtea mediterranea* does fit well to the structural alignment of the other U2 snRNAs. In *Schistosoma mansoni* we found a canonical U12 snRNA, while the sequences of the candidates for minor spliceosomal snRNAs do not fit well to the consensus secondary structure models. Details can be found in the electronic supplement.

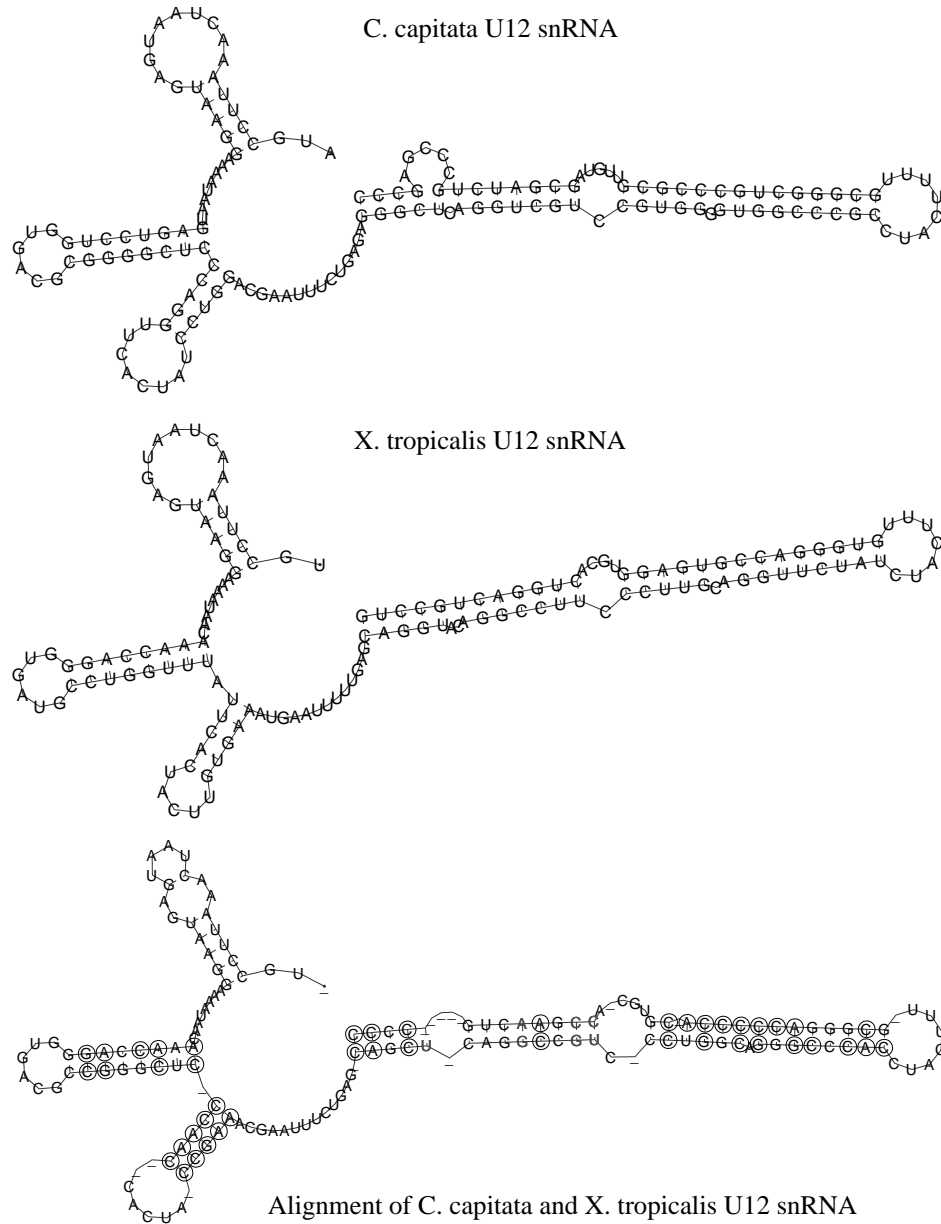


Figure 3.6: Predicted secondary structures of *Capitella capitata*, *Xenopus tropicalis* and an alignment created with RNAalifold of both. Paired circles represent compensatory mutations (e.g. AT  $\rightarrow$  GC), while circles on only one side of a base pair indicate “consistent” mutations (e.g. GU  $\leftrightarrow$  GC).

### 3.2.6 Syntenic Conservation

In order to assess the conservation of the genomic positions of the snRNAs we retrieved the protein coding genes adjacent to the 31 human snRNAs (8 U1, 3

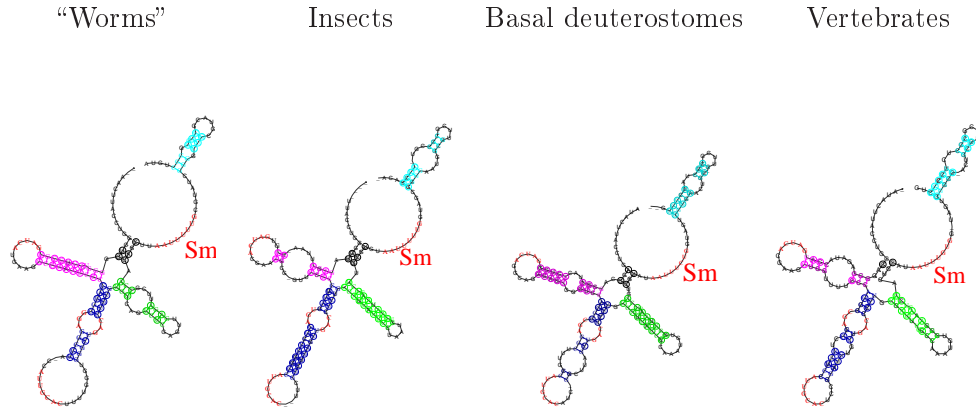


Figure 3.7: Secondary structure prediction of U1 snRNA, folded by *RNAalifold*. From left to right: protostomia without insects, insects, deuterostomes without vertebrates, vertebrates. Red: Conserved sequences in all organisms, which possibly bind to proteins. Sm binding site marked separately.

U2, 2 U4, 5 U5, 7 U6, 1 U11, 1 U12, 3 U4atac and 1 U6atac) and compared the position of their homologs in 14 vertebrate genomes (teleosts, frog, chicken, platypus, opossum, rodents, cow, dog, and chimp) with the 234 snRNA genes that were found in these genomes. We found syntenic conservation of snRNA and flanking genes in only 36 cases, of which 20 belong to the human-chimp comparison. Only 9 of the 31 human snRNA preserve synteny with adjacent genes in the mouse genome, while 22680 annotated human genes give rise to 21480 adjacent pairs that have adjacent homologs in the mouse. Furthermore, only a single pair is conserved between human and opossum and no syntenic conservation can be traced back further in evolutionary history, while large syntenic blocks are conserved across chordata [202]. Including the pseudogenes increases the numbers of conserved pairs to 499 of 1609. Again most of these (453) are human/chimp pairs. The data clearly show that snRNA locations are not syntenically conserved, i.e., snRNA behave like mobile elements in their genomic context.

### 3.2.7 Pseudogenes

As mentioned above, snRNAs are frequently the founders of families of pseudogenes. This is a property that they share with most other small RNA classes such as 7SL RNA, Y RNA, tRNAs etc. Such families of pseudogenes are easily recognized as a by-product of *blast*-based homology searches as a large set of hits with intermediate *E*-values. Fig. 3.9 summarizes such data, more details are provided in the electronic supplement.

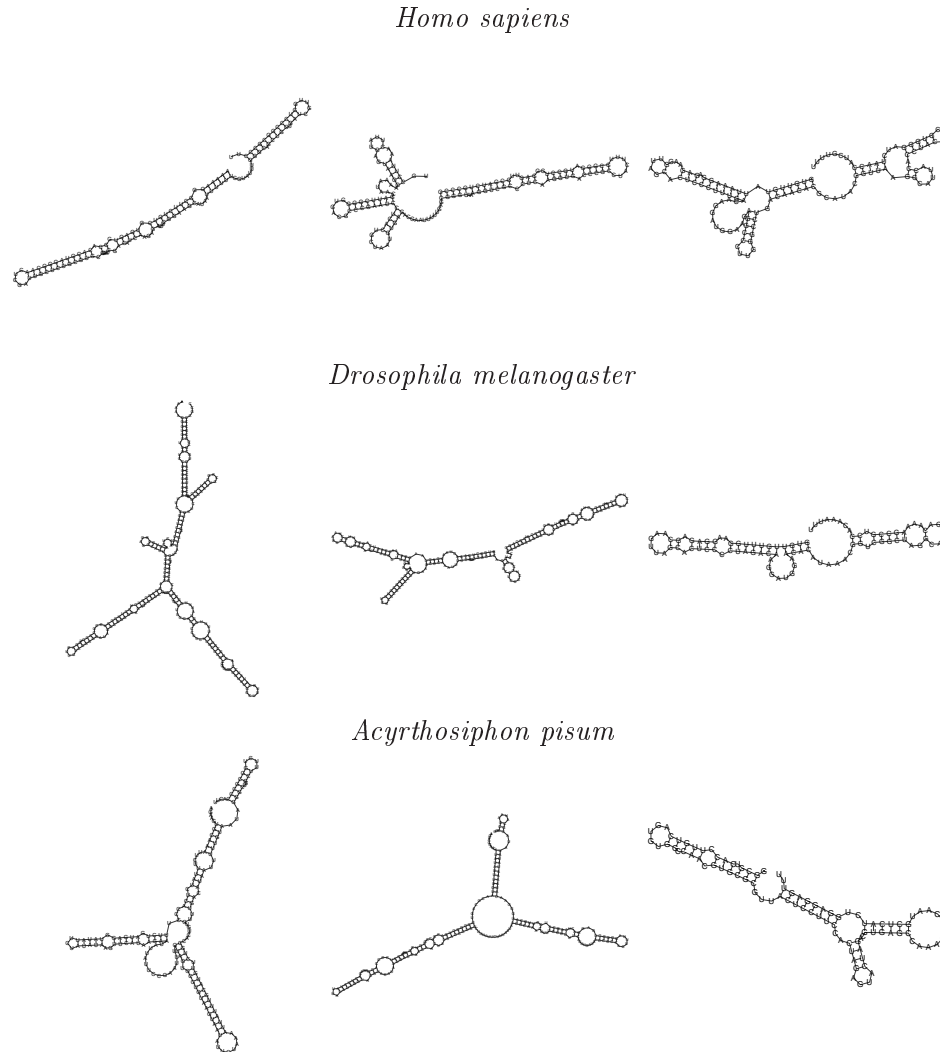


Figure 3.8: Secondary structures of U11 (left), U12 (center), U6atac (right) in *Acyrthosiphon pisum*, *Drosophila melanogaster* and *Homo sapiens*. Drosophilids derived far from all other minor spliceosome structures (e.g. human). Moreover, *Acyrthosiphon pisum* built an autonomous structure group for all minor snRNAs.

Spliceosomal snRNA pseudogenes families are very unevenly distributed across distinct phylogenetic groups and have clearly arisen in independent burst multiple times across animal evolution. Within deuterostomes, almost all sequenced genomes, with the notable exception of teleosts and chicken, contain at least one large family of snRNA-derived pseudogenes.

The genus *Caenorhabditis* shows no pseudogenes, whereas other nematods show nearly such a high number of pseudogenes as primates. Annelids, molluscs and

plathelminths behave similarly. The *Trichoplax adhaerens* genome, on the other hand, contains a single copy of each of the nine spliceosomal snRNAs.

### 3.2.8 Discussion

We have reported here on a comprehensive computational survey of spliceosomal snRNA in all currently available metazoan genomes. We thus provide a comparable and nearly complete collection of animal snRNA sequences. The dense taxon sampling allowed us to verify homology of candidate sequences. Both the major and the minor spliceosome are present in almost all metazoan clades, nematodes (and possibly *Oikopleura*) being the only notable exception. For many of the metazoan families we report here the first evidence on their spliceosomal RNAs.

Using restrictive filtering of the candidates by both secondary structure and canonical promoter structure leaves us with a high-quality data set that was then used to construct secondary structure models. This is useful in particular for the snRNAs of the minor spliceosome for which very few sequences are reported in databases; indeed, the **Rfam 7.0** [86] lists only the U11 and U12 families with a meager set of seed sequences from few model organisms. The sequence and secondary structure data compiled in this study provide a substantially improved databases and set the stage for systematic searches of even more distant homologs.

The analysis of the genomic distribution of snRNAs reveals that discernible paralogs are not uncommon within genera or families. However, no dramatically different paralogs have been found. Spliceosomal snRNAs are prone to spawning large pseudogene families, which arose independently in many species. They behave like mobile genetic elements in that they barely appear in syntenic positions as measured by their flanking genes. While in some genomes snRNAs appear in tandem and/or associated with 5S rRNA genes, these clusters are not conserved over longer evolutionary time-scales. Taken together, the data are consistent with a dominating duplication-deletion mechanism of concerted evolution for the genomic evolution and proliferation of snRNA. This behavior of snRNAs is similar in particular to tRNAs, albeit the copy number of snRNAs is typically much smaller. Recent studies have demonstrated that snoRNAs behave like mobile genetic elements that spread via retroposition [203, 204]. Their mode of expression from spliced-out introns, however, restricts the functional copies predominantly to introns of the same host gene, with only occasional translocations to different carriers, see e.g. [26]. Spliceosomal RNAs, in contrast, appear to freely spread across the genome when they appear as multicopy genes.



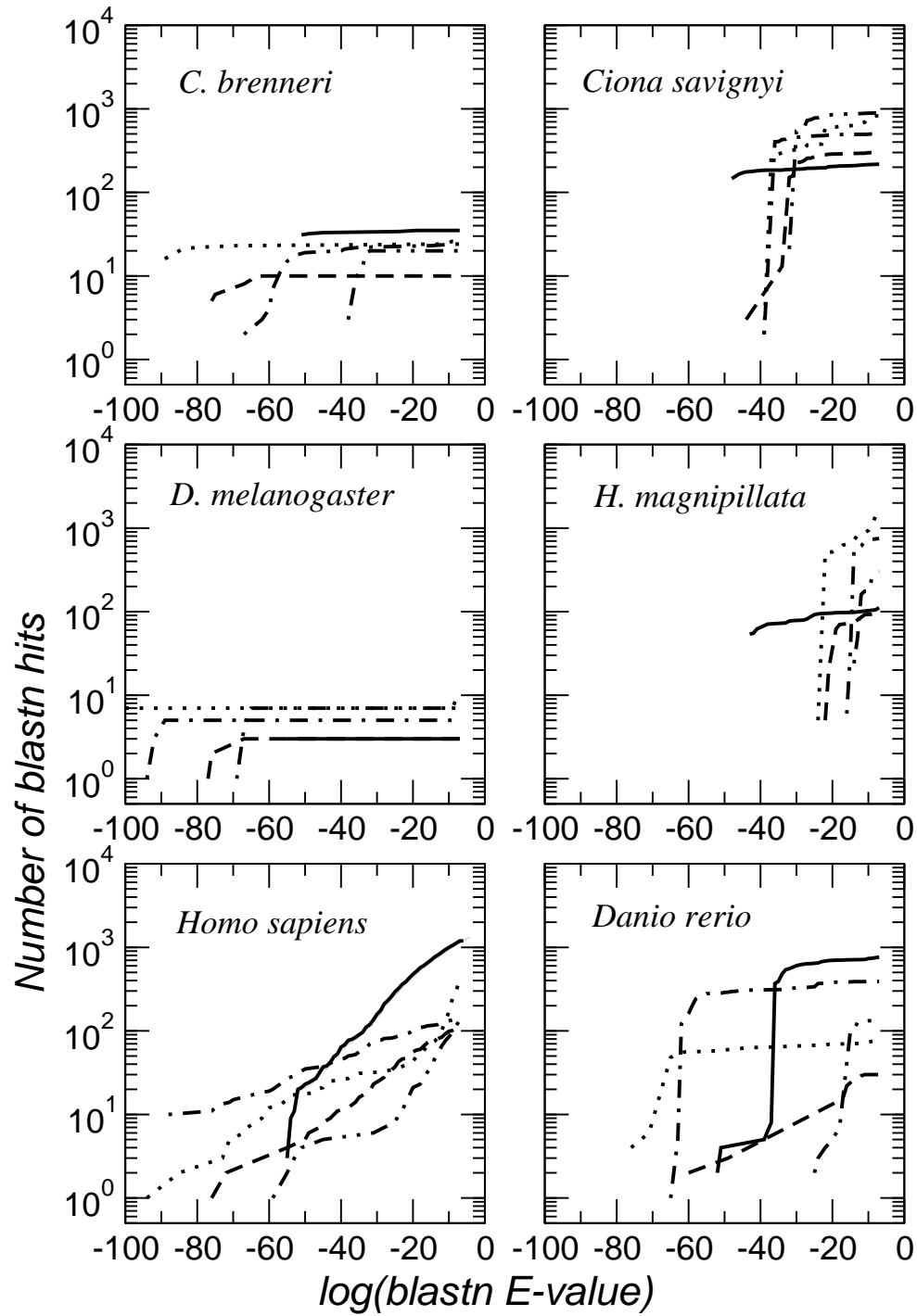


Figure 3.9: Double-logarithmic plot of the number of blast hits versus cut-off  $E$ -value for 6 different genomes. Pseudogene families appear as a slowly increasing curve, while genes without a “cloud” of pseudogene have a flat distribution for  $E < 10^{-5}$ . Dashdotted line – U1; dotted line – U2; dashed line – U4; dashdotdotted line – U5; continuous line – U6.

### 3.3 *trans*-splicing with splice leader RNAs

The standard free energy parameters used in RNA folding algorithms are measured at 37°C and high salt concentrations [70]. Since enthalpy parameters are available separately [205], the parametrization of folding algorithms can be adjusted to physiologically more meaningful temperature values, and option that is provided by the commonly used secondary structure prediction software such as `mfold` [206] and the `Vienna RNA Package` [107]. Systematic effects of temperature on the outcome of secondary structure predictions were already discussed in [207]. It may come as a surprise, therefore, that most computations studies into conserved secondary structures are performed with the default parameters. Here, we use the highly divergent structure predictions for spliced leader (SL) RNAs that can be found in the literature as a the subject of a case study.

Many eukaryotes have two fundamental modes of spliceosomal splicing. *Cis*-splicing is the excision of introns. In *trans*-splicing, on the other hand, a short *leader* sequence is transferred to the 5' end of a (typically protein coding) mRNA, which is usually processed from a polycistronic transcript. This leader contains the 5' hyper-modified cap structure necessary for translational initiation [208]. In all cases described so far, the leader sequence is derived from small non-coding RNAs, the SL RNAs [147, 209]. These molecules share a common organization, Fig. 3.10, and functionality. They provide a short exonic leader sequence with a 5' hyper-modified cap and they play an active role in the spliceosome-catalyzed processing by virtue of binding to the Sm protein. Although SL RNAs are found in wide range of eukaryotic phyla, they are conspicuously absent in many major clades suggesting a complex evolutionary history.

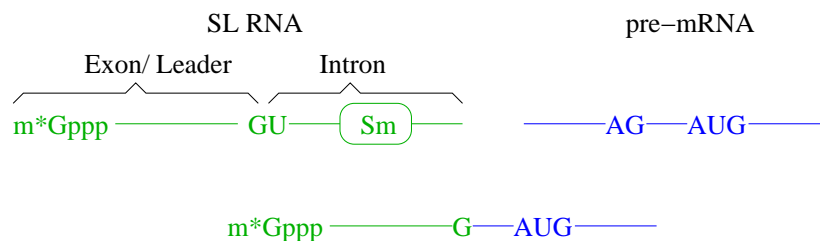


Figure 3.10: Schematic drawing of a typical SL RNA

The first SL RNAs were discovered in kinetoplastids a quarter of a century ago [210, 211]. A few years later, related RNAs were found in *Euglena gracilis* [212]. The first metazoan examples were the nematode *Caenorhabditis elegans* [213] and in platyhelminth *Schistosoma mansoni* [124]. Many more examples were soon found in related species (see Tab. C.1), but it took until the turn of the millennium

before SL RNAs were discovered in additional metazoan phyla (cnidaria [214], tunicates [215, 216], rotifera [217]), and in dinoflagellates [148, 218]. In some species, multiple divergent copies of the SL RNA have been reported [149, 212, 217, 219], and some groups of species harbour two or more clearly distinct types of SL RNAs [149, 214, 220]. Of these SL1 and SL2 are distinguished also in the Rfam [86]. On the other hand, several model organism do not seem to utilize *trans*-splicing: despite substantial efforts, no evidence for *trans*-splicing could be gathered for *Drosophila melanogaster*, *Saccharomyces cerevisiae*, and *Arabidopsis thaliana* [147, 209], and no evidence for *trans*-splicing has been reported in vertebrates despite the availability of extensive transcriptomics data.

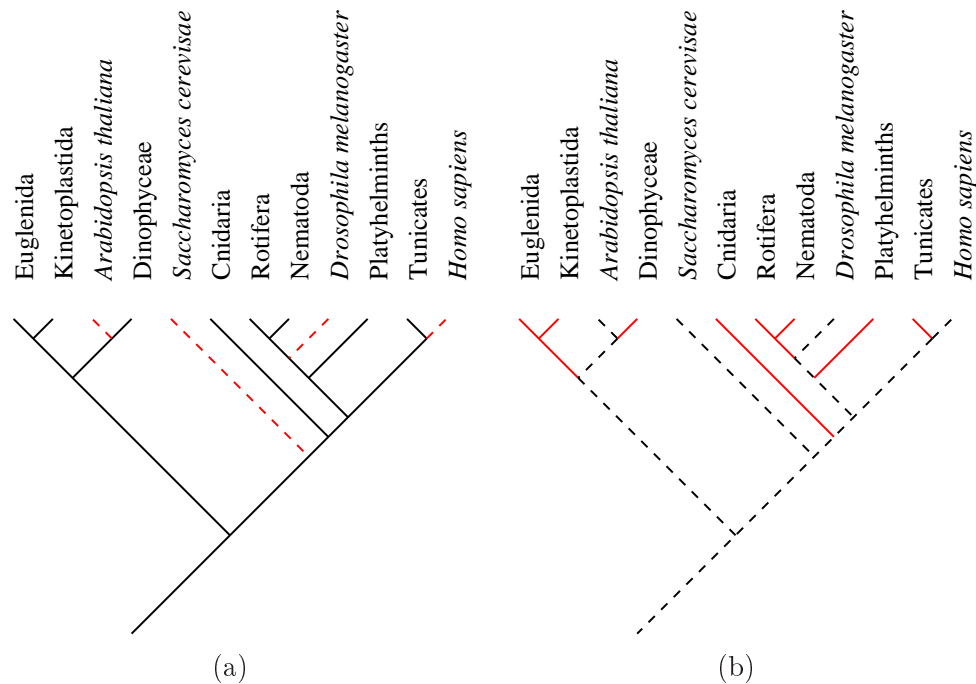


Figure 3.11: Evolution of SL RNAs ([209, 214]). (a) SL RNAs have a common eukaryotic ancestor. Solid black line – Presence of SL RNAs since last common ancestor; Dashed red line – Possible loss of SL RNAs. (b) SL RNAs derived independently in seven clades, which we believe to be improbable. Dashed black line – Absence of SL RNAs; Solid red line – gain of SL RNAs.

The unexpectedly scattered distribution of SL *trans*-splicing across the phylogeny of Eukarya has prompted interest in the evolution of *trans*-splicing already two decades ago. To-date the two major competing hypotheses (reviewed in [147, 209, 214]) still remain unresolved:

1. SL RNAs have a common origin (Fig. 3.11a). Based on the fact, that all SL RNAs have the same function of resolving Pol II transcripts to monocistronic mRNAs.

2. SL RNAs arose on multiple occasions (Fig. 3.11b).

The first hypothesis is supported primarily by the functional and mechanistic similarities of SL-*trans*-splicing, while the failure to detect sequence homology between SL RNAs from different phyla and the apparent disparity of SL RNA secondary structure are quoted in support of the second hypothesis, see e.g. [221].

Here, we re-evaluate the secondary structure models from the published literature. To this end, we not only consider the minimum free energy structure but also suboptimal structures with comparable energies. Furthermore, we include the temperatures at which the organisms in question thrive into our analysis.

### 3.3.1 Re-evaluation of secondary structures

Tab. C.1 (Supplemental Material) summarizes the published secondary structures. Together with the unconstrained predictions at ambient temperatures, they fall into 10 structural classes, which are compiled in the header of Tab. 3.4. Secondary structures for each SL were computed both without constraints and with constraints corresponding to these 10 structural classes. The temperature parameter was always adjusted to each organism's optimal ambient temperature. The optimal structures in the three structural classes with the lowest energies are listed in Table C.1, additional structural alternatives can be found in Supplemental Material Website<sup>2</sup>.

Many of the previously published structures were obtained using `mfold` with standard parameters (i.e., a temperature of  $T = 37^{\circ}\text{C}$ ), which is lethal for most of the organisms in question, in particular almost all the unicellular ones [222]. In several cases, which we will discuss in detail in the next paragraphs, our analysis deviates drastically from the published data. Small corrections and differences at the sequence level are briefly mentioned in the Methods section.

**Dinoflagellata.** The *K. brevis* SL RNA was reported in [148] with a donor splice site that is not consistent with the EST data from the same work. Moreover, the reported structure is energetically unfavourable at all temperatures, with and without additional constraints that forces the Sm binding site to be unpaired. The structural analysis makes it likely, furthermore, that this SL RNA is 24 nt shorter than the published sequence, consistent with proposed  $A_5$  termination signal [148]. Completely different models of dinoflagellate SL RNAs are proposed in [218]. For both *K. micrum* and *P. piscicida* the Sm binding site is shown upstream of the

---

<sup>2</sup>[www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-009](http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-009)

splice site instead of downstream. If correct, this would indicate major differences in the organization of the *trans*-splicing machinery. However, the Sm proteins and U1 snRNA seem to be conserved in dinoflagellates [223]. The *P. minimum* sequence is reported without Sm binding site. A simple sequence alignment of this sequence with the SL RNA of *K. brevis* shows that the SL RNA is clearly conserved among these alveolates. Thus, either the published *K. brevis* is much too long, or the dinoflagellate SL RNA sequences from the work of [218] are truncated. Adding 4 nt on the 5' side and 79 nt on the 3' side to the published sequence from the genomic DNA available from GenBank (*EF143079.1*), we easily obtained structure models that conform to the common organization of SL RNAs of other phyla. For *P. piscicida* (*EF143082.1*) and *P. minimum* (*EF143084.1*) flanking regions were not available, consequently we re-evaluate this sequences only with previously published sequences. Due to their short sequences they are not listed in Tab. 3.4 and Tab. C.1. Detailed alignments can be found at the Supplemental Material Website.

**Euglenida.** Stem IV of the published *Entosiphon* SL RNA [178] has a positive folding energy under all parameter settings, strongly suggesting that this substructure is not formed. For *E. sulcatum* we identified a possible alternative Sm binding site, which would suggest that this SL RNA would be 16 nt shorter, see Tab. C.1. The extended sequence then folds similarly to the known secondary structures in other phyla.

**Sm Binding Motif.** In most of the previous publications, structures were computed with the constraint of an externally unpaired Sm-binding site. However, most of the sequences can fold in a hairpin structure in which (most of) the Sm-binding site is located in an accessible loop, as reported e.g. for *Euglena* [212] and *Hydra* [214]. We do not know, whether SL RNAs occur permanently *in vivo* with Sm-proteins binding to their Sm-binding site. Therefore we report structures with and without a constraint on the Sm-binding site in Tab. C.1.

### 3.3.2 Phylum specific alignments

The SL RNAs are fairly well conserved as sequence level within each of the 7 phyla. Alignments can be found at the Supplement Material Website. An exception is the *T. brucei* SL RNA, which has a long insertion upstream of the Sm-binding site.

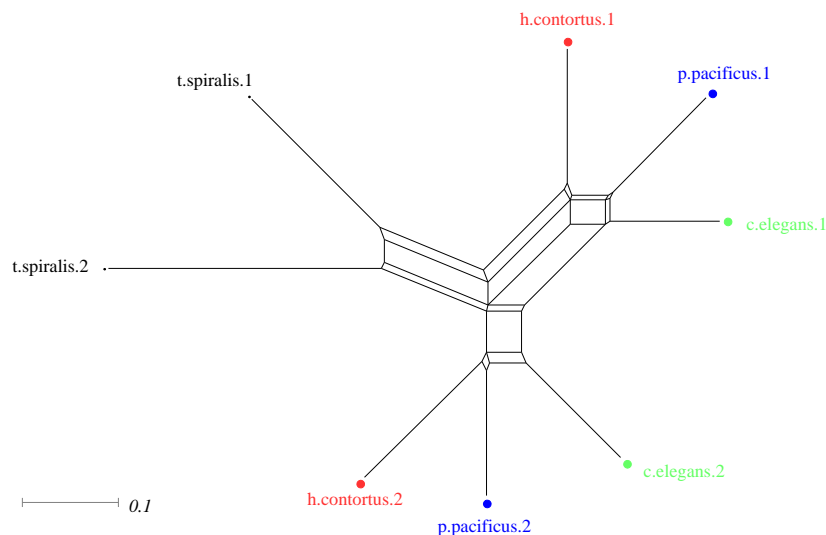


Figure 3.12: Neighbor-net analysis confirms that nematode SL1 and SL2 RNAs form distinct paralog groups, with the possible exception of *Trichinella*.

```
# STOCKHOLM 1.0
c.elegans.1      GGTTT.....AA.TTACCCAAGTTT....GAG
p.pacificus1    GGTTT.....AA.TTACCCAAGTTT....GAG
h.contortus1    GGTTT.....AA.TTACCCAAGTTT....GAG
c.elegans2      GGTTT..TA.....ACCCA.GTTACT.CAAG
h.contortus2    GGTTT..TA.....ACCCA.GTAICT.CAAG
p.pacificus2    GGTTTAT.....ACCCA.GTAICT.CAAG
a.ricciael      GGCTTATTACACTTA.CCAAG.....AG
philodina       GGCTTATTACACTTA.CCAAG.....AG
t.spiralis      GG..TAT.....TTA.CCA.G.ATCTAAAAG
//
```

Figure 3.13: A sequence feature shared between rotifera and nematoda SL RNAs.

The two paralogous SL RNA in *Hydra* are probably recent duplicates, sharing a highly conserved 17 nt long block in the 3' region. Neither paralog has a recognizable homolog in *Nematostella vectensis*.

Most nematodes have multiple SL RNAs. Within Rhabditina, there are clearly discernible paralog groups SL1 and SL2 [224]. The SL RNAs of *Trichinella* (clade Dorylaimia) [149] do not fit well into this scheme, however (Fig. 3.12).

A sequence-based alignment between two adjacent phyla shows no conserved regions, as expected. However, aligning just exonic parts reveals some similarities between rotifers and nematods, Fig. 3.13.

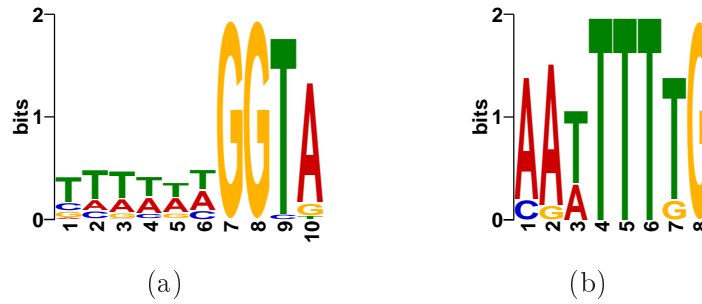


Figure 3.14: Consensus of (a) Donor splice site and (b) Sm binding site of SL RNA for all seven clades.

### 3.3.3 Ubiquitous Sequence Features

Not surprisingly, the donor splice-site,  $G|GU$ , and the U-rich Sm-binding site are well conserved within each phylum and can be detected easily using *Meme* [225] in the complete data set, Fig. 3.14). No other sequence similarities have been reported previously, and a sequence alignment does not pick up any additional motifs. We therefore examined all sequences for common properties, such as positions of pyrimidines or strong pairing bases (T and C). A fully comparison of the IUPAC-code consensus sequences of the 7 phyla identifies several weak sequence features, Fig. 3.15. However, we compared these results exemplarily for *Schistosoma mansoni* with 300 shuffled sequences via *ushuffle* dinucleotide shuffling [226]. They show hardly any similarities with functional SL RNAs. Details can be found at the Supplement Material Website.

- Upstream of SL RNA stem I, there is an A/C-rich region containing an occasional T. We denote this region by  $H^*$  since  $H = \{A, C, T\}$ . This G-poor sequence is conform to the proposed initiator sequences UNCU in euglenoids [221],  $YA^{+1}NU/AYYY$  generally observed in metazoans [227],  $YYHBYA^{+1}ACU$  described for trypanosomes [228] and  $CA^{+1}AUCUC$  in *K. brevis* [148].
- The loop and 3' part of the first hairpin are depleted in C. This may be associated with constraints associated with the splice-site and/or the binding affinity between SL RNA and mRNA.
- The 5' part of the first hairpin (most of the exon) shows a lack of G, explained by the pairing with the  $D = \{A, G, T\}$  region mentioned above. This H-block is less well-conserved than the 3' D-block due to the possibility of G-U pairing. A succession of A-U pairs followed by U-A pairs in the outer part of stem I (just “below” the splice site) was reported in [221] for euglenoids. In several other clades such a stringent pattern is not visible, however.

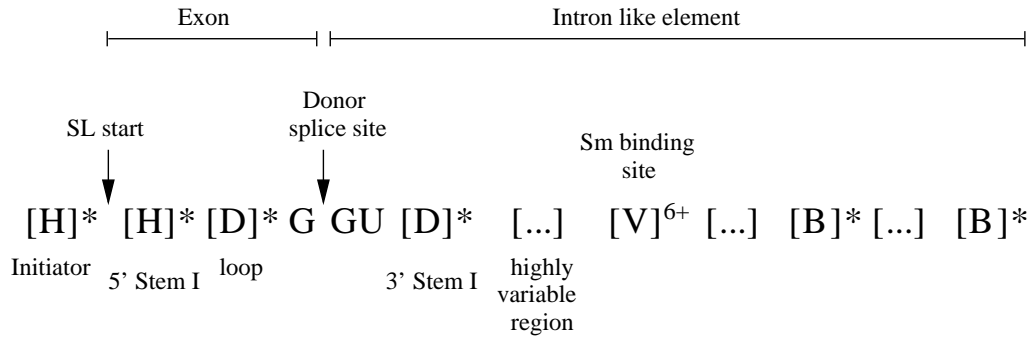


Figure 3.15: Common sequence features of SL-RNAs: Initiator lacking of G, a variable length of Stem I, which 5' part consists mainly of A,C,T followed by a loop, which for SL RNA- $\alpha$  contains mainly T and for SL RNA- $\beta$  A,G. The Donor Splice site is located downstream of the loop in the hairpin with the highly conserved GGU-motif. Stem I continuous nearly without any C. After a highly variable region, the known Sm binding site with the common motif AATTTTGG and a possible unpaired region a last part often structured as a hairpin shows no A to be paired. [X]\* – A cluster of X.

- The loop of stem I shows a clear minority of Y= {C,T} and S= {C,G}. Depending on the A/U-ratio, two subtypes can be distinguished (see below).
- The donor-splicing site is highly conserved with the sequence G|GU. It is always located downstream of loop I.
- The region between stem I and the Sm-binding site is highly variable not only between but also within each phylum.
- The SM binding site consists of a highly conserved D-region. The common pattern is AAUUUUUGG, with the sole exception of *Oikopleura dioica*.
- Stems downstream of the Sm binding site show a highly conserved C,G-rich B= {C,G,U} stem. The loop in contrast is A-rich.

Taken together, we find recognizable sequence constraints covering almost the entire SL RNA gene.

The loop region of stem I clearly distinguishes two sub-types of SL RNAs. In class  $\alpha$ , the loop consists mostly of Us, while the loop in class  $\beta$  is essentially free of Us. In most metazoans with more than one SL RNA (e.g. *C. elegans*, *H. contortus*, *P. pacificus*, *T. spiralis*, and *S. mediterranea*) both types are present. The two rotifers *A. ricciae* and *Philodina sp.*, as well as *C. intestinalis* have only type  $\beta$ , while otherwise type  $\alpha$  appears to be prevalent. In the cnidarian *Hydra* the classification remains ambiguous.



### 3.3.4 Secondary Structure Analysis












Tab. 3.4 summarizes the folding behavior of SL RNAs when constrained to conform to each of the ten structural classes. The most prominent observations can be summarized as follows:

1. Most euglenid SL RNA folds into 4 hairpins. However, other phyla such as cnidaria, rotifers, or tunicates (due to the shorter sequence) never fold into such a structure.
2. The SL RNAs of all species can fold into a single hairpin including both the donor splice site and a Sm binding site. Except for *Oikopleura*, however, this structure is never energetically preferred.
3. Stem I upstream of the Sm-binding site is highly conserved. All SL RNAs except that of *Oikopleura* can form this structure. In most cases, this structure is also thermodynamically highly favoured, (see Tab. 3.4, 3rd and 4th column).
4. For the Sm-binding site either a completely unbound external structure or a mostly unbound location within a hairpin loop were discussed. Both structural models are plausible, the hairpin variant is always energetically favorable.
5. In most SL RNAs, stem I folds can attain two different hairpins, see below (Fig. 3.16).

Interestingly, the sequence underlying stem I can form two alternative structures, Fig 3.16. Since the sequences are highly divergent between phyla, it is very unlikely to observe the same pair of structural alternatives throughout the entire data set. We therefore conclude that the conformational change between the two alternatives is likely required for SL RNA function. The published *Ciona intestinalis* sequence had to be extended by 16nt to obtain the same result as all other SL RNAs. This does not conflict with the work of [215].

Secondary structure alignments show common features within each phylum, as well as for rotifers and nematods together. Weak signals for conserved structure elements were obtained by aligning all protostomia together, Fig. 3.17. The corresponding alignments are compiled at the Supplement Material Website.

Table 3.4: Secondary structures, their  $MFE_{min}$  for all constraint folded  $MFE_{cons}$  and their ratio of constraint folding to minimum MFE ( $MFE_{cons}/MFE_{min}$ ).

Organism	T	$MFE_{min}$											
<i>E. gracilis</i>	29	-35.02	-35.02 1.00	-23.38 0.67	-20.51 0.59	-6.22 0.18	-22.52 0.64	-27.15 0.78	-21.54 0.62	-31.50 0.90	-29.27 0.84	-21.94 0.63	
<i>P. curvicauda</i>	22	-41.51	-41.51 1.00	-34.19 0.82	-24.35 0.59	-11.65 0.28	-32.71 0.79	-32.86 0.79	-23.18 0.56	-34.98 0.84	-35.13 0.85	-25.45 0.61	
<i>C. acus</i>	22	-54.10	-54.10 1.00	-40.10 0.74	-27.01 0.50	-13.84 0.26	-46.80 0.87	-42.92 0.79	-33.50 0.62	-49.63 0.92	-40.73 0.75	-28.97 0.54	
<i>R. costata</i>	23	-42.69	-42.69 1.00	-32.29 0.76	-17.57 0.41	-6.61 0.15	-27.23 0.64	-24.78 0.58	-14.38 0.34	-32.01 0.75	-30.09 0.70	-19.69 0.46	
<i>M. pellucidum</i>	23	-35.73	-35.73 1.00	-28.32 0.79	-15.28 0.43	-7.66 0.21	-	-22.26 0.62	-20.05 0.56	-23.67 0.66	-27.55 0.77	-17.75 0.50	
<i>E. sulcatum</i> long	25	-35.98	-	-	-	-	-35.98 1.00	-33.35 0.93	-14.13 0.39	-	-	-	
* <i>E. sulcatum</i> short	25	-35.07	-16.47 0.47	-19.29 0.55	-11.84 0.34	-7.15 0.20	-	-35.07 1.00	-33.86 0.97	-	-33.66 0.96	-27.13 0.77	
<i>T. cruzi</i>	23	-44.05	-34.93 0.79	-38.23 0.87	-32.84 0.75	-24.96 0.57	-33.17 0.75	-38.58 0.88	-42.24 0.96	-34.75 0.79	-44.05 1.00	-43.74 0.99	
<i>T. vivax</i>	23	-53.64	-46.90 0.87	-53.64 1.00	-43.94 0.82	-22.16 0.41	-	-	-42.79 0.80	-48.25 0.90	-51.40 0.96	-48.61 0.91	
<i>T. brucei</i>	23	-75.05	-60.45 0.81	-75.05 1.00	-34.05 0.45	-15.95 0.21	-48.89 0.65	-45.54 0.61	-44.64 0.59	-55.73 0.74	-71.99 0.96	-43.42 0.58	
<i>L. collosoma</i>	28	-27.43	-	-27.43 1.00	-21.07 0.77	-15.60 0.57	-	-23.54 0.86	-22.69 0.83	-	-21.85 0.80	-22.53 0.82	
<i>C. fasciculata</i>	20	-30.83	-25.40 0.82	-30.83 1.00	-29.69 0.96	-21.44 0.70	-21.20 0.69	-23.13 0.75	-30.10 0.98	-	-30.76 1.00	-29.62 0.96	
<i>L. enriettii</i>	36	-24.72	-12.68 0.51	-17.26 0.70	-11.00 0.44	-10.36 0.42	-	-	-24.72 1.00	-	-	-	
<i>K. brevis</i>	20	-45.34	-38.06 0.84	-45.34 1.00	-40.31 0.89	-22.47 0.50	-	-	-33.03 0.73	-	-21.04 0.46	-22.77 0.50	
* <i>K. brevis</i> 3'SM	20	-50.23	-41.21 0.82	-39.88 0.79	-22.74 0.45	-19.30 0.38	-45.41 0.90	-50.23 1.00	-39.31 0.78	-28.14 0.56	-26.02 0.52	-16.17 0.32	
<i>K. micrum</i> long	20	-54.66	-	-53.29 0.97	-53.56 0.98	-26.06 0.48	-	-23.93 0.44	-30.62 0.56	-	-47.97 0.88	-54.66 1.00	
<i>Hydra-A</i>	18	-18.06	-	-10.96 0.61	-9.26 0.51	-6.60 0.37	-	-	-18.06 1.00	-	-12.63 0.70	-14.35 0.79	
<i>Hydra-B1</i>	18	-18.80	-	-	-18.80 1.00	-7.63 0.41	-	-	-12.19 0.65	-	-	-18.03 0.96	
<i>Hydra-B2</i>	18	-18.03	-	-18.03 1.00	-11.79 0.65	-8.20 0.45	-	-15.86 0.88	-17.23 0.96	-	-14.70 0.82	-10.85 0.60	
<i>Hydra-B3</i>	18	-24.27	-	-11.62 0.48	-8.82 0.36	-8.20 0.34	-	-22.17 0.91	-20.88 0.86	-	-24.27 1.00	-14.93 0.62	
<i>A. ricciae</i>	24	-46.48	-	-	-36.14 0.78	-6.76 0.15	-	-	-	-	-	-46.48 1.00	
<i>Philodina sp.</i>	20	-49.82	-	-	-42.28 0.85	-7.77 0.16	-	-	-	-	-	-49.82 1.00	
<i>C. elegans 1</i>	20	-44.73	-30.04 0.67	-43.54 0.97	-28.75 0.64	-12.27 0.27	-28.25 0.63	-33.80 0.76	-12.27 0.27	-	-44.73 1.00	-31.67 0.71	
<i>C. elegans 2</i>	20	-58.38	-43.52 0.75	-57.14 0.98	-47.88 0.82	-23.35 0.40	-33.79 0.58	-32.85 0.56	-34.06 0.58	-52.20 0.89	-58.38 1.00	-49.20 0.84	
<i>T. spiralis 1</i>	36	-30.15	-20.92 0.69	-27.80 0.92	-21.08 0.70	-8.34 0.28	-	-17.41 0.58	-19.06 0.63	-	-30.15 1.00	-20.71 0.69	
<i>T. spiralis 2</i>	36	-19.51	-	-19.51 1.00	-8.40 0.43	-7.91 0.41	-	-17.78 0.91	-11.74 0.60	-	-18.27 0.94	-11.81 0.61	
<i>P. pacificus 1</i>	20	-54.23	-19.16 0.35	-50.76 0.94	-36.91 0.68	-13.79 0.25	-38.17 0.70	-36.34 0.67	-37.31 0.69	-54.03 1.00	-54.23 1.00	-37.30 0.69	
<i>P. pacificus 2</i>	20	-51.68	-48.85 0.95	-48.98 0.95	-39.48 0.76	-19.94 0.39	-29.28 0.57	-32.14 0.62	-38.51 0.75	-42.26 0.82	-51.68 1.00	-37.19 0.72	
<i>H. contortus 2</i>	25	-41.15	-33.91 0.82	-39.91 0.97	-35.12 0.85	-17.74 0.43	-17.52 0.43	-17.81 0.43	-41.15 1.00	-34.69 0.84	-39.26 0.95	-40.79 0.99	
<i>S. mansoni</i>	28	-31.43	-	-23.53 0.75	-26.65 0.85	-21.63 0.69	-	-26.41 0.84	-21.63 0.69	-	-	-31.43 1.00	
<i>F. hepatica</i>	16	-57.02	-32.80 0.58	-53.75 0.94	-47.25 0.83	-32.89 0.58	-	-41.96 0.74	-41.80 0.73	-	-57.02 1.00	-52.15 0.91	
<i>S. mediterranea-1</i>	22	-44.16	-32.07 0.73	-43.87 0.99	-37.88 0.86	-21.96 0.50	-15.99 0.36	-28.24 0.64	-26.57 0.60	-31.91 0.72	-44.16 1.00	-41.19 0.93	
<i>E. multilocularis</i>	35	-42.20	-33.35 0.79	-36.34 0.86	-26.78 0.63	-19.78 0.47	-	-42.20 1.00	-37.35 0.89	-	-37.37 0.89	-28.69 0.68	
<i>C. intestinalis</i>	21	-14.98	-	-	-5.60 0.37	-5.71 0.38	-	-	-14.98 1.00	-	-	-	
<i>O. dioica</i> SM1	20	-14.24	-	-	-	-14.24 1.00	-	-	-	-	-	-	
<i>O. dioica</i> SM2+9	20	-16.35	-	-0.82 0.05	-16.35 1.00	-14.24 0.87	-	-	-	-	-	-	

0    0.6-0.7    0.8-0.9    0.9-1.0

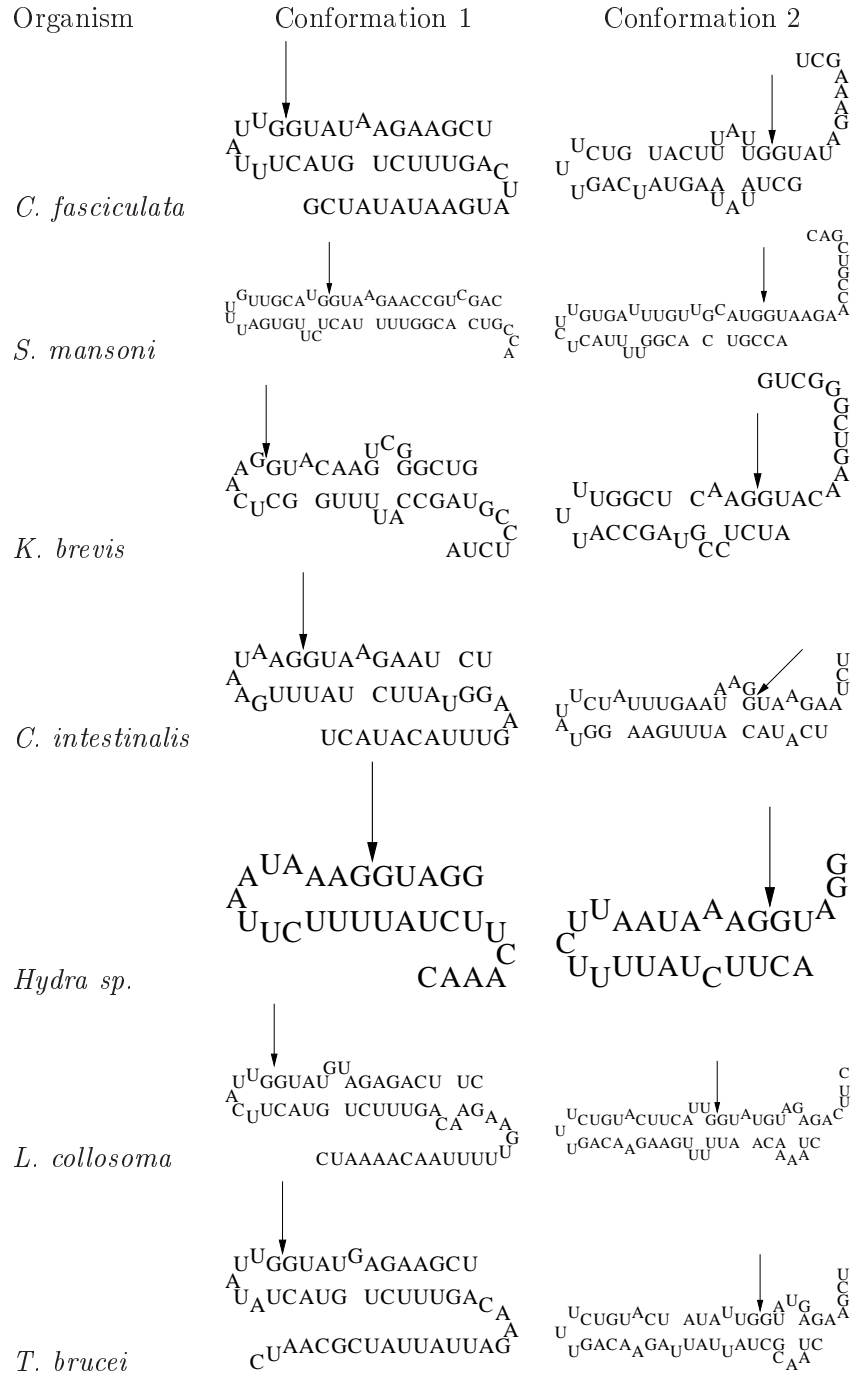


Figure 3.16: Two alternative secondary structures of stem I can be formed in all phyla. In case of *C. intestinalis* 16 nucleotides were added to the 5' end of the published sequence.

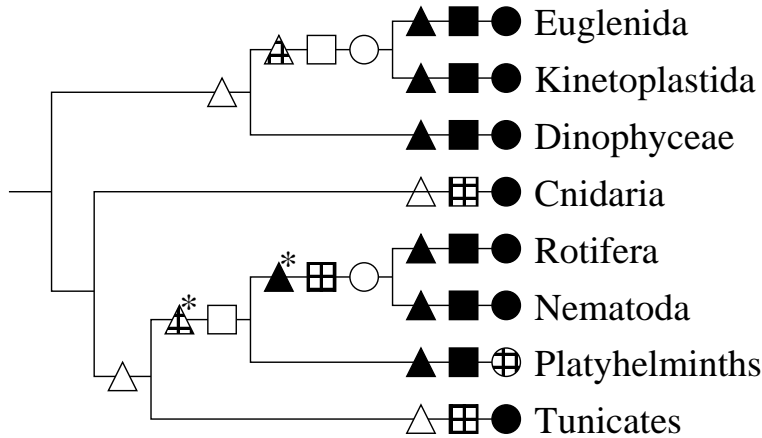


Figure 3.17: Sequence and structure similarities obtained by standard alignment programs *ClustalW* (● – full sequence, ■ – exon only) and the structural alignment tool *locarnate* (▲ – full sequence). Filled symbols indicate similarities, empty symbols indicate that there are no obvious common features; \* – alignable with nematods SL1 only.

### 3.3.5 Discussion

The re-evaluation of the available SL RNA data across Eukarya shows that they share more features than previously recognized. Besides faint sequence similarities, in particular their secondary structures fit a coherent picture when ambient temperature and thermodynamically plausible secondary structures are taken into account. Taken together, the evidence suggests not only a common function but also a common mechanism. SL RNAs share:

1. The relative positioning of the splice-donor site and Sm-binding site is the same.
2. There is a weak but recognizable shared sequence pattern, suggesting common descent or common selection pressures.
3. Structural similarities between SL RNA are much greater than recognized in previous work when natural ambient temperatures are taken into account.
4. Unexpectedly, all SL RNAs share the possibility of two alternative conformations of stem I, suggesting that the structural transition between the two states is involved in SL RNA function.

While logically possible, it seems quite unlikely that *trans*-splicing arose *de novo* several times to give rise to SL RNAs that always share the same sequence and structure constraints. After all, these similarities suggest that they interact in the same way with the same partners. Of course, our analysis does not provide a defini-

tive proof for a common origin of *trans*-splicing, with frequent losses throughout the eukaryotic tree [215]. It put additional burden, however, on the independent innovation hypothesis, which will need to explain why *trans*-splicing originates many times in such a way that it appears to use the same molecular interactions in each case.

The absence of SL *trans*-splicing is plausibly established only in a few organisms such as mammals, fruit-flies, yeast, or *Arabidopsis*. For these, the transcriptome is known sufficiently well to rule out with near certainty that there are any unrelated mRNAs that share a common leader sequence of unclear origin. In general, however, the picture is less clear. In analogy to the basal eukaryotes *Giardia lamblia* [229, 230] and *Trichinella spiralis* [231], and even baker's yeast [232], all of which have functional spliceosomes but only a handful of spliceosomal introns, SL *trans*-splicing might just have escaped detection in some clades.

The lack of obvious sequence similarity among SL RNAs also does not make good argument against homology. Mutation studies in kinetoplastids and nematodes showed that much of the sequence and structure can be disrupted without consequence to function in *trans*-splicing [209, 233]. The observed rapid evolution at the sequence level thus does not come unexpected. This property is shared with several other functionally crucial ncRNAs such as telomerase RNA [234, 235] and 7SK RNA [236, 237], which so far also have been found only in a rather scattered collection of clades.

If one accepts a common origin of SL RNAs, their structural evolution must have followed one of the three scenaria outlined in Fig. 3.18: (1) Most ancestral SL RNAs contain 4 hairpins close to the Sm binding site, Fig. 3.18a, with subsequent simplifications in some clades. (2) Alternatively, the structural complexity may have increased, Fig. 3.18b. (3) A maximum parsimony analysis in which we interpret the hairpins as characters, also points to a structurally fairly simple ancestor, Fig. 3.18c. The inferred ancestral states in these scenaria should help with constructing descriptors for homology search of SL RNAs.

On the methodological level, our case study shows that environmental factors, in this case ambient temperature, is a confounding variable that can have a substantial impact on outcome of computational secondary predictions and thus on the subsequent construction and interpretation of structural consensus models.

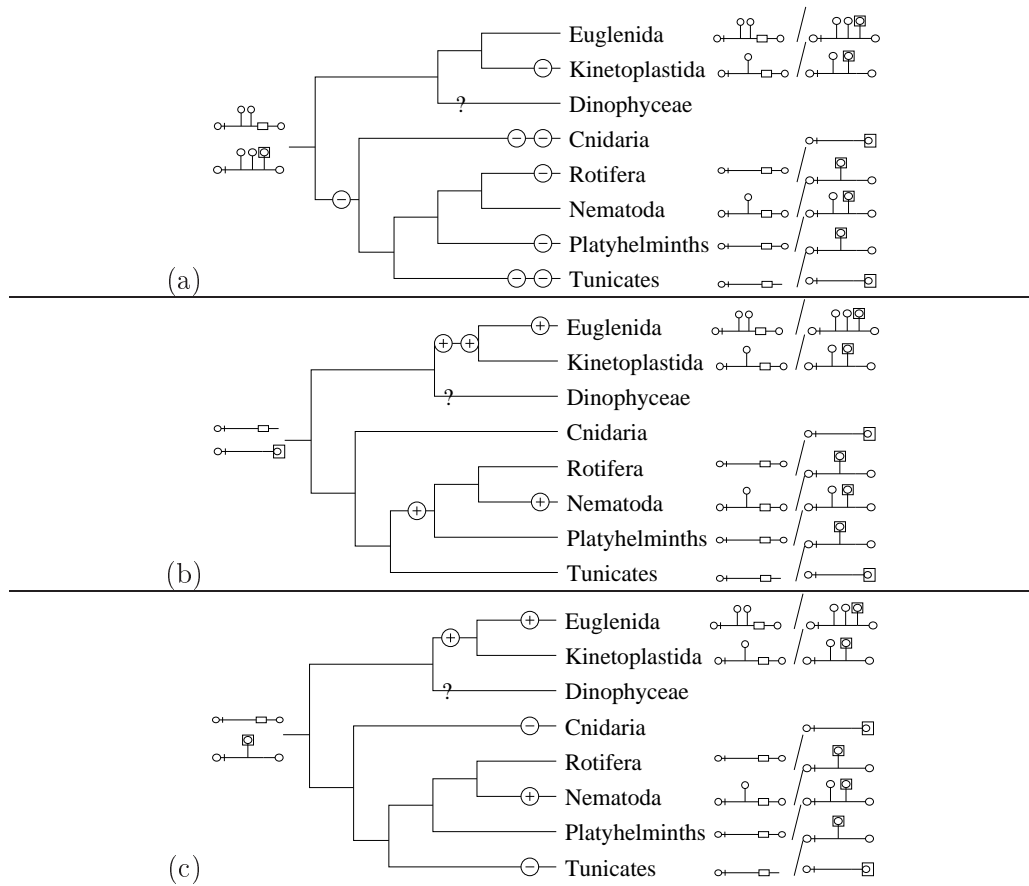


Figure 3.18: Three alternative scenarios for the evolution of SL RNA secondary structure. (a) Ancestral state with 4 hairpins, (b) ancestral state containing only features that are still contained in all present-day SL RNAs, (c) most parsimonious scenario minimizing the number of hairpin gain/loss events.

### 3.4 SmY RNAs

The SmY RNAs are a family of small nuclear RNAs found in Nematoda species. The first SmY RNA was discovered in purified *Ascaris lumbricoides* spliceosome preparations, as well as a second RNA called SmX that is not detectably homologous to SmY [238]. Twelve SmY homologs were identified computationally in *Caenorhabditis elegans*, and ten in *Caenorhabditis briggsae* [239]. Several transcripts from these SmY genes were cloned and sequenced in a systematic survey of small non-coding RNA transcripts in *C. elegans* [240]. SmY RNAs are about 70-90 nucleotides long, with a conserved consensus binding site for the Sm protein, a shared component of spliceosomal snRNPs [238, 239]. In *C. elegans*, SmY RNAs copurify in a complex with Sm, SL75p, and SL26p proteins, while the better-characterized *C. elegans* SL1 trans-splicing snRNA copurifies in a complex with Sm, SL75p and SL21p (a paralog of SL26p) [239]. Loss of function of either SL21p or SL26p individually causes only a weak cold-sensitive phenotype, whereas knockdown of both is lethal, as is a SL75p knockdown. Based on these results, the SmY RNAs are believed to have a function in trans-splicing.

To date, SmY RNAs have been described in *C. elegans*, *C. briggsae*, and *A. lumbricoides*. The range of species possessing SmY RNAs has not been well characterized. Here we report the results of a comprehensive computational characterization of SmY RNA genes in available genome sequences.

#### 3.4.1 Initial SmY sequences.

Thirteen identified SmY sequences are in public DNA databases: *Ascaris lumbricoides* SmY RNA [238] and twelve SmY RNAs from *Caenorhabditis elegans* [239] (Tab. 3.4.1). Full length 5' and 3' ends for all these sequences are experimentally determined [19, 238–240], with three exceptions. SmY-12 was obtained as a partial 3'-truncated sequence [19], and SmY-4 and SmY-7 are predicted from sequence similarity [239].

SmY-2 and SmY-3 have also been identified with slightly different transcript sizes and called C/D small nucleolar RNAs Ce135 (72nt) and Ce96 (98nt) by Zemmann *et al.* [241], who criticized Deng *et al.* [240] for classifying these sequences as “small nuclear RNA like”. Our analysis agrees with MacMorris *et al.* [239] in assigning these as SmY small nuclear RNA homologs, and we have used the transcript sequences deposited by Deng *et al.* [240].

In two cases, we modified a sequence from the accessioned version. We added 8 nt of genomic sequence to the 3' end of the truncated SmY-12 sequence to make

Table 3.5: Previously published SmY RNA sequences. (a) misclassified as small nucleolar RNA; (b) accessions conflict on exact size/sequence, used sequence reported by Deng *et al.*<sup>3</sup>; (c) experimentally determined 78 nt sequence from Deng *et al.*<sup>3</sup> includes 5' G not encoded by WS150 genome; used CESC 77 nt version; (d) reported database sequence is on incorrect strand; (e) accession reports partial 73 nt 3'-truncated sequence, we inferred an additional 3' 8 nt from genomic sequence; CESC: The *C. elegans* Sequencing Consortium. <sup>a</sup> misclassified as small nucleolar RNA; <sup>b</sup> accessions conflict on exact size/sequence, used sequence reported by [240]; <sup>c</sup> experimentally determined 78nt sequence from [240] includes 5' G not encoded by WS150 genome; used CESC 77nt version; <sup>d</sup> reported database sequence is on incorrect strand; <sup>e</sup> accession reports partial 73nt 3'-truncated sequence, we inferred an additional 3' 8nt from genomic sequence. CESC: The *C. elegans* Sequencing Consortium.

Name	Alternative Names	Accession numbers	length (nt)	References
<i>Ascaris lumbricoides</i>				
SmY		U52372.1	72	[238]
<i>Caenorhabditis elegans</i>				
SmY-1	CeN32, C33A12.22	AY948626.1, NR_003443.1	77	[240]
SmY-2	CeN25-1, C33A12.21 <sup>a</sup> , Ce135 <sup>a</sup>	AY948618.1, NR_003442.1 <sup>a</sup> , DQ789540.1 <sup>a</sup>	77 <sup>b</sup>	[240, 241]
SmY-3	CeN25-2, D1086.14 <sup>a</sup> , Ce96 <sup>a</sup>	AY948619.1, NR_003469.1 <sup>a</sup> , DQ789534.1 <sup>a</sup>	82 <sup>b</sup>	[240, 241]
SmY-4	D1086.16	NR_003471.1	81	CESC
SmY-5	CeN25-3, D1086.15	NR_003470.1, AY948620.1	77 <sup>c</sup>	CESC
SmY-6	CeN25-5	AM286190.1	83	[19]
SmY-7	Y73B6BL.46	NR_003463.1	82 <sup>d</sup>	CESC
SmY-8	CeN31, Y45F10B.19	AY948625.1, NR_003460.1	79	[240]
SmY-9	CeN25-7	AM286192.1	77	[19]
SmY-10	CeN112, Y45F10B.20	AY948610.1, NR_003461.1	90	[240]
SmY-11	CeN25-4, Y57G11C.55	AY948621.1, NR_003462.1	78	[240]
SmY-12	CeN25-6	AM286191.1	81 <sup>e</sup>	[19]

it conform to our full-length consensus model. We reversed the orientation of SmY-7, because the accessioned version is in the incorrect (antisense) orientation.

### 3.4.2 Homology searches and a representative seed alignment.

Starting from the sequences in Tab. 3.4.1, we conducted a number of different iterative searches, using a combination of Blast [242] and Infernal 1.0 [243] [<http://infernal.janelia.org>] to identify SmY RNA homologs in a variety of genome sequences. Putative homologs were identified in the following 13 nematode genome sequence assemblies: *Caenorhabditis elegans*, *Caenorhabditis briggsae*, *Caenorhabditis remanei*, *Caenorhabditis japonica*, *Caenorhabditis brenneri*, *Pristionchus pacificus*, *Haemonchus contortus*, *Meloidogyne incognita*, *Meloidogyne hapla*, *Heterodera glycines*, *Brugia malayi*, *Ascaris suum* and *Trichinella spiralis*

68 sequences were selected to be representative of the family. Starting from automated Infernal alignments, a multiple alignment was assembled and manually refined by structure and sequence conservation to form a curated seed alignment suitable for the Rfam database [244]. A Stockholm format text file of this align-



ment is provided in the Supplementary Material (`SmY_seed.stk`).

We used manual comparative sequence analysis to deduce a consensus secondary structure, and also independently predicted a consensus structure using the program `Locarnate` [90]. The two structure predictions largely agree with each other, and with a consensus structure previously published by MacMorris *et al.* [239]. The manual comparative analysis was favored where details differed. Fig. 3.19 shows the predicted consensus secondary structure, together with a summary of the extensive base-pair covariation evidence in the seed alignment that supports it.

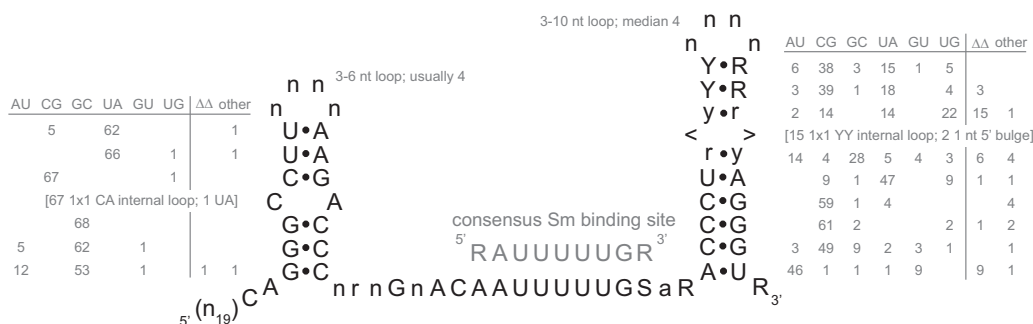


Figure 3.19: Consensus SmY RNA structure, with tables annotating the number of compensatory base pair substitutions, base pair deletions ( $\Delta\Delta$  column), and noncompensatory substitutions (“other” column) observed in the 68 sequences of the representative seed alignment, in support of the structure prediction. The most highly conserved residues are shown as upper case letters in the structure. The sequence at the 5’ end of SmY RNAs is highly variable; the consensus is shown here as  $n_{19}$ , but it varies in both length and sequence in individual SmY sequences.

We did a retrospective analysis to establish the support for each individual sequence’s probable homology to the rest of the family, which confirmed that each sequence is supported by significant ( $< 1 \times 10^{-4}$ ) `Blast` or `Infernal` E-values when searched against phylogenetically independent subsets of the seed alignment (using the `-Z` option of both programs to calculate E-values for an effective search space size of 200 MB), with four exceptions. Four distantly related SmY sequences are predicted in Tylenchid nematode species – two paralogs in *Heterodera glycines*, and one SmY each in the related species *Meloidogyne hapla* and *M. incognita*. The assignment of these sequences as SmY homologs is supported by borderline `Infernal` E-values (0.01-0.001) to more than one `Infernal` model built of other independent SmY sequence subsets, and by the fact that they share the expected pattern of conservation, including base pair covariations consistent with stem 2.

### 3.4.3 Phylogenetic diversity of SmY RNAs.

In the system used by the Rfam RNA database, a consensus **Infernal** statistical model is built from a stable, curated seed alignment, and this consensus model is used to automatically identify and annotate homologs in genome sequences. The seed should be sufficiently representative that this single model identifies all known homologs. We used an **Infernal** model of the 68-sequence seed alignment to search the 13 nematode genomes. This search identifies 155 loci with E-values  $< 0.001$ , and these loci include all the sequences we gathered in our initial iterative searches. An annotated table of all these loci is provided in the Supplementary Material (`SmY_all.tbl`).

Each of these loci was examined in detail. All appear to be plausible SmY homologs based on their overall pattern of conservation. Eight candidates appear to be artifacts of underassembled contigs in draft genomes. We annotated 26 as putative pseudogenes based on significant local deviations from the expected consensus (such as disruption of one of the stems) and/or the lack of an upstream proximal sequence element (PSE), a conserved transcriptional control motif generally found upstream of small nuclear RNAs [103], including SmY RNAs [240]. We annotated the remaining 121 loci as putative SmY RNA genes. Our gene/pseudogene labeling is only a best guess; for non-coding RNAs, it is generally not possible to unambiguously distinguish pseudogenes from genes by computational analysis.

In *C. briggsae*, we assigned 11 SmY genes and 1 pseudogene. Nine of these eleven genes were previously identified and named *cbSmY-1* through *cbSmY-9* [239]; we retained these names, though our analysis revises the predicted 5' and 3' ends of the genes. An additional locus named *cbSmY-10* by MacMorris *et al.* [239] does not appear to us to be an SmY RNA homolog. We detected two additional *C. briggsae* SmY genes, which we named *cbSmY-11* and *cbSmY-12* to be consistent with MacMorris *et al.* [239]. In all other species, we have not assigned gene names, but rather have identified putative SmY loci by their assembly contig name and sequence coordinates.

Fig. 3.20 shows the phylogenetic distribution of SmY RNA genes and pseudogenes. The SmY family has undergone a large paralogous expansion in *Caenorhabditis* and *Pristionchus* species, with copy numbers of 10-26, compared to 1-4 copies in other nematode genomes. Many of these paralogs within a species are more related to each other than to any homolog in another species, suggesting independent paralogous expansions and/or evolutionary turnover (balancing gene loss and paralogous duplication) in these lineages. An extreme case of apparently recent expansion is *Pristionchus*, where most SmY RNAs have 100% identical

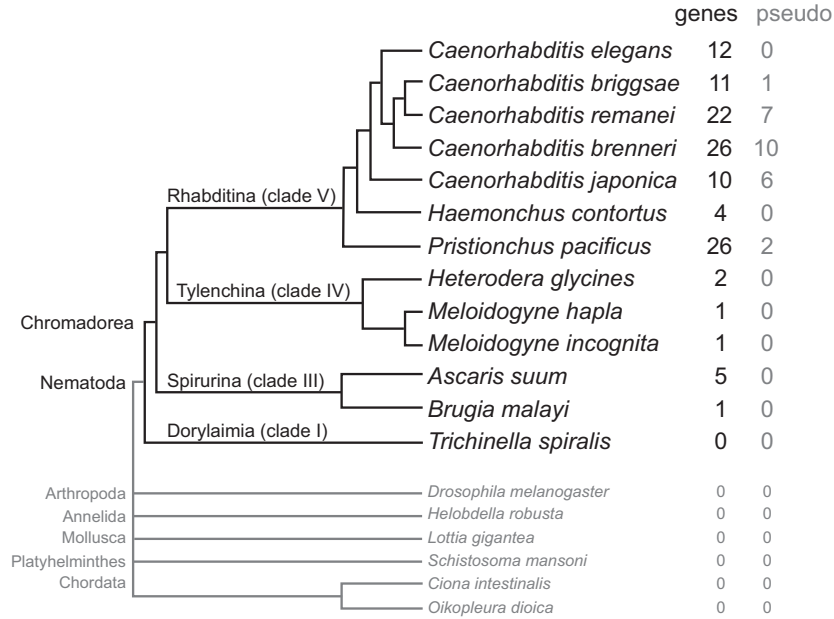


Figure 3.20: Phylogenetic distribution of 147 identified SmY RNA homologs (not shown are another 8 hits that were attributed to underassembled contigs in draft genomes). The species phylogeny is represented as a cladogram (branch lengths are arbitrary), combining the *Caenorhabditis* species phylogeny from Sudhaus and Kiontke [245] with the phylogeny of phylum Nematoda from Blaxter *et al.* and Mitreva *et al.* [246, 247]. To our knowledge there is no sequenced genome yet from a representative of clade II (Enoplia) in the Nematoda. At the root, the relationship of Nematoda to other metazoan phyla is shown as a multifurcation, because most of these relationships remain in some doubt.

paralogs. Only one SmY locus appears to be syntenically conserved among the five *Caenorhabditis* species, with a single copy in *C. japonica* and two copies in the other four species corresponding to *C. elegans* SmY-1 and SmY-2. Rapid turnover of paralogs is a common feature of multicopy structural RNA genes; similar features are seen for tRNA gene families, for example in Lander *et al.* [248].

We also used this model to search for SmY homologs in six non-nematode genomes representing other phyla. We chose genome sequence assemblies of the trematode *Schistosoma mansoni* (TIGR, unversioned, 15 May 2007) [249] and the urochordates *Ciona intestinalis* (JGI, v2.0, Oct 2002) [250] and *Oikopleura dioica* (Genoscope, v3.0, Sept 2006) because these metazoans employ spliced leader trans-splicing [147]. The leech *Helobdella robusta* (JGI, v1.0, July 2007), the snail *Lottia gigantea* (JGI, v1.0, August 2006), and the fruit fly *Drosophila melanogaster* (BDGP, v5.10, July 2008) were chosen as additional representative outgroups to the phylum *Nematoda*. No *Infernal* hit with an E-value better than 0.01 was identified.

### 3.4.4 Discussion

SmY RNA appears to be associated with trans-splicing and spliceosome proteins in *Caenorhabditis elegans* and *Ascaris*, but unlike the trans-spliced leader RNAs SL1 and SL2, it apparently does not contribute a spliced leader sequence to mRNAs. What does SmY RNA do, then? MacMorris *et al.* hypothesized that the role of SmY RNA may be in recycling spliceosome proteins after SL RNAs are consumed in the trans-splicing reaction [239]. They proposed a specific model in *C. elegans* in which the stem-loop 2 sequence of one SmY RNA, SmY-10, base-pairs to SL1 RNAs (which are encoded by a tandem array of about 110 near-identical genes), while stem-loop 2 of the other SmY RNAs base pairs to SL2 RNAs (which are encoded by 18-20 dispersed genes with significant sequence variation). MacMorris' model suggests that the diversification of SmY RNA gene copies (accompanied by sequence variations in stem-loop 2, the more variable stem) may be driven by the diversification of SL2 RNA genes. Although we have not conducted a detailed joint comparative analysis of SL RNAs and SmY RNAs, the results of our SmY RNA survey are broadly in accordance with this model's expectations. SL2 RNAs have as yet only been identified in Rhabditina species, whereas SL1 RNAs have been found throughout the other species of Chromadorea [251]. We find the largest proliferation of paralogous SmY RNA genes in species that have SL2 genes, and smaller numbers of SmY RNAs in species that only have SL1.

We were not able to identify any SmY RNA homologs in the more distantly related Dorylaimid species *Trichinella spiralis*, which has a noncanonical and polymorphic family of SL1-like trans-spliced RNAs [149], but this negative result is inconclusive. The SmY homologs we identified in clade IV Tylenchid nematodes are at the detection limit of the **Infernal** software (and well beyond **Blast**'s limits), so it may be that *Trichinella* SmY homologs exist but are too diverged to be detectable with our methods. The same caution applies to our inability to identify SmY RNAs outside the nematode phylum.

By eye, we do note one suggestive similarity outside nematodes. The SmY RNA structure strongly resembles the proposed structure of a herpesvirus HSUR3 RNA, one of five U snRNAs expressed by herpesvirus saimiri [252]. Like SmY RNA, HSUR3 is a small (75 nt) RNA proposed to have a consensus Sm binding site flanked by two stem-loops of similar length and loop size as the SmY stem-loops, including a C:A mismatch in stem 1. However, an **Infernal** model of SmY does not assign a significant homology score to HSUR3. We note this suggestive visual similarity because the function of the herpesvirus U RNAs remains unknown, and perhaps there is a useful link to the role of the SmY RNAs.

### 3.5 U7 RNA

The U7 snRNA is the smallest polymerase II transcript known to-date, with a length ranging from only 57nt (sea urchin) to 70nt (fruit-flies). Its expression level of only a few hundred copies per cell in mammals is at least three orders of magnitude smaller than the abundance of other snRNAs. It is part of the U7 snRNP, which plays a crucial role in the 3' end processing of histone mRNAs [253]. Replication-dependent histone genes are the only known eukaryotic protein-coding mRNAs that are not polyadenylated ending instead in a conserved stem-loop sequence, see Fig. 3.21 for details and [1] for a recent review.

Beyond metazoan animals, non-polyadenylated histone genes have been described in the algae *Chlamydomonas reinhardtii* and *Volvox carteri* [254], and *Dictyostelium discoideum* has a homolog of the histone RNA hairpin-binding protein HBP/SLBP (*DictyBase DDB0169192*). It appears that replication-dependent histone genes are the only mRNAs that are processed in this way [255].

The 5' region of the U7 snRNA is complementary to the "Histone downstream element" (HDE), located just downstream of the conserved hairpin. The interaction of the U7 snRNP with the HDE is crucial for the correct processing of the histone 3' elements [253]. The 3' part of the U7 is occupied by a modified binding domain for Sm-proteins consisting of a characteristic sequence motif followed by a conserved stem-loop secondary structure motif, see e.g. [256]. U7 snRNA binds five of the seven Sm-proteins that are present in spliceosomal snRNAs, while the D1 and D2 subunits are replaced by the Sm-like proteins Lsm10 and Lsm11 [257–259]. This difference is likely to be associated with the differences in the Sm-binding sequence. Recently, the U7 snRNP has not only received considerable attention from a structural biology point of view, see e.g. [260, 261], but it has also been investigated as a means of modifying splicing dys-regulation. In particular, U7 snRNA-derived constructs which target a mutant dystrophin gene were explored as a gene-therapy approach to Duchenne muscular dystrophy [262, 263].

Given the attention received by histone RNA 3' end processing and the protein components of the U7 snRNP, it may come as a surprise that the U7 snRNA itself has received little attention in the last decades. In fact, the only two experimentally characterized mammalian U7 RNAs are those of mouse [264–267] and human [253, 268], while most of the earliest work on U7 snRNPs concentrated on the sea urchin *Psammechinus miliaris* [269–272] and two *Xenopus* species [273–275]. More recently, the U7 RNA sequences have been reported for *Drosophila melanogaster* [276] and fugu [171].

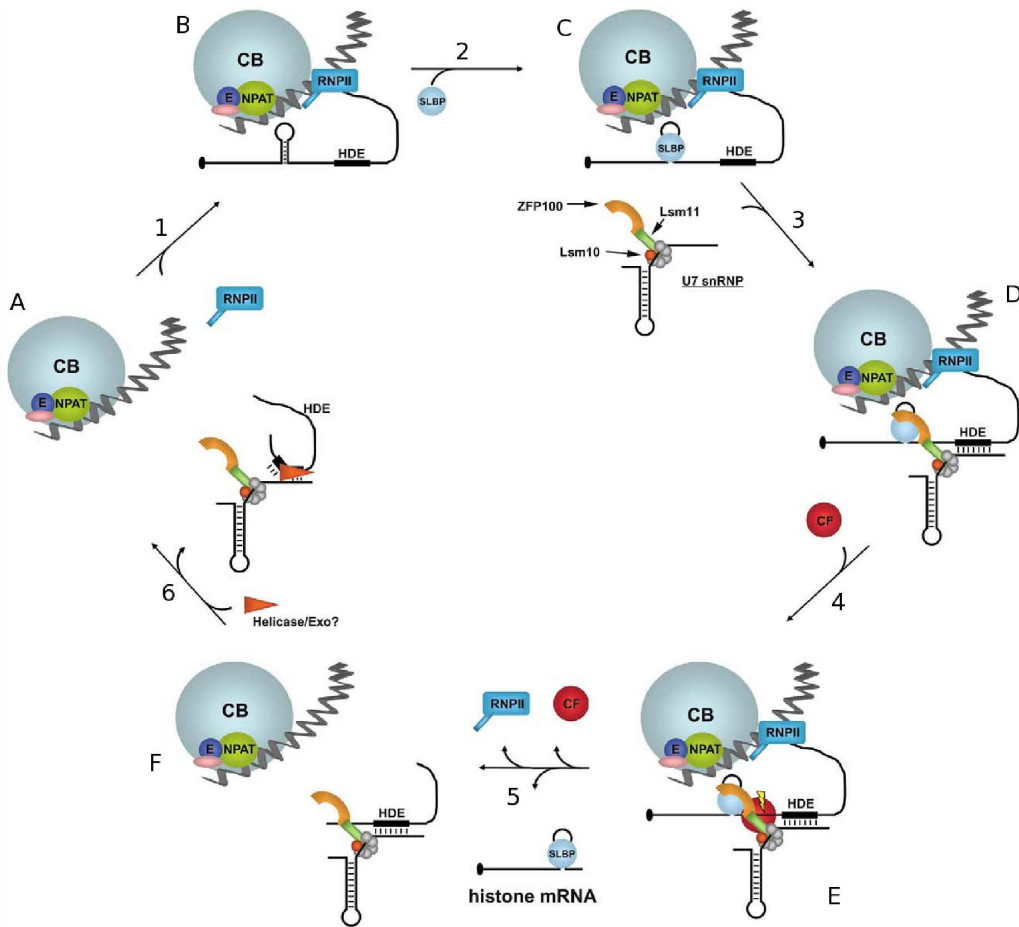


Figure 3.21: The pathway of mammalian histone pre-mRNA biosynthesis, modified from [1]. The histone genes are located near Cajal bodies (CBs), which are likely to play a role in histone pre-mRNA processing (A). Recruitment of ribonucleoprotein II (RNP II) and transcription, (1,B). SLBP probably binds to the histone pre-mRNA during transcription (2), and then recruits the U7 snRNP (3). The U7 snRNP protein, ZFP100, interacts with the SLBP/stem-loop complex and helps stabilizing the binding of U7 snRNA to the HDE (D). This complex recruits an unknown cleavage factor (CF, 4), resulting in the production of the mature histone mRNA, which remains bound to SLBP. Histone pre-mRNA processing, like cleavage and polyadenylation, is linked to transcription termination (5). U7 snRNP is afterwards compounded to its original functional unit (6).

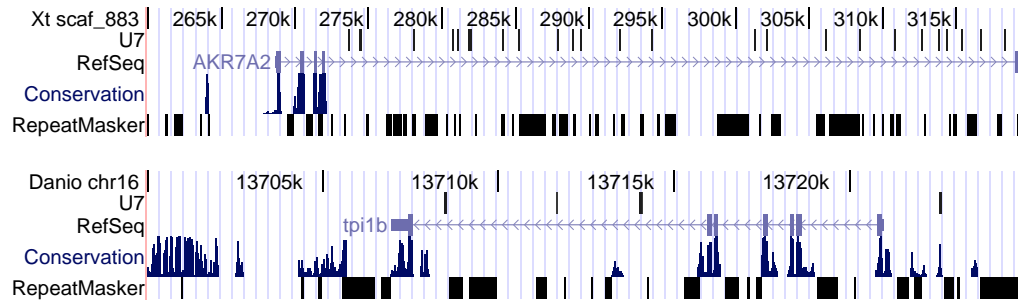


Figure 3.22: Clusters of U7snRNA genes in *Xenopus* and zebrafish taken from the UCSC genome browser. The “U7” track shows `blat` matches of the U7 snRNA sequences; “RepeatMasker” refers to annotated repetitive sequence elements; the “RefSeq” track shows the intron/exon structure of protein-coding genes; the “Conservation” panel displays `phastcons` score measuring sequence conservation across vertebrates. We refer to the data track description at <http://genome.ucsc.edu/> for technical details. `tpi1b` – triphosphate isomerase 1b.

We are aware of only two studies that considered U7 snRNA from a bioinformatics point of view. In [277], the U7 snRNA is used as an example for the application of `Construct` to compute consensus secondary structures, and [26] briefly reports on a `Blast` based homology search which uncovered candidate sequences for chicken and two teleost fishes.

The U7 snRNP-dependent mode of histone end processing is a metazoan innovation [1, 257]. Nevertheless, the most recent release of the `Rfam` database [86] [Version 8.0; Feb. 2007] lists sequences from eukaryotic protozoa, plants, and even bacteria. This discrepancy prompted us to critically assess the available information on U7 snRNAs.

### 3.5.1 *Bona fide* U7 snRNA Sequences

The results of the `Blast`-based searches are summarized in Tab. 3.6. In most species only a single gene with clear snRNA-like upstream elements was found. In addition `Blast` identified several pseudogenes. Clusters of U7 snRNAs as previously described for sea urchin and frog were otherwise only found in zebrafish, Fig. 3.22.

The short length and the substantial divergence of the U7 snRNA sequences make it impossible to distinguish functional U7 snRNAs from pseudogenes based on the U7 sequence alone. To make this distinction, it is necessary to analyze the flanking

Table 3.6: Trusted U7 snRNA sequences.  $\psi$  gives the number of paralog loci, most likely U7 pseudogenes, defined by a Blast *E*-value less than 0.001 compared to the functional copy. CAF-1 refers to the genome freezes provided by the *Drosophila Comparative Genomics Consortium*. These sequences were retrieved from <http://rana.lbl.gov/drosophila/caf1.html> in December 2006. The *Drosophila melanogaster* sequence is the one used by the UCSC browser (Release 4; Apr. 2004, UCSC version dm2). The sea urchin Genome BCM\_Spur\_v2.1 was obtained from [ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Spurpuratus/fasta/Spur\\_v2.1/linearScaff](ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Spurpuratus/fasta/Spur_v2.1/linearScaff).

Species	Assembly	Sequence	from	to	ori	DB ID	$\psi$	
<i>Mus musculus</i>	ensembl 43	Chr.6	124706844	124706905	-	ENSMUSG00000065217	27	
<i>Rattus norvegicus</i>	ensembl 43	Chr.X	118163804	118163865	-	ENSRNOG00000034996	31	
<i>Rattus norvegicus</i>	ensembl 43	Chr.4	160870934	160870995	-	ENSRNOG00000035016	31	
<i>Homo sapiens</i>	ensembl 43	Chr.12	6923240	6923302	+	ENSG00000200368	91	
<i>Macaca mulatta</i>	ensembl 43	Chr.11	7125496	7125557	+	ENSMMSG00000027525	95	
<i>Otolemur garnettii</i>	PreEnsembl 43	scaffold_102959	117572	117633	-	—	0	
<i>Oryctolagus cuniculus</i>	ensembl 43	GeneScaffold_1693	111485	111546	+	—	3	
<i>Procapra capensis</i>	NCBI TRACE	175719230	275	336	+	—	—	
<i>Loxodonta africana</i>	ensembl 43	scaffold_60301	4254	4314	-	—	2	
<i>Echinops telfairi</i>	ensembl 43	GeneScaffold_2204	10742	10803	+	ENSETEG00000020899	57	
<i>Felis catus</i>	ensembl 43	GeneScaffold_69	192907	192968	+	—	7	
<i>Canis familiaris</i>	ensembl 43	Chr.27	41131749	41131810	-	ENSCAFG00000021852	2	
<i>Myotis lucifugus</i>	PreEnsembl 43	scaffold_168837	32294	32356	-	—	0	
<i>Equus caballus</i>	PreEnsembl 43	scaffold_58	7463562	7463623	+	—	0	
<i>Bos taurus</i>	ensembl 43	Chr.5	10349126	10349187	-	AAFC03061782	8	
<i>Tursiops truncatus</i>	NCBI TRACE	194072802	598	659	+	—	—	
<i>Dasypus novemcinctus</i>	ensembl 43	GeneScaffold_1944	24469	24530	+	—	16	
<i>Spermophilus tridec.</i>	PreEnsembl 43	scaffold_139061	45428	45489	-	—	0	
<i>Erinaceus europaeus</i>	ensembl 43	GeneScaffold_2232	5133	5194	+	—	30	
<i>Monodelphis domestica</i>	ensembl 43	Un	131411333	131411393	+	ENSMODG00000022029	1	
<i>Gallus gallus</i>	ensembl 43	Chr.1	80484148	80484212	+	ENSGALG00000017891	1	
<i>Taeniopygia guttata</i>	NCBI TRACE	TGAB-afg09c06.b1	683	748	-	—	—	
<i>Anolis carolinensis</i>	NCBI TRACE	G889P8207RM16.T0	106	171	-	—	—	
<i>Xenopus tropicalis</i>	ensembl 43	scaffold_883	Cluster ~ 20 copies from 272500 to end					
<i>Xenopus laevis</i>	GenBank	X64404	Cluster (partial)					
<i>Xenopus borealis</i>	GenBank	Z54313	Cluster (partial)					
<i>Danio rerio</i>	ensembl 43	Chr.16	Cluster: 4 copies at 13708000 ... 13723000					
<i>Takifugu rubripes</i>	ensembl 43	scaffold_205	229679	229736	+	—	0	
<i>Tetraodon nigroviridis</i>	ensembl 43	Chr.8	9059483	9059541	+	—	(1)	
<i>Gasterosteus aculeatus</i>	ensembl 43	groupXX	11616333	11616392	-	—	0	
<i>Oryzias latipes</i>	ensembl 43	Chr.16	17393002	17393059	+	—	0	
<i>Strongylocentrotus p.</i>	BCM_Spur_v2.1	Cluster: 2 sequences each on scaffolds 83935 and 88560						
<i>Psammecinus miliaris</i>	GenBank	Cluster 5 genes, 1 sequenced M13311.1						
<i>Drosophila melanogaster</i>	UCSC	3L	3577355	3577425	+	CR33504	0	
<i>Drosophila ananassae</i>	CAF-1	CH902618.1	9849345	9849414	-	—	0	
<i>Drosophila erecta</i>	CAF-1	CH954178.1	6292889	6292959	+	—	1	
<i>Drosophila grimshawi</i>	CAF-1	CH916366.1	10347991	10348062	+	—	1	
<i>Drosophila mojavensis</i>	CAF-1	CH933809.1	2924982	2925053	-	—	1	
<i>Drosophila persimilis</i>	CAF-1	CH479328.1	89311	89383	-	—	0	
<i>Drosophila pseudoobscura</i>	CAF-1	CH379070.2	5738714	5738786	+	—	1	
<i>Drosophila simulans</i>	CAF-1	CM000363.1	3136652	3136582	-	—	1	
<i>Drosophila virilis</i>	CAF-1	CH940647.1	4512836	4512907	-	—	1	
<i>Drosophila willistoni</i>	CAF-1	CH964101.1	1418210	1418280	+	—	0	
<i>Drosophila yakuba</i>	CAF-1	CM000159.2	4146836	4146905	+	—	0	



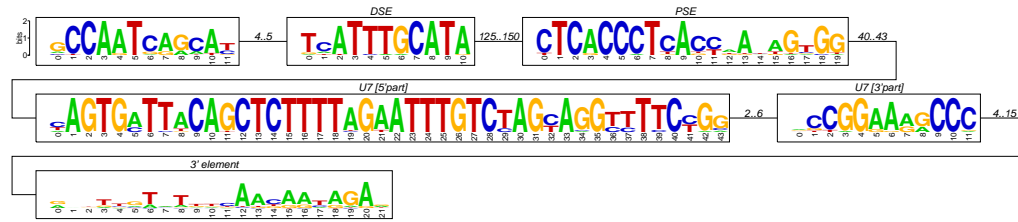


Figure 3.23: Conserved elements in functional U7 snRNA gene. Consensus pattern of the amniote sequences from Tab. 3.6. The classical distal sequence elements (DSE), proximal sequence elements (PSE), and 3'elements of pol-II spliceosomal RNA genes are clearly discernible. The U7 sequence itself is interrupted by a short variable region with substantial length-variation.

sequences as well. *Bona fide* snRNA genes are accompanied by characteristic promoter elements [93, 278]. Fig. 3.23 displays the consensus sequence motives of the presumably functional amniote U7 RNAs.

In the human and mouse, several pseudogenes have been described in detail in addition to the functional genes [267, 279]. Notably, several variant U7 RNA sequences from human HeLa cells were reported in [268]. This might indicate that the human genome, in apparent contrast to mouse, also contains more than one functional U7 snRNA gene, or that some of the pseudogenes are transcribed at low levels. Tab. 3.6 therefore lists the number of U7-associated loci obtained by **Blast** searches that use the presumably functional gene from the same species as query. This number can be fairly large in some mammalian lineages, reaching almost 100 loci in primates. In contrast, in most species there are only a few U7-associated sequences, most of which are readily recognizable as retrogenes by virtue of poly-A tails.

In several genomes we were not able to find an unambiguous candidate for a functional U7 snRNA, although we found sequences that clearly derive from U7 but are not accompanied by a recognizable PSE. Examples include *Sorex araneus* and platypus. Most likely, these **Blast** hits are pseudogenes, although many of them are annotated with **ensembl** gene IDs. This annotation derives from sequence homology with the examples stored in the **Rfam** database. In Fig. 3.24 and Tab. 3.6 we compile the results of our **Blast**-based homology search, which contains only sequences which are either experimentally known to be expressed or which are predicted to be functional genes based on the presence of conserved upstream elements.

Separate multiple sequence alignments of Amniots, Teleosts, frog, sea urchins, and flies reveal strong conservation of the Sm-binding motif, consisting of the



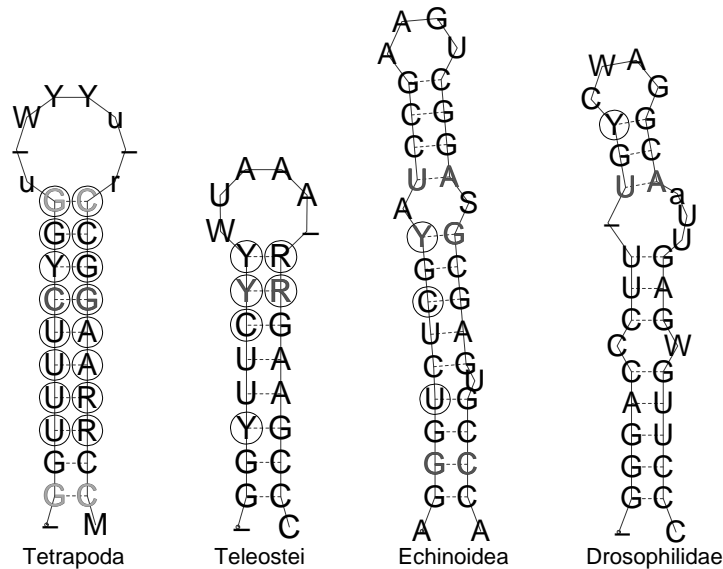


Figure 3.25: Comparison of U7 hairpin structures. Consensus secondary structures are computed using `RNAalifold` using the manual improved alignments of tetrapods, teleost fishes, sea urchins, and fruit-flies, respectively. Circles indicate consistent and compensatory mutations which leave the structure intact. Gray letters indicate that one or two of the aligned sequences cannot form the base pair.

these features as anchors, one obtains the alignment in Fig. 3.24, which highlights the differences between major clades. Notable variations within the vertebrates are in particular the A-rich 5' and the reduced stem in teleosts, and their A-rich sequence in the hairpin loop. The hairpin region is very poorly conserved at sequence level between vertebrates, sea urchins, and flies, although its structural variation is limited in essence to the length of the stem and a few short interior loops or single-nucleotide bulges.

### 3.5.2 More Distant Homologs?

The U7 snRNA sequences evolve rather fast. Together with the short sequence length, this limits the power of sequence-based approaches to distant homology search. The consensus pattern in Fig. 3.24 indicates quite clearly that such methods are bound to fail outside the four groups with experimentally known sequences (tetrapoda, teleosts, echinoderms, fruit-flies). Indeed, both `Blast` and `Fragrep` did not provide additional candidates that could be unambiguously classified as U7 snRNAs based on sequence information alone.

The comparison of the U7 hairpins in the different clades, Fig. 3.25, reveals significant differences in the secondary structures of invertebrates and vertebrates: vertebrates have smaller stem-loop structures with smaller or no interior loops or bulges. The stem in teleosts, furthermore, is systematically shorter than in tetrapods. These structural differences between clades has to be taken into account for homology search. In fact, as a consensus rule, we can only deduce that the stem-loop structure has a total of 8-15 base pairs, that it is nearly symmetric, and that it is enclosed by an uninterrupted stem of length at least 5 with two GC pairs at its base.

Even combined with with the conserved sequence motives in the 5' part of the molecule, this yields only a rather loose definition of a U7. Release 8.0 of the **Rfam** database [86] lists several sequences in its U7 RNA section that are surprising. Neither contained in the literature nor contained in the manually curated U7 “seed-set”, these candidate sequences have been found using a homology search based on **Infernal** [281] and the seed alignment. While the *Danio rerio* sequences are identical with the sequences we identified in work starting from the much closer homolog in *fugu*, the candidates reported for *Caenorhabditis elegans*, and *Girardia tigrina* raise serious doubts. The *Caenorhabditis elegans* sequence, although ostensibly well conserved in comparison with the deuterostome sequences, has no recognizable homologs in any one of the other three sequenced *Caenorhabditis* species, (*C. briggsae*, *C. remanei*, *C. brenneri*). The *Girardia tigrina* sequence is located in the 3' UTR of the *DthoxE-Hox* gene (**X95413**). Both sequences furthermore do not share even the core UUUNUC of the consensus Sm-binding motive.

Several additional candidates were reported for higher plants and even bacteria. Higher plants apparently do not have the replication-dependent metazoan-style histone 3' end processing machinery [1, 257], and bacteria do not even have proper histones. It is very unlikely that these sequences are real U7 snRNAs. No conclusive argument can be given at this point for the few isolated U7 snRNAs candidates listed in the **Rfam** database. These examples show once again that at least for very short ncRNAs, the results from homology searches have to be taken with caution, in particular when they are not corroborated by additional supporting evidence.

The poor sequence conservation between major groups highlighted in Fig. 3.24 suggest that purely sequence-based homology searches have little chance of success in insect or basal deuterostome genomes. Indeed, neither **Blast** nor **Fragrep** found convincing candidates. We therefore resorted to structure-based approaches and explicitly included the PSE in the search procedure. We used **rnabob** (Section 2.1.3) with a non-restrictive pattern to find plausible initial candidates, which were then manually compared with the alignment in Fig. 3.24. The most plausi-

```

# |<Histone-binding-region>|. |<--Sm-->|...<<<<<. <<<<. <<<<.....>>>>.....>>>.....
Homo .....CAGTG. TTACAGCTCTTTTGAATTTGTCTAGTA..GGCTT. TCT. GGC. TTTT. ACC. GGA. AA. GCCCCT.
Mus .....AAGTG. TTACAGCTCTTTTGAATTTGTCTAGCA..GGTTT. TCT. GAG. . TTCG. . GTC. . GGA. AA. ACCCCT.
Xenopus_l .....AAGTG. TTACAGCTCTTTTGAATTTGTCTAGCC..GGTTT. TTA. G. . . . TCT. . . . G. . TTG. GA. GCCACA.
Takifugu .....AGGAATGATT. GCTCTTTAGATATTTCTCTAGTA..GGCTT. TTC. . . . ATACA. . . . GAG. AA. GCCCCCT
Petromyzon-c1 .....ATTGAGGATCTTTGAC. TTTTGTCTTTGTGTGGTGACC. . . . .GAAA. . . . .GGAGC. ACC. . . .
Branchiostoma-c1 .....ACTGG. TAAC. GCTCTTTTAC. CTTTATCCGCG. . . .GGGTA. A. . . . .CCT. . . . .T. TA. TCCGTA.
Branchiostoma-c2 .....GAGTG. TAAC. GTTCTTTTAC. CTTTATCCGCG. . . .GGGTA. . . . .ACCTA. . . . .TA. TCCGTT.
Psammechinus_1 .....ATCTTTCA. AGTTTCTCTAGAA. GGGTCT. CGCGTCCG. AAGT. CGGA. GCGC. AGTGCCCAAC
Bombyx_mori-c1 TCCATCAAT. ATGTTCTATCTTTTA. . . .ATTTATCGAAAA. CGGTCA. AG. A. . . .ACTAGTC. . . .G. CT. TG. GCC. . . .
Bombyx_mori-c2 AAGATTTG. GTGTGTAATCTTTAACTGTTATCTTTTG. CGGTAGG. . . .T. AGCGGTTGGCT. . . . .CT. GCC. . . .
Dr_melanogaster ATGAAAAT. TTTTATCTCTTTGA. AATTTGTCTTGGT. . . .GGGACCCTT. . . .TGT. CTAG. GCA. TT. GAGTGT. TCCCGTT
# |<Histone-binding-region>|. |<--Sm-->|...<<<<<. <<<<. <<<<.....>>>>.....>>>.....

```

Figure 3.26: Best candidates from searches with `rnabob` in the lamprey *Petromyzon marinus*, *Branchiostoma floridae*, and *Bombyx mori*. In addition to the putative U7 RNA sequence shown here, these candidate sequences also have a putative PSE associated with them.

ble candidates are shown in Fig. 3.26, albeit none of them is unambiguous. No convincing candidates were found in the mosquito *Anopheles gambiae* and in the honeybee *Apis mellifera*.

### 3.5.3 Discussion

Since U7 snRNA has its primary function in histone 3' maturation it is virtually certain that this class of non-coding RNAs is restricted to metazoan animals — after all, the process in which they play a crucial role is unknown outside multicellular animals. With its length of 70nt or less, U7 snRNA is the smallest known pol-II transcript. Each of its three major domains, the histone binding region, the Sm-binding sequence, and the 3' stem-loop structure exhibit substantial variation in both sequence and structural details, as can be seen from the detailed sequence alignments (Fig. 3.24) and the structural models of the terminal stem-loop structure (Fig. 3.25). As a consequence, our computational survey not only compiled a large number of previously undescribed U7 homologs from vertebrates and drosophilids, but also stresses the limits of current approaches to RNA homology search.

While `Blast` already fails to unambiguously recognize teleost fish homology from mammalian queries and *vice versa*, even more sophisticated (and computationally expensive) methods have limited success when applied to basal deuterostomes or insect genomes. On the other hand, not only the limited sensitivity of current approaches poses a problem. Conversely, the most sensitive methods are fooled by false positives, as exemplified by the plant and bacterial sequences in `Rfam`.

In summary, thus, this study calls both for more experimental data on U7 snRNAs – Which, if any, of our U7 candidate sequence in lamprey, silk worm, are really U7 snRNAs in these species? – and for improved bioinformatics approaches for homology search that can deal with such small and rapidly evolving genes.

## 3.6 Introns in Insects

A large portion of the transcriptional output of eukaryotic genomes consists of “mRNA-like non-coding RNAs” (mlncRNAs) [14, 15]. These transcripts are capped, polyadenylated and often spliced (sometimes alternatively spliced) just like protein-coding mRNAs, but lack discernible open reading frames. These mlncRNAs are typically much larger than the “house-keeping” RNAs such as transfer (t)RNA, small nuclear (sn)RNAs, small nucleolar (sno)RNAs and they do not seem to have well-conserved secondary structures.

Here, we present a new approach to identify intron-containing mlncRNAs from genomic sequence data alone. Our method exploits characteristic evolutionary signatures of conserved introns. The rationale behind this approach is driven by the observation that intron positions are generally well conserved both in protein-coding and non-coding RNA genes [9, 282–284].

The assumption underlying our approach is that a functional pair of donor (5′) and acceptor (3′) splice sites will be retained over long evolutionary time-scales only if (i) the locus is transcribed into a functional transcript, and (ii) accurate intron removal is necessary to produce a functional transcript. Thus, conserved introns can be employed to determine the presence of a functional transcript directly from comparative genomics data. The advantage of this approach is that we do not need to make any assumption of the transcript itself.

We applied this intron-based approach to 15 insect genomes and reliably predicted novel mlncRNAs. We show that these mlncRNAs are largely unstructured and often not associated with significant sequence conservation, implying that they cannot be predicted by existing methods. Our screen also identified unannotated protein-coding genes and provides a refinement of several gene structures by identifying introns in incomplete coding or untranslated regions. Experimental verification succeeded for 18 of 29 tested predictions. Furthermore, we showed that conserved introns imply conserved expression of the surrounding transcript in other species.

### 3.6.1 Computational identification of spliced RNAs in *Drosophila* genomes

Our approach consists of three steps. Firstly, we predicted introns in individual insect genomes. Secondly, we used genome-wide alignments to identify orthologous introns, defined here as introns that are independently predicted in at least two

genomes and where both donor and acceptor sites are exactly aligned. Thirdly, we compiled a set of evolutionary signatures that are characteristic for introns with conserved splice sites and use machine learning to distinguish between real and false intron predictions. These steps are illustrated in Fig. 3.27 and are detailed below.

We chose *Drosophila* as a model system to test our approach for several reasons: (i) There is a sufficient number of sequenced insect genomes, which allows comparative genomics methods to annotate features such as protein-coding genes, structured RNAs, and regulatory motifs with high accuracy [285–288]. (ii) The majority (54%) of introns in *D. melanogaster* is not longer than 81 nt, a natural cutoff between long and short introns [289, 290]. (iii) The short introns in *D. melanogaster* contain basically all the information needed to identify them in pre-mature transcripts [289], in contrast to most mammalian introns [291, 292].

We observed that most positive samples exhibit a poor sequence conservation in the middle of the intron, while numerous negatives show an atypical high conservation (Fig. 3.28). This pattern is expected because the middle part of an intron usually contains unconstrained sequence [293]. Moreover, positive samples show some length variation between species [294], while negatives rarely do.

To combine these features into a single decision (real intron vs. false prediction), we trained a Support Vector Machine (SVM) using randomly selected  $\sim 95\%$  of our set for training (22,278 positives and 111,530 negatives). Details of the methods are available at [295].

### 3.6.2 Novel spliced transcripts

We used the SVM to evaluate the 342,785 predictions without an overlap to annotated protein-coding transcripts on the same strand to uncover novel introns and therefore novel transcripts. Using a stringent probability threshold of 0.95, we predict 369 introns. We searched ESTs and non-coding FlyBase transcripts and found 131 (35.5%) introns where both splice sites are transcript-confirmed, with the rest (238 cases, 64.5%) being currently unconfirmed. Of these 238 unconfirmed introns, 44 (18%) are supported by ESTs in other *Drosophila* species. This indicates that our approach is successful in uncovering spliced transcripts. Figure 3.29 shows examples of confirmed introns belonging to the 5' UTR of a gene, to an intronic antisense transcript, to a potentially tissue-specific ncRNA and to a structured ncRNA that represents a precursor for short interfering (si)RNAs [296].

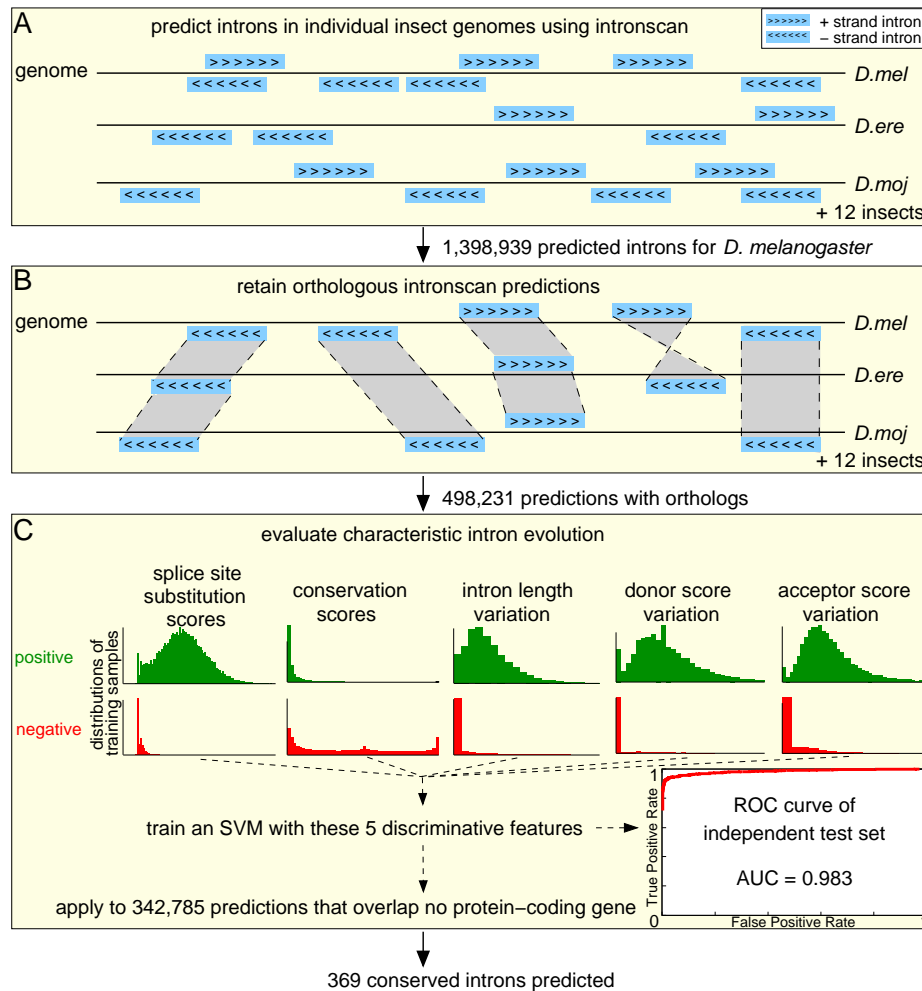


Figure 3.27: Overview of the computational intron prediction procedure. (A) Introns are predicted using *intronscan* on both strands of the *D. melanogaster* genome, yielding a total of ~1.4 million predictions. Independent *intronscan* predictions in the other insect genomes were made.

(B) Only those *D. melanogaster* intron predictions are retained that have an orthologous prediction in at least one additional genome.

(C) A Support Vector Machine (SVM) classifier based on five features is used to distinguish positive (real introns) and negative training samples (false predictions). These features measure characteristic splice site substitutions, sequence conservation in the middle part of introns, and variation of the intron length, donor and acceptor score between species. As indicated by the distributions, these features are highly discriminative for positive and negative samples. Using this classifier we predict 369 conserved introns.



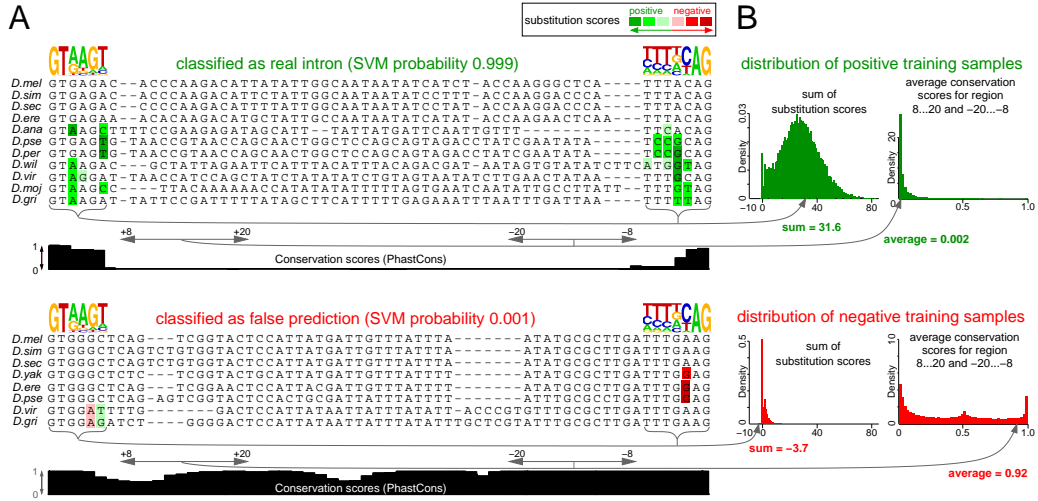


Figure 3.28: Evaluating characteristic intron evolution. (A) Two predicted introns with orthologous intronscan predictions in other species are shown. The prediction on top exhibits several substitutions in the splice site regions that are characteristic for real introns (e.g. C to T substitutions at acceptor position -3). Furthermore, this prediction has a low sequence conservation within the intron (average PhastCons scores for the region +8...+20 and -20...-8 is only 0.002). This prediction gets a high probability for being a real intron (0.999). In contrast, the prediction at the bottom has substitutions that are inconsistent with intron evolution (e.g. A to G substitution at acceptor position -3) and it exhibits conservation throughout the intron (average PhastCons score 0.92). The SVM probability for being a real intron is consequently low (0.001). Positive substitution scores are shown in shades of green, negatives in shades of red. Substitution scores are only considered for the donor (positions +2...+6) and acceptor splice site (positions -7...-3). Note that the substitution scores are specific for each pair *D. melanogaster* with another species, thus the same substitution with respect to different species can get different scores. (B) The distribution of the summed substitution scores (left) and the average conservation scores (right) show a substantial difference between our positive and negative samples. The position of the values of the introns from panel A are indicated. For a better visualization, the y-axis for positive and negative samples has a different scale.

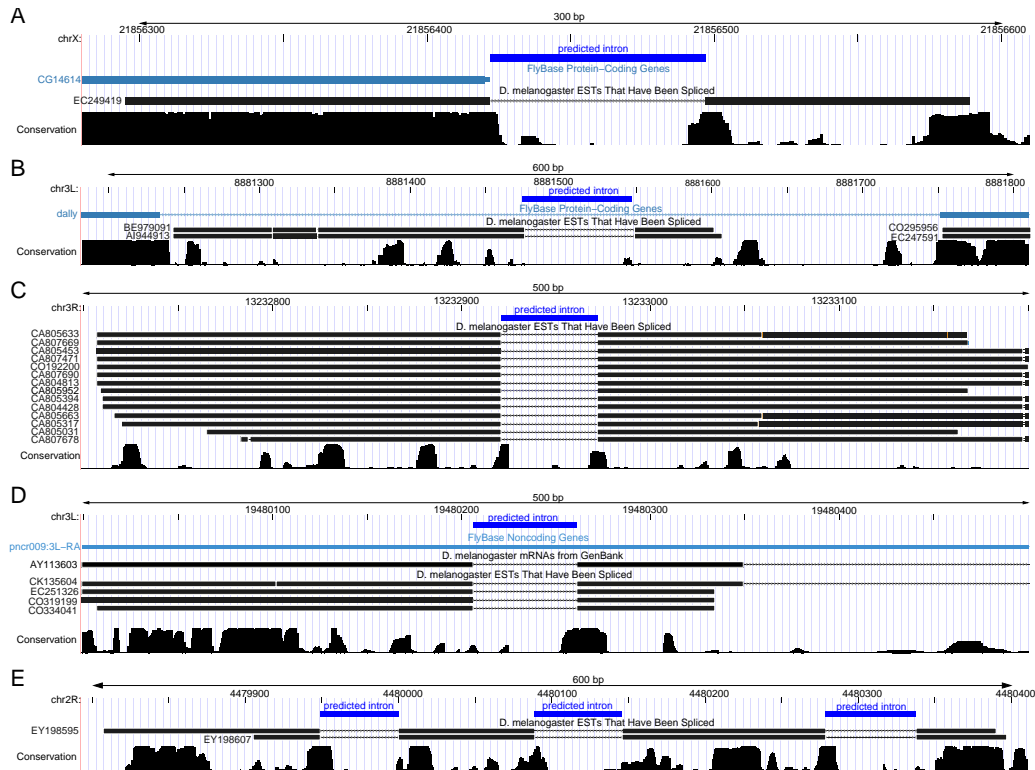


Figure 3.29: Examples of transcript-confirmed intron predictions. (A) A predicted intron is located in the 5' UTR of the protein-coding gene *CG14614*, whose current 5' UTR annotation consists of only 2 nt. (B) Example of a predicted intron that belongs to a transcript overlapping an intron of *dally* in the antisense direction. (C) Example of a predicted intron that belongs to a potentially tissue-specific non-coding RNA, as 13 of the 14 supporting ESTs originate from a salivary gland library (ESG01). (D) A predicted intron that overlaps a non-coding FlyBase transcript (*pncr009*) that has no intron annotation. *pncr009* was found to be a structured precursor for small interfering RNAs [296]. (E) Example of a 'cluster' of three introns within ~400 nt. All three introns are predicted with probability > 0.999 and belong to a potentially coding gene (*blastx* hits in several *Drosophila* species). Examples B-E illustrate that our approach finds introns which are located in regions of low sequence conservation, indicated by low PhastCons conservation scores up- and downstream of the intron. Modified UCSC genome browser [297] screenshots were used to make this figure.

### 3.6.3 Novel spliced non-coding RNAs

29 of 129 introns (22.48%), which are considered to be *bona fide* mlncRNAs, have predicted orthologous introns in species outside the *Drosophila* subgenus (*D. virilis*, *D. mojavensis*, *D. grimshawi*), which indicates exon-intron structure conservation over 63 My of evolution [298].

In contrast to the non-coding RNAs identified in [299], our 129 introns are flanked by regions of rather low sequence conservation (average PhastCons scores for the 100 nt up- and downstream flanks: 0.25). Note that this is no indication that the predictions are not real. Indeed, the seven unconfirmed introns that we experimentally verified (see below) show an even lower flank conservation (average 0.21). A large fraction of these 129 introns overlap coding genes in antisense direction (41 of 129; 32%); however, this is not surprising given that almost half of the *D. melanogaster* genome is covered by exons and introns of coding genes and the fact that many genes overlap each other on opposite strands [300].

### 3.6.4 Novel mlncRNAs are mostly unstructured

Our screen identified two introns located in known mlncRNAs with extensive secondary structures (pncr009, CR32205; Fig. 3.29D) that function as siRNA precursors [296]. To test if our predictions are associated with conserved secondary structures, we applied RNAz to the regions flanking the 129 introns. We obtained 2 (1.6%) predictions of conserved secondary structures. Since RNAz has a certain false-positive rate, we used two control sets to test for enrichment or depletion of conserved structures. 5,000 randomly selected genome regions and their shuffled versions show a highly similar percentage of RNAz hits. Together with the observation that >98% of these mlncRNAs are not associated with conserved secondary structures, this indicates that our method mostly predicts unstructured mlncRNAs, which cannot be identified by RNAz and related methods.

### 3.6.5 Experimental verification of predicted mlncRNAs

Our collaborators from Halle used RT-PCR with primers designed to flank the predicted intron to validate expression of the corresponding transcripts in five different developmental stages of *D. melanogaster*: embryo, larva, pupa, male, and female. We counted as a positive verification only those introns where the transcript is spliced and sequencing confirms the correctness of both splice sites.

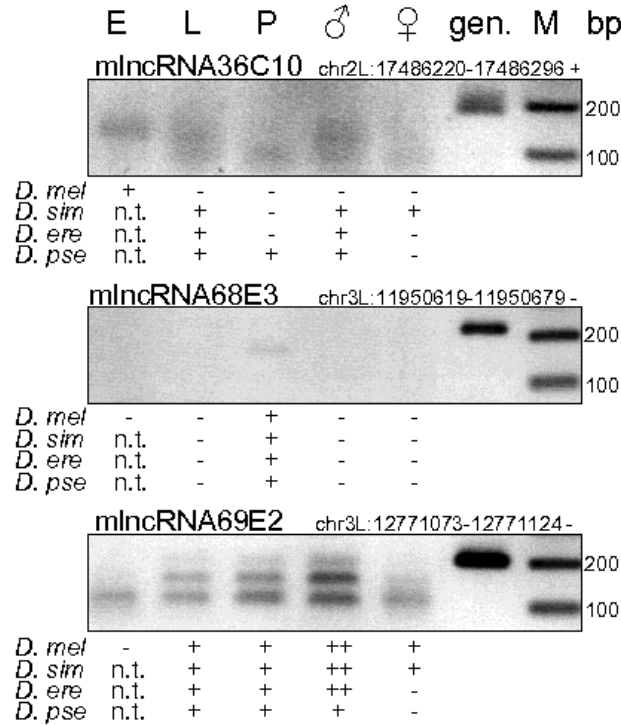


Figure 3.30: Experimentally verified introns in mlncRNA transcripts. The expression of the spliced transcript was tested in embryo (E), larva (L), pupa (P), male and female stages. Ethidium bromide stained agarose gels show the RT-PCR results for *D. melanogaster*. Expression data of the orthologous transcripts in *D. simulans* (*D.sim*), *D. erecta* (*D.ere*) and *D. pseudoobscura* (*D.pse*) is shown below the *D. melanogaster* (*D.mel*) data. Genomic DNA (*gen.*) was used as a PCR control and size was measured according to a 100 bp Ladder (*M*). PCR products were verified by sequencing. +/++ = expressed; - = no band; n.o. = no orthologous intron; n.t. = not tested; We used + and ++ to indicate weaker and stronger expression in different stages. Detailed methods for this experiment can be viewed in [295]. Experiments are done by our collaborators Sandro Lein, Claudia Nickel and Gunter Reuter from Halle, Germany.

We tested 12 introns that likely belong to mlncRNAs and could verify seven (58%) of them (Examples can be viewed in Fig. 3.30). We named these seven mlncRNAs according to their genomic location (cytogenic band). The expression level of all transcripts is low, consistent with previous findings of low expression levels of mlncRNAs [40]. Only two of the seven mlncRNAs can be found in all five tested conditions. The other five show variation in the expression pattern during development, which suggests that their expression is controlled. For example, one mlncRNA is found only in embryos and another mlncRNA shows only a weak expression at the pupal stage MlncRNA69E2 shows two bands on the gel due to usage of an alternative acceptor splice site (Fig. 3.31) and the predicted intron

corresponds to the longer transcript.

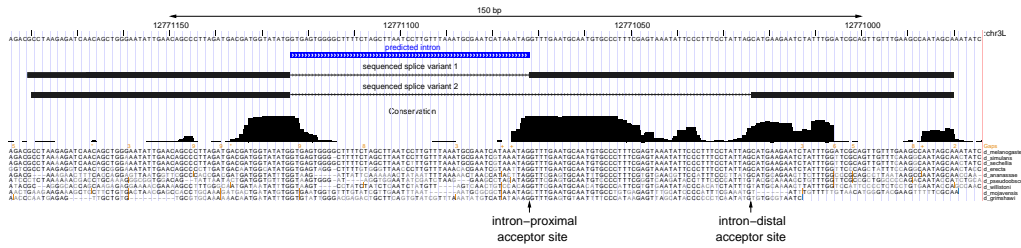


Figure 3.31: MlncRNA69E2 is alternatively spliced. MlncRNA69E2 produces two transcripts differing by the use of the acceptor site, visible as two bands in Fig. 3.30. While the proximal acceptor that corresponds to the predicted intron is deeply conserved, the distal acceptor located 48 nt downstream is only conserved in the *melanogaster* subgroup. Conserved sequence parts that are excluded in the shorter splice variant might suggest that both transcript differ in their function.

Overall, we verified 62% (18 of 29) of our tested predictions. As in all transcriptomic studies, this percentage represents a lower bound as we miss transcripts expressed in other conditions or at expression levels below our sensitivity.

### 3.6.6 Discussion

We describe here a method that predicts intron-containing transcripts by making use of evolutionary characteristics of conserved introns and the observation that introns rarely turnover or shift with respect to the exons. It is important to note that we solely use intron information for predictions. This allows us to identify (i) protein-coding transcripts (including untranslated regions thereof) as well as mncRNAs, (ii) transcripts without conserved secondary structures and (iii) transcripts without evolutionary conserved sequences (see Fig. 3.29). The latter property is important as functional ncRNAs do not necessarily have significantly conserved sequences [301]. For example, the XIST ncRNA has a clear function in X-chromosome inactivation in mammals [302], however a comparison of human and mouse XIST reveals a low overall sequence identity [283].

While our approach is unbiased with respect to these characteristics, it has limitations. Our method predicts only a partial transcript structure, which in general will have to be completed by experimental approaches such as full-length cDNA sequencing. However, gene prediction algorithms that predict only the CDS and high throughput transcriptomic techniques suffer from the same problem. Here, we focused on short introns in *D. melanogaster*, consequently transcripts containing

exclusively longer introns cannot be predicted. It remains unclear whether longer introns and whether short introns in other species are predictable in a similar way.

Furthermore, we currently classify introns with a conserved intron body as false, because the great majority of real introns shows no sequence conservation in the middle. Thus, introns overlapping other functional elements such as putative promoter elements, introns that are miRNA precursors [303, 304], or retained introns that encode a protein domain [305] are unlikely to be predicted.

Apart from the motivation to identify novel mlncRNAs, we aimed at predicting putatively functional mlncRNAs as opposed to transcriptional noise. Despite the observation that our predictions are generally not associated with strong sequence conservation, the detection of a conserved intron indicates that the exon-intron structure is under purifying selection and that the failure to correctly excise the intron likely affects the function of the transcript. Consistent with this, mlncRNA sequences, their splice sites and promoters show reduced substitution, insertion and deletion rates indicative of purifying selection [284]. Furthermore, we showed that conserved introns imply that the respective transcripts are expressed in other flies. While conserved exon-intron structure and conserved expression indicate function, the specific functional aspects of these mlncRNAs have to be addressed in future studies.

## Chapter 4

# Highly divergent structured ncRNAs

Most of the ncRNAs detected in chapter 3 are predictable by sequence conservation, calculated mainly by `Blast` and `GotohScan`. However, the selection pressure of ncRNAs lies in the structure instead of the primary sequence. Therefore we used programs, such as `Infernal`, `RNAmotif`, `rnabob`, `Fragrep` and various programs from the `Vienna RNA Package` to predict highly divergent structured ncRNAs, such as U3 snoRNA (Sec. 4.1), RNase MRP and P (Sec. 4.2), 7SK RNA (Sec. 4.3) and telomerase RNA (Sec. 4.4). All these RNA families vary by multiples of their lengths. They show short or no sequence conservation and invent or delete larger stems, at first sight in a completely random way, but of course all caused by nature. Using up to 30 different programs (Chap. 2) for the identification of a single possible homologous sequence implies a time consuming development of multiple alignments by hand for their verification.

## 4.1 U3 RNA

The U3 snoRNA is an exceptional member of the box C/D subclass. It is much longer than typical box C/D snoRNAs and does not direct chemical modifications. Instead, it acts as an RNA-chaperone mediating structural changes of the pre-rRNA to establish the correct conformation endonuclease cleavage [306]. Together with two other snoRNAs, mammalian U8 and U13 [307], it shares some features with snRNAs. For instance, human U3 snoRNA, has a hypermethylated 2,2,7-trimethylguanosine (TMG) cap at their 5' end [308]. U3 snoRNAs are processed from primary transcripts with a rather particular promoter, which may represent the fusion of two promoter systems: Homo1 D-box and TATA-box [309]. In the first stage, a 3'-extended precursor with a mono-methylated cap, which is then trimmed at the 3' end. In several fungi species, e.g. *Saccharomyces cerevisiae* and *Hansenula wingei* [49, 50] this precursor is spliced. In the final step, the TMG cap is formed [310, 311].

Across eukaryotes, the length of U3 varies by more than a factor of three from 143nt in *Trypanosoma* to 442nt in *Candida glabrata*. Its sequence is highly variable apart from several short highly conserved boxes denoted by A', A, C', B, C, and D, where C and D define the membership in box C/D class of snoRNAs. Due to their poor sequence conservation it is a non-trivial problem to establish the homology of snoRNAs over large evolutionary distances, e.g. between a mammalian and a yeast sequence.

The U3 snoRNA is highly structured and exhibits several conserved structural domains [312–314]. Due to its pivotal function in rRNA maturation, the U3 snoRNAs is believed to be ubiquitously present in Eukaryotes. The latest release of Rfam, v.9.1, [244] reports 141 sequences. We report here on a comprehensive search for homologous sequences in the more than 230 eukaryotic genomes. Using this extended data bases, which covers most major clades, we construct refined secondary structure models for most major clades and provide an overview of the variation of U3 structure.

### 4.1.1 Homology search

Within the 242 genomes investigated in this study and available EST-databases, we found a total of 238 U3 snoRNA homologs. In particular, our search was successful in 91 of 101 metazoan U3 snoRNAs. Negative results in several lophotrochozoa and cnidarian genomes are probably caused by the incompleteness of these genome projects.



Table 4.1: Summary of the homology-based survey. We list the number of genomes with at least one detected U3 snoRNA. Rfam-version 9.1 is used as reference. Since the Rfam alignments contain U3 sequences for which complete genomes are not publicly available, we list the intersection of the known sequences with the collection of genomes interrogated in this study (marked by a \*). Abbreviations: Met- Metazoa; Pla - Plants; Fun - Fungi; Oth - Other Eukaryots

	Met.	Pla.	Fun.	Oth.	Sum
Rfam <sub>seed</sub> *	9	8	2	1	20
Rfam <sub>all</sub> *	61	16	29	4	110
EST -DB	-	12	3	2	17
<b>Novel U3</b>	<b>30</b>	<b>20</b>	<b>28</b>	<b>19</b>	<b>97</b>
Rfam <sub>all</sub>	66	18	53	4	141
<b>All U3</b>	<b>96</b>	<b>38</b>	<b>81</b>	<b>23</b>	<b>238</b>

Among fungi, we found the U3 in 52 of the 53 ascomycota. In contrast, among the other 16 fungal genomes, an unambiguous homolog was identified only in *Phakospora pachyrhizi* and *Batrachochytrium dendrobatidis*. For six additional basidiomycota and the microporidian *Anthospora locustae* only tentative candidates were identified. These candidates exhibit the conserved boxes but have highly variable distances between the boxes, which might indicate insertion domains and/or additional introns.

Across viridiplantae we found 19 U3 snoRNAs out of 28 available (partly partially finished) genomes. The only plant seed sequence, which was not recovered from the incomplete genomic data is that of *Triticum aestivum* **X63065** [315]. 12 additional sequences we retrieved from EST databases. No plausible U3 sequence was identified in prasinophyceae and rhodophyta. In Heterokonta, we found a homolog in *Hyaloperonospora parasitica* in addition to the three known *Phytophthora* sequences.

Starting from the known *Trypanosoma brucei* sequence **X57047** [316], we were able to identify U3 snoRNA in all available genomes of kinetoplastida. Furthermore, we found U3 snoRNAs in all three available genomes of ciliates (*Tetrahymena thermophila*, *Paramecium tetraurelia*, *oxytricha trifallax*). Within Apicomplexa we found all expected U3 snoRNAs except for the genus *Paramecium*.

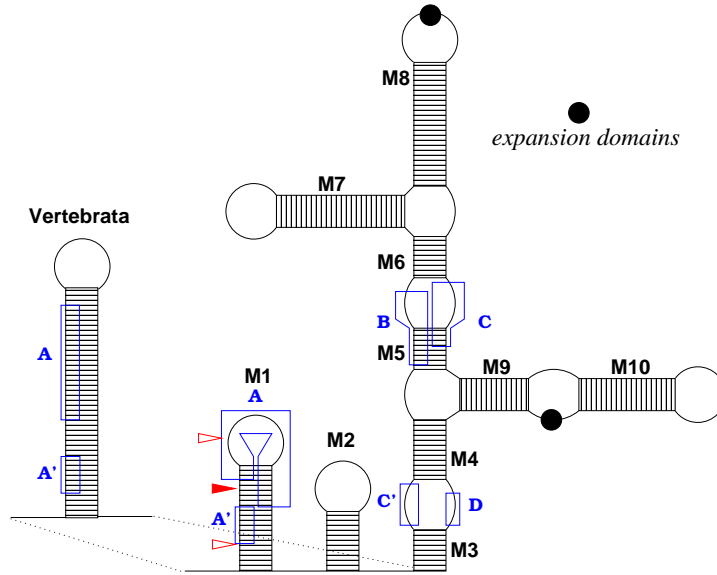
Besides the previously known U3 snoRNA of *Dictyostelium discoideum* **V00190** [317], no other unambiguous candidates were found in basal eukaryotes. Vague U3 candidates for *Plasmodium* and *Toxoplasma gondii* are listed in the supplement.

The sequences are compiled in the Supplemental Material and are submitted to Rfam together with this contribution.

**Secondary Structure Alignments** Separate alignments of the U3 snoRNA sequences have been produced for Metazoans, Fungi, Plants and Other Eukaryots. These can be found in the Supplemental Material. Based on these data, secondary structure models were constructed and then combined to the consensus shown in Fig. 4.1.

With the inclusion of the newly-detected sequences we find that U3 snoRNA structures are quite a bit more diverse than suggested by the **Rfam** seed alignment. This prompted us to propose a new numbering scheme for the helices, Fig. 4.1. In particular, we observe several major deviations from the consensus structure:

- (a) There are several major expansions. The platyhelminth *Echinococcus multilocularis* has expanded the M8 from 15nt (e.g. nematoda) to 91nt. *Candida glabrata* even invented a new stem between M10 and M9 with 49nt.
- (b) M7, which is specific to fungi, is shown as an unstructured region in the in the **Rfam** alignments. In saccharomycotina and some pezizomycotina (sordariomycetes and leotiomycetes) this stem varies from 6bp to 27bp without any recognizable sequence conservation or conservation of the positions of loops and bulges. Since closely related species do not show conserved splice-donor and splice-acceptor motifs, we argue that M7 is indeed a part of the U3 snoRNA.
- (c) The **Rfam** alignment starts with stem M3, omitting the 5' end of the molecule, presumably in order to allow for a structural alignment of vertebrate sequences with other U3 snoRNAs. Secondary structure prediction on vertebrate U3 sequences yields strong support for a large stem-loop structure including Box A' and A. However this model does not fit most other eukaryots (invertebrate animals, fungi, and plants), where clear support for a two-hairpin motif is found. Here, Box A located in the loop of stem M1. The structure of stem M2 is almost perfectly conserved in structure. On the other hand non-vertebrates (including fungi and plants) show clear signals for two short hairpins, whereas box A appears in the loop region of M1. M2 is mostly perfectly conserved in structure and for phylogenetically close related organisms even in their sequence. In the case of invertebrates, binding energies are low, and the diversity of low energy structures computed by **RNASubopt** points at a very flexible region.
- (d) Kinetoplastids show drastically reduced U3 snoRNAs, which lack stem M2, as well as M6-M10.
- (e) The additional sequence data allow a rather clear distinction between the stems M5, M6, and M8.



Element		Metazoa		Fungi					Plants		Low Euk.	
new	old	Vert	Deut	Basi+ Taphr	Sacch	Euro	Dothi	Sorda +Leo	Strepto	Chloro	Kineto +Api	Others
M1	1a	•	•	•	•	•	•	•	•	•	•	•
M2	1b	•	•	•	•	•	•	•	•	•	•	•
M3	-	•	•	•	•	•	•	•	•	•	•	•
M4	5	•	•	•	•	•	•	•	•	•	•	•
M5	-	•	•	•	•	•	•	•	•	-	•	• <sup>‡</sup>
M6	-	•	•	•	-	•	•	•	•	•	•	-
M7	4	-	-	-	•	-	-	•	-	-	-	-
M8	2	•	•	•	•	•	-	•	•	•	•	-
M9	-	•	• <sup>†</sup>	•	•	•	•	•	•	•	•	-
M10	3	•	•	•	•	•	•	•	•	•	•	-
Intron		-	-	-	Sacc.sp.	✓	??	TNM	-	-	-	-

Figure 4.1: Secondary structure model of U3 snoRNA for eukaryots. Boxes A', A, C', B, C, D are indicated as boxes. ► indicates splice sites for subgroups of fungi. For details we refer to the given text. Vert – Vertebrata, Deut – Deuterostomes without Vertebrata, Basi – Basidiomycota, Taphr – Taprinomycota, Sacch – Saccharomycotina, Euro – Eurotiomycetidae, Dothi – Dothideomycetes, Sorda – Sordariomycetes, Leo – Leotiomycetes, Strepto – Streptophyta, Chloro – Chlorophyta, Kineto – Kinetoplastida, Api – Apicomplexa, Sacc.sp – *Saccharomyces sp.* only, TNM – *Trichoderma reesei*, *Neurospora sp.*, *Magnaporthe grisea*, † – not in Diptera, ‡ – might be also M6 are extremely short M8, ?? – Possible Intron.

### 4.1.2 Introns in U3 snoRNA genes

Introns in U3 snoRNA genes have been described in the literature for *S. cerevisiae* [49] and *H. wingei* [50]. Over all, the introns in the U3 snoRNA genes are evolutionarily very flexible. For example, there are *Kluyveromyces* species with and without introns [321]. Therefore, we examined all fungi U3 snoRNAs for introns and found a surprising absence/presence pattern, Fig. 4.2. For all *Saccharomyces sp.* we found an intron located as described previously at 14th nucleotide of U3 snoRNA, directly upstream of box A. For other saccharomycotina we found no intron. For sordariomycetes we found three genera with introns that are phylogenetically interspersed lineages without introns: *Trichoderma reesei*, *Neurospora sp.*, and *Magnaporthe griseae*. Introns are also present in all eurotiomycetidae except *Ascosphaera apis*. In *Uncinocarpus reesii* and *Coccidioides sp.*, the intron was located within the loop of M1 and thus within box A. In contrast, *Histoplasma capsulatum* and *Paracoccidioides brasiliensis* have the intron at the typical position, i.e., after the 14th nucleotide, just upstream of the A box. *Stagonospora nodorum* and *Alternaria brassicicola* might contain an intron at the 8th nucleotide (upstream of box A'), since there is a 5' and 3' splice site. Stem M1 can be formed with and without the the possible intron. Our data are insufficient to decide whether the U3 snoRNA in these to species is spliced or not.

### 4.1.3 Promoters of U3 snoRNAs

Metazoan U3 snoRNAs have snRNA-like promoters with a very well-conserved proximal sequence element (PSE). In several cases, there is also a canonical TATA-box, although most metazoans exhibit only a weak or no TATA-box. Closely related species may show differ in this respects: For instance, *Anopheles gambiae* and *Bombyx mori* have no TATA-box, which is present in *Apis mellifera*. Beside U3 snoRNA in higher plants showing a clear PSE element and TATA box, for *Chlamydomonas* and *Saccharomyces cerevisiae* neither of the two boxes were located. On the other hand for *Schizosaccharomyces pombe* a reasonable PSE and a clear TATA box were obtained. For details we refer to the Supplemental Material.

### 4.1.4 Multiplicity of U3 snoRNA genes

Many genomes contain multiple copies of the U3 snoRNA. Fig. 4.3 summarizes the data, for which species with poor genome assemblies and unassembled shotgun traces not taken into account. While the U3 snoRNA is frequently a single-copy

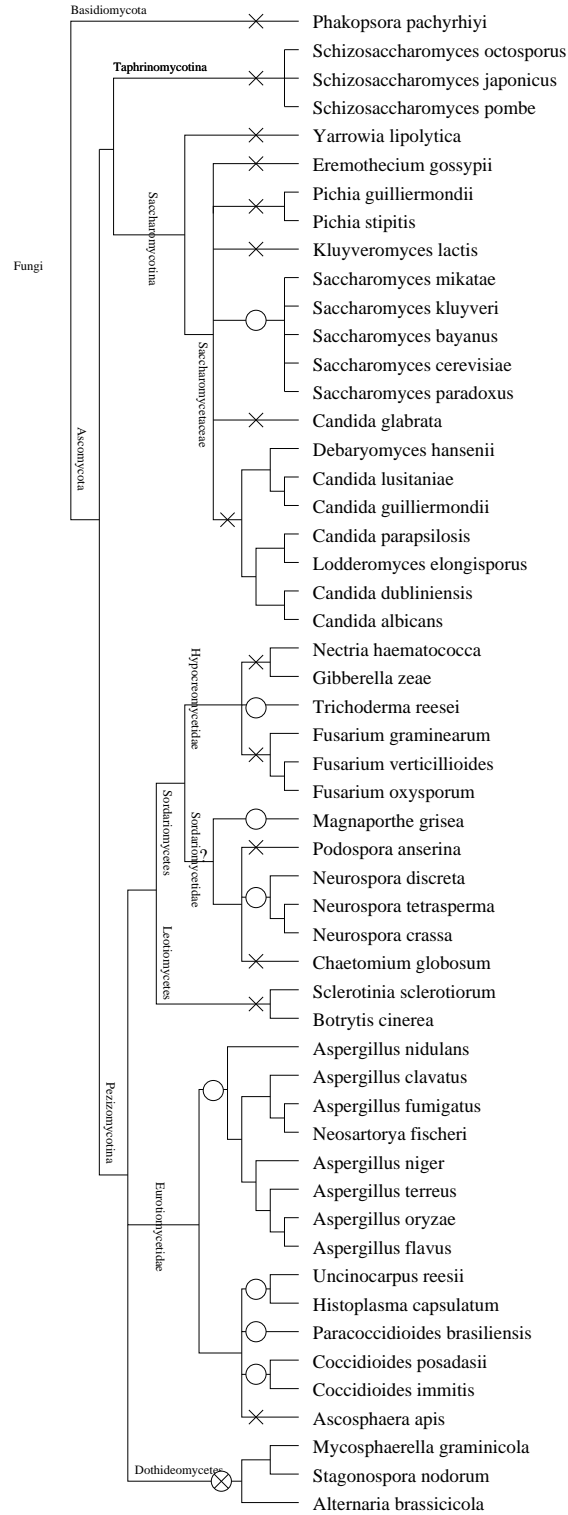


Figure 4.2: Overview of U3 snoRNAs found in fungi with (circle) and without (X) intron. For Dothideomycetes the absence/presence of introns is unclear. Phylogeny taken in combination from NCBI and [318–320].

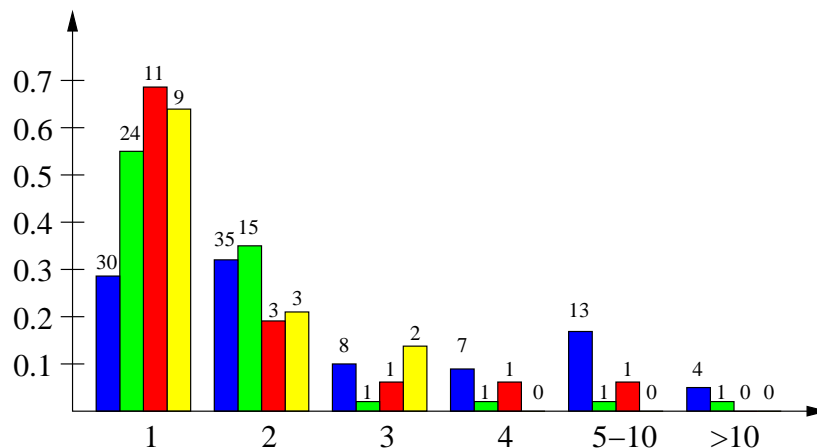


Figure 4.3: Distribution of copy numbers of U3 snoRNA genes. Metazoa (blue), Fungi (green), Plants (red), Other Eukaryots (yellow). Numbers above the bars indicate the number of genomes included in each data point.

gene, metazoa tend to have a few copies. No obvious paralog groups are recognizable suggesting that multiple U3 copies are subject to concerted evolution.

#### 4.1.5 Concluding Remarks

We have conducted a comprehensive survey of U3 snoRNAs. Our data confirm that U3 snoRNAs are (nearly) ubiquitously present in Eukaryota, although there are several basal lineages for which direct evidence is still missing. In particular, no credible candidate sequences were identified in *Giardia* and *Trichomonas*. Given the high variability of both sequence and secondary structure, however, we strongly suspect that our search methods were simply not sensitive enough. Experimental verification of some of the highly derived candidate sequences would extend the seed set and help to construct more general descriptors. In several cases, in particular many of the missing metazoa, the incompleteness of the currently available genomes is likely to blame for our failure to find a U3 homolog.

Secondary structure analysis shows a much larger structural variability than expected, with several lineage-specific expansion domains. This conforms recent surveys of other ncRNA families (telomerase RNA, RNase P and MRP, snRNAs, 7SK [51, 123, 235, 236]). It seems that drastic structural variations are an intrinsic property of ncRNA evolution.

## 4.2 RNase MRP and RNase P

Ribonucleases P (RNase P) and mitochondrial RNA processing (RNase MRP) are ribonucleoprotein complexes that act as endoribonucleases in tRNA and rRNA processing, respectively. Their RNA subunits are evolutionarily related and are involved in the catalytic activity of the enzymes. While it has long been known that RNase P RNA is a ribozyme in bacteria and several archaea, it was demonstrated only recently that eukaryotic RNase P RNA also exhibits ribozyme activity [322]. The main function of RNase P is the generation of the mature 5' ends of tRNAs. See [323] for a recent review of RNase P. In contrast, RNase MRP is eukaryote-specific. It processes nuclear precursor rRNA (cleaving the  $A_3$  site and leading to the maturation of the 5' end of 5.8S rRNA), generates RNA primers for mitochondrial DNA replication, and is involved in the degradation of certain mRNAs.

The phylogenetic distribution of P RNA clearly indicates that it dates back to the Last Universal Common Ancestor [324]. MRP RNA can be traced to the most basal eukaryotes [324] and apparently was part of the rRNA processing cascade of the eukaryotic ancestor [325]. The high similarity of P and MRP RNA secondary structures [326] and similarity of the protein contents and interactions of RNase P and MRP [323, 327] suggest that P and MRP RNAs are paralogs.

RNase P RNA is found almost ubiquitously. Interestingly, so far only MRP RNA has been found in plants including green algae, and red algae [324]. Whether the ancestral P RNA has been lost in these clades or possibly replaced by MRP RNA is unclear. It is also possible that the P RNA sequences are derived from each other that they have escaped detection so far. Despite the highly conserved core structures, P and MRP RNAs can exhibit dramatic variations in size, which mostly arise from large insertions in several "expansion domains" [324, 328]. In eukaryotes, additional P RNAs are often encoded in organelle genomes. Chloroplast P RNA is structurally similar to bacterial type A [329] and exhibits ribozyme activity [20]. Mitochondrial P RNAs, in particular those of fungi, are highly derived and exhibit only a small subset of the conserved structural elements shown in Fig. 4.4, mostly P1, P4, and P18 [330]. Despite its core function in tRNA processing, RNase P appears to be absent in the archaeon *Nanoarchaeum equitans*. Instead, placement of its tRNA gene promoters allows the synthesis of leaderless tRNAs [331].

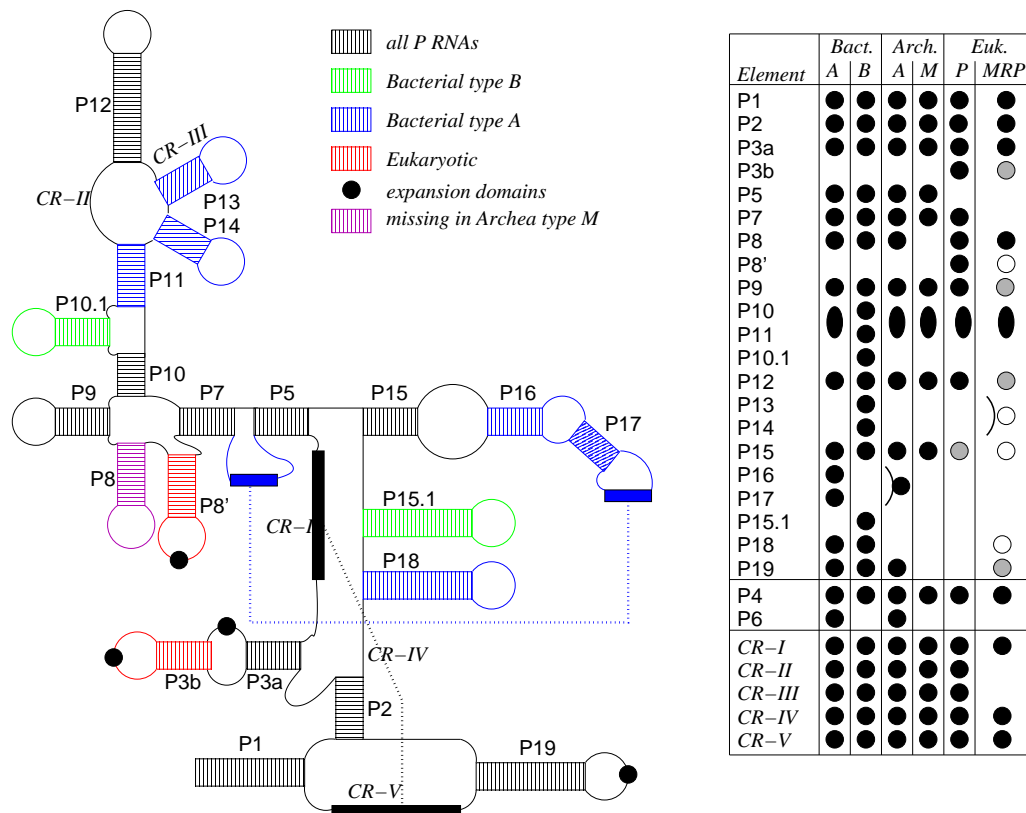


Figure 4.4: Schematic drawing of the consensus structures of P and MRP RNAs. Adapted from [323, 325, 332, 333]. The table indicates the distribution of structural elements. Black circles indicate conserved elements, stems indicated in gray are present in known sequences, open circles refer to elements that are sometimes present.

### 4.2.1 Homology Based Search

We were able to identify 85 RNase MRP's and 87 RNase P's within 98 available sequenced metazoan genomes. Beside elephants RNase P we easily identified based on sequence similarity both related genes in all examined deuterostomes and arthropods. Considering fully sequenced genomes only, we lack both possibly highly derived sequences in the trematode *Schmidtea mediterranea*, cnidarians *Acropora* and *Porites*.

Including eight plant RNase MRP sequences obtained by Blast against EST's of NCBI, we identified 35 plants sequences, of which only five were known recently. Although we know RNase MRP of *Brassica rapa* and *B. oleracea* these sequences were not detected. The reason might be that field and wild mustard are nowadays highly reared and even species genomes differ to an extremely high degree. We



miss higher RNase MRP of higher plants only: *Pinus*, *Triticum*, *Hordeum*, *Lotus* and *Malus*.

Beside some single species we were able to examine all sequences of RNase MRP in available genomes of basal eukaryotes. Recently, the absence of RNase MRP in kinetoplastids and diplomonads was described in [51], although a *Trypanosoma brucei* candidate is presented in [334]. We show alignments including predictions of these sequences in the supplemental material<sup>1</sup>.

Additionally, it was possible to recover most of fungal RNase MRP and RNase P sequences. However, they are not presented in multiple alignments yet, because of their complex structure. This work will be continued.

### 4.2.2 Secondary Structure

The overall secondary structure, absence/presence of single stems and comparisons of stem lengths in RNase MRP and RNase P is shown in Tab. 4.2.

The core pseudoknot P2 and P4 is structurally highly conserved in RNase MRP and RNase P. This part varies little in length similar to P1 represented in any organism with a stem length of about 9 bp. Stem P3a is similarly useful for secondary structure search with a constant length of 4-7 nucleotides in a constant distance to P2 and P4. Beside the bird *Taeniopygia gutatta* and most viridiplantae P3a is extended to P3b. This extension can range from 3 bp in *Petromyzon marinus* RNase MRP to Chlorophyta with 28 bp. For some eukaryotic RNase MRPs the extension of P3a is not a single stem but a forked stem. However the distribution of this P3b.1/P3b.2 system seems to be random. We find MRP sequences of this structure in all phyla: birds, nematocera insects, plants (*C. merlae*) and alveolata (*T. thermophila* and piroplasmida). Interestingly P7 is available in all RNase MRP and RNase P, except Metazoan MRP. Previously, RNase MRP was described to contain no P7. However, examining secondary structures of [324] carefully, P7 is available in all non-metazoan organisms, see Fig. 4.5-a. Considering the observation of P10/12 in RNase P located directly upstream of P7 it is possible to predict a more accurate secondary structure for P8, P9 and P10/12 (Fig. 4.5b-d). The latter stem is predicted with standard programs in numerous variations with a low minimum free energy difference. Therefore, additional information about a possible P7 is highly important to resolve the true secondary structure.

Stem P8' seems to be an invention in RNase P of dipteras only. P8 and P9 of

---

<sup>1</sup>[www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-023](http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-023)

Table 4.2: Structural overview of RNase MRP stems and their length. Indented organisms inherit properties of phyla they belong to, except separately specified values. For RNase MRP non-fungi eukaryotes are listed, for RNase P metazoans are described.

RNase MRP													
Phylum	P1	P2	P3a	P3b.1	P3b.2	P4	P7	P8'	P8	P9	P10-12	P15	P19
Deuterostomes	8-11	6	6-7	4-6	–	7-9	–	–	4	5-8	20-32	–	5-6
Birds					4	5							12-19
<i>T. gutatta</i>	7		5	–	–	4					15		
<i>C. milli</i>									7				15
<i>P. marinus</i>				3									
<i>C. intestinalis</i>													–
<i>O. dioica</i>	5			9									7
Insects	9-12	6-7	6-7	5	–	7-10	–	–	4	4-5	14-24	–	4-7;19-22
<i>A. pisum</i>				3									
<i>P. humanus</i>				12									
Nematocera					4-6								
Brachycera	16									13-21	27-33		
Nematods	7-9	4-7	7	5-6	–	7-8	–	–	4-6	4-5	4	–	–
<i>B. malayi</i>											28		
<i>T. spiralis</i>	12				6						20		
<i>A. suum</i>											17		4
<i>M. incognita</i>										17	12		10
Lophotrochozoa	7-10	4-7	6-7	5-6	–	5-8	–	–	4-5	4-5	19-24	–	5-8
Basal Metazoa	8-11	6-8	6-7	8	–	7-8	–	–	4-5	5	21-24	–	9
<i>N. vectensis</i>				4						9			14
<i>T. adhaerens</i>				10									
Viridiplantae	9-12	6-7	6-7	–	–	6-8	4-5	–	5	4	20-29	–	4-8
Rosids + Asterids			5	6									
<i>S. moellendorffii</i>											13		
Basal Plantae	12-17	6-7	6-7	6-19	–	7-8	3-5	–	4-5	4-5	22-25	–	11-13
Chlorophyta				17-28							15-18		5
<i>C. merolae</i>					16					–	32	28	
Alveolata	11-15	6-7	5-7	9-15	–	6-7	5-6	–	4-6	4	13-19	–	5-10
<i>T. thermophila</i>				5	5					–			–
<i>O. trifallax</i>											37		12
Sarcocystidae				28		9			13	16	23-25		13
<i>Cryptosporidium</i>										19-20	20-22		
<i>C. muris</i>									5	11	30-42		
Piroplasmida				3-4	7-9	9-10							
<i>T. parva</i>													
Haemosporida							3-4		8-9	12			
RNase P													
Phylum	P1	P2	P3a	P3b.1	P3b.2	P4	P7	P8'	P8	P9	P10-12	P15	P19
Deuterostomes	6-11	4-7	4-5	5-11	–	8-14	3-4	–	5	6-8	34-46	–	4-15
<i>C. milli</i>										3			
Amphibia											59-64		
Insects	10-12	5-7	5	5	–	6-11	4	–	5	5-6	27-35	–	5-16
Diptera								6-20					
Nematods	10-12	4-9	5	5-7	–	8-10	3-4	–	3-6	5-7	15-45	7	–
Lophotrochozoa	6-9	5-7	4-6	5	–	8-10	4	–	3-5	5-7	24-41	–	3-7
Basal Metazoa	6-10	6-8	4-6	5-8	–	5-10	3-4	–	3-7	5-6	31-39	–	4-8
Cnidaria										–			

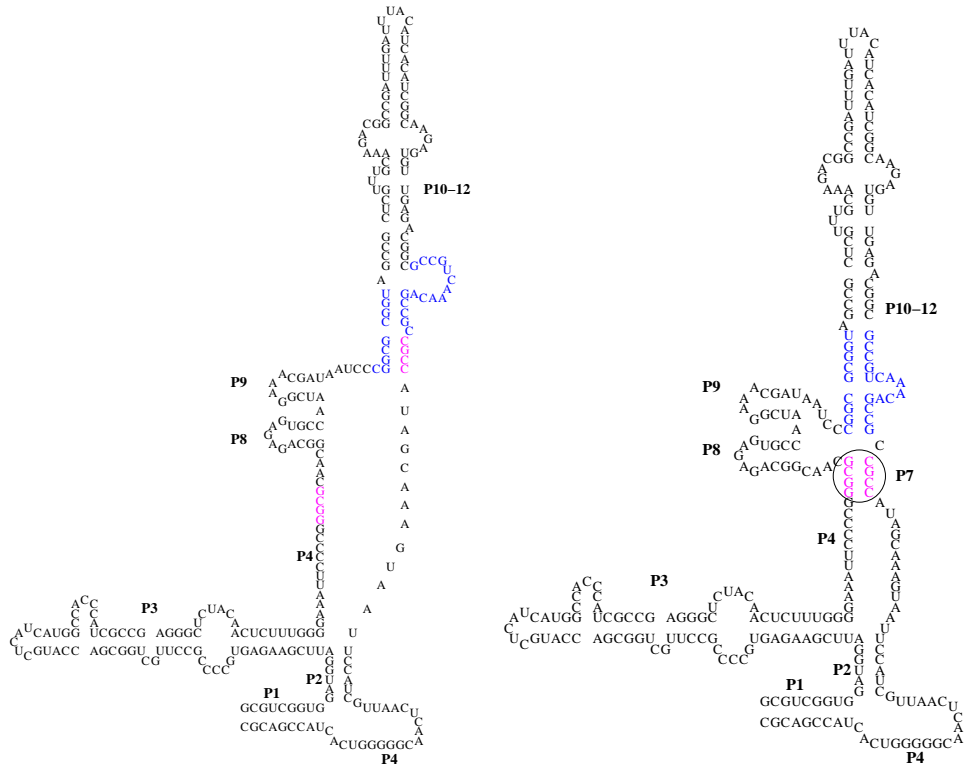
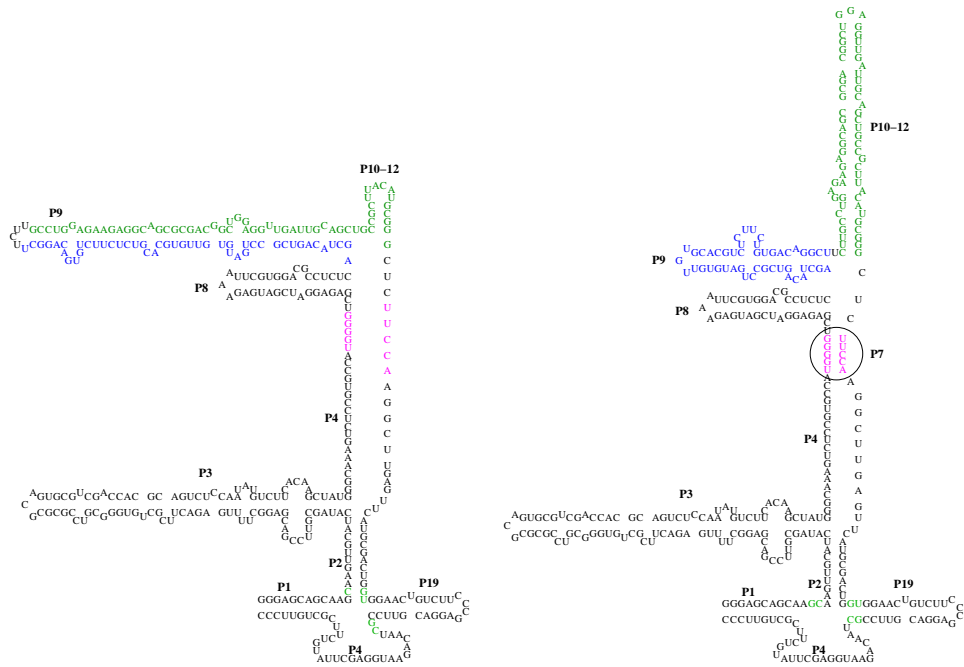
*Thalassiosira pseudonana**Toxoplasma gondii*

Figure 4.5: Secondary structure of RNase MRP obtained by alignments with related species containing P7. Coloured parts are remodelled. Left: Structure model of [324]; Right: Recalculated secondary structure model containing P7.

RNase P has a constant length, with the exception of a complete P9 deletion in cnidarians. However, RNase MRP increased the length of these stems up to fivefold. The most dramatic variations in sequence and structure are part of P10-P12. This stemlength varies from 4 nt (nematods MRP) to 64 nt (amphibians RNase P). In any organism P10/12 of RNase P is longer than its corresponding RNase MRP, usually by factor 1.5. An additional stem between P7 and P2 (P15 or P18) is available for RNase MRP in *C. merolae* and conserved for RNase P in nematods. Finally, stem P19 might be absent for both RNAs or differing in length from 3 nt to 22 nt.

### 4.2.3 Conclusion

Although the prediction of RNase MRP and RNase P and hence the calculation of multiple alignments is not finished yet, fundamental statements may be proposed. RNase P is much more conserved in structure, length of single stems and sequence as RNase MRP. This might be related with its substrate. Transfer-RNAs are very conserved in sequence and length across all known organisms compared to any other ncRNA. On the contrary, rRNAs show evolutionary variation spread all over the molecule.

Generally, genes for RNase MRP and RNase P are appear once per genome. For some very less genomes two copies were observed. Mostly multiple copies are recognized in genomes available on contig level. In a second scenario one copy is mapped to a specific chromosome and the second copy is located within an "Unknown Chromosome" as for *A. gambiae*. Only for *Gallus gallus* we did find two copies located directly next to each other within 500 nt, suggesting a tandem duplication.

Upstream regions (see Fig. 4.6) and poly-T termination signals of RNase MRP and RNase P are well defined and show once more a possible recruitment of Polymerase III.

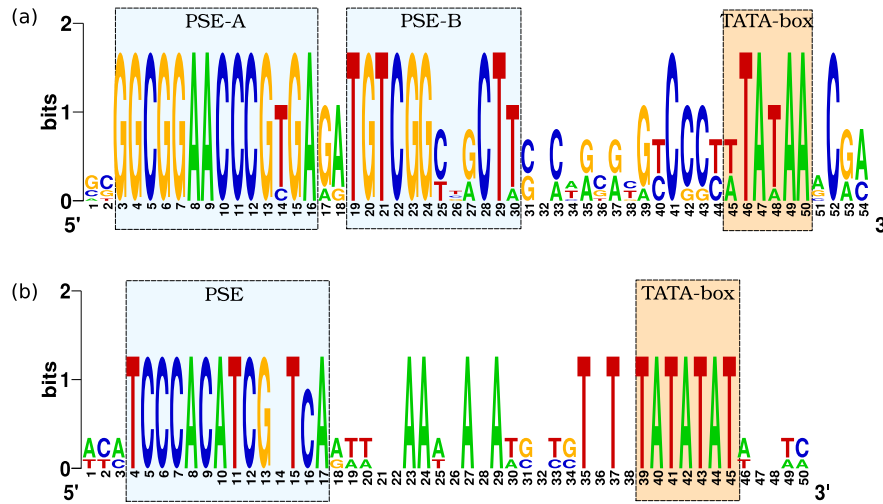


Figure 4.6: Promoter Region of RNase MRP, RNase P and recruitment signals upstream of other polymerase III transcripts, such as U6 snRNA and U6atac snRNA. (a) *Caenorhabditis remanei* upstream sequence including PSE-A, PSE-B and TATA-box, (b) only RNase MRP is predicted in *Arabidopsis thaliana*, with PSE and TATA-box.

### 4.3 7SK RNA

The 7SK snRNA is a highly abundant noncoding RNA in vertebrate cells. The Pol III transcript with a length of about 330nt [335, 336] is highly conserved in vertebrates [337]. Due to its abundance it has been known since the 1960s. Its function as a transcriptional regulator, however, has only recently been discovered. 7SK mediates the inhibition of the general transcription elongation factor P-TEFb by the HEXIM1/2 proteins (also known as CLP1, MAQ1, and EDG1) and thereby represses transcript elongation by Pol II [338–341]. Furthermore, 7SK RNA suppresses the deaminase activity of APOBEC3C and sequesters this enzyme in the nucleolus [342]. A highly specific interaction with LARP7 (La-related protein 7), on the other hand, regulates its stability [338–341, 343–345].

The sequence of the 7SK snRNA is extremely well-conserved across jawed vertebrates. In contrast, the sequence of the lamprey *Lampetra fluviatilis* is highly divergent [337], and invertebrate 7SK RNAs were recently found only using specialized sophisticated homology search techniques [236, 237]. The latter study made extensive use of the fact that the 7SK genes feature a canonical class-3 pol-III promoter structure [346]. Despite considerable efforts, phylogenetic distribution and evolutionary age of 7SK RNA remains uncertain because no homologs have

been found so far e.g. in basal metazoan lineages and in important invertebrate phyla such as Platyhelminthes and Nematoda.

Since 7SK RNA interacts specifically with HEXIM and LARP7, we survey here the phylogenetic distribution of these proteins to determine in which organisms we can also expect a 7SK gene. Since the primary interaction sites with HEXIM and LARP7 are among the few well-conserved features of the invertebrate 7SK snRNAs [237], we re-evaluate and refine the secondary structure model [347]. This in turn forms the basis for the detection of additional invertebrate 7SK RNAs.

### 4.3.1 Phylogenetic Distribution of HEXIM

Homologs of HEXIM were found across metazoan tree, using known HEXIM1 protein sequences and `tblastn`. In particular, we identified clear homologs in the poriferan *Reniera sp.*, the placozoan *Trichoplax adhaerens*, and the cnidarians *Nematostella vectensis* and *Hydra magnapapillata* implying that HEXIM was present in the metazoan ancestor. On the other hand, no homologs were detected in fungi, plants, and the choanoflagellate *Monosiga brevicollis*, suggesting that HEXIM is an animal innovation, Fig. 4.7. Full alignments are available in the Electronic Supplemental Material.

Eutheria are well known to carry two HEXIM paralogs [348]. Marsupials (*Monodelphis domestica*) have clearly recognizable orthologs of both HEXIM1 and HEXIM2. Our search identified a HEXIM2 in all mammals except platypus (*Ornithorhynchus anatinus*). On the other hand, Afrotheria (*Echinops telfairi*, *Loxodonta africana*) and Xenarthra (*Dasyurus novemcinctus*) do not have a copy of HEXIM1. We conclude that HEXIM was duplicated before the divergence of Metatheria and Eutheria, with secondary loss of HEXIM1 in some eutherian clades. Since the phylogenetic relationships of the major Eutherian groups are under intense discussion [349], it remains unclear whether the loss in Afrotheria and Xenarthra was independent, or whether these are sister groups whose ancestor already lost HEXIM1.

HEXIM1 and HEXIM2 are always located very close to each other (from  $\sim 10.000$ nt in *Canis familiaris* up to  $\sim 26.000$ nt in *Myotis lucifugus*) on the same chromosome (where sequence assembly allows such observations). Therefore, we propose that HEXIM2 likely derives from a duplication of HEXIM1. HEXIM2, as well as protostome HEXIM/CLP-1, contains a number of introns (conserved at least from mice to humans, whereas the HEXIM1 gene does not have any, suggesting that HEXIM1 derived from reverse transcription of HEXIM2. Comparing mammalian HEXIM1/2 proteins to HEXIMs of birds, frogs, and fish (Gnathostomes) a much

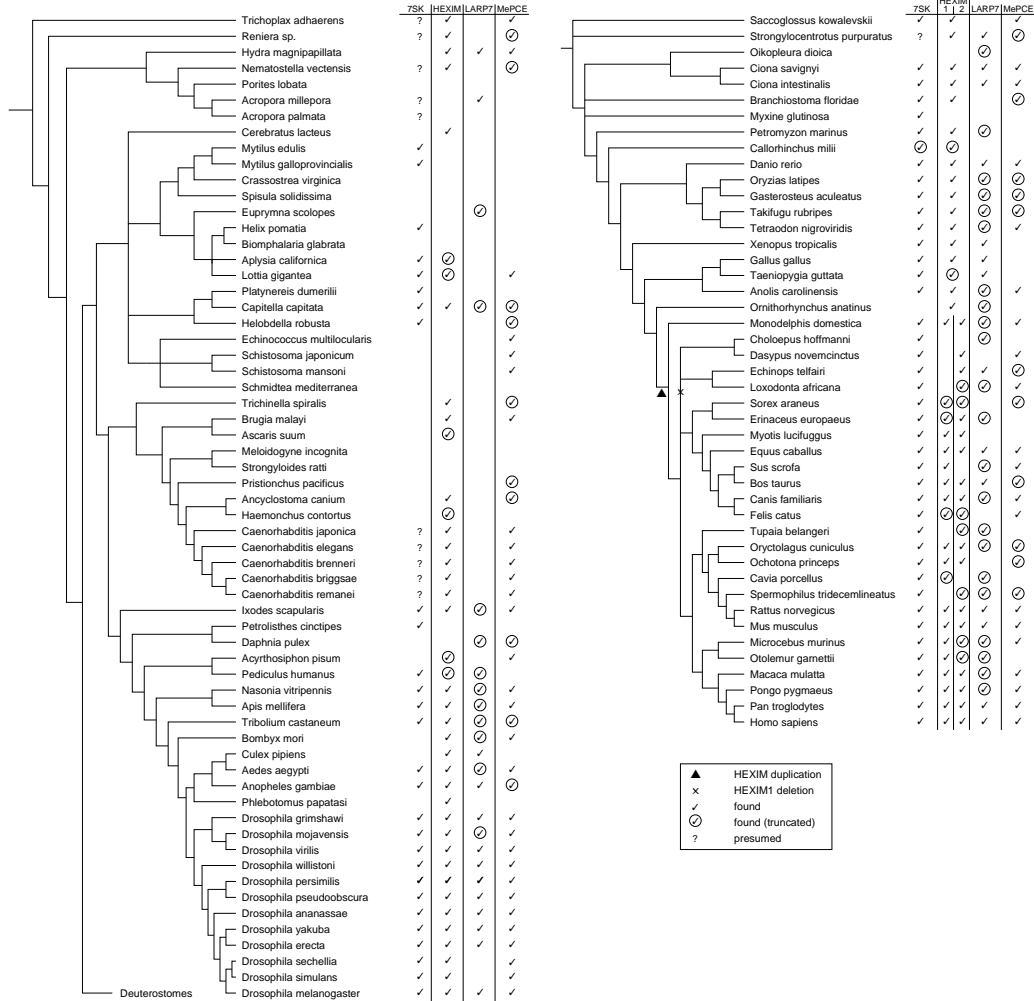


Figure 4.7: Distribution of HEXIM1/2, LARP7, MePCE/BCDIN3, and 7SK RNA. Findings of complete proteins and 7SK RNA respectively are indicated by a tick, incomplete findings by an encircled tick. Cases for which we are not sure whether we have found a true homolog or not are labeled with question marks. Missing data due to unsequenced organisms is marked by a black filled diamond. The black filled triangle indicates the HEXIM duplication event. The secondary HEXIM1 loss for Afrotheria and Xenarthra is marked with a cross within the taxonomic tree. The underlying tree is created from the NCBI taxonomy.

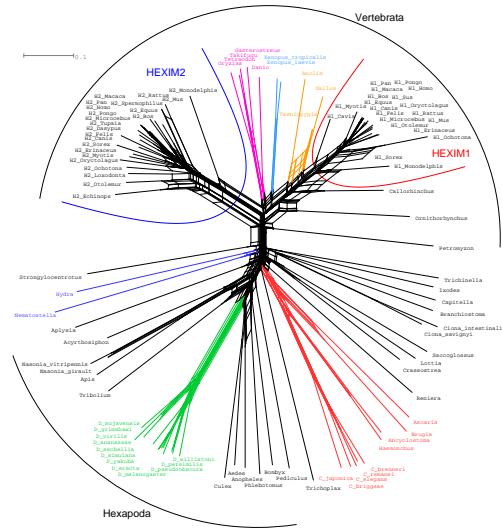


Figure 4.8: NeighborNet of all metazoan HEXIMs created with SplitsTree [2]. The major groups of HEXIM1 and HEXIM2, as well as the close relationship of the protostomia HEXIMs, are very well supported. The split, illustrating the closer relationship of HEXIM of fishes (magenta), amphibians (light blue), and birds (orange) to HEXIM1 of mammals is clearly identifiable. Additional well supported groups are those of nematodes (red), drosophilids (green), and cnidarians (dark blue). Due to their very basal position all other sequences can not be resolved any further. See supplement for ClustalW source alignment.

higher similarity of HEXIM to HEXIM1 is apparently observed. Both, sequence alignments and NeighborNet [2] analyses support this view, Fig. 4.8.

### 4.3.2 Phylogenetic Distribution of LARP7

A local `tblastn` search for LARP7 revealed its existence in all major metazoan phyla, including basal lineages such as porifera, placozoa, and cnidaria. LARP7 of protostomes and deuterostomes are clearly distinguishable and within each group very excellently alignable. The La-domain (PFAM **PF05383**) is located at the C-terminus and the RNA recognition motif, type 1 (RRM1, PFAM **PF00076**) is located downstream of the La-domain. Unambiguous LARP7 homologs were found in species in which we also found 7SK and/or HEXIM, Fig. 4.7, including *Hydra magnapapillata* and *Acropora millepora*. Although LARP7 has a complex gene structure, we provide carefully constructed alignments with sequence and genomic locations of all organisms in the supplemental material. LARP7 is distinguishable to other known LARP families by its La domain (compared to LARP1,2,4,5,6) and has a clearly recognizable RRM1, which is unknown in Larp1,2,6 and poorly conserved in LARP4,5. LARP7 has similarities to LARP3 (=Sjogren syndrome



antigen B (SSB)), containing a similar La-domain, however has a slightly different RRM3 instead of RRM1 (PFAM *PF08777*). In nematods we were able to identify sequences with a LARP7 La domain and an RRM3, which might have taken over the duty of RRM1. Alternatively we detected LARP3. The reason why we were able to identify LARP7 unambiguously in some lophotrochozoans only, is mostly the assembly status on contig layer. LARP7 can be distributed over 10 000 nt (*Anolis carolinensis*), interrupted by various introns.

### 4.3.3 Revised Secondary Structure Model of 7SK RNA

A complete alignment of all 79 known 7SK can be found in the Supplemental Material, unclear candidates are excluded. The expanded collection of sequences provides sufficient information for the construction of a global multiple sequence alignment. In contrast, previous studies [236, 237] were content with local alignments of the best-conserved regions, Fig. 4.9. Based on the new fully alignment, a much more comprehensive consensus structure model can be derived, Fig. 4.10.

A comparison of the structure proposed for the human 7SK RNA based on chemical probing [347] shows that most of our structure model is consistent with the previous proposal. There are, however, several novel features that provide new insights of the function of 7SK RNA. Most parts of the stems M3, M5, M7, and M8 were described previously and correspond to the stems 1, 3, 5, and 6 of [347], Fig. 4.11. Our re-evaluation of the invertebrate data demonstrates that these stems are conserved and can be identified in all organisms.

1. Stem M1 is the best-conserved feature of 7SK RNAs. It is recognizable in all known homologs [236, 237]. Corresponding to stem 1 of [347], it contains the HEXIM binding site, an absolutely conserved helical region with the sequence `GATC:GATC`.
2. The additional stem M4 is highly conserved at the structural level in all organisms, although there is not recognizable sequence similarity.
3. Stem M8, corresponding to stem 6 of [347], is also very well conserved. Therefore this might harbour the LARP7 binding site.
4. Drosophilids have an expansion domain between M4 and M5, which forms a stem-loop structure covering about 90nt. They also have an extended loop M4.
5. M5, corresponding to stem 3 of [347], is not only conserved in its structure but also in its sequence, see Fig. 4.12, which contains the motif `CGNNGC`

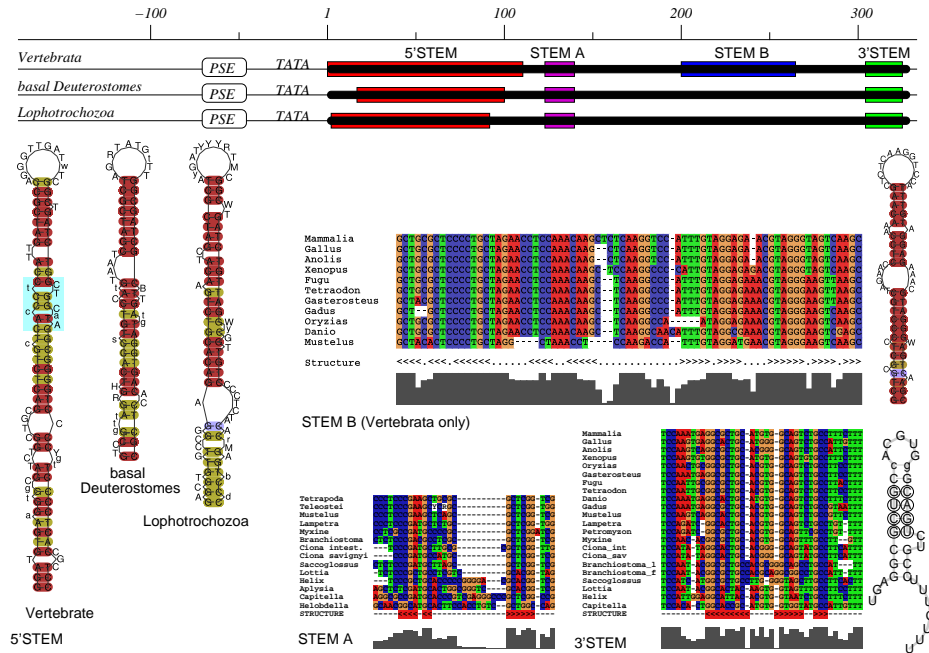


Figure 4.9: Common structural elements of 7SK snRNAs. The top panel schematically compares the location of upstream elements and RNA secondary features. While the structure of the 3' stem is common to all 7SK snRNAs (except for the elongation of the stem by a GC pair in *Branchiostoma* and *Saccoglossus*), there are substantial clade-specific variations in the 5' stem. A common structure, stem B, in the “middle region”, on the other hand, can be found only in vertebrates. With the exception of marginal differences in the small region marked in the vertebrate 5' stem, our consensus model is in complete agreement with previously published structures of vertebrate 7SK snRNAs [338, 347]. Conserved nucleotides in stems are shown in red; ochre color (and circles in the 3' stem, resp.) indicate consistent and compensatory mutations [236].

pairing with GCNNCG in all known 7SK RNAs. The M5 stem is slightly shorter in deuterostomes compared to other metazoa.

6. Most species have an additional stem, M6, located between M4/M5 and M7. It is missing, however, in many insects (drosophilids, *Tribolium*, and *Pediculus*) and in the two *Ciona* species. The absence of conserved sequence motifs suggests that it does not specifically interact with other molecules.
7. The most interesting part of the structure in M2, region 15-25 nt in length located between stems M1 and M3. Surprisingly, it can form three distinct structural alternatives in all known cases, as shown in Figs. 4.10,4.11,4.14.

M2a It can form a local hairpin. This local hairpin is much smaller in vertebrates.

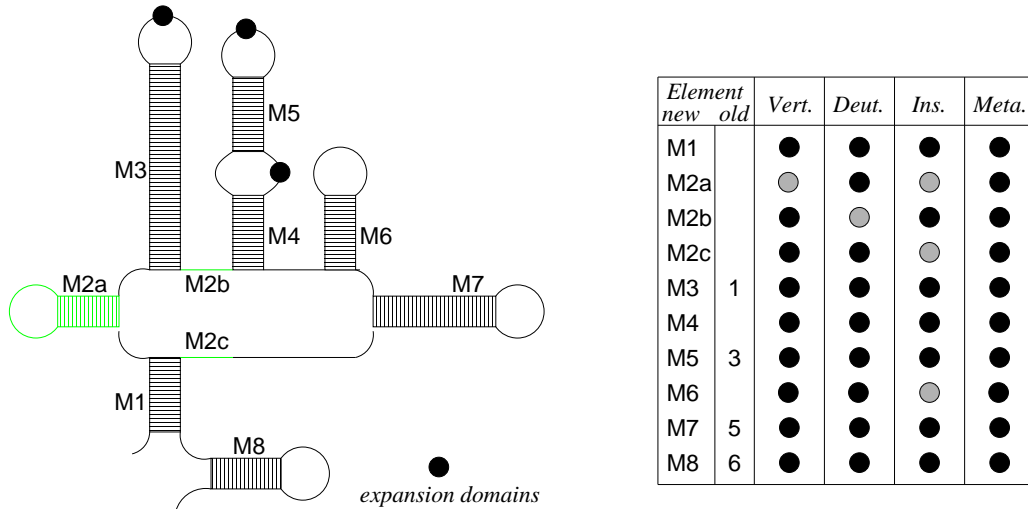


Figure 4.10: Revised secondary structure model of 7SK RNA. M1 to M8 are introduced here. M3 refers to hairpin 1, M5 to stem 3, M7 to stem 5 and M8 to last stem of Wassarman and Steitz [347]. M1 was published recently in Gruber et al. [236]. M6 is not present in *Drosophila*, however this species shows an expansion domain between M5 and M4. Stem M2 has one of three possibilities to basepair. M2a: M2 builds a hairpin as drawn in the picture, which is rudimentary present in vertebrates and absent in *Drosophila sp.* M2b: M2 binds downstream of M3, constructing an extended M3 stem, which is absent in *Ciona*. Therefore in *Ciona* M2 binds upstream to M1 (M2c) extremely well. This extension form of M1 is absent in some insects.

M2b It can binding downstream of M3, as published previously [237], resulting in an extension of stem M3. *Ciona* is the only case in which this structural alternative seems to be absent.

M2c It can bind upstream of M1, resulting in an extension of M1.

The conservation of this flexible arrangement suggests that re-folding the M2 region between the three structural alternatives is part of the core functionality of 7SK, i.e., that 7SK RNA is an RNA switch.

#### 4.3.4 Homology Search for 7SK snRNAs

Due to the high sequence conservation across jawed vertebrates, the 7SK genes of newly sequenced genomes such as *Tupaia belangeri*, *Equus caballus*, *Tribolium castaneum*, *Acyrtosiphon pisum* were easily retrieved by `blast`. Additionally via NCBI-`blast` partially 7SK sequences were obtained for *Platynereis sp.* **CT030666** (EMBL), *Mytilus edulis* **AM880723**, *Mytilus galloprovincialis* **EH663179.1**, *Petrolisthes cinctipes* **CAYF7296.g3**.

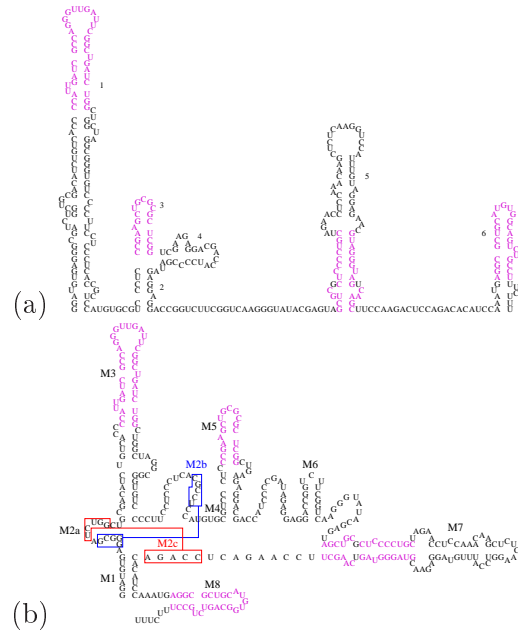


Figure 4.11: (a) Predicted human 7SK RNA of Wassarman and Steitz [347]. (b) Revised prediction of human 7SK. Equally structure parts are coloured purple. Stem 5 of Wassarman and Steitz and M7 show similar energies (RNAfold: -20.32 (stem 5) and -22.00 kcal/mol (M7)).

Homology search was performed by `blast`, `GotohScan` [61], and `Fragrep` [350]. Based on the experience with these approaches, and the previously known 7SK snRNAs, we constructed a specialized automaton to recognize 7SK RNAs.

It combines four separate `rnabob` searches of the target genome, Fig. 4.13, and requires some target-specific training.

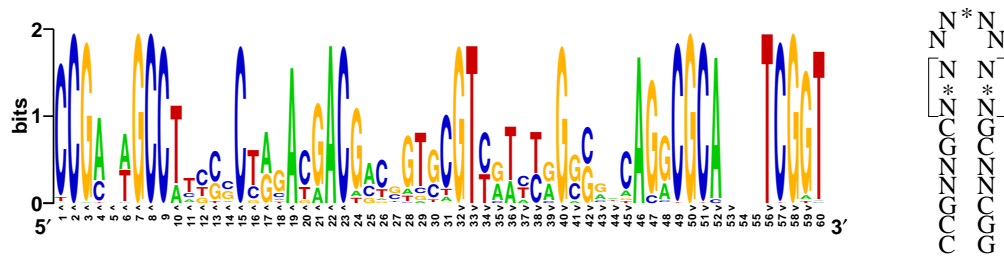


Figure 4.12: Left: Consensus sequence of M5 created by `Weblogo` [74] and expanded by corresponding basepairings. Nucleotide 1 to 8 and 50 to 59 conserved for all 74 known 7SK sequences including lophotrochozoans, arthropods, and deuterostomes. Base pairing from nucleotide 10 to 45 is observed in drosophilids only. R.h.s.: General overview of M5. Brackets indicate stem extension of insects.

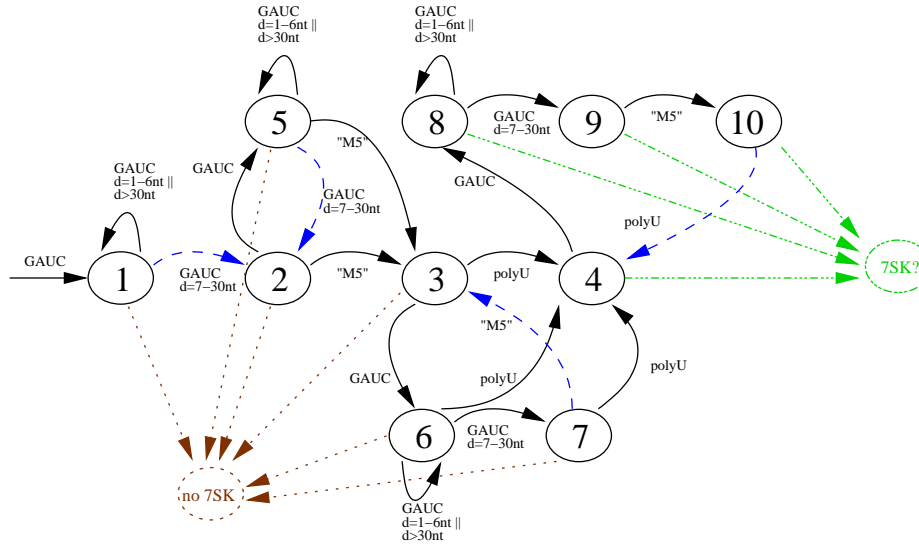


Figure 4.13: 7SK-automaton. **rnabob**-hits for GAUC, "M5" and poly-U within the whole genome or 500nt downstream of potential polymerase III candidates were separately searched. The automaton searched for a correct order and distances between these motifs and discarded all entries if state 4 is not reached with a distance  $d > 500nt$  (brown, dotted) between last GAUC-pair in correct distance (blue, dashed) and actual motif. If all motifs in a possible distance and order are obtained (state 4 was reached) and the actual motif has a distance  $d > 500$  the candidate is assumed to be a potential 7SK candidate (green, dash-dotted).

- **Promoter search.** Promoter sequences were obtained by aligning the 100nt upstream flanking sequences of pol-III transcripts (U3 snoRNA, snRNA U6, snRNA U6atac, RNase MRP and RNase P). For the search in Nematodes, for instance, we used the *C. remanei* PSE motif `GGCGGAACCCGnnnnnTGTCGG`, allowing three mismatches, and searched the UCSC rhabditina alignment, obtaining 92 hits. The 500nt downstream of these hits were extracted and passed to next stage.
- **GATC search** locates the highly conserved pattern GATC.
- **poly-T search** locates stretches of 5 thymidines within in 7nt, which might constitute a terminator signal.
- **Stem M5 search** searches for a GC-rich stem-loop that could constitute stem M5.

The hits obtained in steps 2-4 are sorted by location filtered w.r.t. distance constraints and secondary structure constraints as summarized in Fig. 4.13. In particular, there needs to a stem-loop not more than 20 nt upstream of the terminator, and two of the GATC hits must form an additional hairpin.

In order to assess candidates, we then attempted to incorporate them into the sequence/structure alignment described below. In addition, the promoter regions were compared with those of other known pol-III transcripts of the same organism, in particular U3, U6, U6atac, RNase MRP, and RNase P RNAs.

**Nematoda.** Using the promoter-based approach, we obtained a hit in *C. briggsae* that warranted detailed analysis, Fig. 4.14. The sequence is well-conserved across the genus *Caenorhabditis*. Although it is significantly shorter than other 7SK RNAs, it bears the hallmarks of a true 7SK homolog: (1) M1 is structural highly conserved. (2) It can form all three alternative helices M2a/M2b/M2c (3) M3 contains the highly conserved GATC sequence. (4) M5 is usually a GC-rich stem, (5) A stem-loop structure precedes the poly-T indicative of a pol-III terminator.

**Vague Invertebrate Candidates.** Using the same methods as for nematods we search additionally for all invertebrates with a predictable promoter region and obtained four more 7SK candidates. However these sequences lack at least one of the seven described features. The promoter *Trichoplax adhaerens* is very clear recognizable with RNase MRP, RNase P and snRNA U6 and U6atac. Only 65 regions distributed other the whole genome were found with at most 3 pointmutations. These candidates were observed in detail for 7SK. With this method we provide in the supplemental material vague candidates for *Reniera sp.*, *Nematostella vectensis*, and *Strongylocentrotus purpuratus*. If the latter candidate is a real 7SK it diverged drastically from other deuterostomes. M1 would have an unexpected low MFE, M3 changed in sequence and structure and M4-M7 is not conserved to other deuterostomes. On the other hand the ultraconserved GATC:GATC basepairing is present and a typically polymerase III terminator (poly-T) directly after a hairpin with the proper length of 8nt was found.

### 4.3.5 Discussion

**Phylogenetic Distribution of BCDIM3/MePCE.** Methylphosphate Capping Enzyme (MePCE) is known to be the capping enzyme of 7SK [351]. It was identified by homology based search with BCDIN3 (bicoid-interacting protein 3) of *Drosophila*. Additionally BCDIN3 was identified computationally in plants *Arabidopsis thaliana* and fungi *Schizosaccharomyces pombe*, but not in *Saccharomyces cerevisiae*. Additionally, we were able to identify this protein in *Laccaria bicolor* (*XP\_001879607*).



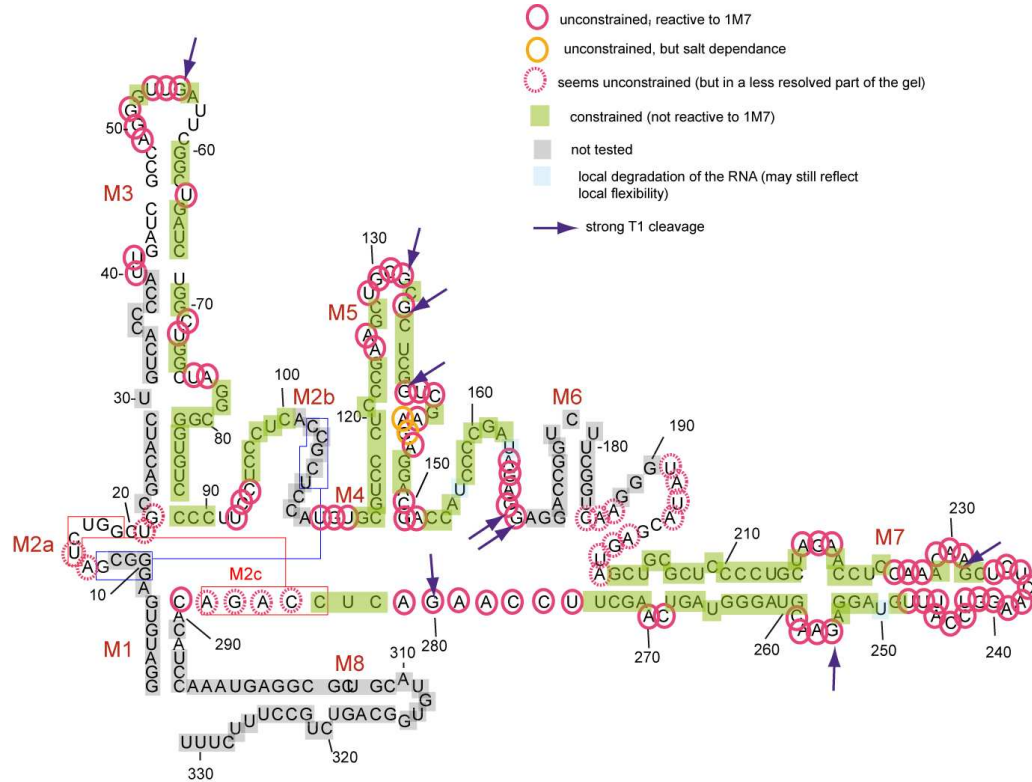


Figure 4.15: Accessibility of single nucleotides in human 7SK RNA, experimentally verified by our collaborators Denise Martinez and Anne-Catherine Dock. Accessible nucleotides are denoted by red circles. Green boxed nucleotides show no reaction to 1M7, due to the fact of basepairing to another part of 7SK RNA or a possible interaction to specific proteins. Secondary structure calculated in this thesis.

the secondary structure properties described above leave little doubt, that we fished the correct 7SK. The *C. elegans* sequence will be verified by Northern Blots of our collaborator Olivier Bensaude.

**Human 7SK RNA.** The human 7SK RNA 3D-structure will be determined at the moment by our collaborators Denise Martinez and Anne-Catherine Dock. For most nucleotides they were able to assign their accessibility (Fig. 4.15).

The basal part of stem M7 shows an expected picture, inaccessible nucleotides (green) are paired to each other, and accessible nucleotides (red circled) are part of internal bulges. The apical part of M7 stem in Fig. 4.15 shows a high possibility of occurring unpaired. Here a larger part, than the five nucleotides of the loop, seem to interact with some other molecule. Analogous, M3 seems to be predicted correctly. Interesting parts like M1, M6 and M8 are not experimentally verifiable



at the moment, however this problem will be addressed by Denise and Anne-Catherine currently. Stem M5 seems to be questionable. Although this part is conserved through all organisms, dehiscent red circled nucleotides, the argument of this stem is, from the computationally point of view, not holdable. Finally, the secondary structure and tertiary structure will need some future experimentally assessment.

## 4.4 Telomerase RNA

In contrast to the circular genomes of prokaryotes, eukaryotes have linear chromosomes. Special mechanisms are necessary to replicate the chromosome ends, the telomers. In almost all species investigated to date, a telomerase enzyme maintains telomere length by adding G-rich telomeric repeats to the ends of eukaryotic chromosomes. Telomerase thus dates back to the origin of eukaryotes. Notable exceptions are diptera including *Anopheles* and *Drosophila*, which use retrotransposons or unequal recombination instead of a telomerase enzyme.

The core telomerase enzyme consists of two components: an essential spliceosomal maturated RNA component [352], which serves as template for the repeat sequence, and the catalytic protein component telomerase reverse transcriptase (TERT). The RNA component varies dramatically in sequence composition and size. Although dozens of telomerase RNAs (usually called TR in vertebrates and TLC-1 in yeasts) have been cloned and sequenced, the known examples were recently restricted to four narrow phylogenetic groups: vertebrates, yeasts, ciliates, and plasmodia. The protein component TERT on the other hand is known in a much wider range of eukaryotes: Invertebrates (nematodes, insects, basal deuterostomes), Fungi (pezizomycotina), Plants, Algae, Kinetoplastids and Basal Eukaryots [353].

Yeast telomerase RNAs appear to be even less well conserved: In [354], only seven short sequence motifs are reported within more than 1.2kb transcripts of *Kluyveromyces* species, and of these only a few are partially conserved in *Saccharomyces*. In fact, *Saccharomyces* and *Kluyveromyces* TLC-1 genes cannot be aligned with each other by standard alignment programs. The same is true for the recently discovered TLC gene of *Schizosaccharomyces pombe* [355, 356]. Yeast snRNA and snoRNA methyltransferase Tgs1 is responsible for TLC1 methylation. The absence of Tgs1 causes changes in telomere length and structure, improved telomeric silencing and stabilized telomeric recombination. [357].

The small ciliate TR genes include a pseudoknot domain that contains an unusual triple-helical segment with an AUU base triplet. This domain is also shared by the vertebrate and yeast telomerase RNAs [358]. Whether such a structure is also present in the computationally predicted TR genes of plasmodia [359] is not yet known.

Although there is a common core structure of all these telomerase RNAs [360], and despite their length of several hundred to almost 2000nt, these RNAs remain a worst case scenario for homology search on sequence and structural level. Indeed,

a survey of vertebrate telomerase RNAs [234] shows dramatic sequence variation with only a few, short, well-conserved sequence patterns separated by regions of highly variable length. The recent discovery of the TR genes of teleost fishes [235] highlights the variability of this molecule, which has acquired several lineage-specific domains, such as the snoRNA domain in vertebrates and the Ku80 binding domain in budding yeast, see Figure 4.16.

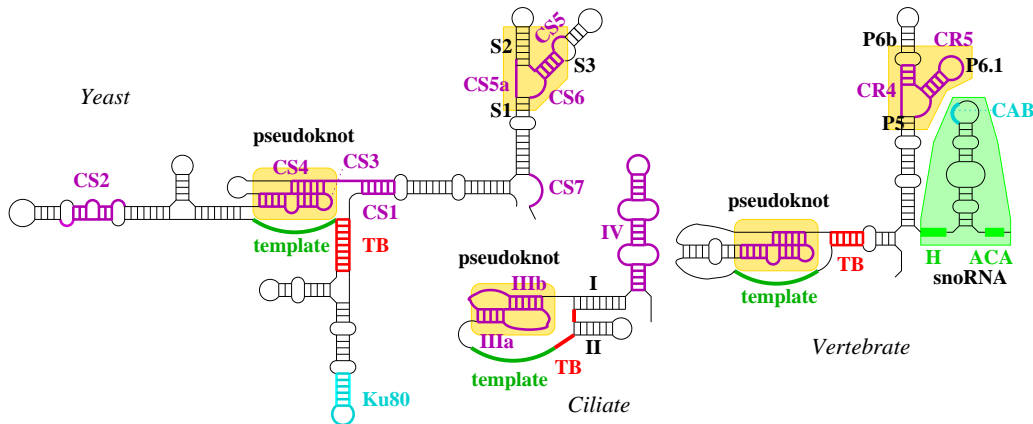


Figure 4.16: Telomerase RNA structures of yeast and human share the topology of the pseudoknot region and a functionally important junction region (S2/S3 and P6b/P6.1 respectively). The template and its boundary element (TB) are highlighted. The yeast structure is a consensus of *Saccharomyces* [361, 362] and *Kluyveromyces* structures [363]. The Ku80 binding domain is specific for *Saccharomyces*. Vertebrate telomerases have a snoRNA domain [364] at their 3' end. This domain carries a Cajal-body localization signal (CAB) [365], which is present in all vertebrates except teleosts [235]. Black regions may vary dramatically in length.

#### 4.4.1 Homology Based Search

Among vertebrates we were able to identify 16 additional sequences with Blast: *Pongo pygmaeus*, *Macaca mulatta*, *Microcebus murinus*, *Spermophilus tridecemlineatus*, *Ochotona princeps*, *Myotis lucifuggus*, *Sorex araneus*, *Echinops telfari*, *Canis familiaris*, *Erinaceus europaeus*, *Procapra capensis*, *Vicugna vicugna*, *Pteropus vampyrus* and partial sequences for *Pan tropicalis*, *Otelemur garnetti* and *Loxodonta africana*.

For *Monodelphis domestica* the telomerase RNA looks very similar to other mammalian sequences, however it has a huge insert between P1 and P4, which forms a stable hairpin.

Alignments in Stockholm format are available at the supplement material<sup>2</sup>.

<sup>2</sup>[www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-022](http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-022)

Table 4.3: Expected number of the template CCCUAA with a length of 8 on both strands in examined organisms.

Organism	Genome Size (bp)	Expected Frequency	Obtained Frequency
<i>S. purpuratus</i>	809 952 877	148 307	170 820
<i>C. instestinalis</i>	141 233 565	25 861	22 330
<i>C. savignyi</i>	255 955 828	46 867	82 776
<i>N. crassa</i>	1 860 657 949	340 697	342 708
<i>N. discreta</i>	556 883 022	101 968	183 461
<i>N. tetrasperma</i>	487 800 222	89 319	133 339

#### 4.4.2 Pipeline for Prediction of Divergent Telomerase Candidates

For detection of non-vertebrate telomerases we developed a pipeline consisting of three steps: (1) Generating Candidate Set (2) Filtering Candidate Set (3) Scoring Candidate Set.

##### Candidate Set Generation

Comparing all known telomerases (ciliates, fungi and vertebrates) no conserved sequence motif is detectable. The structure of all these telomerase RNA sequences is highly variable (Fig. 4.16). The only common structural feature is a pseudoknot of different length located 5-prime of the template in variable distance. However as yet no reliable pseudoknot-finding program exists.

How to find a gene without knowing anything about its length, sequence or structure?

The telomeric repeat region of specific organisms is known [353], the repeat varies from 5nt (insects) to 25nt (*Candida*). For the organisms examined here (*Neurospora sp.*, *Strongylocentrotus centrotus* and *Ciona sp.*) the telomere sequence is 5'-TTAGGG-3'. The reverse complement is believed to be part of telomerase RNA. However the rotation is unknown, therefore within a genome of 809MB size, the template with a length of at least 8 nt is expected to occur on both strands 2 million times by chance, however for *Strongylocentrotus purpuratus* just 170 000 hits were obtained (Tab. 4.3).

### Filter Candidate Set

Each sequence of the candidate set included the template region and 500nt downstream. This set was filtered by the following criteria:

- (a) Candidates with three or more telomere sequences (templates) after each other were removed. From known telomerases we learnt that the template sequence occurs at most 1.5 times. Candidates with more than three templates are believed to be repeats.
- (b) Candidates with more than 10 unknown neighbouring nucleotides (N) were removed.
- (c) Sequences shorter than 100nt were removed. Due to technical reasons: the following filter steps and scoring steps could not be computed, as described below.
- (d) Identical or highly similar sequences were identified by `blastclust` and removed. Most genomes are assembled on scaffold or contig level, consequently multiple copies should be removed.
- (e) Potential protein sequences were removed. This step was performed with `blastx` against all known proteins from NCBI.

### Score Candidate Set

The remaining candidates were scored based on the presence of a telomerase-like pseudoknot. We specifically designed and developed a program, `TR-PK-finder` to detect such pseudoknots. In case of *S. purpuratus* deep sequencing reads were available from Julian Chen. Distances between template and reads were additionally scored as well as the clustersize of the reads. If possibly, a H/ACA stem loop, as known for vertebrates within 1000nt downstream of the template was scored as well.

For a deeper understanding of the pseudoknot scoring function we introduce here `TR-PK-finder` before presenting parameters and results.

**TR-PK-finder:** We developed a specific telomerase pseudoknot finding program (`TR-PK-finder`) mainly based on different `rnabob`-search steps and `RNAfold`-folding steps.

The pseudoknot known from fungi and vertebrates is generally structured as in Fig. 4.17. All known telomerases have a common main structure (blue parts).



The interacting adenine-run consists of 5–11 adenines with at most two point mutations. Known TR sequences suggest that the guanine at the end of the adenine-run can not be substituted in this model. This nucleotide has to follow directly or with a 1nt bulge after the A-run.

Each sequence of the candidate set was allocated to a score-vector  $\mathfrak{M}$  as follows:

$$\mathfrak{M} = \begin{pmatrix} t \\ c \\ a \\ mfe_{rel} \\ hom \\ r_d \\ r_c \\ haca \end{pmatrix} \quad (4.1)$$

with

$$t = \begin{cases} 0, & \text{if } < 4/7 \text{ T in } s2 \\ 1, & \text{if } 4/7 \text{ T in } s2 \\ 2, & \text{if } 5/7 \text{ T in } s2 \\ 3, & \text{if } \geq 6/7 \text{ T in } s2 \end{cases} \quad a = \begin{cases} 0, & \text{if } \leq 4/9 \text{ A in } s6 \\ 1, & \text{if } 5/9 \text{ A in } s6 \\ 2, & \text{if } 6/9 \text{ A in } s6 \\ 3, & \text{if } \geq 7/9 \text{ A in } s6 \end{cases}$$

$$c = \begin{cases} 0, & \text{if no point mutation in } s4 \\ 1, & \text{if one point mutation in } s4 \end{cases}$$

$$mfe_{rel} = \frac{mfe_A + mfe_B}{mfe_C}$$

where  $mfe_A$  is an `RNAfold -C` for the first loop of the pseudoknot,  $mfe_B$  for the second loop and  $mfe_C$  for the whole molecule as described in Fig. 4.17 with A, B and C, respectively.

*hom* is the number of closely related organisms containing a hit by `blast`.

$$r_d = \begin{cases} 1, & \text{if } 200 < d < 300 \\ 2, & \text{if } 100 < d < 200 \\ 3, & \text{if } 50 < d < 100 \\ 4, & \text{if } 0 < d < 50 \\ 0, & \text{else} \end{cases}$$

where  $d$  is the minimal distance between template and reads.

$$r_c = \begin{cases} 1, & \text{if } 500 < c < 1000 \\ 2, & \text{if } 300 < c < 500 \\ 3, & \text{if } 50 < c < 300 \\ 0, & \text{else} \end{cases}$$

where  $c$  is the maximal cluster size.

$$haca = \begin{cases} 1, & \text{if } 500 < h < 1000 \\ 3, & \text{if } 200 < h < 500 \\ 2, & \text{if } 50 < h < 200 \\ 0, & \text{else} \end{cases}$$

where  $h$  is the distance from template to H-box (ANANNA) followed by an ACA-box within the next 200nt with the highest  $haca$  value.

### 4.4.3 Results

#### *Neurospora*

Starting with 342 708 initial candidates, 90% of the *Neurospora crassa* candidates were removed in the filtering steps. The remaining 17 516 were sorted by their scores  $\mathfrak{M}$  (Eq. 4.1). The best 500 candidates were sent to Julian Chen, who examines whether the true telomerase RNA is among the candidates.

#### *Strongylocentrotus*

For *Strongylocentrotus purpuratus* we obtained 23 870 sequences after the filter steps (13.9%). Additionally, we were able to adjust most of the parameters in



Eq. 4.1, since illumina deep sequencing reads of TERT-affinity enriched RNA samples were supported by Julian Chen. After scoring and sorting the subset of candidates, the best 100 sequences were send to the Julian Chen's wet lab.

### *Ciona*

For *Ciona intestinalis* we filtered and calculated 105106 candidates. Just 62 sequences are obtained by Blast (-e 1e-10) in *Ciona savignyi* in combination with available H/ACA-boxes. For this small number of hits we filtered for 200 to 400nt distance between template and H-box and 60 to 80nt between H-box and ACA-box. Currently, the remaining 62 candidates are analyzed in wet labs.



## Chapter 5

# NcRNA Screens in specific Organisms

A detailed annotation of non-protein coding RNAs is typically missing in initial releases of newly sequenced genomes. Here we report on a comprehensive ncRNA annotation of the genomes of *Trichoplax adhaerens* and *Schistosoma mansoni*, the presumably most basal metazoan whose genome has been published to date. Since **Blast** identified only a small fraction of the best-conserved ncRNAs — in particular rRNAs, tRNAs, and some snRNAs — we used a semi-global dynamic programming tool, **GotohScan**, to increase the sensitivity of the homology search. It successfully identified the full complement of major and minor spliceosomal snRNAs, the genes for RNase P and MRP RNAs, the SRP RNA, as well as several small nucleolar RNAs. We did not find any microRNA candidates in *Trichoplax* and vague microRNAs candidates in *Schistosoma* homologous to known eumetazoan sequences.

## 5.1 *Trichoplax adhaerens*

The phylum Placozoa consists of only one recognised species – the marine dweller *Trichoplax adhaerens*. Extensive genetic variation between individual placozoan lineages, however, suggests the existence of different species [366]. The phylogenetic position of the phylum Placozoa has been the subject of contention dating from the 19th century. Originally, Placozoa were regarded to represent the base of Metazoa, later they were seen as derived (secondarily reduced) with sponges being considered to be the most basal metazoans (see e.g. [141, 367] for overview and discussion). Most recently, a basal position among all diploblastic animals has been suggested [368].

*Trichoplax* lacks tissues, organs and any type of symmetry. It is composed of only a few hundred to a few thousand cells. This organism has a simple upper and lower epithelium, which surround a network of fiber cells, and as such has an irregular, three-layered, sandwich-type organisation. Only five different cell-types have so far been described; upper and lower epithelial cells, glands cells, fibre cells, and recently discovered type of small cells that are arranged a relatively evenly spaced pattern within the marginal zone, where upper and lower epithelia meet [369]. It is therefore among the simplest multicellular organism. With 106Mb, the nuclear genome of *Trichoplax adhaerens*, which has recently been completely sequenced [370], is among the smallest animal genomes.

So far, the non-coding RNA complement of Placozoa has not been studied. The genome-wide annotation of non-coding RNAs has turned out to be a more complex and demanding problem than one might think. While a few exceptional classes of RNA genes, first and foremost rRNAs and tRNAs are readily found and annotated by `Blast` and the widely used tRNA detector `tRNAscan-SE` [75], most other ncRNAs are relatively poorly conserved and hard to find within complete genomes. This is in particular true whenever the sensitivity of comparative approaches are limited by large evolutionary distances to the closest well-annotated genomes. The placozoan *Trichoplax adhaerens* is a prime example for this situation.

In this contribution we primarily report on a careful annotation of those *Trichoplax* ncRNA genes that have well-described homologs in other animals. In addition, we describe computational surveys for novel ncRNA candidates. For a subset of the annotated ncRNAs we verify expression to demonstrate that the predicted homologs are functional genes.

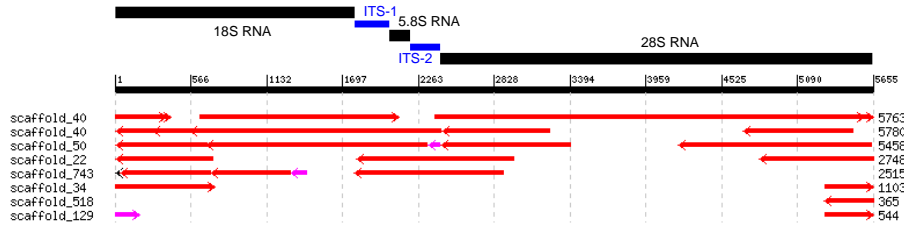


Figure 5.1: *Trichoplax* pre-rRNA cluster reconstructed from previously published sequences *L10828*, *Z22783*, *AY652578* (SSU), *AY303975*, *AY652583* (LSU), *U65478* (internal spacers and 5.8S) and *Triad1* genomic sequence. Blast hits of the pre-rRNA to the *Triad1* genome assembly are shown below as in the JGI genome browser.

### 5.1.1 Results

#### tRNAs

The *Trichoplax* genome contains 49 canonical tRNA genes, a single selenocysteine-tRNA gene and one tRNA pseudogene recognizable by `tRNAscan-SE`.

Interestingly, the *Trichoplax* genome is essentially devoid of tRNA-like sequences.

#### Ribosomal RNAs

In eukaryots, rRNAs (except 5S) are processed from a polycistronic “rRNA operon” which consists of SSU (18S), 5.8S, and LSU (28S) RNAs, two “internal spacers” ITS-1 and ITS-2, and two “external spacers”, reviewed in [371]. *Trichoplax* is no exception, see Fig. 5.1. The rRNA sequences have already received considerable attention in a phylogenetic context, see [366, 372–374]. The pre-rRNA sequence appears in several copies throughout the genome. Somewhat disappointingly, the *Triad1* assembly contains none of them in complete and uninterrupted form. The consensus sequence of the pre-rRNA can be easily constructed starting from the previously published sequences and the five fairly complete genomic loci (on scaffolds 22, 40 (two), 50, and 734) together with a partial copy on scaffold 34. Only the exact ends of the external transcribed spacers remain uncertain. Fig. 5.1 summarizes the `Blast` matches of the pre-rRNA to the *Trichoplax* genome.


The 5S rRNA sequence of *Trichoplax* has long been known [375]. The current genome assembly contains nine 5S RNA genes, one of which is a degraded pseudogene. Interestingly, there are three anti-parallel pairs (two head-to-head, and one tail-to-tail which contains the pseudogene).

## Spliceosomal snRNAs

Previously, nothing was known about placozoan snRNAs. With the exception of the U4atac, the snRNAs were easily found by **Blast**. The U4atac was found by **GotohScan** only. The expression of the U4atac was also verified experimentally, see [61] for details of methods. With the exception of two U6 genes, each snRNA is encoded by a single gene in the *Trichoplax* genome.

Their secondary structures, Fig. 5.2, closely conform to the metazoan consensus [123], with slightly shorter stems II of U11 snRNA and IV of U12 snRNA. The U12 contains an 5nt insert indicated in red in Fig. 5.2.

Table 5.1: Proximal sequence element (PSE) and location of snRNAs in *Trichoplax adhaerens*. The sequence-logo was generated using **aln2pattern** [350].

snRNA	Location	Sequence
U1	-58	. . . . . G . . . GG .
U2	-55	A . . . . . G . G . . . A . .
U4	-57	. . . . . A . . . . .
U5	-57	A . . . . . G . . . GC .
U6.1	-62	. . T . . . . AG . . . . .
U6.2	-62	. . T . . . . AG . . . . .
U4atac	-59	. . . . . AG . . . C .
U6atac	-63	. . . . . AA . . . . .
U11	-59	A . . . . . CA . . . C . G
U12	-60	. . . . . G . G . T . C . .
Sequence logo		
Consensus	-59	CCCATAATTRAAGNNA

In contrast to many other invertebrates, *Trichoplax* snRNAs feature a clearly recognizable proximal sequence element (PSE) see [93, 123], which is easily detected by **MEME** [91, 225], see Tab. 5.1. In line with other species, the PSE element is shared between the pol-II and pol-III transcribed snRNAs. On average the PSE elements differ by 3 nucleotides from the consensus.

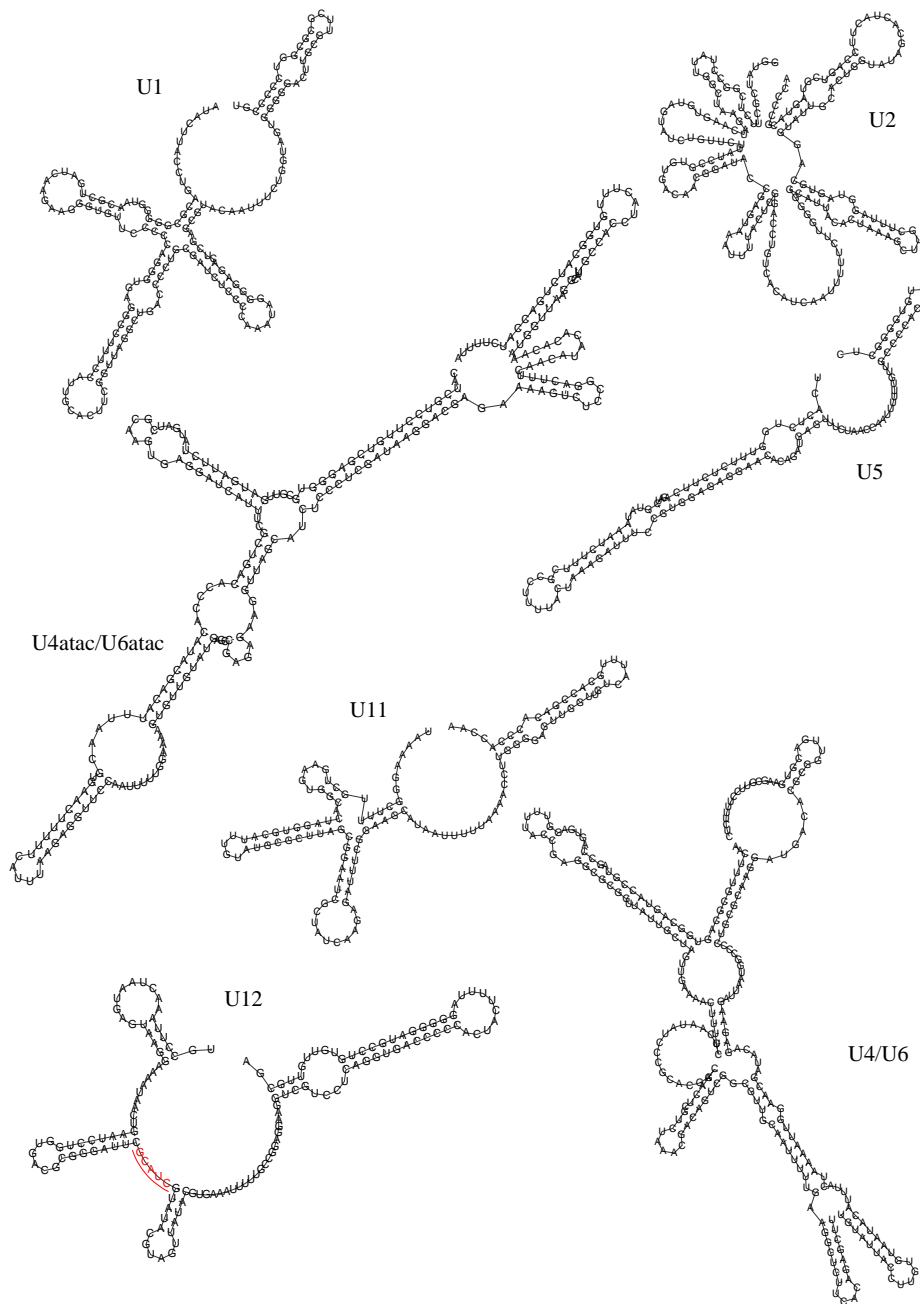


Figure 5.2: RNA secondary structures of major spliceosomal (U1, U2, U4, U5, U6) and minor spliceosomal (U11, U12, U4atac, U5, U6atac) snRNAs. For U4/U6 and U4atac/U6atac the interaction structures computed by means of *RNAcofold* are shown. The 5nt insert (relative to other metazoa) is highlighted in the U12.





the *Trichoplax* MRP candidates share the crucial features with both of them, leaving little doubt that we have indeed identified the true MRP sequence. Fig. 5.4 shows the homology-based secondary structure model.

The signal recognition particle (SRP) binds to the signal peptide emerging from the exit site of the ribosome and targets the signal peptide-bearing proteins to the prokaryotic plasma membrane or the eukaryotic endoplasmic reticulum membrane [376]. Its RNA component, called 7SL or SRP RNA, is well conserved and hence easy to identify by **Blast** comparison starting from the SRP RNA sequences compiled in the SRPDB [377]. The *Trichoplax* SRP RNA is shown in Figure 5.3.

### Small Nucleolar RNAs

The two classes of snoRNAs, box H/ACA snoRNAs and box C/D snoRNAs, are mutually unrelated in both their function (directing two different chemical modifications of single residues in their target RNA) and their structure, reviewed e.g. in [378].

The U3 snoRNA candidate sequence was easily verified by **Infernal**-alignment to the corresponding **Rfam** model, Fig. 5.3. Its expression was verified experimentally. Other snoRNAs were verified by Jana Hertel [61] with **GotohScan** and **snoReport** [82]. Candidates were aligned by hand to **Rfam**-alignments.

Table 5.2: Small nucleolar RNAs in *Trichoplax*. Target sites homologous to the ones in human rRNAs are indicated by an asterisk.

Name	Class	target	conservation	Note
U3	C/D	18S 5-22* 18S 1129-1140*	eukaryots	verified
U18	C/D	28S A740 *	eukaryots	
U36	C/D	18S A615 *	eukaryots	
U76	C/D	28S A1549 *	vertebrates	
U106	C/D	28S A2227?	vertebrates	
U17	H/ACA	†	eukaryots	
U71 ?	H/ACA	?	vertebrates	uncertain
sc.3857:- 103-213(-)	H/ACA	28S U1370 U1884	novel	

†The U17 snoRNA probably targets the 5'externally transcribed spaces (5'ETS), the exact target is still unknown, however [306, 379].

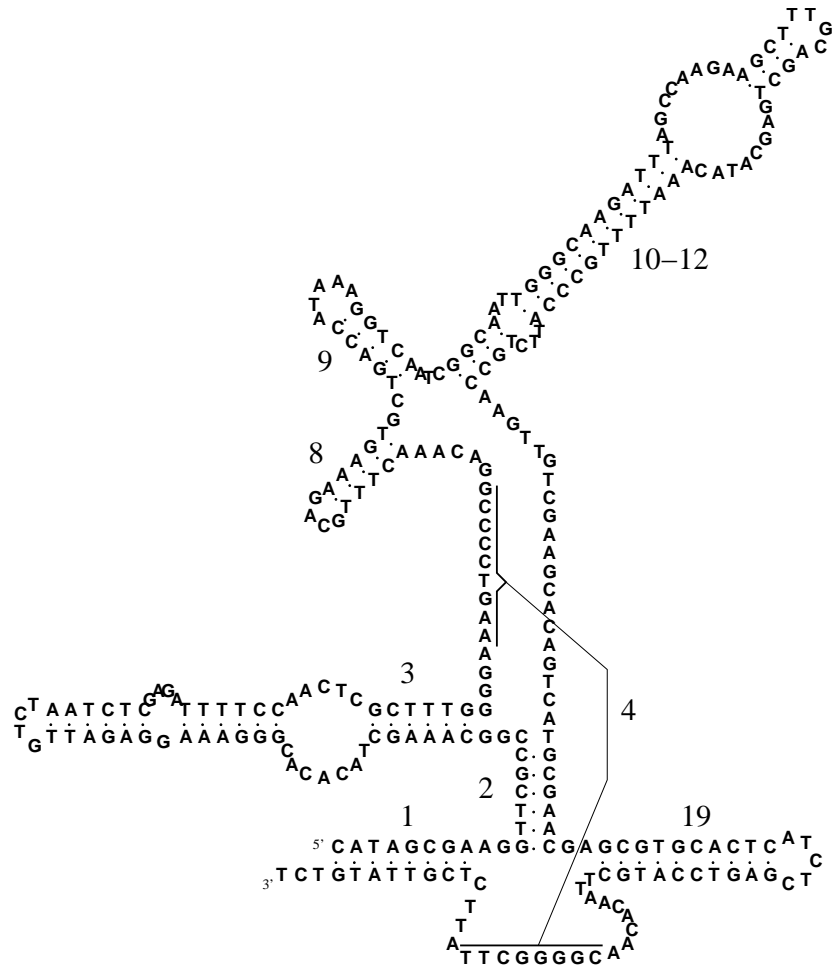


Figure 5.4: Secondary structure of *Trichoplax adhaerens* RNase MRP RNA inferred from the multiple alignment of metazoan RNase MRP RNAs provided in the Electronic Supplement.

This stringent filtering step left 3 H/ACA and 4 C/D snoRNA (plus U3 snoRNA), Table 5.2. The multiple sequence alignments, are provided in the Electronic Supplement.

The putative host genes of the *Trichoplax* snoRNAs are not conserved in human. It is known, however, that snoRNAs can change their genomic location on evolutionary time-scales. For instance, several host gene switches are observed for U17 already within vertebrates [26], see also [203]. Furthermore, several human snoRNA host genes are non-coding (e.g., the GAS5 transcript for U76 and the unnamed host gene of U71) or are poorly described ORFs (such as C20orf199 for snoRNA U106), making it virtually impossible to determine whether they are homologous between human and *Trichoplax*.

### No MicroRNAs

Homology based searches for microRNAs remained unsuccessful employing both **Blast** and **GotohScan** using the complete set of pre-microRNA hairpins listed in **miRBase** (release 12.0) as query. Both short **Blast** hits and weak **GotohScan** signals were analysed. Removing all sequences for which sequence conservation was very poor on the putative mature microRNA sequence and/or the putative precursor did not fold into the characteristic hairpin structure left a single candidate possibly homologous to mir-789. The best-conserved region is located opposite to the annotated mature sequence from *Caenorhabditis* species. Hence this candidate also remains inconclusive.

### *Ab initio* ncRNA Prediction

The use of comparative genomics in *Trichoplax* is limited by the comparably large distance to other sequenced genomes, because most of the genome thus cannot be unambiguously aligned with better understood genomes. We therefore investigated two different genome-wide alignments. In the first screen, we used three species **MultiZ**-alignments [110] of *Trichoplax adhaerens*, and the cnidaria *Hydra magnipapillata* and *Nematostella vectensis*. We used all alignment blocks containing *Trichoplax* and at least one of the two cnidarians.

A second screen was performed using **NcDNAalign** alignments [111] constructed from *Trichoplax adhaerens*, *Porites lobata*, and shotgun traces from *Amphimedon queenslandica*, *Acropora millepora*, *Acropora palmata*, and *Hydra magnipapillata*. This screen was limited to alignment blocks containing *Trichoplax* and at least two other species. As expected, the large evolutionary distances in both screen

Table 5.3: Summary of annotated ncRNAs using RNAz screens of *Trichoplax adhaerens* genome.

	Multiz	NcDNaIalign	known
Aligned DNA (nt)	4837148	135140	—
alignments	35039	744	—
RNAz $p > 0.5$	1416	101	—
FDR random	56% 797	43% 43	—
RNAz $p > 0.9$	751	79	—
FDR	27% 386	15% 15	—
tRNAs	39	35	50+1
5S rRNA	6	8	9
rRNA operon	33+3	43	*
snRNAs	6	4	10
MRP,P,7SL	1	0	3
protein coding	1022	11	96963
repeat elements	66	1	—
total annotated	1211	101	
unannotated with EST	12	0	
without annotation	205	0	

The asterisk (\*) indicates that the rDNA operons appear as series of multiple RNAz hits. *Known* refers to all ncRNAs that have been reported previously and those that have been identified by homology search in this study. FDR – False Discovery Rate.

limit the sensitivity of the comparative approach and preclude the detection of Placozoan-specific ncRNAs.

Both of the differently created alignment sets are screened with RNAz, the corresponding results are compiled in Tab. 5.3. The restrictive NcDNaIalign alignments revealed no novel ncRNAs.

### 5.1.2 Discussion

We have reported here on a comprehensive computational study of non-protein-coding RNA genes in the genome of the placozoan *Trichoplax adhaerens*. We observed that only a limited set of the best-conserved ncRNAs, in particular tRNAs, rRNAs, and a few additional “housekeeping” RNAs are readily found by means of Blast. We used therefore a more sensitive tool, GotohScan, which implements a full semi-global dynamic programming algorithm. Using this method, we were

able to detect homologs of several fast-evolving ncRNAs, including a few box C/D and box H/ACA snoRNAs, the RNase MRP RNA, and the full complement of spliceosomal snRNAs.

In addition to the homology-based annotation, we conducted surveys evolutionary conserved RNA secondary structures using *RNAz* and *RNAmicro*. Reasoned by the large evolutionary distance between *Trichoplax* and other sequenced genomes, the sensitivity of these screens was rather low, however. Nevertheless a handful of novel ncRNA candidates was found.

Due to the small size and slow growth of *Trichoplax adhaerens*, it is hard – if not impossible – to obtain sufficient amounts of RNAs to verify the expression of ncRNA candidates directly by Northern blots. Instead, we used here a PCR-based approach introduced by [380], which requires much smaller quantities of RNA. We did not attempt to validate the entire set of predictions but rather selected a small subset, consisting of a few of the homologs detected by *GotohScan* and a small collection of novel predictions. Due to the small amount of RNA, the sensitivity is still limited. Nevertheless, we unambiguously identified a few previously undescribed *Trichoplax* ncRNAs, namely: U4atac, as a representative of the minor spliceosome; the U3 snoRNA and a putative novel ncRNA on scaffold 3857.

Our computational annotation of the *Trichoplax* genome reveals much of the expected complement of the ncRNA repertoire. Most ncRNAs are single-copy genes or appear in very small copy numbers. This contrasts the situation in many of the higher metazoa, for which more detailed ncRNA annotations are available (e.g. *C. elegans* [381], *Drosophila* [117, 287], and the Rfam-based annotation in mammalian genomes). In particular, the small copy number of tRNAs and other pol-III transcripts is surprising, since these genes appear in dozens or hundreds of copies in many bilaterian genomes.

The lack of microRNAs is surprising at a first glance. While a few orthologous microRNAs — in particular the mir-100 family — are shared between Cnidaria and Bilateria [382, 383], we found no trace of these genes in *Trichoplax*. Neither did we find a homolog of one of the 8 sponge microRNAs [384]. Our analysis is thus consistent with the recent report based on short RNA sequencing [384] that *Trichoplax* does not have microRNAs. The continuing expansion of the repertoire of microRNA and their targets has been associated with both major body-plan innovations as well as the emergence of phenotypic variation in closely related species [382, 383, 385–387]. The microRNA precursors of Cnidaria and Bilateria are imperfectly paired hairpin structures about 80 nt in length. In contrast, the precursors of the recently discovered miRNAs of the sponge *Amphimedon queenslandica* [384] are not orthologous to any of the Cnidarian/Bilaterian microRNA families

and resemble the structurally more diverse and more complex RNAs described in slime-molds [388], algae [389, 390] and plants [391–393]. Under the hypothesis of monophyletic diploplasts, which has recently gained substantial support [368, 394], Placozoa have secondarily lost their ability to produce microRNAs, while sponges have secondarily relaxed the constraints on precursor structures. The complete loss of microRNAs in Placozoa is consistent with the morphological simplicity of *Trichoplax*.

*De novo* predictions of evolutionarily conserved RNAs suggest that the *Trichoplax* genome may have preserved some ncRNAs characteristic to basal metazoans, such as the handful of hairpin structures that are conserved between *Trichoplax* and *Nematostella*. We do not know at this point, however, whether these purely computational signals are expressed *in vivo*, and what their function might be.

Our survey also misses several ncRNA classes that we should expect to be present in *Trichoplax*, in particular telomerase RNA, U7 snRNA (which are involved in histone 3'-end processing [1], the Ro-associated Y-RNAs, the RNA components of the vault complex (the *Trichoplax* genome contains the Major Vault Protein), and possibly also a 7SK RNA. In contrast to microRNAs, however, recent studies have highlighted how difficult it is to identify these particular classes of RNA from genomic DNA: Telomerase RNA evolves so rapidly that — despite its size of over 300nt — it has not been identified so far in any invertebrate species [235]. A similarly fast evolution is observed for the 7SK RNA [236, 237]. Due to their small size and weak sequence constraints, U7 snRNA [395, 396], Y RNAs [397, 398], and vault RNAs [399] are also largely unknown beyond deuterostomes (in some cases Drosophilids or *C.elegans*, where homologs were discovered independently). Our failure to find these genes thus most likely points at the limitations of the currently available homology search methodology rather than at the absence of these RNA classes in the *Trichoplax* genome.

## 5.2 *Schistosoma mansoni*

Most non-vertebrate genome projects have put little emphasis on a comprehensive annotation of ncRNAs. Indeed, most non-coding RNAs, with the notable exception of tRNAs and rRNAs, are difficult or impossible to detect with **Blast**. Hence their annotation is not part of generic genome annotation pipelines. Dedicated computational searches for particular ncRNAs, for example, RNase P and MRP [324, 325] (Section 4.2), 7SK RNAs [236, 237] (Section 4.3), or telomerase RNA [235, 400] (Section 4.4), are veritable research projects in their own right. Despite best efforts, large territory remains uncharted across the animal phylogeny.

Schistosomes belong in an early-diverging group within the Digenea, but are clearly themselves highly derived [401–403]. The flatworms are a long-branch group, suggesting rapid mutation rates (see [404]).

**Schistosome genomes** are comparatively large, estimated at about 300 megabase pairs for the haploid genome of *Schistosoma mansoni* [405]. The other major schistosome species parasitizing humans probably have a genome of similar size, based on the similarity in appearance of their karyotypes [406]. These large sizes may be characteristic of platyhelminth genomes in general: the genome of *Schmidtea mediterranea*, the only other sequenced platyhelminth genome, is even larger, with the current genome sequencing project reporting a size of  $\sim 480$  megabase pairs [76]<sup>1</sup>.

The protein-coding portion of the *Schistosoma* genomes have received much attention in recent years. Published work includes transcriptome databases for both *S. japonicum* [407] and *S. mansoni* [408], characterization of promoters [409, 410], and physical mapping and annotation of protein-coding genes from both the *S. mansoni* and *S. japonicum* genome projects [249]. Recently, a systematic annotation of protein-coding genes in *S. japonicum* was reported [411]. In contrast to other, better-understood, parasites such as *Plasmodium* [119], however, not much is known about the non-coding RNA complement of schistosomes. Only the spliced leader RNA (SL RNA, Section 3.3) of *S. mansoni* [124], the hammerhead ribozymes encoded by the SINE-like retrotransposons Sm- $\alpha$  and Sj- $\alpha$  [412, 413], and secondary structure elements in the LTR retrotransposon *Boudicca* [414] have received closer attention. Ribosomal RNA sequences have been available mostly for phylogenetic purposes [415], and tRNAs have been studied to a limited degree [416].

---

<sup>1</sup><http://genome.wustl.edu/genome.cgi?GENOME=Schmidtea%20mediterranea>

Table 5.4: Non-coding RNA predictions from the sequenced genome of *S. mansoni*.

RNA class	Functional Category	Copy No.	Related references
7SK	Transcription regulation	(1)	This study
Hammerhead ribozymes	Self-cleaving	> 24,000	[412]
miRNA	translation control	4	[385], this study
potassium channel motif	RNA editing	3	[417]
RNase MRP	Mitochondrial tRNA processing	(1)	This study
RNase P	tRNA processing	1	This study
rRNA-operon	Polypeptide synthesis	80 - 105	[418], this study
5S rRNA	Polypeptide synthesis	21	This study
SL RNA	Trans-splicing	6-48	[124], this study
SnoRNA U3	Nucleolar rRNA processing	1	This study
SRP	Protein transportation	12	This study
tRNA	Polypeptide synthesis	663	This study
U1	Splicing	3-34	[123], this study
U2	Splicing	3-15	[123], this study
U4	Splicing	1-19	[123], this study
U5	Splicing	2-9	[123], this study
U6	Splicing	9-55	[123], this study
U11	Splicing	1	This study
U12	Splicing	1-2	[123], this study
U4atac	Splicing	1	This study
U6atac	Splicing	1	This study

In this section we give a comprehensive overview of the evolutionary conserved non-coding RNAs in the *S. mansoni* genome. We discuss representatives of 23 types of ncRNAs that were detected based on both sequence and secondary structure homology.

### 5.2.1 Results & Discussion

Structure and homology-based searches of the *S. mansoni* genome revealed ncRNAs from 23 different RNA categories. Table 5.4 lists these functional ncRNA



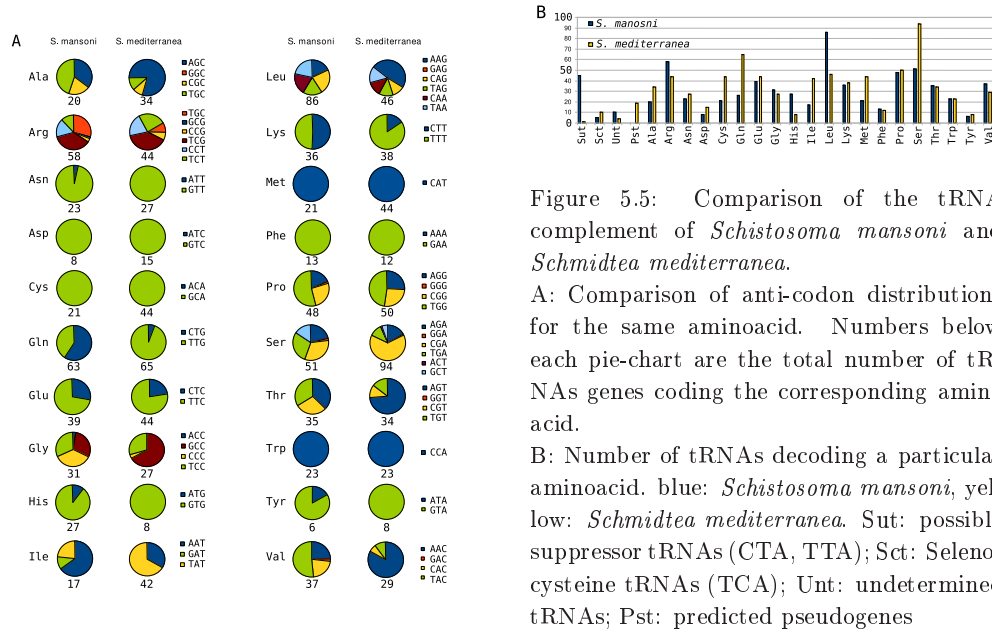


Figure 5.5: Comparison of the tRNA complement of *Schistosoma mansoni* and *Schmidtea mediterranea*.

A: Comparison of anti-codon distributions for the same amino acid. Numbers below each pie-chart are the total number of tRNAs genes coding the corresponding amino acid.

B: Number of tRNAs decoding a particular amino acid. blue: *Schistosoma mansoni*, yellow: *Schmidtea mediterranea*. Sut: possible suppressor tRNAs (CTA, TTA); Sct: Selenocysteine tRNAs (TCA); Unt: undetermined tRNAs; Pst: predicted pseudogenes

category, the number of predicted genes in each category, and references associated with each RNA type. Supplementary **fasta** files containing the ncRNA genes, **bed** files with the genome annotation, and **stockholm-format** alignment files can be accessed at <http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-014>.

## Transfer RNAs

Candidate tRNAs were predicted with **tRNAscan-SE** in the genomes of both *S. mansoni* and *S. mediterranea* (a free-living platyhelminth, used for comparison). After removal of transposable element sequences (see below), **tRNAscan-SE** predicted a total of 663 tRNAs for *S. mansoni* and 728 for *S. mediterranea*. These included tRNAs encoding the standard 20 amino acids of the traditional genetic code, selenocysteine encoding tRNAs (tRNA<sup>sec</sup>) [419] and possible suppressor tRNAs [420] in both genomes. The tRNA<sup>sec</sup> from schistosomes has been characterized, and is similar in size and structure to tRNA<sup>sec</sup> from other eukaryots [421].

The tRNA complements of the two platyhelminth genomes are compared in detail in Figure 5.5.

Homology-based analysis yielded similar, though somewhat less sensitive, results to those of **tRNAscan-SE**. A **Blast** search with **Rfam's** tRNA consensus yielded 617 predicted tRNAs compared to the 663 predictions made by **tRNAscan-SE**.

### Ribosomal RNAs

As usual in eukaryotes, the 18S, 5.8S, and 28S genes are produced by RNA polymerase I from a tandemly repeated polycistronic transcript, the ribosomal RNA operon. The *S. mansoni* genome contains about 90-100 copies [418, 422] which are nearly identical at sequence level, because they are subject to concerted evolution [166]. The repetitive structure of the rRNA operons causes substantial problems for genome assembly software [423]. In order to obtain a conservative estimate of the copy number, we retained only partial operon sequences that contained at least two of the three adjacent rRNA genes. We found 48 loci containing parts of 18S, 5.8S, and 28S genes, 32 loci covering 18S and 5.8S rRNA, and 57 loci covering 5.8S and 28S rRNAs (Figures 5.6-A, 5.7-A). Adding the copy numbers, we have not fewer than 80 copies (based on linked 18S rRNAs) and no more than 137 copies (based on linked 5.8S rRNA). The latter is probably an overestimate due to the possibility that the 5.8S rRNA may be contained in two scaffolds. The copy number of rRNA operons is thus consistent with the estimate of 90-100 from hybridization analysis [418]. For comparisons we examined the *S. japonicum* genome for rRNAs and yielded 90 rRNA clusters located on 88 scaffolds. 18S and 5.8S rRNA was obtained in 36 scaffolds, whereas 32 scaffolds contained 5.8S and 28S, in 22 cases complete operons were detected (Figures 5.6-B, 5.7-B).

The 5S rRNA is a polymerase III transcript that has not been studied in *S. mansoni* so far. We find 21 copies of the 118nt long 5S rRNA. Four of these copies are located within a 3000nt cluster on *Scaffold010519*, Figure 5.8.

### Spliceosomal RNAs and Spliced Leader RNA

Spliceosomes, the molecular machines responsible for most splicing reactions in eukaryotic cells, are ribonucleoprotein complexes similar to ribosomes [424] as described in Section 3.2. By homology search we found 34 U1, 15 U2, 19 U4, 9 U5, and 55 U6 sequences in the genome assembly. Interpreting all sequences that are identical in short flanking regions as the same, we would retain only 3 U1, 3 U2, 1 U4, 2 U5, and 9 U6 genes [123]. The true copy number in the *S. mansoni* genome is most likely somewhere between these upper and lower bounds. Comparison with *S. japonicum* affirm these predictions, Table 5.5. Secondary structures for these are similar to those of typical snRNAs, Fig. 5.9.

Non-coding RNAs of the minor spliceosome are typically much less conserved, therefore, these RNAs were detectable only by means of *GotohScan* [61] but not with the much less sensitive *Blast* searches. Although U4atac and U6atac are

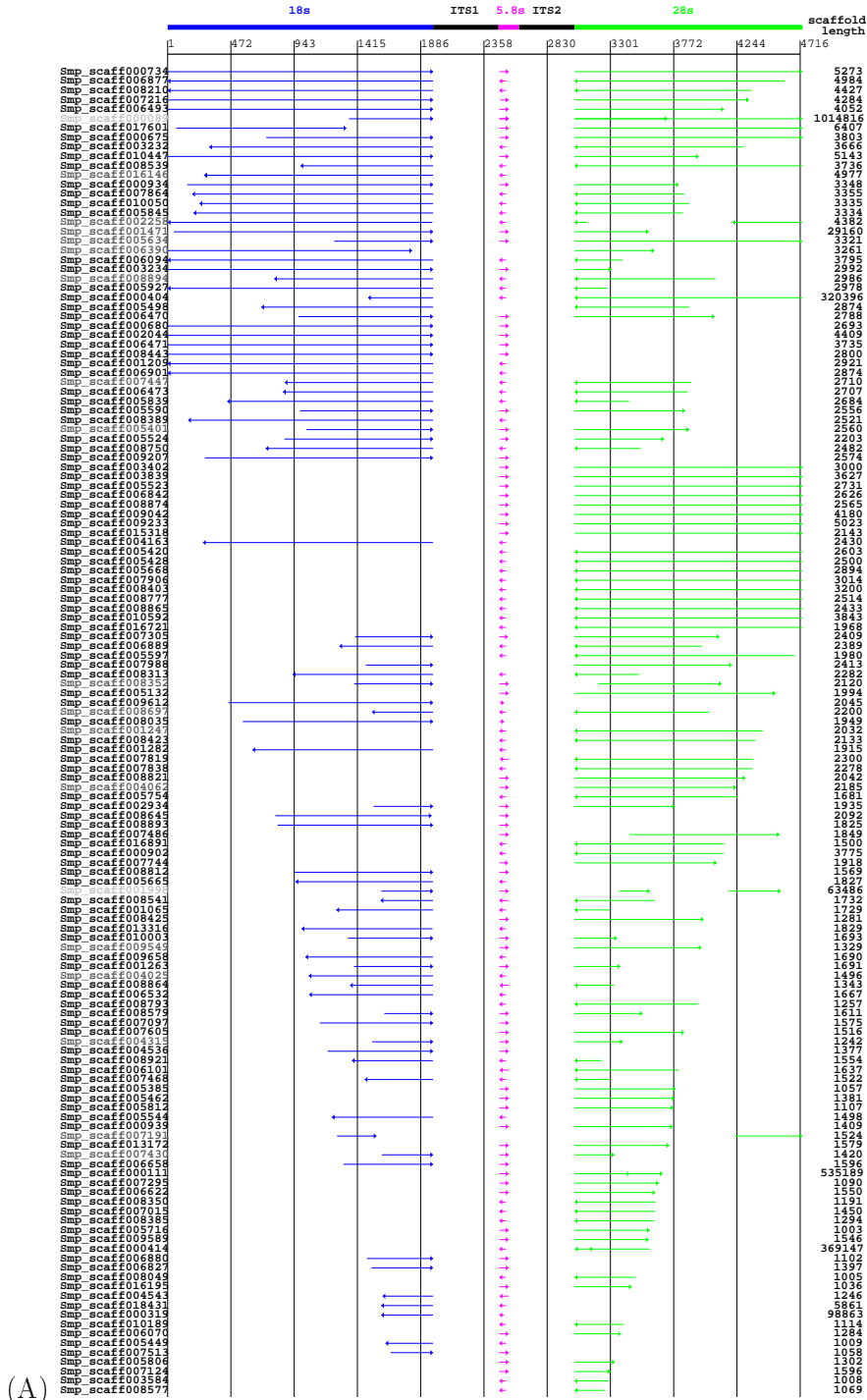


Figure 5.6: Fragments of RNA Operons in (A) *S. mansoni* and (B) *S. japonicum* (next page). Whole and partial pol-I transcribed rRNA operons. Scale representation of portions of scaffolds that include either whole rRNA operons or fragments including 18S and 5.8S or 28S and 5.8S. Right-facing arrows represent plus-strand transcripts; left-facing arrows represent minus-strand transcripts. Scaffold names are shown in the far left column; names in light gray have large runs of unknown nucleotides ("N's"), names in dark gray have smaller runs of unknown nucleotides. Scaffold lengths are shown in the far right column. Top line: scale drawing of the whole pol-I transcribed rRNA operon. Second line: scale, in nt. Scaffold lines: blue arrows: 18S regions, pink arrows: 5.8S regions, green arrows: 28S regions.



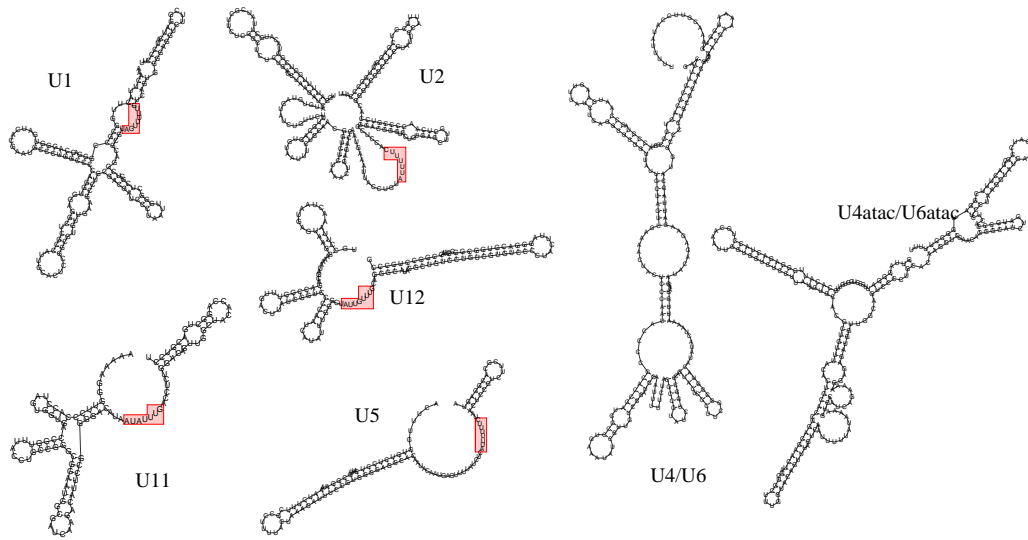


Figure 5.9: Secondary structures of the 9 snRNAs and the interaction complexes of U4/U6 and U4atac/U6atac, respectively.

Table 5.5: Copy number of snRNAs in *Schistosoma japonicum* (black) and *Schistosoma mansoni* (blue). U1, U2, U4, U5 and U6 are major spliceosomal non-coding RNAs, whereas U11, U12, U4atac and U6atac act for the minor spliceosome. breco<sub>ba</sub> – Method referring to a pipeline: **blast**, reduce output by combining hits, built consensus sequence, **blast** again and verify by alignment.

snRNA	"breco <sub>ba</sub> "		structure-alignment	different up regions	pse/tata element	
U1	12	34	9	3	2	3
U2	89	15	63	1	1	3
U4	11	19	6	1	1	1
U5	70	9	24	1	1	22
U6	19	55	12	10	2	9
U11	0	0	1	1	1	1
U12	0	0	1	1	0	1
U4atac	0	0	1	1	1	1
U6atac	0	0	2	2	1	1

quite diverged compared to known homologs, they can be recognized based on both secondary structure and conserved sequence motifs. Furthermore, the U4atac and U6atac sequences can interact to form the functionally necessary duplex structure shown in Fig. 5.9.

An analysis of promoter sequences showed that the putative snRNA promoter motifs in *S. mansoni* are highly derived. Only one of the two U12 genes exhibits a clearly visible snRNA-like promoter organization.

The Spliced Leader (SL) RNA (Section 3.3) is one of the very few previously characterized ncRNAs from *S. mansoni* [124]. The 90nt SL RNA, which was found in a 595nt tandemly repeated fragment (accession number *M34074*). Using **Blast**, we identified 54 SL RNA genes. These candidates, along with 100nt flanking sequence, were aligned using **ClustalX**, revealing 6 sequences with aberrant flanking regions, which we suspect to be pseudogenic. The remaining sequences are 43 identical copies and 5 distinct sequence variants. A secondary structure analysis corroborates the model of [124], according to which the *S. mansoni* SL RNA has only two loops, with an unpaired Sm binding site (Figure 5.10). This coincides with the SL RNA structure of Rotifera [146], but is in contrast to the SL RNAs in most other groups of eukaryots, which exhibit single or triple stem-loop structures [425], see section 3.3. A **Blast**-search against *S. mansoni* EST data confirms that the 5' part of the SL is indeed *trans*-spliced to mRNAs.

### SRP RNA and Ribonuclease P RNA

Signal recognition particle (SRP) RNA, also known as 7SL RNA, is part of the signal recognition particle, a ribonucleoprotein that directs packaged proteins to their appropriate locations in the endoplasmic reticulum. Although one of the protein subunits of this ribonucleoprotein was cloned in 1995 [426], little is known about the other subunits or the RNA component in *S. mansoni*. We found eight probable candidates for the SRP RNA, with one almost canonical sequence (Figure 5.11). For *S. japonicum* we found four (Figure 5.11b-left) and two more (Figure 5.11b-right) candidates with point mutations which may influence their function. By comparing the sequences of these two organisms the latter SRP RNA seems to be a pseudogene.

The RNA component of Ribonuclease P (RNase P) is the catalytically active part of this enzyme that is required for the processing of tRNA precursors [427, 428], as described in section 4.2. We found one classic RNase P RNA in the *S. mansoni* genome using both **GotohScan** and **rnabob** with the eukaryotic (“nuclear”) **Rfam** consensus sequence for RNase P as search sequence.

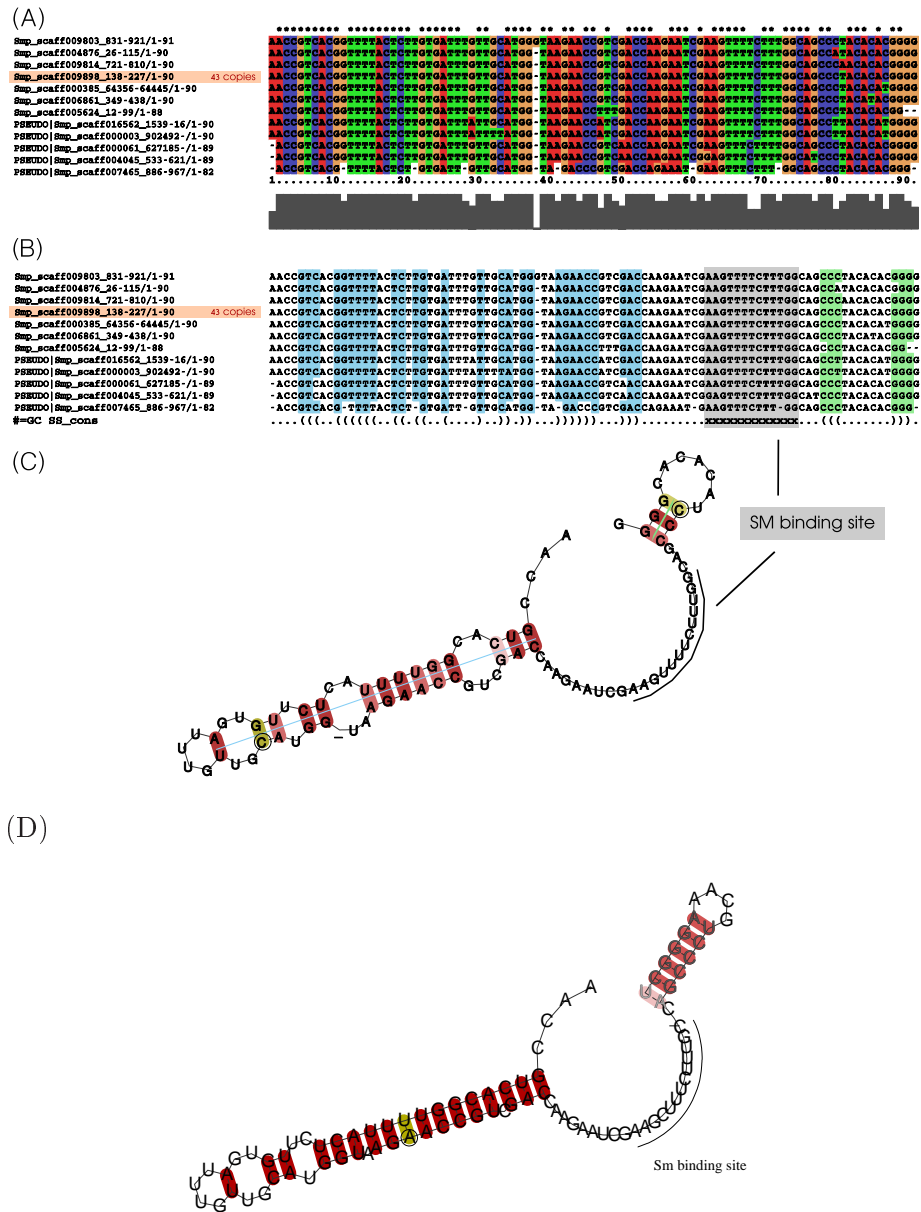


Figure 5.10: SL RNA sequences and structure. A) Clustal alignments of SL RNA candidates and putative pseudogenes. All sequences are single-copy except for the salmon- highlighted sequence, which represents a cluster of 43 copies. B) Alignments in Emacs ralee mode, with structural elements highlighted. Consensus secondary structure is represented at the last line of the alignment. Blue and green highlight: base-paired regions. Grey highlight: the Sm binding site. C) Secondary structure predicted by RNAalifold with the constraint that the Sm binding site must be unpaired. For a full alignment, including flanking regions, see <http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/08-014>. Secondary structure analysis of these candidates revealed structural conservation and thermodynamic stability indicating a likely ncRNA. Like [124] we found that the *S. mansoni* SL RNA has only two loops, with an unpaired Sm binding site, whereas most other SL RNAs have a triple stem-loop structure. D) For comparison secondary structure for *S. japonicum*.

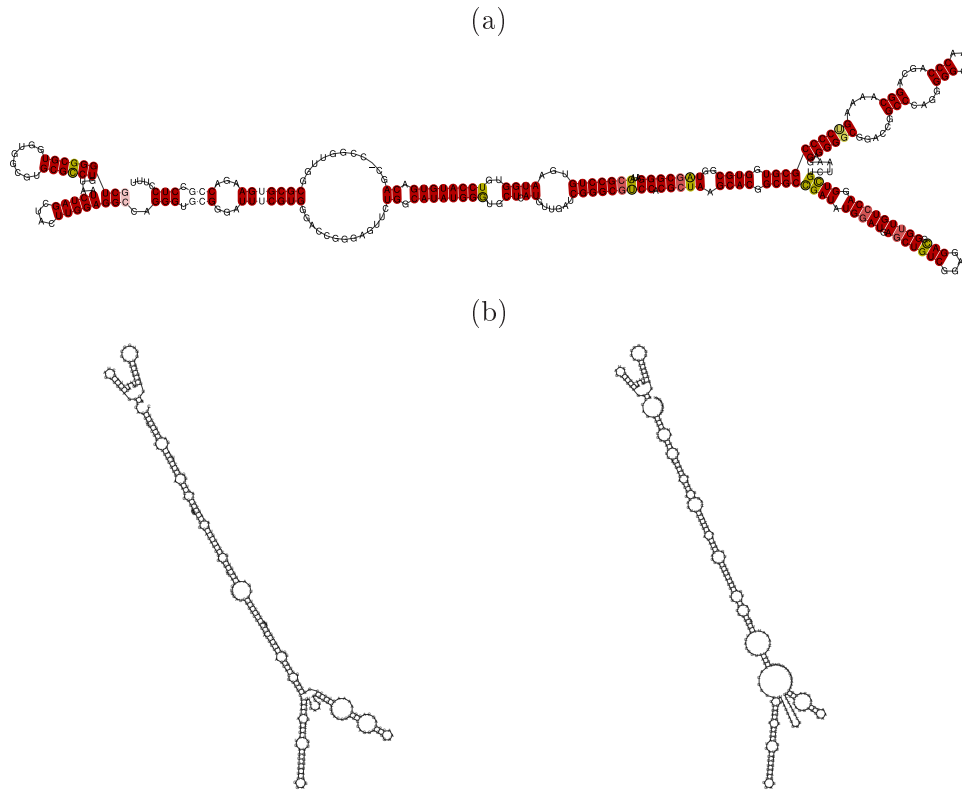


Figure 5.11: SRP RNA. (a) Secondary structure for the predicted *S. mansoni* signal recognition particle. Red nucleotides represent base pairs with conserved nucleotides across different species in the alignment. Yellow nucleotides represent positions with a high level of point mutations in different species, but with conserved secondary structure (compensatory mutations). Alignments are also available as supplementary files. (b) Secondary structure of 4 copies SRP (left) and 1 copy SRP (right) obtained in *S. japonicum*.

## MicroRNAs

So far, no microRNA has been verified experimentally in *S. mansoni*. The presence of four protein-coding genes encoding crucial components of the microRNA processing machinery (Dicer, Argonaut, Drosha, and Pasha/DGCR8) [429, 430], and the presence of Argonaut-like genes in both *S. japonicum* [431] and *S. mansoni* (detected by `tblastn` in EST data), strongly suggests that schistosomes have a functional microRNA system. Indeed, most recently five miRNAs were found by direct cloning for *S. japonicum* that are also conserved in *S. mansoni* [432]: *let-7*, *mir-71*, *bantam*, *mir-125*, and a single schistosome-specific microRNA. The precursor sequences, however, are quite diverged from the consensus of the homologous genes in Bilateria.





Using bioinformatics methods (See chapter 2) we were able to find only four miRNAs, see Figure 5.12.

The small number of recognizable microRNAs in schistosomes is in strong contrast to the extensive microRNA complement in *S. mediterranea*, indicating massive loss of microRNAs relative to the planarian ancestor. This may be a consequence of the parasitic lifestyle of the schistosomes.

### Small Nucleolar RNAs

Small nucleolar RNAs play essential roles in the processing and modification of rRNAs in the nucleolus [433, 434], as shown before (sec. 4.1,5.1). Both major classes, the box H/ACA and the box C/D snoRNAs are relatively poorly conserved at the sequence level and hence are difficult to detect in genomic sequences. This has also been observed in a recent ncRNA annotation project of the *Trichoplax adhaerens* genome [61], see section (5.1). The best-conserved snoRNA is the atypical U3 snoRNA (see also section 4.1), which is essential for processing of the 18S rRNA transcript into mature 18S rRNA [48]. In the current assembly of the *S. mansoni* genome we find six U3 loci, but they are also identical in the flanking sequences, suggesting that in fact there is only a single U3 gene. No unambiguous homologue was detected for any of the other known snoRNAs.

A *de novo* search for snoRNAs (see methods for details) resulted in 2610 promising candidates (1654 box C/D and 956 box H/ACA), listed in the Electronic Supplement. All these predictions exhibit highly conserved sequence boxes as well as the typical secondary features of box C/D and box H/ACA snoRNAs, respectively.

### Other RNA Motifs

Two examples of relatively well-known schistosome non-coding RNAs are the hammerhead ribozyme motifs within the Sm- $\alpha$  and Sj- $\alpha$  SINE-like elements [412, 413]. A **Blast** search of the hammerhead ribozyme motif from the **Rfam** database resulted in 24,447 candidates. While high, this number is not surprising considering the generally high copy number of SINE elements; previously, the copy number for Sm- $\alpha$  elements in the *S. mansoni* genome was estimated to exceed 10,000 [412]. The potassium channel RNA editing signal is another structured RNA element that was described previously [417]. We found three copies of the gene for this signal in the *S. mansoni* genome assembly. U7 RNA was not examined in *S. mansoni*, however in *S. japonicum* a reasonable candidate was observed. Figure 5.13 shows a possible interaction with the histone downstream element.



telomerase proteins (Smp\_066300 and Smp\_066290) and has the same telomeric repeat sequences as many other metazoan animals [435]. Telomerase RNAs are notoriously difficult to find (see section 4.4), as they are highly divergent among different species, varying in both size and sequence composition [436].

### 5.2.2 Conclusions

We have described here a detailed annotation of “housekeeping” ncRNAs in the genome of the parasitic planarian *Schistosoma mansoni*. Limited to the best conserved structured RNAs, our work nevertheless uncovered important genomic features such as the existence of a schistosome-specific SINE family derived from tRNA-Gln-TTG [437]. Our data furthermore establish the presence of a minor spliceosome in schistosomes and confirms spliced-leader *trans*-splicing.

Platyhelminths are known to be a fast-evolving phylum [438]. It is not surprising therefore that in particular the small ncRNAs are hard or impossible to detect by simple homology search tools such as **Blast**. Even specialized tools have been successful in identifying only the better conserved genes such as tRNA, microRNAs, RNase P RNA, SRP RNA. Notoriously poorly conserved families, such as snoRNAs, mostly escaped detection.

The description of several novel and in many case quite derived ncRNAs contributes significantly to the understanding of the evolution of these RNA families. The schistosome ncRNA sequences, furthermore, are an important input to subsequent homology search projects, since they allow the construction of improved descriptors for sequence/structure-based search algorithms. Last but not least, the ncRNA annotation track is an important contribution to the genome-wide annotation dataset. It not only completes the protein-based annotation but also helps to identify annotation errors, e.g. cases where putative proteins are annotated that overlap rRNA operons or other ncRNAs.

## Chapter 6

# Conclusion

We have discussed here various strategies for homology-based identification of ncRNAs in eukaryots. Within the last decade the number of known ncRNAs increased enormously and the biological impact and enquiry is gigantic. In the current version of **Rfam** (version 9.1) 1372 families are described, of which most are snoRNAs and miRNAs. Unfortunately no magical allround-ncRNA homology based search tool exists so far. Some genes, such as snRNAs, are relatively conserved in sequence and structure. However, the majority of ncRNAs, particularly longer structured ncRNAs, such as 7SK or telomerase RNA, vary in sequence and structure so extensively that we have to discuss the distance between the theoretical approach of homology search by conserved sequence and structure elements and the practical use of existing programs. At present, the success of computational ncRNA identification is constrained to the conservation degree of the functional gene and an adequate number of links between the query and target genome. But even in wet labs the proof of homology between e.g. *TLC-1* of *S. cerevisiae* and *S. pombe* can be performed just by their functions. It is at present impossible to align these sequences, neither on sequence nor on structure level. To prove the homology in a bioinformatician's way by common descent, more telomerase RNAs phylogenetically located between these two fungi ("missing links") are needed.

Using existing programs with default parameters retaining their direct output would be negligent. Instead, for each ncRNAs it is important to use the appropriate program with adjusted parameters. Suboptimal candidates should be investigated carefully in detail. Specialities, such as introns, expansion/invention/deletion of domains, multiple copies, etc., must be examined separately.

---

Although each ncRNA showed its own peculiarities we were able to predict and annotate these ncRNAs with a variety of programs in a wide eukaryotic range. Consequently, we were able to reconstruct evolutionary incidences for each of these groups.

We investigated in detail the evolutionary history of *cis*-splicing, through the nine spliceosomal snRNA families (U1, U2, U4, U5, U6, U11, U12, U4atac, and U6atac) across the completely or partially sequenced genomes of metazoan animals. Representatives of the five major spliceosomal snRNAs were found in all genomes. None of the minor spliceosomal snRNAs were detected in nematodes and in the shotgun traces of *Oikopleura dioica*, while in all other animal genomes at most one of them was missing. Although snRNAs are present in multiple copies in most genomes, distinguishable paralog groups are not stable over long evolutionary times, although they appear independently in several clades. In general, animal snRNA secondary structures are highly conserved, albeit in particular U11 and U12 in insects exhibit dramatic variations. While in some genomes snRNAs appear in tandem and/or associated with 5S rRNA genes, these clusters are not conserved over longer evolutionary time-scales. An analysis of genomic context of snRNAs revealed that they behave like mobile elements, exhibiting very little syntenic conservation. Taken together, the data are consistent with a dominating duplication-deletion mechanism of concerted evolution for the genomic evolution and proliferation of snRNAs.

The structures attained by RNA molecules did not only depend on their sequence but also on environmental parameters such as their structure. So far, this effect has been largely neglected in bioinformatics studies. We showed that structural comparisons can be facilitated and more coherent structural models can be obtained when differences in environmental parameters are taken into account. We re-evaluated the secondary structures of the spliced leader RNAs from the seven eukaryotic phyla in which SL RNA *trans*-splicing has been described. By adjusting structure prediction to natural growth temperatures and considering energetically similar secondary structures we observe striking similarities among Euglenida, Kinetoplastida, Dinophyceae, Cnidaria, Rotifera, Nematoda, Platyhelminthes, and Tunicata that could not be explained easily by the independent innovation of SL RNAs in each of these phyla.

We were not able to detect any dependencies between minor spliceosomal snRNAs and SL RNAs as indicated by [209], see Tab. 6.1.

Table 6.1: Presence and absence of spliced leader RNAs and minor spliceosomal snRNAs [123, 144, 209, 425].

Taxon	SL RNA	Minor Spliceosome
Euglenozoa	+	-
Plants	-	+
Dinophyceae	+	?
Ascomycota	-	-
Cnidaria	+	+
Rotifera	+	?
Nematoda	+	-/?
Insects	-	+
Platyhelminthes	+	+
Tunicates	+	+
Vertebrates	-	+

SmY RNAs copurify in a small snRNP complex related to SL1 and SL2 involved in *trans*-splicing. We described a comprehensive computational analysis of SmY RNA homologs found in currently available genome sequences. We identified homologs in all sequenced nematode genomes within the class Chromadorea. However, we were unable to identify homologs in a more distantly related nematode species, *Trichinella spiralis* (class Dorylaimia) and non-nematoda phyla. MacMorris *et al.* hypothesized that the role of SmY RNA might be in recycling spliceosome proteins after SL RNAs are consumed in the *trans*-splicing reaction [239]. MacMorris' model suggests that the diversification of SmY RNA gene copies (accompanied by sequence variations in stem-loop 2, the more variable stem) may be driven by the diversification of SL2 RNA genes. Although we have not conducted a detailed joint comparative analysis of SL RNAs and SmY RNAs, the results of our SmY RNA survey are broadly in accordance with this model's expectations.

We described a computational search for functional U7 snRNA genes throughout vertebrates which included the upstream sequence elements characteristic for snRNAs transcribed by pol-II. Based on the results of this search, we discussed the high variability of U7 snRNAs in both sequence and structure and we reported on an attempt to find U7 snRNA sequences in basal deuterostomes and non-Drosophilid insect genomes based on a combination of sequence, structure, and promoter features. Due to the extremely short sequence and the high variability in both sequence and structure, no unambiguous candidates were found. This part of the thesis calls for both, more experimental data on U7 snRNA as well as improved

---

bioinformatics approaches for homology search that can deal with such small and rapidly evolving genes.

Beside phylogenetic searches of ncRNAs performing the step of processing, we demonstrated that a genome-wide comparative genomic approach searching for short conserved introns is capable of identifying conserved transcripts with a high specificity. Predicted mlncRNAs were even confirmed in wet labs. As conserved introns indicate both purifying selection on the exon-intron structure and conserved expression of the transcript in related species, the novel mlncRNAs are good candidates for functional transcripts.

We saw a comprehensive computational survey resulting in U3 sequences for more than 90 additional eukaryotes. This extended data basis is used to improve the secondary structure models and to investigate in detail the structural variation of U3 snoRNAs, which are much more extensive than previously thought. Many fungal U3 genes in addition contain introns. U3 promoters are snRNA-like but show substantial variations even between related species.

Only two years ago 7SK RNA was considered as a highly conserved vertebrate innovation. We discovered poorly conserved homologs in several insects and lophotrochozoans. This implies a much earlier evolutionary origin. The mechanism of 7SK function requires interaction with the proteins HEXIM and LARP7. Here, we presented a comprehensive computational analysis of these two proteins in metazoa, and we extended the collection of 7SK RNAs by several additional candidates. Furthermore, we derive an improved secondary structure model of 7SK RNA, which shows that the structure is quite well-conserved across animal phyla despite the extreme divergence at sequence level.

We predicted several ncRNAs, which are known to be highly divergent from their homologous, however these candidates are proved in wet labs by our collaborators. For 7SK RNA *Caenorhabditis* candidates are verified by Olivier Bensaude, which until recently were believed not to exist in nematods. We predicted RNase MRP in *Giardia lamblia*, which will be verified by Astrid Schön. This sequence plays a key role for the origin and evolution of RNase MRPs. Finally, we were even able to predict telomerase RNA candidates in *Strongylocentrotus purpuratus*, *Neurospora* fungi and *Branchiostoma*, which Julian Chen is currently examining in detail.

We used a variety of techniques and time to gain experience with ncRNAs. Finally, we were able to screen complete genomes for all reported ncRNAs, with all their specialities. We reported on a comprehensive ncRNA annotation of the genome *Trichoplax adhaerens* and *Schistosoma mansoni*. Since **Blast** had identified only a small fraction of the best-conserved ncRNAs (in particular rRNAs, tRNAs and



some snRNAs) we used `GotohScan` to increase the sensitivity of the homology search. We successfully identified the full complement of major and minor spliceosomal snRNAs, the genes for RNase P and MRP RNAs, the SRP RNA, as well as several small nucleolar RNAs. We confirmed five miRNAs in *S. mansoni* and none in *Trichoplax*. We provided candidates for 7SK, however could not annotate expected ncRNAs, such as telomerase RNA, vault RNA or Y RNA. Interestingly in the most basal metazoan genome (*T. adhaerens*) most ncRNAs, including the pol-III transcripts, appeared as single-copy genes or with very small copy number.

Moreover, for all examined ncRNAs, complete multiple alignments in Stockholm format created by combining various programs (at least `ClustalW`, `Locarnate`, `RNAfold`, `RNAsubopt`, `RNA duplex`) are available on our supplemental material pages and included in `Rfam`.

## Promoter Elements

Since most of the non-coding RNAs are transcribed by Polymerase III a reliable indication for a candidate being an *in vivo* functional gene is the promoter sequence upstream of transcription start site and a poly-U transcription termination signal directly downstream of the transcribed DNA part. Polymerase II is recruited by less restrict promoter upstream elements, since they commonly lack the TATA box.

There are two main ways for promoter recognition. On the one hand we can observe *gene specific* promoter elements. These might be internal regulators, such as A-box or B-box of tRNAs, or upstream regulators such as the commonly known TATA-box [439]. To forecast *species specific* promoters is another approach to qualify candidates. For instance we saw that the proximal sequence element of vertebrates is completely different as for insects, nematods or any other clade. However, closely related organisms share similarities between their PSE. We mostly examined 100 nt upstream of transcription start site and up to 20 nt downstream of a gene. An extensive study with a wider range including internal promoters or downstream located enhancers would give more insight of regulation, specificity and could give more possibilities of evidence to computational prediction of a gene's functionality.

If there exists a bioinformatician lab pet for homology search, it would be *Trichoplax adhaerens*. This very basal metazoan animal has a fairly small genome, hardly any pseudogenes, long structured non-coding RNAs without any large extensions or deletions in sequence or structure and last but not least a homogenous proximal sequence element and TATA-box. Searching with 16 nt PSE and a down-

---

stream located TATA-box as query in *Trichoplax* with one pointmutation, only 28 hits remain as polymerase III transcripts. This will be a way to observe unknown functional non-coding RNAs in future work.

Some of the ncRNAs, namely tRNA, 5S RNA, SRP RNA, ALU-repeat family derived from SRP RNA and viral RNA (VA RNA) contain internal promoters. Interestingly, these ncRNAs are the oldest, which are believed to exist since LUCA. RNase P, which is also common in all known organisms and all other ncRNAs, believed to originated later, has external promoters. Therefore RNase P might have been originated later than other ancestral ncRNAs of LUCA.

The evolution of promoters has to be examined in detail in the future. However, one hypothesis might be, that in the RNA world ncRNAs needed to keep their promoters within the gene, because they lack a large context of its gene. Only after establishing the DNA as a storage for RNAs it was possible to emerge external enhancers to a strongly dependent model, such as external promoters. Beside tRNAs and other ncRNAs, which kept their internal promoters all over the time due to the fact that they have additionally a structural function, ncRNAs with external promoters could be, according to this hypothesis, much more variable in their evolutionary development.

## Evolution of Secondary Structures

We have seen a wide variety of ncRNAs: Some, such as the very old tRNAs are highly conserved in their structure. Apart from intron carrying tRNAs, adding only a few nucleotides to tRNAs is usually interpreted as a non-functional transcript. Comparing this to more recent innovations, such as RNase MRP, we observe extensions from 7 nt to more than 300 nt in only one stem or even whole insertions and deletions of stems. In the case of Telomerase RNA we do not even know a common structure. Is there a correlation between the first appearance of a ncRNA and its variability? To find a measure would be another future assignment. To base this measure on sequence patterns and mutation rates would not describe the nature of ncRNAs. On the other hand highly conserved regions, mostly interacting with DNA, RNA or proteins, have to be considered mandatory.

To measure the variability of a secondary structure might dependent on the length and its standard deviation. There is a tendency of longer molecules to be more variable, however, for the very short variable SL RNAs or the long conserved SRPs this is against the rule. Describing a ncRNAs variability with formulae depending on sequence and structure, one has to consider protein interactions. Some highly variable ncRNAs might not interact with proteins as much as conserved ones.

However in general one observes that directly interacting parts (protein binding sites, RNA-RNA interaction parts or RNA-DNA interaction sites) usually appear as short but highly conserved motifs. Indirect functional parts are represented in all families of ncRNAs as variable structure elements. Hairpins can have different nucleotides, different lengths.

Another fact is that the very conserved major snRNAs are believed to exist earlier than the minor spliceosomal RNA components, the latter show much more variations in their structure. Similar for RNase RNAs: RNase P can be dated back to LUCA and is much more conserved in structure and sequence than RNase MRP, which is believed to originate from RNase P. On the other hand this might be an artefact, due to the fact that more variable sequences are much harder to date back as far as highly conserved sequences.

## Concluding Remarks

Existing computational programs miss the possibility of searching for sequence motifs or hairpins with insertions of an extra internal nucleotide or small internal bulges, respectively, at *unknown* positions. Another problem of existing programs is the trade between all query hits, the run time and the size of the output. `rnabob` is very fast, however returns one candidate, only. Other programs, such as `RNAmotif` or `Erpin`, return multiple candidates, but have the problem of handling the output. Another general problem of existing programs is the lack of testing hairpins for their possible existence. Many predicted hairpins theoretically bind to each other, considering the sequence of nucleotides, however, from the MFE point of view, they would not form hairpins.

Prediction Tools searching with information of 3D structure exist only for very specific problems and molecules. On the one hand, bioinformaticians work hard on pseudoknot prediction tools, on the other hand, new programs considering the stereochemical arrangement of longer hairpins by e.g. position weight matrices considering di- or trinucleotides of unpaired regions might be elementary for the function of especially ncRNAs.

With this thesis we were able to answer many basic evolutionary questions about ncRNAs. However, the world of ncRNA exhibits numberless questions.



# Appendix A

## Alternativ Alignment Listener

### A.1 ComposAlign

Evolution and Selection shape the phenotype and genotype of an organism in a unique way. Homologous sequences are derived from a common ancestor by a sequence of selective changes and diverge over time. Multiple selective constraints on a genomic sequence constrain evolution and result in interesting structures, e.g. modularization. Evolutionarily shaped structures become discernible when sequences derived from a common ancestor are aligned. The result as well as the method is called “alignment”. The data structure is a matrix, which is not only highly informative and story-telling for a biological expert but also patterned in a sometimes aesthetic way. Some patterns are visible when one of the numerous visualization tools is applied [86, 88, 440, 441].

Nevertheless, the modular and structured nature of much music has struck many as providing opportunities to understand genomic data by translating it to sound [442–444]. However, only a few trials have been made to use music to convey the patterns to the interested party [445–448]. All of them focus on single DNA or protein sequences. Early attempts transposed DNA sequences directly to music [444]. The assignment of two notes to each of the four characters (4 nucleotides) allowed for some flexibility to arrange notes to musical themes. Sonification of protein sequences offered a larger set of initial characters (20 amino acids) but was even more constrained and suffered from the creation of a monotonous string of notes without musical depth. Consideration of further properties [449–452] of characters or groups of characters and mathematical derivation based upon this additional information resulted in more exiting music but blurred the underlying information. A tool called `gene2music` [447] can be used for automated conversion

of protein-coding sequences to music. It maps the 20 amino acids on 13 chords, grouping chemically similar characters together while the chord duration is dependent on the frequency of the underlying codon. One system, PROMUSE [445] deals with sonification of amino acid features as well as structural information and the similarity between related proteins along the sequences. This similarity between proteins and genomic sequences results from common ancestry and light variation and is of central importance to studies in evolution and genomics.

Presentation of highly complex, multidimensional data requires far more channels to transport information than can be handled in the visual channel alone. Visualization and animation are fairly well developed, however, research on the transport of information via sonification is only recently gaining some interest [453]. Surprisingly, the complexity of the information transported by the audio channel is usually low, even though musical compositions for entertainment or artistic purposes show highly complex structures. In a multi-media setting, Lodha et al. [448] show that sonification can be efficient in disambiguating data in cases where visual presentation alone would be unclear. However, a direct comparison of the efficiency in auditory or visual information uptake is hard to perform. We can expect, however, that the perception of data via sonification and visualization is conceptually very different. Whether this can be beneficial for data presentation is an area we wish to continue to exam.

In this contribution we describe COMPOSALIGN, the first prototype for alignment sonification that translates genomewide aligned data into a musical composition. Such an acoustic representation requires a unique mapping of alignment information onto musical features. While some mapping is easy to frame, we strive for a intuitive mapping that is easy to perceive and also lives up to the demand to be artistic, pleasant and interesting.

## Methods

### Mapping

The main focus of our approach is to sonify the presence and absence of characters in the alignment such that their assignment to the corresponding sequence/species is clear. For simplicity, we assume that sequences are from different species, which allows us to refer to “different sequences” as “different species”. However, the sources of the sequences is irrelevant for our theoretical framework. Therefore we have chosen the following mapping, formalized as follows:

*A musical motif or pattern* is an arrangement of notes played in one measure.

Given a set  $\mathbb{S}$  of species, a set  $\mathbb{I}$  of instruments and a set  $\mathbb{P}$  of (different) patterns, we assign to each species a particular instrument which plays a particular pattern. Therefore, we define an injective function  $f : \mathbb{S} \rightarrow \mathbb{A}$  with  $\mathbb{A} = \{(x, y) \mid x \in \mathbb{I} \text{ and } y \in \mathbb{P}\}$  which is the cartesian product of the sets  $\mathbb{I}$  and  $\mathbb{P}$ . Moreover we assign to *every* species  $S \in \mathbb{S}$  a value  $f(S)$ . Thus it holds  $|\mathbb{S}| \leq |\mathbb{A}|$ , since  $f$  is injective. Many mappings  $\mathbb{S} \rightarrow \mathbb{A}$  full-fill the requirement that each species  $S \in \mathbb{S}$  is determined and distinguishable from another species by its values  $f(S)$ . The remaining degrees of freedom can be used to include additional information such as the phylogenetic relationship of the species. Therefore, we assign instruments to species such that the relationships among the instruments reflect the relationship among species. However, this assignment is done by hand since the relatedness for instruments is a matter of perception. The usage of two independent features  $(x, y) \mid x \in \mathbb{I} \text{ and } y \in \mathbb{P}$  to encode the species allows us to handle alignments with up to 144 species and to represent two-dimensional phylogenetic information as returned by `splitstree` [2].

Given a sequence  $s$  we consider  $n$  units  $u_1, \dots, u_n$  which are, in particular, subsequences of  $s$  such that  $\bigcup_{i=1}^n u_i \subseteq s$ . Biologically, these units are referred to as characters in general, “genes” in this contribution. Moreover the units  $u_1, \dots, u_n$  are ordered, such that  $u_i$  occurs before  $u_j$  whenever  $i \leq j$ .

Each unit  $u_i$  can be absent, i.e. “0”, or present, “+” or “-” if directed.

We are now able to define the following matrix  $A$ , also called *alignment*.

$$A_{i,j} = \begin{cases} + & , \text{ if } u_i \text{ appears in species } S_j \text{ in } + \text{ orientation} \\ - & , \text{ if } u_i \text{ appears in species } S_j \text{ in } - \text{ orientation} \\ 0 & , \text{ else} \end{cases}$$

This means that all entries  $A_{i,j} \neq 0$  for a fixed  $i$  are homologous. As explained we have assigned to every species a particular instrument playing a particular pattern. In general, an instrument and the corresponding pattern  $f(S_j)$  assigned to species  $S_j$  plays during time interval  $i$  whenever unit  $u_i$  appears in species  $S_j$ , i.e.  $A_{i,j} \neq 0$ . Whether  $f(S_j)$  sounds or not is only dependent on  $A_{i,j}$  and independent of  $A_{k,l}$  with  $k \neq i, l \neq j$ . However, three parameters can be set to enhance particular information.

**Orientation parameter.** This parameter indicates whether a pattern is played forwards or backwards, depending on the orientation of the occurring unit. To be more precise let  $f(S_j) = (I, P)$  and let unit  $u_i$  occur in species  $S_j$ . Then pattern  $P$

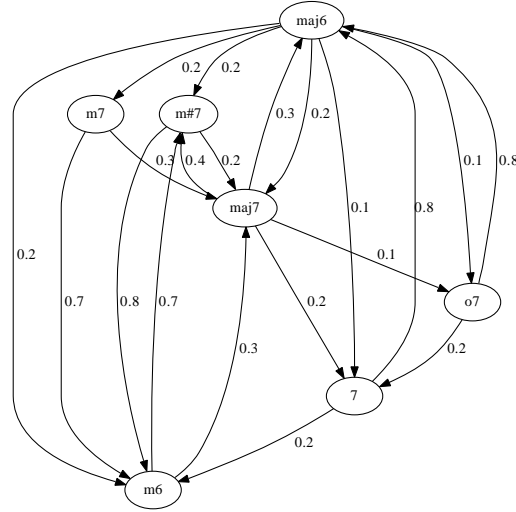


Figure A.1: Transposition probabilities between Markov states: maj6 – Major tonics sixth, m#7 – Minor tonics seventh, m7 – Minor diminished seventh, maj7 – Major tonics seventh, 7 – Major dominant, m6 – Minor tonics sixth, o7 – Minor dominant seventh.

is played forwards or backwards, whenever  $A_{i,j} = "+"$  or  $A_{i,j} = "-"$ , respectively. As a default  $A_{i,j} \neq 0$  is set to  $A_{i,j} = "+"$ .

**Compression parameter.** We distinguish two ways of playing the patterns. In general patterns are played such that every note is played separately in order of their appearance in the pattern. If we switch on this parameter and unit  $u_i$  is present in *all* species then for all species  $S \in \mathbb{S}$  the chosen instruments are playing the first notes of each of the respective patterns  $f(S)$  as one chord.

**Probabilistic parameter.** This parameter allows the possibility to alter the harmony for all patterns. A *transposition* of a pattern moves all notes up or down in pitch by a constant number of semitones. We transpose every pattern whenever unit  $u_i$  is present in *all* species. The transposition is chosen by a probability, depending on the pitch of the current pattern (Figure A.1). Thus a transposition maps a pattern  $P_j$  to pattern  $P'_j$ , which defines the new  $P_j$ . This process is well-known as *first-order Markov chain*.

### Invertability of the Mapping

Clearly, it is desirable to introduce a mapping that is not only able to translate information to music but that also provides a unique way to retrieve the information from the acoustic representation.



If we switch off all parameters it is easy to see that we can determine the species  $S_i$  by their values  $f(S_i)$  since  $f : \mathbb{S} \rightarrow A' \subseteq \mathbb{A}$  with  $A' = \{f(S) \mid S \in \mathbb{S}\}$  is a bijective function.

**Orientation parameter – Induced Constraints.** If we want to distinguish figure out what we hear for a particular sequence  $S \in \mathbb{S}$  it must be possible to distinguish whether  $f(S)$  is played forwards or backwards. Thus no symmetric patterns are allowed. Moreover, it is not allowed to have patterns  $P, P' \in \mathbb{P}$  such that playing  $P$  backwards sounds just like  $P'$  in forward direction and vice versa.

**Compression parameter – Induced Constraints.** If some unit occurs in all species  $S \in \mathbb{S}$  then for all  $S$  the corresponding  $f(S)$  the first note of each pattern is used to play a single chord. Thus, the first note of each pattern must consist of notes that are in the underlying harmony of the pattern and may not be non-chord tones. As a consequence, species information encoded in  $\mathbb{P}$  is lost during compression. However, we argue that the qualitative information “presence in all species given” is sufficient in most cases.

**Probabilistic parameter – Induced Constraints.** This parameter requires more restrictions on instrument and pattern usage if we want to distinguish different species  $S$  by listening to their respective values  $f(S)$ . We will denote  $f_1(S)$  and  $f_2(S)$ , resp., as the instrument and the pattern of  $S$ , resp.

We can distinguish two cases. First for all species  $S \neq S'$  holds that the instrument are unequal ( $f_1(S) \neq f_1(S')$ ). Then we can ignore pattern, since each species is determined by its instrument.

If some species  $S$  and  $S'$  have the same instruments we have to distinguish them by their particular pattern. Thus it is not allowed that any transposition of  $f_2(S)$  leads to  $f_2(S')$  or a transposition of  $f_2(S')$ . In addition, if we have switched on the orientation parameter we must ensure that no transposition leads to a symmetric pattern. The latter case will never occur since no pattern is symmetric and by definition of the term transposition.

## Implementation

Our program COMPOSALIGN consists of a back end for the composition of the music using COMMON MUSIC [454] which runs in Gauche Scheme [455]. COMMON MUSIC is a valuable toolbox for algorithmic composition and also for outputting

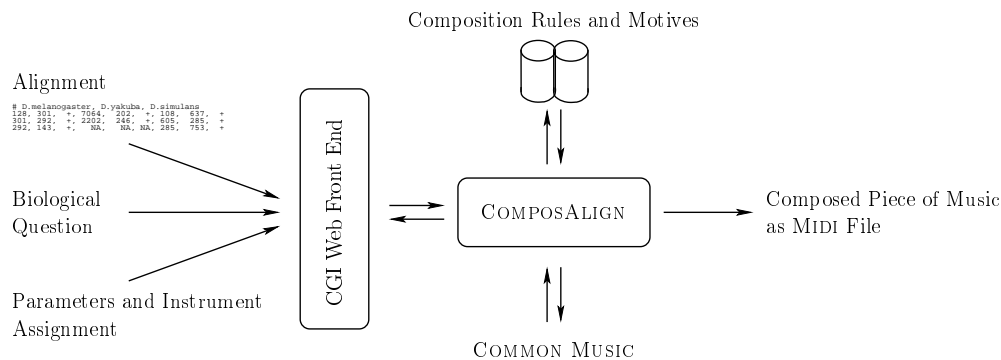


Figure A.2: Data flow diagram of COMPOSALIGN. An alignment (input data), a biological question, and parameter settings and mapping are given to COMPOSALIGN via the front-end [www2.bioinf.uni-leipzig.de/cgi-bin/ComposAlign](http://www2.bioinf.uni-leipzig.de/cgi-bin/ComposAlign). Using a list of prepared motives and mapping rules a piece of music is composed.

MIDI data. It allows for a high level description of the compositional elements and convenient definition of the transformation process due to the expressive power of SCHEME. Additionally, there is a web front-end written in Haskell [456] acting as a CGI program<sup>1</sup>, which allows easy usage without the need to install additional software. The data flow is depicted in Figure A.2.

The user can upload an input file. After the initial analysis of the file and automatic selection of settings the user has the opportunity to change various parameters. Among these are the selection of the reference sequence and the assignment of musical instrument and motives to the individual sequences. The default settings are the ones discussed in this paper, however, depending on the biological question, a different assignment might be optimal.

The alignment data are transformed to music based on the settings. For this purpose, an appropriate SCHEME file is generated which is in turn processed by COMMON MUSIC to create a MIDI file. The SCHEME contains the collection of motives, the rules for the composition, and the mapping of the species to any of the twelve motives and available instruments. The user can listen to or download the generated piece of music.

**Input.** Unfortunately, there is no standardized format for genomewide alignment data – other than nucleotide and protein sequence alignments – that is general enough to handle complex characters such as genes with several attributes. For our purpose we choose the number of attributes  $c = 3$ . We therefore decided to use

<sup>1</sup><http://www2.bioinf.uni-leipzig.de/cgi-bin/ComposAlign/>

```
# D.melanogaster, D.yakuba, D.simulans
319128, 448301, +, 697064, 742202, +, 376108, 476237, +
448301, 468292, +, 742202, 770246, +, 501605, 521285, +
468292, 470143, +, NA, NA, NA, 521285, 522753, +
2651106, 2690081, +, 7722449, 7772786, -, 2682081, 2724631, +
2690081, 2724012, +, 7687085, 7722449, -, 2724631, 2760070, +
2724012, 2868216, +, 7493667, 7687085, -, 2760070, 2909374, +
2868216, 2878317, +, 721765, 722722, +, 2909374, 2922484, +
```

Table A.1: Input example for COMPOSALIGN with three species.

our own input format, a custom comma separated ASCII file type. It can contain lines of comments at the beginning of the file starting with a “#” symbol. The first comment line is interpreted as a list of sequence IDs tagging the  $m$  sets of columns. All other lines are data lines and are supposed to contain exactly  $c \cdot m$  columns separated with comma, where  $m$  is the number of sequences/species or channels in general. Each block of  $c$  columns contains the genomic start and end positions and an indicator for the direction, “+” for forward or “-” for reverse. If the gene is not present in a sequence, NA is used as the value for all  $c$  fields. In the present implementation, the start positions of the reference sequence are used to sort the rows. Sonification of directional information can be turned on or off. In principle, it is therefore easy to use any tabular data with absence/presence information for sonification with COMPOSALIGN. An example input file and the corresponding output files can be found in the supplemental material at <http://www2.bioinf.uni-leipzig.de/cgi-bin/ComposAlign/>.

## Application to Gene Annotation Alignments

We have chosen for our particular species the particular instruments, connected with particular patterns, see figure A.3 and table A.4. For our particular chosen assignment  $f$  all restrictions for patterns using parameters are fulfilled. Thus, we can uniquely determine in which species  $S$  a unit occurs just listening to  $f(S)$ .

We used the gene annotations and gene correspondences from the 12 sequenced Drosophilid genomes as input (see supplementary file `all.R3.dir.map`) [457]. The input is a matrix  $(c \cdot m) \times n$  with  $n$  rows for  $n$  genes and  $c$  columns for each of the  $m$  species. The genes are treated as independent characters and are either present (denoted by their coordinates) or absent (denoted by a gap character, here “NA”). The order of the genes is of biological relevance, since the order reflects the genomic order in a reference species (here *Drosophila melanogaster*). A single gene can have many properties, e.g. similarity to other genes in the same species, distance to

its neighboring genes, length etc, of which  $c$  will be specified in the input file. Here, we used the position and the relative orientation of the genes. This means that the genes of the reference sequence are assigned a “+” (forward) orientation while the identical and inverted orientation in aligned sequences are assigned a “+” (forward) or “-” (reverse) orientation, respectively.

We attempted to sonify data of this kind in a flexible way. Figure A.3 shows an initial selection of musical motives developed to be assigned to each organism’s gene. These motives were designed so that they could be placed in various registers. They were also created with varied contours and rhythms to aid in them being individually perceivable in a musical texture.

Next, we decided to assign a motive and an instrument to each species. We wanted to have the instrumentation reflect something of the relative closeness of each species. This closeness is part of a biologist’s expert knowledge and reflected in the tree in Figure A.4.

Of the 12 *Drosophila* species, five are very closely related – *D. melanogaster*, *D. simulans*, *D. sechellia*, *D. yakuba* and *D. erecta*. One of them, *D. melanogaster*, is the model organism and reference species, which we placed in a continuous motive in the piano as this provided the basis for the rest of the music. We thus looked to place the other four in strings and woodwinds so as to provide some similarity but also enough timbral and register difference so they could be distinguished (Figure A.4).

In our first trials we translated the alignment file for these species to music by simply mapping each gene to a measure of music. If the gene existed in the corresponded species, either in forward or reverse direction, the motive in the instrument would play, if not the instrument would rest. To include one more piece of pertinent data we also considered the direction of the gene. In this case we decided to simply reverse the motive if the gene was reversed. Being that there are 345 genes in an input file and each measure represented 2 seconds of music based on the tempo we selected, this created a sonification about 11.5 minutes long. Also, since the music did not change harmony, each motive was simply played repeatedly. See supplementary material for example files.

### Employing Compression and Stochasticity

We addressed the issue of overly repeating patterns in two ways. To shorten the length of the sonification and focus on the areas of interest, we decided to “compress” the results by simply playing a *tutti* chord in a quarter or eight-note rhythm whenever a gene was present in all species.

**A**

**B**

Figure A.3: Panel A shows the 12 motifs in forward orientation. Panel B shows the assignment of instruments to the transposed motifs from panel A. The transpositions are based on appropriate instrument ranges. E.g., motive 1 is transposed up two octaves to sound in a more typical flute range. When motive 2 is set to clarinet it is transposed up an octave in order for it to be perceptible when other instruments are sounds.

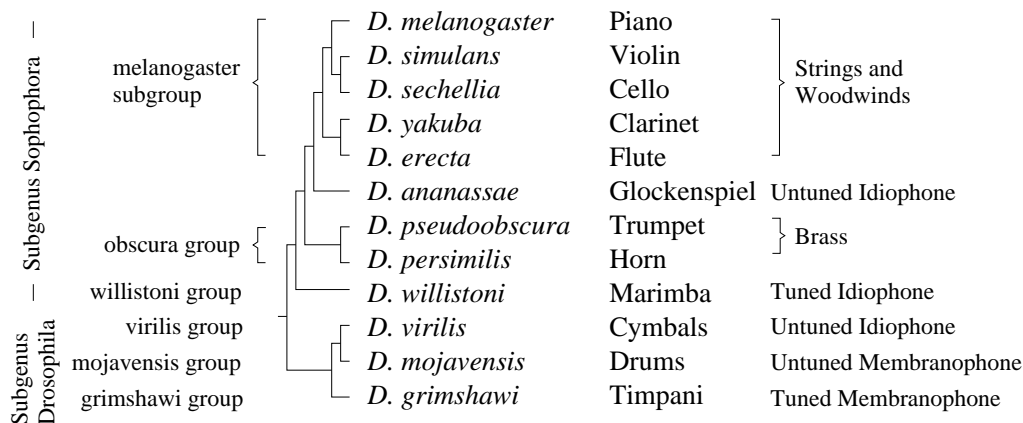


Figure A.4: Mapping of fly species to instruments. The tree on the left-hand side represents the topology of the phylogenetic tree [457]. Branch lengths are arbitrary.

The second change we made was to allow for the possibility of altering the harmony for each measure. Using a simple first-order Markov chain based on some tonal harmonic progressions a new harmony was stochastically chosen before playing either a measure or a chord of a gene present in all species. The motive would be realized in the new harmony, thus providing some pleasant musical variation.

## Results

For a determination of the results we wanted to compare the complexity of the biological data with the perceived complexity of their musical representation. We did this through development of a simple user testing scenario in which we could ask a listener (usually not a musician) if they have been able to gain something from the sonification that would otherwise be difficult to observe from raw input data. Here we show how sonification can present the data, such that answers to biological questions become intuitive.

The following analysis of COMPOSALIGN is based on impressions of 50 non-musician test persons.

**Number of Organisms/Instruments** Depending on the education in the arts of the test persons, up to 12 instruments were distinguished. For most people it was possible to determine up to 6 organisms/instruments. If COMPOSALIGN should be used for 12 species/instruments the majority of people need to be trained to more clearly differentiate the instruments used or we might be able to utilize other

types of instrumental sounds and even create synthesized sounds which would be more easily identified by untrained users.

In the case of 2 or 3 sequences users found it easy to hear which genes were present in which sequence. With just 2 or 3 different instruments and motives, the composition is already musically pleasing. Nevertheless, the untrained listener's ability to resolve the presence/absence pattern decreased rapidly with the number of different instruments and/or motives playing in one measure. In cases where the input contains more than 6 sequences only evolutionary changes, in terms of presence/absence of genes that involve groups of sequences, were found easy to hear.

During the test, persons had to concentrate on a specific instrument and tried to observe the presence/absence of this instrument at a specific time point, most of them easily found the correct solution independently of the number of instruments played concurrently.

**Markov chains.** The introduction of changes in harmony based on the local context improved the artistic value of the output and the listeners attention span. Apart from this aesthetic effect, it also helped emphasized the changes in the presence/absence pattern from one gene to the next. All participants had the impression of a much more interesting piece of music, if the Markov chain was included.

**Conserved genes – compressed chords.** To emphasize conservation, meaning that a character is present in all sequences, we play a *tutti* chord in a quarter or eight-note rhythm. While this sets the presence of  $m$  and  $m - 1$  sequences clearly apart from each other, it also causes a time compression and allows the user to focus on the data where the absence/presence patterns are more informative from a biological perspective. In the context of larger patterns this makes it easier to estimate the amount of characters present in all sequences or sequence groups.

All persons tested the program were enthusiastic after including Markov chain and compressed chords in the variability of the program. The outcome was described much more “happier”, “interesting, irregular”, “less crowded”, “rhythmically interesting” and “dramatic”.

Both the feedback on the compressed chords and the Markov chain harmonic progressions provides an intriguing result in that certain choices that were made largely for aesthetic reasons also appear to make the sonification more legible to users.

**Orientation of a gene.** The asymmetry of the individual motives, some of which are clearly ascending, is an essential attribute to sonify a character's direction information. To do so, we use the forward and reversed motives for the "+" and "-" orientation, respectively. The character of the motives allows the user still to identify the mirrored motives as belonging to the same motive.

The results sound pleasant, however most test persons found it difficult to follow which motives were reversed when several instruments played at the same time. It is unclear if the ear needs some training only or if it might be necessary to explore other strategies which may help in communicating this information.

**Mapping** Using different settings we expected to find combinations that might sound unpleasant. Given an uncommon combination of instruments (e.g. drums, marimba and trumpet) most people found the outcome to be surprisingly rich in character and interesting. When various outputs for the same data file were heard, they all seemed to emphasize the underlying structure in the data. This shows that the motives fit together nicely in any combination, always returning a balanced piece of music which reflects the structure of the data.

The test persons were also asked to listen to two pieces of music and determine what biological information was different (different input files and same mapping function) or if the mapping function changed (same input file and different mapping function). Most people correctly answered these questions.

**Overall impressions** COMPOSALIGN draws its power from the motive design and mapping rules that are modular and flexible. Also biological sequence alignments are particularly suited for sonification since individual elements of information become blurred in a composition when researcher's become more interested in the overall picture (e.g. alignments with many sequences or frequent changes in the absence/presence pattern from one row to the next).

Taking into account the many mapping permutations, a large number of pieces of music can be obtained from a single data file. At the same time, it is possible to answer different biological questions while maintaining a pleasant aesthetic experience.

Some overall comments of the test persons: "surprisingly harmonic on large parts", "a lot of things are good to hear, I get some feeling for the alignments", "to hear biological features becomes hard with more than 10 instruments", "surprisingly fun,



musical motifs are memorable”, “excellent abstraction to the biological information, nice opportunity to listen to nature.”

## Conclusion and Future Work

Our tool COMPOSALIGN is the only existing tool for alignment sonification.

Existing sonification methods for single biological sequences map each individual nucleotide on single notes or chords. It was obvious that this approach would not work well for alignments where multiple sequences are present and multiple notes or chords would sound at the same time. In particular, since the absence/presents and the assignment of the present characters to their origin is of central importance. We therefore decided to map one character to a measure. This had mainly two effects. First, it added the necessary degrees of freedoms to encode more information and still allowed us to take compositional aspects into account and make it sound pleasant. Second, it stretched the information onto a larger time interval, allowed organized presentation of the information with a measure and therefore insured that the information was easy to perceive.

The presented framework and results urge us to ask two major questions: (1) Will it be possible to sonify nucleotide alignments (with annotation) based on the framework presented in this contribution? The mapping of a character to a measure seems promising. However, the definition of “character” and whether single nucleotides or higher order features, e.g. conserved regions, structural or functional elements, shall be treated as characters has a significant impact on the mapping and the biological interpretation of the results. It might turn out that music is a suitable medium to convey information on different levels of resolution at the same time. This leads us immediately to the second question: (2) Can sonification outperform the currently dominating visualization? If not, is sonification able to transport a certain kind of information better than visualization? The omnipresence of visualization might suggest a better performance in all respects. However, to perform a fair test, a competitive sonification tool first needs to be developed.

Based on the experience gained during our project, we intend to construct a mapping for nucleotide alignments that allows us to add different kinds of additional/contextual information (e.g. lengths of characters, distance between characters, higher order annotation, phastcons score). An interactive interface shall allow the user to edit the parameters on runtime and display the scores and alignment in flying windows. This shall allow the interested user to play (with) his/her alignment.



## Appendix B

# Genomes and Accessionnumbers

### B.1 Sources of Used RNA Sequences

Database	Version	Download Location
General Databases		
NCBI	–	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>
UCSC	–	<a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a>
Flybase	–	<a href="http://flybase.org/">http://flybase.org/</a>
EBI	–	<a href="http://www.ebi.ac.uk/">http://www.ebi.ac.uk/</a>
General NcRNA Databases		
Rfam	Version 8.1	<a href="http://www.sanger.ac.uk/">http://www.sanger.ac.uk/</a>
Rfam	Version 9.1	<a href="http://rfam.janelia.org/">http://rfam.janelia.org/</a>
NonCode	v2.0	<a href="http://www.noncode.org/">http://www.noncode.org/</a>
Specific Databases		
Telemerase	–	<a href="http://telomerase.asu.edu/">http://telomerase.asu.edu/</a>
SnoBase	Version 3	<a href="http://www-snorna.biotoul.fr/">http://www-snorna.biotoul.fr/</a>
MirBase	Release 13.0	<a href="http://microrna.sanger.ac.uk/">http://microrna.sanger.ac.uk/</a>
P-Database	Release 12	<a href="http://www.mbio.ncsu.edu/RNaseP/home.html">http://www.mbio.ncsu.edu/RNaseP/home.html</a>
MRP/P Collection	–	<a href="http://bio.lundberg.gu.se/p_mrp/">http://bio.lundberg.gu.se/p_mrp/</a>

## B.2 Sources of SL NcRNAs

### SL RNA

#### Query Sequences

<i>Ciona intestinalis</i>	[215]
<i>Oikopleura dioica</i>	[216]
<i>Caenorhabditis elegans</i>	[213]
<i>Ascaris</i>	[458]
<i>Wucheria bancrofti</i>	[459]
<i>Haemonchus contortus</i>	[460]
<i>Pristionchus pacificus</i>	[461]
<i>Trichinella spiralis</i>	[149]
<i>Schistosoma mansoni</i>	[124]
<i>Fasciola hepatica</i>	[462]
<i>Echinococcus multilocularis</i>	[463]
<i>Schmidtea mediterranea</i>	[464]
<i>Philodina sp.</i>	[217]
<i>Adineta ricciae</i>	[217]
<i>Hydra sp.</i>	[214]
<i>Euglena gracilis</i>	[212]
<i>Entosiphon sulcatum</i>	[178]
<i>Cyclidiopsis acus</i>	[221]
<i>Phacus curvicauda</i>	[221]
<i>Rhabdomonas castata</i>	[221]
<i>Menoidium pellucidum</i>	[221]
<i>Trypanosoma cruzi</i>	[210, 211]
<i>T. vivax</i>	[210, 211]
<i>T. brucei</i>	[210, 211]
<i>Leptomonas collosoma</i>	[210, 211]
<i>Leishmania enriettii</i>	[465]
<i>Crithidia fasciculata</i>	[466]
<i>Bodo caudatus</i>	[467]
<i>Karenia brevis</i>	[148]
<i>Karlodinium micrum</i>	[218]
<i>Pfiesteria piscicida</i>	[218]
<i>Prorocentrum minimum</i>	[218]

Sequences obtained by homology based search

<i>Ascaris</i>	<b><i>AB022045.1</i></b>
<i>Loa</i>	<b><i>U31638.1</i></b>
<i>Mansonella</i>	<b><i>AJ279033.1</i></b>
<i>Acanthocheilonema</i>	<b><i>U31646.1</i></b>
<i>Onchocerca</i>	<b><i>M37737.1</i></b>
<i>Foleyella</i>	<b><i>AJ250988.1</i></b>
<i>Setaria</i>	<b><i>AF282181.1</i></b>
<i>Toxocara</i>	<b><i>U65503.1</i></b>
<i>Enterobius</i>	<b><i>AY234784.1</i></b>
<i>Nippostrongylus</i>	<b><i>EB185208.1</i></b>
<i>Meloidogyne</i>	<b><i>CN443291.1</i></b>
<i>Haemonchus</i>	<b><i>CA994732.1</i></b>
<i>Teladorsagia</i>	<b><i>CB043522.1</i></b>
<i>Echinostoma</i>	<b><i>U85825.1</i></b>
<i>Bdelloidea</i>	<b><i>AY823993.1</i></b>
<i>Herpetomonas</i>	<b><i>AY547489.1</i></b>
<i>Phytomonas</i>	<b><i>AF243335.1</i></b>
<i>Wallaceina</i>	<b><i>AY547488.1</i></b>

### B.3 7SK sequences

<i>Homo sapiens</i>	<b><i>X05490, X04236</i></b> , [335, 347, 468, 469]
<i>Mus musculus</i>	<b><i>M63671</i></b> [470],
<i>Rattus norvegicus</i>	<b><i>K02909</i></b> [471],
<i>Takifugu rubripes</i>	<b><i>AJ890104</i></b> , [171, 338],
<i>Tetraodon nigroviridis</i>	<b><i>AJ890103</i></b> , [338],
<i>Danio rerio</i>	<b><i>AJ890102</i></b> , [338],
<i>Gallus gallus</i>	<b><i>AJ890104</i></b> , [338]

### B.4 FTP Sites of Genome Assemblies

Species	Code	Download Source	Download Date
Homo sapiens	hsa	ftp://ftp.ncbi.nih.gov/genomes/H_sapiens/Assembled_chromosomes	14.09.2006
Pan troglodytes	ptr	http://hgdownload.cse.ucsc.edu/goldenPath/panTro2/bigZips/chromFa.tar.gz	12.08.2008
Pongo pygmaeus	ppy	ensembl	03.07.2008
Macaca mulata	mac	ensembl	08.09.2008
Otolemur garnettii	oga	ensembl	03.07.2008
Microcebus murinus	mmr	ensembl	03.07.2008
Mus musculus	mmu	NCBI	05.05.2008
Rattus norvegicus	rno	http://www.hgsc.bcm.tmc.edu	10.11.2005
Spermophilus tridecemlineatus	str	ensembl	04.07.2008
Cavia porcellus	cpo	ensembl	03.07.2008
Ochotona princeps	opr	ensembl	04.07.2008
Oryctolagus cuniculus	ocu	ensembl	24.09.2008
Tupaia belangeri	tbe	ensembl	03.07.2008
Felis catus	fca	ensembl	03.07.2008
Canis familiaris	cfa	ftp://ftp.ensembl.org/pub/current_fasta/canis_familiaris/dna/	18.07.2007
Bos taurus	bta	ftp://ftp.ensembl.org/pub/current_fasta/bos_taurus/dna/	29.09.2008
Sus scrofa	ssc	ftp://ftp.sanger.ac.uk/pub/S_scrofa/assemblies/PreEnsembl_Sscrofa8/	17.03.2009
Equus caballus	eca	ensembl	03.07.2008
Myotis lucifugus	mlu	ensembl	03.07.2008
Erinaceus europaeus	eeu	ensembl	03.07.2008
Sorex araneus	sar	ensembl	03.07.2008
Loxodonta africana	laf	http://www.broad.mit.edu/ftp/pub/assemblies/mammals/elephant/loxAfr1/	27.06.2005
Echinops telfairi	ete	http://www.broad.mit.edu/ftp/pub/assemblies/mammals/tenrec/echTel1/	13.03.2006
Dasyptes novemcinctus	dno	http://www.broad.mit.edu/ftp/pub/assemblies/mammals/armadillo/dasNov1/	02.12.2008
Choloepus hoffmanni	cho	http://genome.wustl.edu/pub/organism/Other_Vertebrates/Choloepus_hoffmanni/assembly/Choloepus_hoffmanni-1.0/output/	03.06.2008
Monodelphis domestica	mdo	ftp://ftp.ensembl.org/pub/current_monodelphis_domestica/data/fasta/dna/	18.07.2007
Ornithorhynchus anatinus	oan	ensembl	09.04.2007
Anolis carolinensis	acr	http://www.broad.mit.edu/ftp/pub/assemblies/reptiles/lizard/AnoCar1.0/	28.11.2008
Taeniopygia guttata	tgu	NCBI	05.07.2008
Gallus gallus	gga	http://genome.wustl.edu/pub/organism/Other_Vertebrates/Gallus_gallus/assembly/Gallus_gallus-2.1/output/chromosomes/	31.10.2007
Xenopus tropicalis	xtr	USCS	31.10.2007
Tetraodon nigroviridis	tni	ensembl	05.06.2008
Takifugu rubripes	tru	http://genome.jgi-psf.org/Takru4/Takru4.download.ftp.html	05.11.2006
Oryzias latipes	ola	ensembl	05.08.2008
Gasterosteus aculeatus	gac	UCSC	31.10.2007
Danio rerio	dre	ensembl	03.11.2007
Callorhynchus mili	cmi	http://esharkgenome.imcb.a-star.edu.sg/resources.html	27.02.2007
Petromyzon marinus	pma	ftp://genome.wustl.edu/pub/organism/Other_Vertebrates/Petromyzon_marinus/assembly/Petromyzon_marinus-3.0/output/	10.05.2007
Branchiostoma floridae	bfl	ftp://ftp.jgi-psf.org/pub/JGI_data/Branchiostoma_floridae/v1.0/Branchiostoma_floridae_v2.0.assembly.fasta.gz	10.12.2008
Ciona intestinalis	cin	ftp://ftp.jgi-psf.org/pub/JGI_data/Ciona/v2.0/	06.05.2008
Ciona savignyi	csa	http://www.broad.mit.edu/cgi-bin/annotation/ciona/download_license.cgi	18.07.2007
Oikopleura dioica	odi	http://www.genoscope.cns.fr/externe/Download/Projets/Projet_HG/data/assembly/unmasked/	03.09.2008
Strongylocentrotus purpuratus	spu	ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Spurpuratus/fasta/Spur_v2.1/linearScaffolds	31.01.2007
Saccoglossus kowalevskii	sko	ftp://ftp.ensembl.org/pub/traces/saccoglossus_kowalevskii/fasta	22.10.2007

<i>Drosophila melanogaster</i>	dme	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/dm3/bigZips/chromFa.tar.gz">http://hgdownload.cse.ucsc.edu/goldenPath/dm3/bigZips/chromFa.tar.gz</a>	14.05.2008
<i>Drosophila simulans</i>	dsi	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droSim1/bigZips/chromFa.tar.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droSim1/bigZips/chromFa.tar.gz</a>	14.05.2008
<i>Drosophila sechellia</i>	dse	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droSec1/bigZips/scaffoldFa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droSec1/bigZips/scaffoldFa.gz</a>	14.05.2008
<i>Drosophila erecta</i>	der	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droEre2/bigZips/droEre2.fa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droEre2/bigZips/droEre2.fa.gz</a>	14.05.2008
<i>Drosophila yakuba</i>	dya	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droYak2/bigZips/chromFa.tar.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droYak2/bigZips/chromFa.tar.gz</a>	14.05.2008
<i>Drosophila ananassae</i>	dan	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droAna3/bigZips/droAna3.fa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droAna3/bigZips/droAna3.fa.gz</a>	10.05.2008
<i>Drosophila pseudoobscura</i>	dps	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/dp4/bigZips/dp4.fa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/dp4/bigZips/dp4.fa.gz</a>	14.05.2008
<i>Drosophila persimilis</i>	dpe	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droPer1/bigZips/scaffoldFa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droPer1/bigZips/scaffoldFa.gz</a>	14.05.2008
<i>Drosophila willistoni</i>	dwi	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droWil1/bigZips/droWil1.fa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droWil1/bigZips/droWil1.fa.gz</a>	10.05.2008
<i>Drosophila virilis</i>	dvi	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droVir3/bigZips/droVir3.fa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droVir3/bigZips/droVir3.fa.gz</a>	14.05.2008
<i>Drosophila mojavensis</i>	dmo	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droMoj3/bigZips/droMoj3.fa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droMoj3/bigZips/droMoj3.fa.gz</a>	14.05.2008
<i>Drosophila grimshawi</i>	dgr	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/droGri2/bigZips/droGri2.fa.gz">http://hgdownload.cse.ucsc.edu/goldenPath/droGri2/bigZips/droGri2.fa.gz</a>	14.05.2008
<i>Phlebotomus papatasi</i>	ppp	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/phlebotomus_papatasi/fasta.phlebotomus_papatasi.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/phlebotomus_papatasi/fasta.phlebotomus_papatasi.001.gz</a>	17.03.2008
<i>Anopheles gambiae</i>	aga	<a href="ftp://ftp.ensembl.org/pub/current_fasta/anopheles_gambiae/dna/">ftp://ftp.ensembl.org/pub/current_fasta/anopheles_gambiae/dna/</a>	31.05.2008
<i>Aedes aegypti</i>	aae	<a href="ftp://ftp.ensembl.org/pub/current_fasta/aedes_aegypti/dna/">ftp://ftp.ensembl.org/pub/current_fasta/aedes_aegypti/dna/</a>	02.04.2008
<i>Culex pipiens</i>	cpj	<a href="ftp://ftp.vectorbase.org/public_data/organism_data/cpipiens/Geneset/cpipiens.TRANSSCRIPTS-CpipJ1.1.fa.gz">ftp://ftp.vectorbase.org/public_data/organism_data/cpipiens/Geneset/cpipiens.TRANSSCRIPTS-CpipJ1.1.fa.gz</a>	21.03.2008
<i>Bombyx mori</i>	bmo	<a href="http://silkworm.swu.edu.cn/silkdb/doc/download.html">http://silkworm.swu.edu.cn/silkdb/doc/download.html</a>	15.05.2008
<i>Tribolium castaneum</i>	tca	<a href="http://hgdownload.cse.ucsc.edu/goldenPath/triCas2/bigZips/">http://hgdownload.cse.ucsc.edu/goldenPath/triCas2/bigZips/</a>	07.05.2008
<i>Apis mellifera</i>	ame	<a href="ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Amelifera/fasta/Amel20060310-freeze/">ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Amelifera/fasta/Amel20060310-freeze/</a>	31.03.2008
<i>Nasonia girault</i>	ngi	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/nasonia_girault/">ftp://ftp.ncbi.nih.gov/pub/TraceDB/nasonia_girault/</a>	17.03.2008
<i>Nasonia vitripennis</i>	nvi	<a href="ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Nvitripennis/fasta/Nvit_1.0/linearized_sequence/">ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Nvitripennis/fasta/Nvit_1.0/linearized_sequence/</a>	31.03.2008
<i>Pediculus humanus</i>	phu	<a href="ftp://ftp.vectorbase.org/public_data/organism_data/phumanus/Genome/">ftp://ftp.vectorbase.org/public_data/organism_data/phumanus/Genome/</a>	01.04.2008
<i>Acyrtosiphon pisum</i>	api	<a href="ftp://ftp.ensembl.org/pub/traces/acyrtosiphon_pisum/fasta/*">ftp://ftp.ensembl.org/pub/traces/acyrtosiphon_pisum/fasta/*</a>	21.02.2007
<i>Daphnia pulex</i>	dpu	<a href="ftp://ftp.ensembl.org/pub/traces/daphnia_pulex/fasta">ftp://ftp.ensembl.org/pub/traces/daphnia_pulex/fasta</a>	10.11.2007
<i>Ixodes scapularis</i>	isc	<a href="ftp://ftp.ensembl.org/pub/traces/ixodes_scapularis/fasta/*">ftp://ftp.ensembl.org/pub/traces/ixodes_scapularis/fasta/*</a>	21.02.2007
<i>Caenorhabditis remanei</i>	cre	<a href="http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_remanei/assembly/Caenorhabditis_remanei-15.0.1/output/">http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_remanei/assembly/Caenorhabditis_remanei-15.0.1/output/</a>	10.11.2007
<i>Caenorhabditis briggsae</i>	cbr	<a href="http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_briggsae/assembly/Caenorhabditis_briggsae-1.0/output/chromosomes/">http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_briggsae/assembly/Caenorhabditis_briggsae-1.0/output/chromosomes/</a>	10.11.2007
<i>Caenorhabditis brenneri</i>	cbe	<a href="http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_PB2801/assembly/Caenorhabditis_PB2801-6.0.1/output/">http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_PB2801/assembly/Caenorhabditis_PB2801-6.0.1/output/</a>	03.06.2008
<i>Caenorhabditis elegans</i>	cel	<a href="ftp://ftp.wormbase.org/pub/wormbase/genomes/elegans/sequences/dna/">ftp://ftp.wormbase.org/pub/wormbase/genomes/elegans/sequences/dna/</a>	23.05.2008
<i>Caenorhabditis japonica</i>	cja	<a href="http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_japonica/assembly/Caenorhabditis_japonica-3.0.2/output/*gz">http://genome.wustl.edu/pub/organism/Invertebrates/Caenorhabditis_japonica/assembly/Caenorhabditis_japonica-3.0.2/output/*gz</a>	03.06.2008
<i>Haemonchus contortus</i>	hco	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/haemonchus_contortus">ftp://ftp.ncbi.nih.gov/pub/TraceDB/haemonchus_contortus</a>	16.05.2008
<i>Ancylostoma caninum</i>	acn	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/ancylostoma_caninum">ftp://ftp.ncbi.nih.gov/pub/TraceDB/ancylostoma_caninum</a>	16.05.2008
<i>Pristionchus pacificus</i>	ppa	<a href="http://genome.wustl.edu/pub/organism/Invertebrates/Pristionchus_pacificus/assembly/Pristionchus_pacificus-5.0.1/output/">http://genome.wustl.edu/pub/organism/Invertebrates/Pristionchus_pacificus/assembly/Pristionchus_pacificus-5.0.1/output/</a>	03.06.2008
<i>Strongyloides ratti</i>	sra	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/strongyloides_ratti">ftp://ftp.ncbi.nih.gov/pub/TraceDB/strongyloides_ratti</a>	16.05.2008
<i>Meloidogyne incognita</i>	min	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/meloidogyne_incognita">ftp://ftp.ncbi.nih.gov/pub/TraceDB/meloidogyne_incognita</a>	16.05.2008
<i>Ascaris suum</i>	asu	<a href="http://www.nematode.net/FTP/wgs_ftp/*WGS">http://www.nematode.net/FTP/wgs_ftp/*WGS</a>	16.05.2008
<i>Brugia malayi</i>	bma	<a href="ftp://ftp.ncbi.nih.gov/genomes/Brugia_malayi/FASTA/">ftp://ftp.ncbi.nih.gov/genomes/Brugia_malayi/FASTA/</a>	01.04.2008
<i>Trichinella spiralis</i>	tsp	<a href="ftp://ftp.ensembl.org/pub/traces/trichinella_spiralis/fasta">ftp://ftp.ensembl.org/pub/traces/trichinella_spiralis/fasta</a>	10.11.2007
<i>Schistosoma mansoni</i>	sma	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma_mansoni/genome/Assembly-v3.1/">ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma_mansoni/genome/Assembly-v3.1/</a>	29.06.2007
<i>Schistosoma haematobium</i>	sha	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma_haematobium/Shaem.tar.gz">ftp://ftp.sanger.ac.uk/pub/pathogens/Schistosoma_haematobium/Shaem.tar.gz</a>	07.11.2008
<i>Schistosoma japonicum</i>	sja	<a href="ftp://down:lsbi@lifecenter.sgst.cn:2121/subjectData/schistosoma/sjc_mRNA.zip">ftp://down:lsbi@lifecenter.sgst.cn:2121/subjectData/schistosoma/sjc_mRNA.zip</a>	03.09.2008
<i>Schmidtea mediterranea</i>	sme	<a href="http://genome.wustl.edu/pub/organism/Invertebrates/Schmidtea_mediterranea/assembly/Schmidtea_mediterranea-3.1/output/">http://genome.wustl.edu/pub/organism/Invertebrates/Schmidtea_mediterranea/assembly/Schmidtea_mediterranea-3.1/output/</a>	27.05.2008
<i>Echinococcus multilocularis</i>	emu	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/Echinococcus/">ftp://ftp.sanger.ac.uk/pub/pathogens/Echinococcus/</a>	07.11.2008

Helobdella robusta	hro	<a href="http://genome.jgi-psf.org/Helro1/Helro1.download.ftp.html">http://genome.jgi-psf.org/Helro1/Helro1.download.ftp.html</a>	26.10.2008
Capitella sp	cca	<a href="ftp://ftp.jgi-psf.org/pub/JGI_data/Capitella/v1.0">ftp://ftp.jgi-psf.org/pub/JGI_data/Capitella/v1.0</a>	11.11.2007
Lottia gigantea	lgi	<a href="ftp://ftp.ensembl.org/pub/traces/lottia_gigantea/fasta">ftp://ftp.ensembl.org/pub/traces/lottia_gigantea/fasta</a>	16.11.2007
Aplysia californica	aca	<a href="ftp://ftp.ensembl.org/pub/traces/aplysia_californica/fasta/*wgs.fasta.gz">ftp://ftp.ensembl.org/pub/traces/aplysia_californica/fasta/*wgs.fasta.gz</a>	28.08.2008
Biomphalaria glabrata	bgl	<a href="ftp://ftp.ensembl.org/pub/traces/biomphalaria_glabrata/fasta/*">ftp://ftp.ensembl.org/pub/traces/biomphalaria_glabrata/fasta/*</a>	16.11.2007
Euprymna scolopes	esc	<a href="ftp://ftp.ensembl.org/pub/traces/euprymna_scolopes/fasta/">ftp://ftp.ensembl.org/pub/traces/euprymna_scolopes/fasta/</a>	02.11.2007
Spisula solidissima	sso	<a href="ftp://ftp.ensembl.org/pub/traces/spisula_solidissima/fasta/*">ftp://ftp.ensembl.org/pub/traces/spisula_solidissima/fasta/*</a>	21.02.2007
Cerebratulus lacteus	cla	<a href="ftp://ftp.ensembl.org/pub/traces/cerebratulus_lacteus/fasta/*">ftp://ftp.ensembl.org/pub/traces/cerebratulus_lacteus/fasta/*</a>	21.02.2007
Acropora palmata	apa	<a href="ftp://ftp.ensembl.org/pub/traces/acropora_palmata/fasta">ftp://ftp.ensembl.org/pub/traces/acropora_palmata/fasta</a>	11.06.2007
Acropora millepora	ami	<a href="ftp://ftp.ensembl.org/pub/traces/acropora_millepora/fasta">ftp://ftp.ensembl.org/pub/traces/acropora_millepora/fasta</a>	11.06.2007
Porites lobata	plo	<a href="ftp://ftp.ensembl.org/pub/traces/porites_lobata/fasta/">ftp://ftp.ensembl.org/pub/traces/porites_lobata/fasta/</a>	16.11.2007
Nematostella vectensis	nve	<a href="http://genome.jgi-psf.org/Nemvel1/Nemvel1.home.html">http://genome.jgi-psf.org/Nemvel1/Nemvel1.home.html</a>	15.05.2007
Hydra magnipapillata	hma	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/hydra_magnipapillata/">ftp://ftp.ncbi.nih.gov/pub/TraceDB/hydra_magnipapillata/</a>	11.06.2007
Reniera spez	rsp	<a href="ftp://ftp.ensembl.org/pub/traces/reniera_sp_jgi_2005/fasta/*">ftp://ftp.ensembl.org/pub/traces/reniera_sp_jgi_2005/fasta/*</a>	26.11.2007
Trichoplax adhaerens	tad	<a href="http://genome.jgi-psf.org/Triad1/Triad1.download.ftp.html">http://genome.jgi-psf.org/Triad1/Triad1.download.ftp.html</a>	09.12.2007
Alternaria brassicicola	fabr	<a href="http://genome.wustl.edu/pub/organism/Fungi/Alternaria_brassicicola/assembly/Alternaria_brassicicola-1.0/output">http://genome.wustl.edu/pub/organism/Fungi/Alternaria_brassicicola/assembly/Alternaria_brassicicola-1.0/output</a>	04.07.2008
Stagonospora nodorum	fsno	<a href="http://www.broad.mit.edu/annotation/genome/stagonospora_nodorum.2/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/stagonospora_nodorum.2/MultiDownloads.html</a>	04.07.2008
Mycosphaerella graminicola	fmgf	<a href="http://genome.jgi-psf.org/Mycgr1/Mycgr1.download.ftp.html">http://genome.jgi-psf.org/Mycgr1/Mycgr1.download.ftp.html</a>	04.07.2008
Ascospaera apis	faap	<a href="ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Aapis/Aapis-01Jun2006-contigs">ftp://ftp.hgsc.bcm.tmc.edu/pub/data/Aapis/Aapis-01Jun2006-contigs</a>	04.07.2008
Coccidioides immitis	feim	<a href="http://www.broad.mit.edu/annotation/genome/coccidioides_group/MultiHome.html">http://www.broad.mit.edu/annotation/genome/coccidioides_group/MultiHome.html</a>	04.07.2008
Coccidioides posadasii	fcpo	<a href="http://www.broad.mit.edu/annotation/genome/coccidioides_group/MultiHome.html">http://www.broad.mit.edu/annotation/genome/coccidioides_group/MultiHome.html</a>	04.07.2008
Histoplasma capsulatum	fhca	<a href="http://www.broad.mit.edu/annotation/genome/histoplasma_capsulatum/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/histoplasma_capsulatum/MultiDownloads.html</a>	04.07.2008
Paracoccidioides brasiliensis	fpbr	<a href="http://www.broad.mit.edu/annotation/genome/paracoccidioides_brasiliensis/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/paracoccidioides_brasiliensis/MultiDownloads.html</a>	04.07.2008
Uncinocarpus reesii	fure	<a href="http://www.broad.mit.edu/annotation/genome/uncinocarpus_reesii.3/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/uncinocarpus_reesii.3/MultiDownloads.html</a>	04.07.2008
Neosartorya fischeri	fnfi	<a href="http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html</a>	04.07.2008
Aspergillus clavatus	facl	<a href="http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html</a>	04.07.2008
Aspergillus flavus	fafl	<a href="http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html</a>	04.07.2008
Aspergillus fumigatus	fafu	<a href="http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html</a>	04.07.2008
Aspergillus nidulans	fani	<a href="http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html</a>	04.07.2008
Aspergillus niger	fang	<a href="http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html</a>	04.07.2008
Aspergillus oryzae	faor	<a href="http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html</a>	04.07.2008
Aspergillus terreus	fate	<a href="http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/aspergillus_group/MultiDownloads.html</a>	04.07.2008
Botrytis cinerea	fbci	<a href="http://www.broad.mit.edu/annotation/genome/botrytis_cinerea.2/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/botrytis_cinerea.2/MultiDownloads.html</a>	04.07.2008
Sclerotinia sclerotiorum	fssc	<a href="http://www.broad.mit.edu/annotation/genome/sclerotinia_sclerotiorum.2/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/sclerotinia_sclerotiorum.2/MultiDownloads.html</a>	04.07.2008
Chaetomium globosum	fcgo	<a href="http://www.broad.mit.edu/annotation/genome/chaetomium_globosum.2/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/chaetomium_globosum.2/MultiDownloads.html</a>	04.07.2008
Neurospora crassa	fnrc	<a href="http://www.broad.mit.edu/annotation/genome/neurospora/assets/neurospora_crassa_7.fasta.gz">http://www.broad.mit.edu/annotation/genome/neurospora/assets/neurospora_crassa_7.fasta.gz</a>	20.06.2008
Neurospora discreta	fndi	ncbi	20.06.2008
Neurospora tetrasperma	fnre	ncbi	20.06.2008
Podospora anserina	fpan	<a href="http://podospora.igmors.u-psud.fr/download.html">http://podospora.igmors.u-psud.fr/download.html</a>	04.07.2008
Magnaporthe grisea	fmgf	<a href="http://www.broad.mit.edu/annotation/genome/magnaporthe_grisea/Downloads.html">http://www.broad.mit.edu/annotation/genome/magnaporthe_grisea/Downloads.html</a>	04.07.2008
Fusarium oxysporum	ffox	<a href="http://www.broad.mit.edu/annotation/genome/fusarium_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/fusarium_group/MultiDownloads.html</a>	04.07.2008
Fusarium graminearum	ffgr	<a href="http://www.broad.mit.edu/annotation/genome/fusarium_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/fusarium_group/MultiDownloads.html</a>	04.07.2008
Fusarium verticillioides	ffve	<a href="http://www.broad.mit.edu/annotation/genome/fusarium_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/fusarium_group/MultiDownloads.html</a>	04.07.2008
Trichoderma reesei	ftre	<a href="http://genome.jgi-psf.org/Trire2/Trire2.download.ftp.html">http://genome.jgi-psf.org/Trire2/Trire2.download.ftp.html</a>	04.07.2008
Gibberella zeae	fgze	<a href="ftp://ftp.ncbi.nih.gov/genomes/Fungi/Gibberella_zeae">ftp://ftp.ncbi.nih.gov/genomes/Fungi/Gibberella_zeae</a>	04.07.2008
Nectria haematococca	fhae	<a href="http://genome.jgi-psf.org/Necha2/Necha2.download.ftp.html">http://genome.jgi-psf.org/Necha2/Necha2.download.ftp.html</a>	04.07.2008
Candida albicans	fcad	<a href="http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html</a>	04.07.2008
Candida dubliniensis	fcdu	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/Candida/dubliniensis">ftp://ftp.sanger.ac.uk/pub/pathogens/Candida/dubliniensis</a>	04.07.2008
Candida parapsilosis	fcpa	<a href="http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html</a>	04.07.2008



<i>Candida guilliermondii</i>	fegu	<a href="http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html</a>	04.07.2008
<i>Candida lusitanae</i>	felu	<a href="http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html</a>	04.07.2008
<i>Candida glabrata</i>	fcgl	<a href="http://cbi.labri.fr/Genolevures/download/CAGL_chromosomes.php">http://cbi.labri.fr/Genolevures/download/CAGL_chromosomes.php</a>	04.07.2008
<i>Yarrowia lipolytica</i>	fyli	<a href="http://cbi.labri.fr/Genolevures/download/sequence/">http://cbi.labri.fr/Genolevures/download/sequence/</a>	04.07.2008
<i>Debaryomyces hansenii</i>	fdha	<a href="http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html</a>	04.07.2008
<i>Lodderomyces elongisporus</i>	fiel	<a href="http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/candida_albicans/MultiDownloads.html</a>	04.07.2008
<i>Saccharomyces cerevisiae</i>	fsce	<a href="ftp://ftp.ncbi.nih.gov/genomes/Saccharomyces_cerevisiae/">ftp://ftp.ncbi.nih.gov/genomes/Saccharomyces_cerevisiae/</a>	17.03.2009
<i>Saccharomyces bayanus</i>	fsba	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharomyces_bayanus/fasta.saccharomyces_bayanus.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharomyces_bayanus/fasta.saccharomyces_bayanus.001.gz</a>	07.07.2008
<i>Saccharomyces kluyveri</i>	fskl	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharomyces_kluyveri/fasta.saccharomyces_kluyveri.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharomyces_kluyveri/fasta.saccharomyces_kluyveri.001.gz</a>	07.07.2008
<i>Saccharomyces mikatae</i>	fsmi	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharomyces_mikatae/fasta.saccharomyces_mikatae.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharomyces_mikatae/fasta.saccharomyces_mikatae.001.gz</a>	07.07.2008
<i>Saccharomyces paradoxus</i>	fspa	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharomyces_paradoxus/fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharomyces_paradoxus/fasta*</a>	07.07.2008
<i>Saccharomyces degradans</i>	fsde	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharophagus_degradans_2-40/fasta.saccharophagus_degradans_2-40.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/saccharophagus_degradans_2-40/fasta.saccharophagus_degradans_2-40.001.gz</a>	07.07.2008
<i>Kluyveromyces lactis</i>	fkla	<a href="http://cbi.labri.fr/Genolevures/download/sequence/">http://cbi.labri.fr/Genolevures/download/sequence/</a>	04.07.2008
<i>Pichia stipitis</i>	fpst	<a href="ftp://ftp.ncbi.nih.gov/genomes/Fungi/Pichia_stipitis/">ftp://ftp.ncbi.nih.gov/genomes/Fungi/Pichia_stipitis/</a>	04.07.2008
<i>Pichia guilliermondii</i>	fpgu	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/pichia_guilliermondii/fasta.pichia_guilliermondii.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/pichia_guilliermondii/fasta.pichia_guilliermondii.001.gz</a>	07.07.2008
<i>Eremothecium gossypii</i>	fego	<a href="ftp://ftp.ncbi.nih.gov/genomes/Fungi/Eremothecium_gossypii">ftp://ftp.ncbi.nih.gov/genomes/Fungi/Eremothecium_gossypii</a>	04.07.2008
<i>Schizosaccharomyces japonicus</i>	fsja	<a href="http://www.broad.mit.edu/annotation/genome/schizosaccharomyces_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/schizosaccharomyces_group/MultiDownloads.html</a>	04.07.2008
<i>Schizosaccharomyces octosporus</i>	fsoc	<a href="http://www.broad.mit.edu/annotation/genome/schizosaccharomyces_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/schizosaccharomyces_group/MultiDownloads.html</a>	04.07.2008
<i>Schizosaccharomyces pombe</i>	fspo	<a href="http://www.broad.mit.edu/annotation/genome/schizosaccharomyces_group/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/schizosaccharomyces_group/MultiDownloads.html</a>	04.07.2008
<i>Phaeerochaete chrysosporium</i>	fpch	<a href="http://genome.jgi-psf.org/Phchr1/Phchr1.download.ftp.html">http://genome.jgi-psf.org/Phchr1/Phchr1.download.ftp.html</a>	04.07.2008
<i>Postia placenta</i>	fppl	<a href="ftp://ftp.jgi-psf.org/pub/JGI_data/Postia_placenta/v1.0/Postia_placenta.fasta.gz">ftp://ftp.jgi-psf.org/pub/JGI_data/Postia_placenta/v1.0/Postia_placenta.fasta.gz</a>	07.07.2008
<i>Laccaria bicolor</i>	fbli	<a href="ftp://ftp.jgi-psf.org/pub/JGI_data/Laccaria_bicolor/laccaria.fasta.gz">ftp://ftp.jgi-psf.org/pub/JGI_data/Laccaria_bicolor/laccaria.fasta.gz</a>	06.07.2008
<i>Coprinus cinereus</i>	fecu	<a href="http://www.broad.mit.edu/annotation/genome/coccidioides_group/MultiHome.html">http://www.broad.mit.edu/annotation/genome/coccidioides_group/MultiHome.html</a>	04.07.2008
<i>Cryptococcus neoformans</i>	fcne	<a href="http://www.broad.mit.edu/annotation/genome/cryptococcus_neoformans.2/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/cryptococcus_neoformans.2/MultiDownloads.html</a>	04.07.2008
<i>Sporobolomyces roseus</i>	fsro	<a href="http://genome.jgi-psf.org/Sporo1/Sporo1.download.ftp.html">http://genome.jgi-psf.org/Sporo1/Sporo1.download.ftp.html</a>	04.07.2008
<i>Phakopsora pachyrhizi</i>	fpfa	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/phakopsora_pachyrhizi">ftp://ftp.ncbi.nih.gov/pub/TraceDB/phakopsora_pachyrhizi</a>	07.07.2008
<i>Puccinia graminis tritici</i>	fpgr	<a href="http://www.broad.mit.edu/annotation/genome/puccinia_graminis.3/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/puccinia_graminis.3/MultiDownloads.html</a>	04.07.2008
<i>Ustilago maydis</i>	fuma	<a href="http://www.broad.mit.edu/annotation/genome/ustilago_maydis.2/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/ustilago_maydis.2/MultiDownloads.html</a>	04.07.2008
<i>Allomyces macrogynus</i>	fama	NCBI	07.07.2008
<i>Batrachochytrium dendrobatidis</i>	fbde	<a href="http://www.broad.mit.edu/annotation/genome/batrachochytrium_dendrobatidis.3/download/?sp=EASupercontigs-Fasta&amp;sp=SBD_JEL423&amp;sp=S.zip">http://www.broad.mit.edu/annotation/genome/batrachochytrium_dendrobatidis.3/download/?sp=EASupercontigs-Fasta&amp;sp=SBD_JEL423&amp;sp=S.zip</a>	07.07.2008
<i>Spizellomyces punctatus</i>	fspu	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/spizellomyces_punctatus/fasta.spizellomyces_punctatus.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/spizellomyces_punctatus/fasta.spizellomyces_punctatus.001.gz</a>	07.07.2008
<i>Antonospora locustae</i>	falo	<a href="http://gmod.mbl.edu/perl/site/antonospora01?page=download">http://gmod.mbl.edu/perl/site/antonospora01?page=download</a>	04.07.2008
<i>Encephalitozoon cuniculi</i>	fecu	<a href="ftp://ftp.ncbi.nih.gov/genomes/Fungi/Encephalitozoon_cuniculi">ftp://ftp.ncbi.nih.gov/genomes/Fungi/Encephalitozoon_cuniculi</a>	04.07.2008
<i>Rhizopus oryzae</i>	fror	<a href="http://www.broad.mit.edu/annotation/genome/pyrenophora_tritici_repentis.3/MultiDownloads.html">http://www.broad.mit.edu/annotation/genome/pyrenophora_tritici_repentis.3/MultiDownloads.html</a>	04.07.2008
<i>Phycomyces blakesleeana</i>	fpbl	<a href="ftp://ftp.jgi-psf.org/pub/JGI_data/Phycomyces_blakesleeana/assembly/v1.0/Phybl1_scaffolds.fasta.gz">ftp://ftp.jgi-psf.org/pub/JGI_data/Phycomyces_blakesleeana/assembly/v1.0/Phybl1_scaffolds.fasta.gz</a>	07.07.2008
<i>Acanthamoeba castellanii</i>	aces	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/acanthamoeba_castellanii">ftp://ftp.ncbi.nih.gov/pub/TraceDB/acanthamoeba_castellanii</a>	11.05.2007
<i>Entamoeba histolytica</i>	ehi	<a href="ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/e_histolytica/whole_genome_sequencing/HISTOLYTICA.SINGLE-TONS.seq">ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/e_histolytica/whole_genome_sequencing/HISTOLYTICA.SINGLE-TONS.seq</a>	06.07.2008
<i>Dictyostelium discoideum</i>	ddi	<a href="http://dictybase.org/db/cgi-bin/dictyBase/download/download.pl?area=blast_databases&amp;ID=dicty_chromosomal.gz">http://dictybase.org/db/cgi-bin/dictyBase/download/download.pl?area=blast_databases&amp;ID=dicty_chromosomal.gz</a>	04.07.2008
<i>Physarum polycephalum</i>	ppo	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/physarum_polycephalum/fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/physarum_polycephalum/fasta*</a>	07.07.2008
<i>Giardia lamblia</i>	gla	<a href="http://www.giardiadb.org/common/downloads/release1.1/GlambliGenomic_GiardiaDB-1.1.fasta">http://www.giardiadb.org/common/downloads/release1.1/GlambliGenomic_GiardiaDB-1.1.fasta</a>	06.07.2008
<i>Trichomonas vaginalis</i>	tva	<a href="ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_vaginalis/whole_genome_sequencing/T.vaginalis_Scaffolds_20050331.fasta.gz">ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_vaginalis/whole_genome_sequencing/T.vaginalis_Scaffolds_20050331.fasta.gz</a>	02.10.2007
<i>Emiliana huxleyi</i>	ehu	<a href="ftp://ftp.jgi-psf.org/pub/JGI_data/Emiliana_huxleyi/assembly/v1.0/Emihu1_scaffolds.fasta.gz">ftp://ftp.jgi-psf.org/pub/JGI_data/Emiliana_huxleyi/assembly/v1.0/Emihu1_scaffolds.fasta.gz</a>	07.07.2008
<i>Naegleria gruberi</i>	ngr	<a href="ftp://ftp.jgi-psf.org/pub/JGI_data/Naegleria_gruberi/assembly/v1.0/Naegr1_scaffolds.fasta.gz">ftp://ftp.jgi-psf.org/pub/JGI_data/Naegleria_gruberi/assembly/v1.0/Naegr1_scaffolds.fasta.gz</a>	07.07.2008
<i>Leishmania braziliensis</i>	lbr	<a href="http://www.sanger.ac.uk/Projects/L_braziliensis/">http://www.sanger.ac.uk/Projects/L_braziliensis/</a>	23.04.2007
<i>Leishmania infantum</i>	lin	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/L_infantum/DATASETS/LinJwholegenome_20080508.v3.0a.fasta">ftp://ftp.sanger.ac.uk/pub/pathogens/L_infantum/DATASETS/LinJwholegenome_20080508.v3.0a.fasta</a>	04.07.2008
<i>Leishmania major</i>	lma	<a href="ftp://ftp.sanger.ac.uk/pub/databases/L_major_sequences/DATASETS/LmjFwholegenome_20070731_V5.2.fasta">ftp://ftp.sanger.ac.uk/pub/databases/L_major_sequences/DATASETS/LmjFwholegenome_20070731_V5.2.fasta</a>	05.07.2008

Trypanosoma brucei	tbr	<a href="ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_brucei/annotation_dbs/">ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_brucei/annotation_dbs/</a>	16.06.2007
Trypanosoma congolense	tco	<a href="ftp://ftp.sanger.ac.uk/pub/databases/T.congolense_sequences/May2007_phusion_assembly/Tcongo_phusion_scaffs.fas.gz">ftp://ftp.sanger.ac.uk/pub/databases/T.congolense_sequences/May2007_phusion_assembly/Tcongo_phusion_scaffs.fas.gz</a>	06.07.2008
Trypanosoma cruzi	tcr	<a href="http://teruzidb.org/teruzidb/">http://teruzidb.org/teruzidb/</a>	14.05.2007
Trypanosoma vivax	tvi	<a href="http://www.sanger.ac.uk/Projects/T_vivax/">http://www.sanger.ac.uk/Projects/T_vivax/</a>	18.07.2007
Paramecium tetraurelia	pte	<a href="http://www.genoscope.cns.fr/externe/Francais/Projets/Projet_FN/data/assembly/unmasked/Ptetraurelia_V2.1.fasta">www.genoscope.cns.fr/externe/Francais/Projets/Projet_FN/data/assembly/unmasked/Ptetraurelia_V2.1.fasta</a>	15.05.2007
Tetrahymena thermophila	tth	<a href="ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_thermophila/Assemblies_and_Sequences/Assembly_ttg_2.1_Dec-2007.fasta">ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_thermophila/Assemblies_and_Sequences/Assembly_ttg_2.1_Dec-2007.fasta</a>	04.07.2008
Oxytricha trifallax	otr	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/oxytricha_trifallax">ftp://ftp.ncbi.nih.gov/pub/TraceDB/oxytricha_trifallax</a>	07.07.2008
Plasmodium falciparum	pfa	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/plasmodium_falciparum/">ftp://ftp.ncbi.nih.gov/pub/TraceDB/plasmodium_falciparum/</a>	06.07.2008
Plasmodium knowlesi	pkn	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/plasmodium_knowlesi">ftp://ftp.ncbi.nih.gov/pub/TraceDB/plasmodium_knowlesi</a>	07.07.2008
Plasmodium vivax	pvi	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/plasmodium_vivax_sai-1">ftp://ftp.ncbi.nih.gov/pub/TraceDB/plasmodium_vivax_sai-1</a>	07.07.2008
Plasmodium reichenowi	pre	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/plasmodium_reichenowi">ftp://ftp.ncbi.nih.gov/pub/TraceDB/plasmodium_reichenowi</a>	07.07.2008
Plasmodium berghei	pbe	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/P_berghei/version2/BERG.contigs_111007.fasta">ftp://ftp.sanger.ac.uk/pub/pathogens/P_berghei/version2/BERG.contigs_111007.fasta</a>	07.07.2008
Plasmodium gallinaceum	pga	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/gallinaceum/P_gallinaceum.phusion_supercontigs.180705">ftp://ftp.sanger.ac.uk/pub/pathogens/Plasmodium/gallinaceum/P_gallinaceum.phusion_supercontigs.180705</a>	07.07.2008
Theileria parva	tpa	<a href="ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_parva/annotation_dbs/*1con">ftp://ftp.tigr.org/pub/data/Eukaryotic_Projects/t_parva/annotation_dbs/*1con</a>	07.08.2008
Theileria annulata	tan	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/T_annulata/TANN.contigs.fasta.092304">ftp://ftp.sanger.ac.uk/pub/pathogens/T_annulata/TANN.contigs.fasta.092304</a>	07.07.2008
Babesia bigemina	bbi	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/babesia_bigemina.fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/babesia_bigemina.fasta*</a>	07.07.2008
Cryptosporidium hominis	chm	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/cryptosporidium_hominis.fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/cryptosporidium_hominis.fasta*</a>	07.07.2008
Cryptosporidium muris	cmu	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/cryptosporidium_muris.fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/cryptosporidium_muris.fasta*</a>	08.07.2008
Cryptosporidium parvum	cpa	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/cryptosporidium_parvum.fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/cryptosporidium_parvum.fasta*</a>	08.07.2008
Eimeria tenella	etn	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/genome/assemblies/assembly_2007_05_08.gz">ftp://ftp.sanger.ac.uk/pub/pathogens/Eimeria/tenella/genome/assemblies/assembly_2007_05_08.gz</a>	08.07.2008
Neospora caninum	nca	<a href="ftp://ftp.sanger.ac.uk/pub/pathogens/Neospora/caninum/NEOS.contigs.version1/NEOS.contigs.version1.0.fasta">ftp://ftp.sanger.ac.uk/pub/pathogens/Neospora/caninum/NEOS.contigs.version1/NEOS.contigs.version1.0.fasta</a>	07.07.2008
Toxoplasma gondii gt1	tgo	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/">ftp://ftp.ncbi.nih.gov/pub/TraceDB/</a>	07.07.2008
Monosiga brevicollis	mbr	<a href="http://genome.jgi-psf.org/Monbr1/Monbr1.download.ftp.html">http://genome.jgi-psf.org/Monbr1/Monbr1.download.ftp.html</a>	04.07.2008
Phaeodactylum tricornutum	hptr	<a href="http://genome.jgi-psf.org/Phatr2/Phatr2.download.ftp.html">http://genome.jgi-psf.org/Phatr2/Phatr2.download.ftp.html</a>	04.07.2008
thalassiosira pseudonana	htps	<a href="http://genome.jgi-psf.org/Thaps3/Thaps3.home.html">http://genome.jgi-psf.org/Thaps3/Thaps3.home.html</a>	25.06.2007
phytophthora ramorum	hpra	v1.1 (August 2004)	02.08.2007
phytophthora sojae	hpso	jgi	04.12.2005
phytophthora infestans	hpin	<a href="http://www.broad.mit.edu/annotation/genome/phytophthora_infestans/assets/phytophthora_infestans_1.fasta.gz">http://www.broad.mit.edu/annotation/genome/phytophthora_infestans/assets/phytophthora_infestans_1.fasta.gz</a>	07.07.2008
Hyaloperonospora parasitica	hhpa	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/hyaloperonospora_parasitica.fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/hyaloperonospora_parasitica.fasta*</a>	07.07.2008
Ectocarpus siliculosus	hesi	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/ectocarpus_siliculosus">ftp://ftp.ncbi.nih.gov/pub/TraceDB/ectocarpus_siliculosus</a>	07.07.2008
Malus x domestica	pmdo	<a href="http://genomics.msu.edu/fruitdb/analyses/apple_v4_clustered.fsa">genomics.msu.edu/fruitdb/analyses/apple_v4_clustered.fsa</a>	23.05.2007
BAC Lotus japonicus	plja	<a href="http://www.plantgdb.org/download/Download/xGDB/LjGDB/LjBAC160.bz2">http://www.plantgdb.org/download/Download/xGDB/LjGDB/LjBAC160.bz2</a>	07.07.2008
Glycine max	pgma	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/glycine_max.fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/glycine_max.fasta*</a>	07.07.2008
Phaseolus vulgaris	ppvu	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/phaseolus_vulgaris.fasta.phaseolus_vulgaris.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/phaseolus_vulgaris.fasta.phaseolus_vulgaris.001.gz</a>	28.10.2008
Medicago truncatula	pmtr	<a href="ftp://ftpmips.gsf.de/plants/medicago/MT_2_0/Mt2.0_pseudomolecule.tar.gz">ftp://ftpmips.gsf.de/plants/medicago/MT_2_0/Mt2.0_pseudomolecule.tar.gz</a>	07.07.2008
Populus trichocarpa	pptr	<a href="http://genome.jgi-psf.org/Poptr1_1/Poptr1_1.download.ftp.html">genome.jgi-psf.org/Poptr1_1/Poptr1_1.download.ftp.html</a>	23.05.2007
Ricinus communis	prco	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/ricinus_communis">ftp://ftp.ncbi.nih.gov/pub/TraceDB/ricinus_communis</a>	07.07.2008
Arabidopsis thaliana	path	<a href="http://www.plantgdb.org/XGDB/download.php?GDB=At">http://www.plantgdb.org/XGDB/download.php?GDB=At</a>	06.07.2008
Brassica oleracea	pbol	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/brassica_oleracea.fasta.brassica_oleracea.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/brassica_oleracea.fasta.brassica_oleracea.001.gz</a>	07.07.2008
BAC Brassica rapa	pbra	<a href="http://www.plantgdb.org/download/Download/xGDB/BrGDB/BrGDBbac154.bz2">http://www.plantgdb.org/download/Download/xGDB/BrGDB/BrGDBbac154.bz2</a>	06.07.2008
BAC Gossypium hirsutum	pghi	<a href="http://www.plantgdb.org/download/Download/xGDB/GhGDB/GHGDB.sql.bz2">http://www.plantgdb.org/download/Download/xGDB/GhGDB/GHGDB.sql.bz2</a>	07.07.2008
Solanum lycopersicum	psly	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/solanum_lycopersicum.fasta.solanum_lycopersicum.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/solanum_lycopersicum.fasta.solanum_lycopersicum.001.gz</a>	07.07.2008
Solanum tuberosum	pslu	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/solanum_tuberosum.fasta.solanum_tuberosum.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/solanum_tuberosum.fasta.solanum_tuberosum.001.gz</a>	28.10.2008
Nicotiana benthamiana	pnbe	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/nicotiana_benthamiana.fasta.nicotiana_benthamiana.001.gz">ftp://ftp.ncbi.nih.gov/pub/TraceDB/nicotiana_benthamiana.fasta.nicotiana_benthamiana.001.gz</a>	28.10.2008
Vitis vinifera	pvvi	<a href="http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/assembly/goldenpath/unmasked/">http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/assembly/goldenpath/unmasked/</a>	14.11.2008
Oryza sativa	posa	<a href="http://rapdownload.lab.nig.ac.jp/">http://rapdownload.lab.nig.ac.jp/</a>	05.07.2008
Hordeum vulgare (BAC)	phvu	<a href="http://www.plantgdb.org/download/Download/xGDB/HvGDB/HvGDBbac157.bz2">http://www.plantgdb.org/download/Download/xGDB/HvGDB/HvGDBbac157.bz2</a>	07.07.2008

Triticum aestivum (BAC)	ptae	<a href="http://www.plantgdb.org/download/Download/xGDB/TaGDB/TaGDBbac154.bz2">http://www.plantgdb.org/download/Download/xGDB/TaGDB/TaGDBbac154.bz2</a>	07.07.2008
Sorghum bicolor	psbi	<a href="http://www.plantgdb.org/download/Download/xGDB/SbGDB/SBgenome.bz2">http://www.plantgdb.org/download/Download/xGDB/SbGDB/SBgenome.bz2</a>	07.07.2008
Zea mays	pzma	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/zea_mays/fasta*">ftp://ftp.ncbi.nih.gov/pub/TraceDB/zea_mays/fasta*</a>	07.07.2008
Pinus taeda	ppta	<a href="ftp://ftp.ncbi.nih.gov/pub/TraceDB/pinus_taeda">ftp://ftp.ncbi.nih.gov/pub/TraceDB/pinus_taeda</a>	07.07.2008
Selaginella moellendorffii	psmo	<a href="ftp://ftp.jgi-psf.org/pub/JGI_data/Selaginella_moellendorffii/v1.0/Selmo1_assembly_scaffolds.fasta.gz">ftp://ftp.jgi-psf.org/pub/JGI_data/Selaginella_moellendorffii/v1.0/Selmo1_assembly_scaffolds.fasta.gz</a>	07.07.2008
Physcomitrella patens	pppa	<a href="http://genome.jgi-psf.org/Phypa1_1/Phypa1_1.download.ftp.html">genome.jgi-psf.org/Phypa1_1/Phypa1_1.download.ftp.html</a>	23.05.2007
Chlamydomonas reinhardtii	acre	<a href="http://genome.jgi-psf.org/Chlre3/Chlre3.download.ftp.html">http://genome.jgi-psf.org/Chlre3/Chlre3.download.ftp.html</a>	23.05.2007
Volvox carteri	avca	<a href="http://genome.jgi-psf.org/Volca1/">http://genome.jgi-psf.org/Volca1/</a>	02.07.2007
Micromonas pusilla	ampu	<a href="http://genome.jgi-psf.org/MicpuC2/MicpuC2.download.ftp.html">http://genome.jgi-psf.org/MicpuC2/MicpuC2.download.ftp.html</a>	04.07.2008
Ostreococcus lucimarinus	aolu	<a href="http://genome.jgi-psf.org/Ost9901_3/Ost9901_3.download.ftp.html">genome.jgi-psf.org/Ost9901_3/Ost9901_3.download.ftp.html</a>	25.06.2007
Ostreococcus tauri	aota	<a href="ftp://ftp.jgi-psf.org/pub/JGI_data/Ostreococcus_tauri/Otauri.fasta.gz">ftp://ftp.jgi-psf.org/pub/JGI_data/Ostreococcus_tauri/Otauri.fasta.gz</a>	07.07.2008
Cyanidioschyzon merolae	acme	<a href="http://merolae.biol.s.u-tokyo.ac.jp/download/complete_chromosomes.txt">http://merolae.biol.s.u-tokyo.ac.jp/download/complete_chromosomes.txt</a>	04.07.2008
Aureococcus anophagefferens	aaan	<a href="http://genome.jgi-psf.org/Auran1/Auran1.download.ftp.html">http://genome.jgi-psf.org/Auran1/Auran1.download.ftp.html</a>	04.07.2008

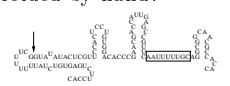
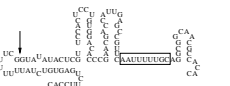
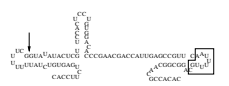

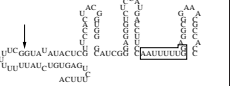

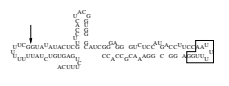
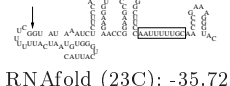
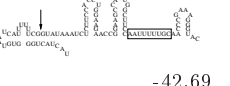
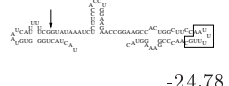
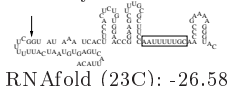

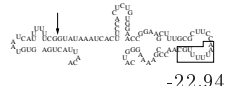
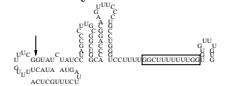
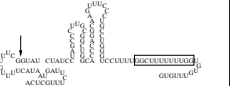
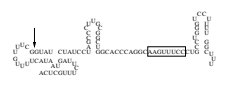
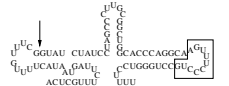
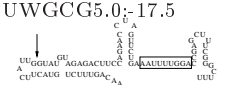

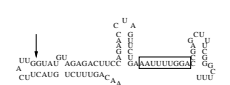
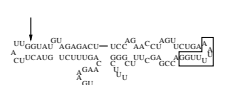
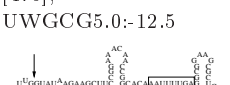


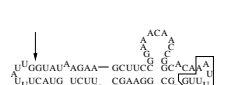
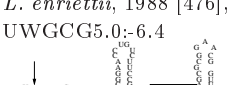
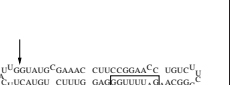
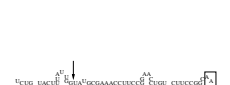



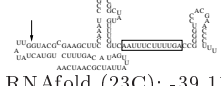
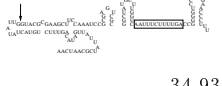
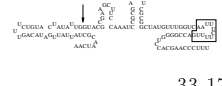
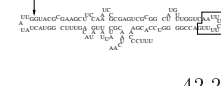
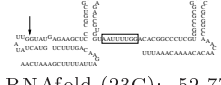


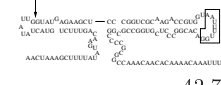
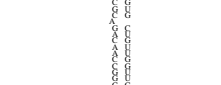
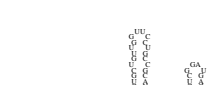






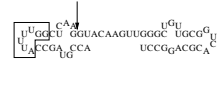
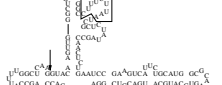

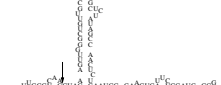
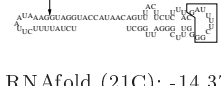
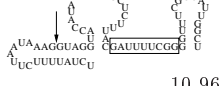

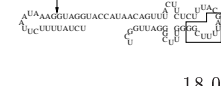
# Appendix C

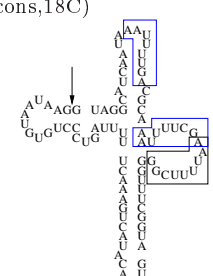

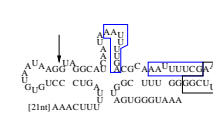
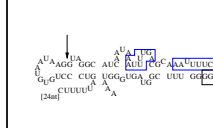
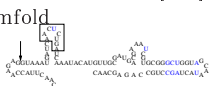






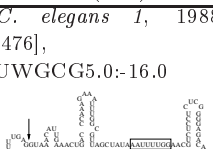

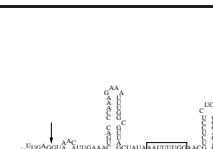
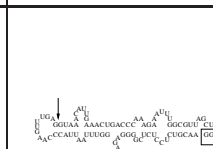
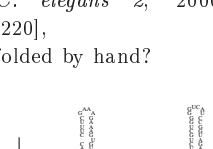



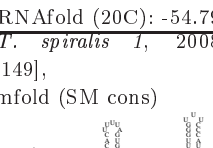
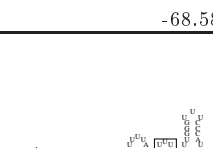
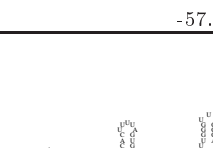
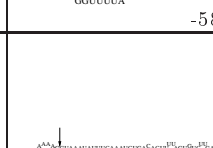
## Secondary Structures of SL-RNAs

Table C.1: Sequences, secondary structures, and folding energies  $\Delta G(\text{kcal/mol})$  of known SL RNAs. Donor splice site (arrow) and Sm-binding site (box) are marked. **Left column:** Structures proposed in the literature. **Right column:** Alternative structural models proposed in this work. Abbreviations: UWGCG – University of Wusconsin Genetics Computer Group; \* – Recalculated; *T* – natural ambient temperatur of organism; Blue nucleotides (*Euglena*, Rotifera) indicate mutations within known SL RNA alignments; Blue Box (*Hydra*) – alternative SM-binding sites; Green Arrow (*K. brevis*) indicates erroneus splice site from the literature; Sequences, constraints and drawings are available at [www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-009](http://www.bioinf.uni-leipzig.de/Publications/SUPPLEMENTS/09-009)

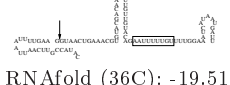
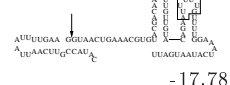
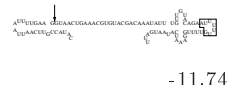
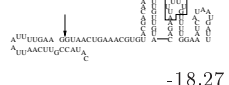
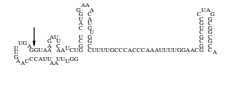
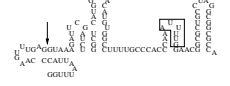

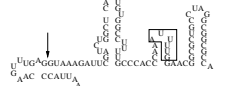


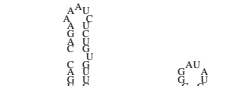

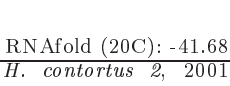
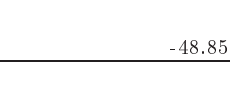
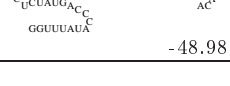
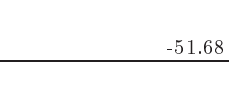
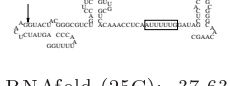

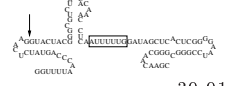
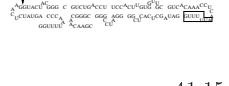
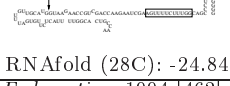
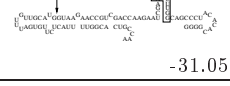
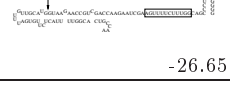
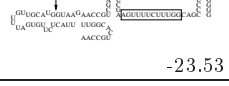
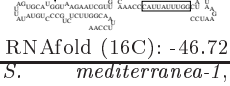
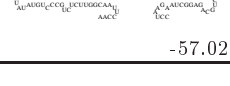

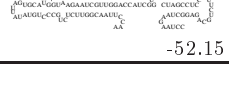
Published SL RNA	T	Alternative possible structures at organisms temperature		
<p><i>E. gracilis</i>, 1991 [212], folded by hand</p> <p>RNAfold (29C): -26.52</p>	29C [472]	<p>Additional Constraint: locarna-output</p> <p>-35.02</p>	<p>-32.30</p>	<p>-28.04</p>
<p><i>E. gracilis</i>, 1999 [178], folded by hand?</p> <p>RNAfold (29C): -23.49</p>	29C [472]	<p>Additional Constraint: locarna-output</p> <p>-32.59</p>	<p>-33.69</p>	<p>-28.54</p>

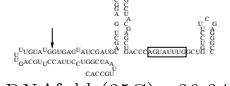
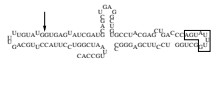
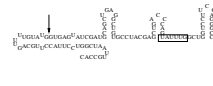
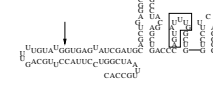
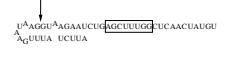
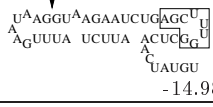
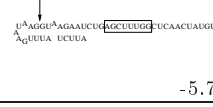
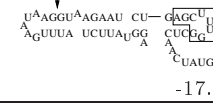

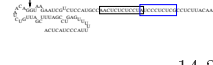
Published SL RNA	T	Alternative possible structures at organisms temperature		
<p><i>P. curvicauda</i>, 2000 [221], folded by hand?</p>  <p>RNAfold (22C): -34.62</p>	22C [473]	 <p>-41.51</p>	 <p>-32.85</p>	
<p><i>C. acus</i>, 2000 [221], folded by hand?</p>  <p>RNAfold (22C): -48.37</p>	22C	 <p>-55.20</p>	 <p>-54.10</p>	 <p>-42.92</p>
<p><i>R. costata</i>, 2000 [221], folded by hand?</p>  <p>RNAfold (23C): -35.72</p>	23C [474]	 <p>-42.69</p>	 <p>-24.78</p>	
<p><i>M. pellucidum</i>, 2000 [221], folded by hand?</p>  <p>RNAfold (23C): -26.58</p>	23C	 <p>-35.73</p>	 <p>-22.94</p>	
<p><i>E. sulcatum</i>, 1999 [178], folded by hand?</p>  <p>RNAfold (25C): -33.00</p>	25C [475]	 <p>-33.55</p>	<p>16 nt shorter different SM-binding site</p>  <p>-34.10</p>	 <p>-33.55</p>
<p><i>L. collosoma</i>, 1988 [476], UWGCG5.0:-17.5</p>  <p>RNAfold (28C): -24.00</p>	28C [477]	<p>suboptimal stem I</p>  <p>-27.43</p>	 <p>-24.00</p>	 <p>-20.04</p>
<p><i>C. fasciculata</i>, 1988 [476], UWGCG5.0:-12.5</p>  <p>RNAfold (20C): -30.83</p>	20C [478]	<p>suboptimal stem I</p>  <p>-30.10</p>	 <p>-30.83</p>	 <p>-31.40</p>
<p><i>L. enriettii</i>, 1988 [476], UWGCG5.0:-6.4</p>  <p>RNAfold (36C): -16.26</p>	36C [479]	 <p>-20.40</p>	 <p>-24.72</p>	 <p>-20.87</p>

Published SL RNA	T	Alternative possible structures at organisms temperature		
<p><i>T. cruzi</i>, 1988 [476], UWGGC5.0:-24.1</p>  <p>RNAfold (23C): -39.11</p>	23C [480]	 <p>-34.93</p>	 <p>-33.17</p>	 <p>-42.24</p>
<p><i>T. vivax</i>, 1988 [476], UWGGC5.0:-26.5</p>  <p>RNAfold (23C): -52.77</p>	23C [480]	 <p>-46.90</p>	 <p>-53.64</p>	 <p>-42.79</p>
<p><i>T. brucei</i>, 1988 [476], UWGGC5.0:-42.3</p>  <p>RNAfold (23C): -67.58</p>	23C [480]	 <p>-60.45</p>	 <p>-53.11</p>	<p>alternative SM site</p>  <p>-57.98</p>
<p><i>K. brevis</i>, 2007 [148], mfold (SM cons)</p>  <p>RNAfold (20C): -46.29</p>	20C [481]	24 nt 3' removed		
		 <p>-58.15</p>	 <p>-39.64</p>	 <p>-46.43</p>
<p><i>K. micrum</i>, 2007 [218], mfold3.1.2 (SM cons,20C)</p>  <p>RNAfold (20C): -27.68</p>	20C	79 nt 3' added		
		 <p>-66.47</p>	 <p>-50.85</p>	 <p>-63.97</p>
<p><i>Hydra-A</i>, 2001 [214], mfold2.3 (SM,Donor cons,18C)</p>  <p>RNAfold (21C): -14.37</p>	21C [482]	 <p>-10.96</p>	 <p>-9.26</p>	 <p>-18.06</p>

Published SL RNA	T	Alternative possible structures at organisms temperature		
<p><i>Hydra-B</i>, 2001 [214], mfold2.3 (SM,Donor cons,18C)</p>  <p>ACGGAAAAAACCGGUAAA RNAfold (21C): -32.62</p>	21C [482]	 <p>-18.03</p>	 <p>-22.17</p>	 <p>-20.88</p>
<p><i>A. ricciae</i>, 2005 [217], mfold</p>  <p>RNAfold (24C): -45.31</p>	24C [483]	 <p>-36.14</p>	0	 <p>-46.28</p>
<p><i>Philodina sp.</i>, 2005 [217], mfold</p>  <p>RNAfold (20C): -51.41</p>	20C [484]	 <p>-42.28</p>	 <p>-42.21</p>	 <p>-49.82</p>
<p><i>C. elegans 1</i>, 1988 [476], UWGCG5.0:-16.0</p>  <p>RNAfold (20C): -38.14</p>	20C [485]	 <p>-44.73</p>	 <p>-43.54</p>	 <p>-33.80</p>
<p><i>C. elegans 2</i>, 2000 [220], folded by hand?</p>  <p>RNAfold (20C): -54.79</p>	20C [485]	 <p>-68.58</p>	 <p>-57.14</p>	 <p>-58.38</p>
<p><i>T. spiralis 1</i>, 2008 [149], mfold (SM cons)</p>  <p>RNAfold (36C): -29.49</p>	36C [486]	 <p>-30.15</p>	 <p>-27.80</p>	 <p>-19.06</p>



Published SL RNA	T	Alternative possible structures at organisms temperature			
<p><i>T. spiralis</i> 2, 2008 [149], mfold (SM cons)</p>  <p>RNAfold (36C): -19.51</p>	36C [486]	 <p>-17.78</p>	 <p>-11.74</p>	 <p>-18.27</p>	
<p><i>P. pacificus</i> 1, 2003 [461]</p>  <p>RNAfold (20C): -45.98</p>	20C [487]	 <p>-57.87</p>	 <p>-50.76</p>	 <p>-54.23</p>	
<p><i>P. pacificus</i> 2, 2003 [461]</p>  <p>RNAfold (20C): -41.68</p>	20C [487]	 <p>-48.85</p>	 <p>-48.98</p>	 <p>-51.68</p>	
<p><i>H. contortus</i> 2, 2001 [460]</p>  <p>RNAfold (25C): -37.63</p>	25C [488]	 <p>-47.38</p>	 <p>-39.91</p>	 <p>-41.15</p>	
<p><i>S. mansoni</i>, 1990 [124], Zuker:-16.6</p>  <p>RNAfold (28C): -24.84</p>	28C [489]	 <p>-31.05</p>	 <p>-26.65</p>	 <p>-23.53</p>	
<p><i>F. hepatica</i>, 1994 [462], Zuker</p>  <p>RNAfold (16C): -46.72</p>	16C [490]	 <p>-57.02</p>	 <p>-47.25</p>	 <p>-52.15</p>	
<p><i>S. mediterranea</i>-1, 2005 [464], mfold (SM cons)</p>  <p>RNAfold (22C): -43.84</p>	22C [491]	 <p>-44.16</p>	 <p>-32.07</p>	 <p>-41.19</p>	

Published SL RNA	T	Alternative possible structures at organisms temperature		
<p><i>E. multilocularis</i>, 2000 [463], mfold no SM given</p>  <p>RNAfold (35C): -36.34</p>	35C [492]	 <p>-42.20</p>	 <p>-33.35</p>	 <p>-37.37</p>
<p><i>C. intestinalis</i>, 2001 [215], mfold3.0</p>  <p>RNAfold (21C): -5.71</p>	21C [493]	 <p>-14.98</p>	 <p>-5.71</p>	<p>4nt 5' added</p>  <p>-17.06</p>
<p><i>O. dioica</i>, 2004 [216], folded by hand?</p>  <p>RNAfold (20C): -14.19</p>	20C [494]	 <p>-14.24</p>		

# Bibliography

- [1] W. F. Marzluff (2005) *Metazoan replication-dependent histone mRNAs: a distinct set of RNA polymerase II transcripts*. *Curr. Opin. Cell. Biol.*, **17**, 274–280.
- [2] D. H. Huson und D. Bryant (2006) *Application of Phylogenetic Networks in Evolutionary Studies*. *Mol. Biol. Evol.*, **23**, 254–267.
- [3] C. R. Woese (1987) *Bacterial evolution*. *Microbiol Rev*, **51** (2), 221–271.
- [4] N. R. Pace (1997) *A molecular view of microbial diversity and the biosphere*. *Science*, **276** (5313), 734–740.
- [5] M. Di Giulio (1992) *On the origin of the transfer RNA molecule*. *J Theor Biol*, **159** (2), 199–214.
- [6] M. Di Giulio (1995) *Was it an ancient gene codifying for a hairpin RNA that, by means of direct duplication, gave rise to the primitive tRNA molecule?*. *J Theor Biol*, **177** (1), 95–101.
- [7] J. Miyazaki, S. Nakaya, T. Suzuki, M. Tamakoshi, T. Oshima und A. Yamagishi (2001) *Ancestral residues stabilizing 3-isopropylmalate dehydrogenase of an extreme thermophile: experimental evidence supporting the thermophilic common ancestor hypothesis*. *J Biochem*, **129** (5), 777–782.
- [8] S. Prohaska, P. Stadler und D. Krakauer (2009) *Origins and Innovations in Gene Regulation – The Case of Chromatin Regulation*. in preparation.
- [9] F. Rodríguez-Trelles, R. Rosa Tarrío und F. J. Ayala (2006) *Origins and Evolution of Spliceosomal Introns*. *Annu. Rev. Genet.*, **40**, 47–76.
- [10] Y. I. Wolf und E. V. Koonin (2007) *On the origin of the translation system and the genetic code in the RNA world by means of natural selection, exaptation, and subfunctionalization*. *Biology Direct*, **2**, 14.
- [11] B. Boussau, S. Blanquart, A. Neculea, N. Lartillot und M. Gouy (2008) *Parallel adaptations to high temperatures in the Archaean eon*. *Nature*, **456**, 942–946.
- [12] P. Forterre (2002) *The origin of DNA genomes and DNA replication proteins*. *Curr. Op. Microbiol.*, **5**, 525–532.
- [13] P. Forterre (2006) *Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: A hypothesis for the origin of cellular domain*. *Proc. Natl. Acad. Sci. USA*, **103**, 3669–3674.
- [14] The ENCODE Project Consortium (2007) *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project*. *Nature*, **447**, 799–816.
- [15] N. Maeda, T. Kasukawa, R. Oyama, J. Gough, M. Frith, P. G. Engström, B. Lenhard, R. N. Aturaliya, S. Batalov, K. W. Beisel, C. J. Bult, C. F. Fletcher, A. R. Forrest, M. Furuno, D. Hill, M. Itoh, M. Kanamori-Katayama, S. Katayama, M. Katoh, T. Kawashima, J. Quackenbush, T. Ravasi, B. Z. Ring, K. Shibata, K. Sugiura, Y. Take-naka, R. D. Teasdale, C. A. Wells, Y. Zhu, C. Kai, J. Kawai, D. A. Hume, P. Carninci und Y. Hayashizaki (2006) *Transcript Annotation in FANTOM3: Mouse Gene Catalog Based on Physical cDNAs*. *PLoS Genetics*, **2**, e62. Doi:10.1371/journal.pgen.0020062.
- [16] T. Ravasi, H. Suzuki, K. C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M. C. Frith, M. M. Gongora, S. M. Grimmond, D. A. Hume, Y. Hayashizaki und J. S. Mattick (2006) *Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome*. *Genome Res.*, **16**, 11–19.
- [17] P. Kapranov, J. Cheng, S. Dike, D. Nix, R. Duttap Gupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hacker Müller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, G. Madhavan, A. Piccolboni, V. Sementchenko, H. Tammana und T. R. Gingeras (2007) *RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription*. *Science*, **316**, 1484–1488.
- [18] J. R. Manak, S. Dike, V. Sementchenko, P. Kapranov, F. Biemar, J. Long, J. Cheng, I. Bell, S. Ghosh, A. Piccolboni und T. R. Gingeras (2006) *Biological function of unannotated transcription during the early development of Drosophila melanogaster*. *Nat Genet*, **38**, 1151–1158.

- [19] H. He, J. Wang, T. Liu, X. S. Liu, T. Li, Y. Wang, Z. Qian, H. Zheng, X. Zhu, T. Wu, B. Shi, W. Deng, W. Zhou, G. Skogerbø and R. Chen (2007) *Mapping the C. elegans Noncoding Transcriptome With a Whole-Genome Tiling Microarray*. *Genome Res.*, **17**, 1471–1477.
- [20] D. Li, D. K. Willkomm, A. Schön and R. K. Hartmann (2007) *RNase P of the Cyanophora paradoxa cyanelle: a plastid ribozyme*. *Biochimie*, **89**, 1528–1538.
- [21] M. Havilio, E. Y. Levanon, G. Lerman, M. Kupiec and E. Eisenberg (2005) *Evidence for abundant transcription of non-coding regions in the Saccharomyces cerevisiae genome*. *BMC Genomics*, **6**, 93.
- [22] L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis and L. M. Steinmetz (2006) *A high-resolution map of transcription in the yeast genome*. *Proc. Natl. Acad. Sci. USA*, **103**, 5320–5325.
- [23] F. Miura, N. Kawaguchi, J. Sese, A. Toyoda, M. Hattori, S. Morishita and T. Ito (2006) *A large-scale full-length cDNA analysis to explore the budding yeast transcriptome*. *Proc. Natl. Acad. Sci. USA*, **103**, 17846–17851.
- [24] B. T. Wilhelm, S. Marguerat, S. Watt, F. Schubert, V. Wood, I. Goodhead, C. J. Penkett, J. Rogers and J. Bähler (2008) *Dynamic repertoire of a eukaryotic transcriptome surveyed at single-nucleotide resolution*. *Nature*. Epub.
- [25] S. Gottesman (2004) *The small RNA regulators of Escherichia coli: roles and mechanisms*. *Annu. Rev. Microbiol.*, **58**, 303–328.
- [26] A. F. Bompfünnewerer, C. Flamm, C. Fried, G. Fritsch, I. L. Hofacker, J. Lehmann, K. Missal, A. Mosig, B. Müller, S. J. Prohaska, B. M. R. Stadler, P. F. Stadler, A. Tanzer, S. Washietl and C. Wittwer (2005) *Evolutionary Patterns of Non-Coding RNAs*. *Th. Biosci.*, **123**, 301–369.
- [27] W. Gilbert (1986) *The RNA World*. *Nature*, **319**, 618.
- [28] R. F. Gesteland and J. F. Atkins, Herausgeber (1993) *The RNA World*. Cold Spring Harbor Laboratory Press, Plainview, NY.
- [29] S. J. Freeland, R. D. Knight and L. F. Landweber (1999) *Do Proteins Predate DNA?*. *Science*, **286**, 690–692.
- [30] A. Serganov and D. J. Patel (2007) *Ribozymes, riboswitches and beyond: regulation of gene expression without proteins*. *Nat Rev Genet.*, **8**, 776–790.
- [31] S. A. Strobel and J. C. Cochrane (2007) *RNA catalysis: ribozymes, ribosomes, and riboswitches*. *Curr Opin Chem Biol*, **11**, 636–643.
- [32] P. B. Moore and T. A. Steitz (2002) *The involvement of RNA in ribosome function*. *Nature*, **418**, 229–235.
- [33] K. EV (2005) *Orthologs, paralogs, and evolutionary genomics*. *Annu Rev Genet*, **39**, 309–338.
- [34] P. Bertone, V. Stolc, T. E. Royce, J. S. Rozowsky, A. E. Urban, X. Zhu, J. L. Rinn, W. Tongprasit, M. Samanta, S. Weissman, M. Gerstein and M. Snyder (Dec 2004) *Global identification of human transcribed sequences with genome tiling arrays*. *Science*, **306** (5705), 2242–2246.
- [35] D. Kampa, J. Cheng, P. Kapranov, M. Yamanaka, S. Brubaker, S. Cawley, J. Drenkow, A. Piccolboni, S. Bekiranov, G. Helt, H. Tammana and T. R. Gingeras (Mar 2004) *Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22*. *Genome Res*, **14** (3), 331–342.
- [36] A. Matsui, J. Ishida, T. Morosawa, Y. Mochizuki, E. Kaminuma, T. A. Endo, M. Okamoto, E. Nambara, M. Nakajima, M. Kawashima, M. Satou, J. M. Kim, N. Kobayashi, T. Toyoda, K. Shinozaki and M. Seki (Aug 2008) *Arabidopsis transcriptome analysis under drought, cold, high-salinity and ABA treatment conditions using a tiling array*. *Plant Cell Physiol*, **49** (8), 1135–1149.
- [37] T. Akama, K. Suzuki, K. Tanigawa, A. Kawashima, H. Wu, N. Nakata, Y. Osana, Y. Sakakibara and N. Ishii (May 2009) *Whole-genome tiling array analysis of Mycobacterium leprae RNA reveals high expression of pseudogenes and noncoding regions*. *J Bacteriol*, **191** (10), 3321–3327.
- [38] Y. Okazaki, M. Furuno, T. Kasukawa, J. Adachi, H. Bono, S. Kondo, I. Nikaido, N. Osato, R. Saito, H. Suzuki, I. Yamanaka, H. Kiyosawa, K. Yagi, Y. Tomaru, Y. Hasegawa, A. Nogami, C. Schönbach, T. Gojobori, R. Baldarelli, D. P. Hill, C. Bult, D. A. Hume, J. Quackenbush, L. M. Schriml, A. Kanapin, H. Matsuda, S. Batalov, K. W. Beisel, J. A. Blake, D. Bradt, V. Brusica, C. Chothia, L. E. Corbani, S. Cousins, E. Dalla, T. A. Dragani, C. F. Fletcher, A. Forrest, K. S. Frazer, T. Gaasterland, M. Gariboldi, C. Gissi, A. Godzik, J. Gough, S. Grimmond, S. Gustincich, N. Hirokawa, I. J. Jackson, E. D. Jarvis, A. Kanai, H. Kawaji, Y. Kawasaki, R. M. Kedzierski, B. L. King, A. Konagaya, I. V. Kurochkin, Y. Lee, B. Lenhard, P. A. Lyons, D. R. Maglott, L. Maltais, L. Marchionni, L. McKenzie, H. Miki, T. Nagashima, K. Numata, T. Okido, W. J. Pavan, G. Pertea, G. Pesole, N. Petrovsky, R. Pillai, J. U. Pontius, D. Qi, S. Ramachandran, T. Ravasi, J. C. Reed, D. J. Reed, B. Z. Ring, M. Ringwald, A. Sandelin, C. Schneider, C. A. Semple, M. Setou, K. Shimada, R. Sultana, Y. Takenaka, M. S. Taylor, R. D. Teasdale, M. Tomita, R. Verardo, L. Wagner, C. Wahlestedt, Y. Wang, Y. Watanabe, C. Wells, L. G. Wilming, A. Wynshaw-Boris, M. Yanagisawa, I. Yang, L. Yang, Z. Yuan, M. Zavolan, Y. Zhu, A. Zimmer, P. Carninci, N. Hayatsu, T. Hirozane-Kishikawa, H. Konno, M. Nakamura, N. Sakazume, K. Sato, T. Shiraki, K. Waki, J. Kawai, K. Aizawa, T. Arakawa, S. Fukuda, A. Hara, W. Hashizume, K. Imotani, Y. Ishii, M. Itoh, I. Kagawa, A. Miyazaki, K. Sakai, D. Sasaki, K. Shibata, A. Shinagawa, A. Yasunishi, M. Yoshino, R. Waterston, E. S. Lander, J. Rogers, E. Birney, Y. Hayashizaki, FANTOM Consortium und RIKEN Genome Exploration

Research Group Phase I & II Team (Dec 2002) *Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs*. *Nature*, **420** (6915), 563–573.

- [39] T. Imanishi, T. Itoh, Y. Suzuki, C. O'Donovan, S. Fukuchi, K. O. Koyanagi, R. A. Barrero, T. Tamura, Y. Yamaguchi-Kabata, M. Tanino, K. Yura, S. Miyazaki, K. Ikeo, K. Homma, A. Kasprzyk, T. Nishikawa, M. Hirakawa, J. Thierry-Mieg, D. Thierry-Mieg, J. Ashurst, L. Jia, M. Nakao, M. A. Thomas, N. Mulder, Y. Karavidopoulou, L. Jin, S. Kim, T. Yasuda, B. Lenhard, E. Eveno, Y. Suzuki, C. Yamasaki, J. Takeda, C. Gough, P. Hilton, Y. Fujii, H. Sakai, S. Tanaka, C. Amid, M. Bellgard, M. d. e. F. Bonaldo, H. Bono, S. K. Bromberg, A. J. Brookes, E. Bruford, P. Carninci, C. Chelala, C. Couillault, S. J. de Souza, M. A. Debily, M. D. Devignes, I. Dubchak, T. Endo, A. Estreicher, E. Eyraas, K. Fukami-Kobayashi, G. R. Gopinath, E. Graudens, Y. Hahn, M. Han, Z. G. Han, K. Hanada, H. Hanaoka, E. Harada, K. Hashimoto, U. Hinz, M. Hirai, T. Hishiki, I. Hopkinson, S. Imbeaud, H. Inoko, A. Kanapin, Y. Kaneko, T. Kasukawa, J. Kelso, P. Kersey, R. Kikuno, K. Kimura, B. Korn, V. Kuryshv, I. Makalowska, T. Makino, S. Mano, R. Mariage-Samson, J. Mashima, H. Matsuda, H. W. Mewes, S. Minoshima, K. Nagai, H. Nagasaki, N. Nagata, R. Nigam, O. Ogasawara, O. Ohara, M. Ohtsubo, N. Okada, T. Okido, S. Oota, M. Ota, T. Ota, T. Otsuki, D. Piatier-Tonneau, A. Poustka, S. X. Ren, N. Saitou, K. Sakai, S. Sakamoto, R. Sakate, I. Schupp, F. Servant, S. Shery, R. Shiba, N. Shimizu, M. Shimoyama, A. J. Simpson, B. Soares, C. Steward, M. Suwa, M. Suzuki, A. Takahashi, G. Tamiya, H. Tanaka, T. Taylor, J. D. Terwilliger, P. Unneberg, V. Veeramachaneni, S. Watanabe, L. Wilming, N. Yasuda, H. S. Yoo, M. Stodolsky, W. Makalowski, M. Go, K. Nakai, T. Takagi, M. Kanehisa, Y. Sakaki, J. Quackenbush, Y. Okazaki, Y. Hayashizaki, W. Hide, R. Chakraborty, K. Nishikawa, H. Sugawara, Y. Tateno, Z. Chen, M. Oishi, P. Tonellato, R. Apweiler, K. Okubo, L. Wagner, S. Wiemann, R. L. Strausberg, T. Isogai, C. Auffray, N. Nomura, T. Gojobori and S. Sugano (Jun 2004) *Integrative annotation of 21,037 human genes validated by full-length cDNA clones*. *PLoS Biol*, **2** (6).
- [40] T. Ravasi, H. Suzuki, K. C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M. C. Frith, M. M. Gongora, S. M. Grimmond, D. A. Hume, Y. Hayashizaki und J. S. Mattick (2006) *Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome*. *Genome Res*, **16** (1), 11–19.
- [41] S. Cawley, S. Bekiranov, H. H. Ng, P. Kapranov, E. A. Sekinger, D. Kampa, A. Piccolboni, V. Sementchenko, J. Cheng, A. J. Williams, R. Wheeler, B. Wong, J. Drenkow, M. Yamanaka, S. Patel, S. Brubaker, H. Tammana, G. Helt, K. Struhl und T. R. Gingeras (Feb 2004) *Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs*. *Cell*, **116** (4), 499–509.
- [42] A. Hüttenhofer, P. Schattner und N. Polacek (May 2005) *Non-coding RNAs: hope or hype?*. *Trends Genet*, **21** (5), 289–297.
- [43] N. Berteaux, S. Lottin, E. Adriaenssens, F. Van Coppenolle, F. Van Coppenolle, X. Leroy, J. Coll, T. Dugimont und J. J. Curgy (Oct 2004) *Hormonal regulation of H19 gene expression in prostate epithelial cells*. *J Endocrinol*, **183** (1), 69–78.
- [44] P. Carninci, T. Kasukawa, S. Katayama, J. Gough, M. C. Frith, N. Maeda, R. Oyama, T. Ravasi, B. Lenhard, C. Wells, R. Kodzius, K. Shimokawa, V. B. Bajic, S. E. Brenner, S. Batalov, A. R. Forrest, M. Zavolan, M. J. Davis, L. G. Wilming, V. Aidinis, J. E. Allen, A. Ambesi-Impombato, R. Apweiler, R. N. Aturaliya, T. L. Bailey, M. Bansal, L. Baxter, K. W. Beisel, T. Bersano, H. Bono, A. M. Chalk, K. P. Chiu, V. Choudhary, A. Christofels, D. R. Clutterbuck, M. L. Crowe, E. Dalla, B. P. Dalrymple, B. de Bono, G. Della Gatta, D. di Bernardo, T. Down, P. Engstrom, M. Fagiolini, G. Faulkner, C. F. Fletcher, T. Fukushima, M. Furuno, S. Futaki, M. Gariboldi, P. Georgii-Hemming, T. R. Gingeras, T. Gojobori, R. E. Green, S. Gustincich, M. Harbers, Y. Hayashi, T. K. Hensch, N. Hirokawa, D. Hill, L. Huminiecki, M. Iacono, K. Ikeo, A. Iwama, T. Ishikawa, M. Jakt, A. Kanapin, M. Katoh, Y. Kawasawa, J. Kelso, H. Kitamura, H. Kitano, G. Kollias, S. P. Krishnan, A. Kruger, S. K. Kummerfeld, I. V. Kurochkin, L. F. Lareau, D. Lazarevic, L. Lipovich, J. Liu, S. Liuni, S. McWilliam, M. Madan Babu, M. Madera, L. Marchionni, H. Matsuda, S. Matsuzawa, H. Miki, F. Mignone, S. Miyake, K. Morris, S. Mottagui-Tabar, N. Mulder, N. Nakano, H. Nakauchi, P. Ng, R. Nilsson, S. Nishiguchi, S. Nishikawa, F. Nori, O. Ohara, Y. Okazaki, V. Orlando, K. C. Pang, W. J. Pavan, G. Pavesi, G. Pesole, N. Petrovsky, S. Piazza, J. Reed, J. F. Reid, B. Z. Ring, M. Ringwald, B. Rost, Y. Ruan, S. L. Salzberg, A. Sandelin, C. Schneider, C. Schönbach, K. Sekiguchi, C. A. Sempile, S. Seno, L. Sessa, Y. Sheng, Y. Shibata, H. Shimada, K. Shimada, D. Silva, B. Sinclair, S. Sperling, E. Stupka, K. Sugiura, R. Sultana, Y. Takenaka, K. Taki, K. Tammoja, S. L. Tan, S. Tang, M. S. Taylor, J. Tegner, S. A. Teichmann, H. R. Ueda, E. van Nimwegen, R. Verardo, C. L. Wei, K. Yagi, H. Yamanishi, E. Zabarovsky, S. Zhu, A. Zimmer, W. Hide, C. Bult, S. M. Grimmond, R. D. Teasdale, E. T. Liu, V. Brusic, J. Quackenbush, C. Wahlestedt, J. S. Mattick, D. A. Hume, C. Kai, D. Sasaki, Y. Tomaru, S. Fukuda, M. Kanamori-Katayama, M. Suzuki, J. Aoki, T. Arakawa, J. Iida, K. Imamura, M. Itoh, T. Kato, H. Kawaji, N. Kawagashira, T. Kawashima, M. Kojima, S. Kondo, H. Konno, K. Nakano, N. Ninomiya, T. Nishio, M. Okada, C. Plessy, K. Shibata, T. Shiraki, S. Suzuki, M. Tagami, K. Waki, A. Watahiki, Y. Okamura-Oho, H. Suzuki, J. Kawai, Y. Hayashizaki, FANTOM Consortium und RIKEN Genome Exploration Research Group and Genome Science Group (Genome Network Project Core Group) (Sep 2005) *The transcriptional landscape of the mammalian genome*. *Science*, **309** (5740), 1559–1563.
- [45] H. Kawaji, J. Severin, M. Lizio, A. Waterhouse, S. Katayama, K. M. Irvine, D. A. Hume, A. R. Forrest, H. Suzuki, P. Carninci, Y. Hayashizaki und C. O. Daub (Apr 2009) *The FANTOM web resource: from mammalian transcriptional landscape to its dynamic regulation*. *Genome Biol*, **10** (4).
- [46] P. J. French, T. V. Bliss und V. O'Connor (2001) *Ntab, a novel non-coding RNA abundantly expressed in rat brain*. *Neuroscience*, **108** (2), 207–215.
- [47] M. Sawata, H. Takeuchi und T. Kubo (Jul 2004) *Identification and analysis of the minimal promoter activity of a novel noncoding nuclear RNA gene, AncR-1, from the honeybee (Apis mellifera L.)*. *RNA*, **10** (7), 1047–1058.

- [48] A. A. Lukowiak, S. Granneman, S. A. Mattox, W. A. Speckmann, K. Jones, H. Pluk, W. J. Venrooij, R. M. Terns und M. P. Terns (Sep 2000) *Interaction of the U3-55k protein with U3 snoRNA is mediated by the box B/C motif of U3 and the WD repeats of U3-55k*. *Nucleic Acids Res*, **28** (18), 3462–3471.
- [49] A. Mougin, A. Grégoire, J. Banroques, V. Ségault, R. Fournier, F. Brulé, M. Chevrier-Miller und C. Branlant (Nov 1996) *Secondary structure of the yeast *Saccharomyces cerevisiae* pre-U3A snoRNA and its implication for splicing efficiency*. *RNA*, **2** (11), 1079–1093.
- [50] F. Brulé, J. Venema, V. Ségault, D. Tollervey und C. Branlant (Feb 1996) *The yeast *Hansenula wingei* U3 snoRNA gene contains an intron and its coding sequence co-evolved with the 5' ETS region of the pre-ribosomal RNA*. *RNA*, **2** (2), 183–197.
- [51] M. D. López, M. A. Rosenblad und T. Samuelsson (Jul 2009) *Conserved and variable domains of RNase MRP RNA*. *RNA Biol*, **6** (3).
- [52] S. Washietl, I. L. Hofacker und P. F. Stadler (2005) *Fast and reliable prediction of noncoding RNAs*. *Proc. Natl. Acad. Sci. USA*, **102**, 2454–2459.
- [53] S. Washietl und I. L. Hofacker (Sep 2007) *Identifying structural noncoding RNAs using RNAz*. *Curr Protoc Bioinformatics*, **Chapter 12**.
- [54] A. Donath, S. Findeiß, J. Hertel, M. Marz, W. Otto, C. Schulz, P. F. Stadler und S. Wirth (2008) *Non-Coding RNAs*. G. Caetano-Anolles, Herausgeber, *Evolutionary Genomics*. Wiley. In press.
- [55] S. F. Altschul, W. Gish, W. Miller, E. W. Myers und D. J. Lipman (Oct 1990) *Basic local alignment search tool*. *J Mol Biol*, **215** (3), 403–410.
- [56] R. Durbin, S. Eddy, A. Krogh und G. Mitchison (2006) *Biological sequence analysis – Probabilistic models of proteins and nucleic acids*. Cambridge University Press.
- [57] S. Griffiths-Jones (2005) *RALEE—RNA ALignment editor in Emacs*. *Bioinformatics*, **21**, 257–259.
- [58] T. Macke, D. Ecker, R. Gutell, D. Gautheret, D. Case und R. Sampath (Nov 2001) *RNA Motif, an RNA secondary structure definition and search algorithm*. *Nucleic Acids Research*, **29** (22), 4724–4735.
- [59] D. Gautheret und A. Lambert (Nov 2001) *Direct RNA motif definition and identification from multiple sequence alignments using secondary structure profiles*. *J Mol Biol*, **313** (5), 1003–1011.
- [60] I. L. Hofacker, M. Fekete und P. F. Stadler (2002) *Secondary Structure Prediction for Aligned RNA Sequences*. *J. Mol. Biol.*, **319**, 1059–1066.
- [61] J. Hertel, D. de Jong, M. Marz, D. Rose, H. Tafer, A. Tanzer, B. Schierwater und P. F. Stadler (2009) *Non-coding RNA annotation of the genome of *Trichoplax adhaerens**. *Nucleic Acids Res*, **37** (5), 1602–1615.
- [62] W. R. Pearson (1991) *Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms*. *Genomics*, **11**, 635–650.
- [63] U. Roshan, S. Chikkagoudar und D. R. Livesay (2008) *Searching for evolutionary distant RNA homologs within genomic sequences using partition function posterior probabilities*. *BMC Bioinformatics*, **9**, 61.
- [64] O. Gotoh (1982) *An improved algorithm for matching biological sequences*. *J. Mol. Biol.*, **162** (3), 705–708.
- [65] E. P. Nawrocki, D. L. Kolbe und S. R. Eddy (2009) *Infernal 1.0: inference of RNA alignments*. *Bioinformatics*, **25** (10), 1335–1337.
- [66] D. Gautheret, F. Major und R. Cedergren (Oct 1990) *Pattern searching/alignment with RNA primary and secondary structures: an effective descriptor for tRNA*. *Comput Appl Biosci*, **6** (4), 325–331.
- [67] A. Laferrière, D. Gautheret und R. Cedergren (Apr 1994) *An RNA pattern matching program with enhanced performance and portability*. *Comput Appl Biosci*, **10** (2), 211–212.
- [68] S. Eddy (1992-1996) *RNAJOB: a program to search for RNA secondary structure motifs in sequence databases*.
- [69] M. J. Serra und D. H. Turner (1995) *Predicting thermodynamic properties of RNA*. *Methods Enzymol*, **259**, 242–261.
- [70] D. H. Mathews, J. Sabina, M. Zuker und D. H. Turner (May 1999) *Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure*. *J Mol Biol*, **288** (5), 911–940.
- [71] S. Gräf, D. Strothmann und G. Steger (2000) *The Syntax and Semantics of a Language for Describing Complex Patterns in Biological Sequences*. Report Technische Fakultät, Universität Bielefeld.
- [72] S. Gräf, D. Strothmann, S. Kurtz und G. Steger (Jan 2001) *HyPaLib: a database of RNAs and RNA structural elements defined by hybrid patterns*. *Nucleic Acids Res*, **29** (1), 196–198.
- [73] A. Mosig, K. Sameith und P. Stadler (Feb 2006) *Fragrep: an efficient search tool for fragmented patterns in genomic sequences*. *Genomics Proteomics Bioinformatics*, **4** (1), 56–60.

- [74] G. E. Crooks, G. Hon, J. M. Chandonia und S. E. Brenner (2004) *WebLogo: A sequence logo generator*. Genome Research, **14**, 1188–1190.
- [75] T. M. Lowe und S. Eddy (1997) *tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence*. Nucl. Acids Res., **25**, 955–964.
- [76] S. M. C. Robb, E. Ross und A. S. Alvarado (2008) *SmedGD: the Schmidtea mediterranea genome database*. Nucleic Acids Res, **36**, D599–D606.
- [77] M. Regalia, M. A. Rosenblad und T. Samuelsson (Aug 2002) *Prediction of signal recognition particle RNA genes*. Nucleic Acids Res, **30** (15), 3368–3377.
- [78] D. Yusuf, M. Marz, P. F. Stadler und I. L. Hofacker (2009) *Bcheck: a wrapper tool for RNase P RNA gene prediction*. , (in preparation), .
- [79] S. Griffiths-Jones, H. K. Saini, S. van Dongen und A. J. Enright (2008) *miRBase: tools for microRNA genomics*. Nucleic Acids Res., **36**, D154–D158.
- [80] J. D. Thompson, D. G. Higgins und T. J. Gibson (1994) *CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice*. Nucl. Acids Res., **22**, 4673–4680.
- [81] I. L. Hofacker, W. Fontana, P. F. Stadler, L. S. Bonhoeffer, M. Tacker und P. Schuster (1994) *Fast Folding and Comparison of RNA Secondary Structures*. Monatsh. Chem., **125**, 167–188.
- [82] J. Hertel, I. L. Hofacker und P. F. Stadler (2008) *snoReport: Computational identification of snoRNAs with unknown targets*. Bioinformatics, **24**, 158–164.
- [83] L. Lestrade und M. J. Weber (2006) *snoRNA-LBME-db, a comprehensive database of human H/ACA and C/D box snoRNAs*. Nucl. Acids Res., **34**, D158–D162.
- [84] T. M. Lowe und S. R. Eddy (1999) *A Computational Screen for Methylation Guide snoRNAs in Yeast*. Science, **283**, 1168–1171.
- [85] H. Tafer, S. Kehr, J. Hertel und P. Stadler (2009) *RNA-snoop: Efficient target prediction for box H/ACA snoRNAs*. Submitted.
- [86] S. Griffiths-Jones, S. Moxon, M. Marshall, A. Khanna, S. R. Eddy und A. Bateman (Jan 2005) *Rfam: annotating non-coding RNAs in complete genomes*. Nucleic Acids Res, **33** (Database issue), 121–124.
- [87] C. Liu, B. Bai, G. Skogerbø, L. Cai, W. Deng, Y. Zhang, D. B. Bu, Y. Zhao und R. Chen (2005) *NONCODE: an integrated knowledge database of non-coding RNAs*. Nucleic Acids Res., **33**, D112–D115.
- [88] M. A. Larkin, G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson und D. G. Higgins (Nov 2007) *Clustal W and Clustal X version 2.0*. Bioinformatics, **23** (21), 2947–2948.
- [89] B. Morgenstern (1999) *DIALIGN2: improvement of the segment-to-segment approach to multiple sequence alignment*. Bioinformatics, **15**, 211–218.
- [90] W. Otto, S. Will und R. Backofen (2008) *Structure Local Multiple Alignment of RNA. Proceedings of German Conference on Bioinformatics (GCB'2008)*, Band P-136 von *Lecture Notes in Informatics (LNI)*, 178–188. Gesellschaft für Informatik (GI).
- [91] T. L. Bailey, N. Williams, C. Misleh und W. W. Li (2006) *MEME: discovering and analyzing DNA and protein sequence motifs*. Nucleic Acids Res., **34**, W369–W373.
- [92] T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li und W. S. Noble (May 2009) *MEME SUITE: tools for motif discovery and searching*. Nucleic Acids Res.
- [93] N. Hernandez (2001) *Small Nuclear RNA Genes: a Model System to Study Fundamental Mechanisms of Transcription*. J. Biol. Chem., **276**, 26733–26736.
- [94] A. M. Domitrovich und G. R. Kunkel (2003) *Multiple, dispersed human U6 small nuclear RNA genes with varied transcriptional efficiencies*. Nucleic Acids Res., **31**, 2344–2352.
- [95] W. Y. Tarn, T. A. Yario und J. A. Steitz (1995) *U12 snRNAs in Vertebrates: Evolutionary conservation of 5' sequences implicated in splicing of pre-mRNAs containing a minor class of introns*. RNA, **1**, 644–656.
- [96] C. Bark, P. Weller, J. Zabielski und U. Pettersson (1986) *Genes for human U4 small nuclear RNA*. Gene, **50**, 333–344.
- [97] G. R. Kunkel und T. Pederson (1988) *Upstream elements required for efficient transcription of a human U6 RNA gene resemble those of U1 and U2 genes even though a different polymerase is used*. Genes Dev, **2**, 196–204.
- [98] G. M. Korf und W. E. Stumph (1986) *Chicken U2 and U1 RNA genes are found in very different genomic environments but have similar promoter structures*. Biochemistry, **25**, 2041–2047.

- [99] H. S. Bhatthal, Z. Zamrod, T. Tobaru und W. E. Stumph (1995) *Identification of Proximal Sequence Element Nucleotides Contributing to the Differential Expression of Variant U4 Small Nuclear RNA Genes*. J. Biol. Chem., **270**, 27629–27633.
- [100] S. M. Mount, V. Gotea, C.-F. Lin, K. Hernandez und W. Makalowski (2007) *Spliceosomal small nuclear RNA genes in 11 insect genomes*. RNA, **13**, 5–14.
- [101] J. M. Sierra-Montes, S. Pereira-Simon, S. S. Smail und R. J. Herrera (2005) *The silk moth Bombyx mori U1 and U2 snRNA variants are differentially expressed*. Gene, **352**, 127–136.
- [102] B. Stefanovic und W. F. Marzluff (1992) *Characterization of two developmentally regulated sea urchin U2 small nuclear RNA promoters: a common required TATA sequence and independent proximal and distal elements*. Mol Cell Biol, **12**, 650–660.
- [103] J. Thomas, K. Lea, E. Zucker-Aprison und T. Blumenthal (1990) *The spliceosomal snRNAs of Caenorhabditis elegans*. Nucleic Acids Res, **18**, 2633–2642.
- [104] P. Flicek, B. L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, T. Eyre, S. Fitzgerald, J. Fernandez-Banet, S. Gräf, S. Haider, M. Hammond, R. Holland, K. L. Howe, K. Howe, N. Johnson, A. Jenkinson, A. Kähäri, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. J. Vilella, J. Vogel, S. White, M. Wood, E. Birney, T. Cox, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, T. J. P. Hubbard, A. Kasprzyk, G. Proctor, J. Smith, A. Ureta-Vidal und S. Searle (2008) *Ensembl 2008*. Nucleic Acids Res., **36**, D707–D714.
- [105] T. Hubbard, D. Andrews, M. Caccamo, G. Cameron, Y. Chen, M. Clamp, L. Clarke, G. Coates, T. Cox, F. Cunningham, V. Curwen, T. Cutts, T. Down, R. Durbin, X. M. Fernandez-Suarez, J. Gilbert, M. Hammond, J. Herrero, H. Hotz, K. Howe, V. Iyer, K. Jekosch, A. Kahari, A. Kasprzyk, D. Keefe, S. Keenan, F. Kokocinski, D. London, I. Longden, G. McVicker, C. Melsopp, P. Meidl, S. Potter, G. Proctor, M. Rae, D. Rios, M. Schuster, S. Searle, J. Severin, G. Slater, D. Smedley, J. Smith, W. Spooner, A. Stabenau, J. Stalker, R. Storey, S. Trevanion, A. Ureta-Vidal, J. Vogel, S. White, C. Woodwark und E. Birney (2005) *Ensembl 2005*. Nucleic Acids Res., **33**, D447–D453.
- [106] T. Kirsten und E. Rahm (2006) *BioVoice: Mapping-based data intergation in bioinformatics*. U. Leser, F. Naumann und B. Eckman, Herausgeber, *Proceedings of the 3rd International Workshop on Data Integration in the Life Sciences (DILS)*, Band 4075, 124–135. Springer Verlag, Berlin, Heidelberg.
- [107] I. L. Hofacker (Jul 2003) *Vienna RNA secondary structure server*. Nucleic Acids Res, **31** (13), 3429–3431.
- [108] F. Huang, J. Qin, C. Reidys und P. Stadler (2009) *Partition Function and Base Pairing Probabilities for RNA-RNA Interaction Prediction*. , (in preparation), .
- [109] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller und D. Haussler (2006) *Identification and classification of conserved RNA secondary structures in the human genome*. PLoS Comput Biol, **2**, e33.
- [110] M. Blanchette, W. J. Kent, C. Riemer, L. Elnitski, A. F. A. Smit, K. M. Roskin, R. Baertsch, K. Rosenbloom, H. Clawson, E. D. Green, D. Haussler und W. Miller (2004) *Aligning multiple genomic sequences with the threaded blockset aligner*. Genome Res., **14**, 708–715.
- [111] D. Rose, J. Hertel, K. Reiche, P. F. Stadler und J. Hackermüller (2008) *ncDNAign: Plausible Multiple Alignments of Non-Protein-Coding Genomic Sequences*. Genomics, **92**, 65–74.
- [112] S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer und P. F. Stadler (2005) *Mapping of conserved RNA Secondary Structures predicts Thousands of functional Non-Coding RNAs in the Human Genome*. Nature Biotech., **23**, 1383–1390.
- [113] S. Washietl, J. S. Pedersen, J. O. Korbel, A. Gruber, J. Hackermüller, J. Hertel, M. Lindemeyer, K. Reiche, C. Stoesits, A. Tanzer, C. Ucla, C. Wyss, S. E. Antonarakis, F. Denoeud, J. Lagarde, J. Drenkow, P. Kapranov, T. R. Gingeras, R. Guigó, M. Snyder, M. B. Gerstein, A. Reymond, I. L. Hofacker und P. F. Stadler (2007) *Structured RNAs in the ENCODE Selected Regions of the Human Genome*. Gen. Res., **17**, 852–864.
- [114] D. Rose, J. Jörns, J. Hackermüller, K. Reiche, Q. Li und P. F. Stadler (2008) *Duplicated RNA Genes in Teleost Fish Genomes*. J. Bioinf. Comp. Biol. In press.
- [115] K. Missal, D. Rose und P. F. Stadler (2005) *Non-coding RNAs in Ciona intestinalis*. Bioinformatics, **21** S2, i77–i78.
- [116] K. Missal, X. Zhu, D. Rose, W. Deng, G. Skogerbø, R. Chen und P. F. Stadler (2006) *Prediction of Structured Non-Coding RNAs in the Genome of the Nematode Caenorhabditis elegans*. J. Exp. Zool.: Mol. Dev. Evol., **306B**, 379–392.
- [117] D. Rose, J. Hackermüller, S. Washietl, S. Findeiß, K. Reiche, J. Hertel, P. F. Stadler und S. J. Prohaska (2007) *Computational RNomics of Drosophilids*. BMC Genomics, **8**, 406.
- [118] S. Steigele, W. Huber, C. Fried, P. F. Stadler und K. Nieselt (2007) *Comparative Analysis of Structured RNAs in S. cerevisiae Indicates a Multitude of Different Functions*. BMC Biology, **5v**, 25.



- [119] T. Mourier, C. Carret, S. Kyes, Z. Christodoulou, P. P. Gardner, D. C. Jeffares, R. Pinches, B. Barrell, M. Berriman, S. Griffiths-Jones, A. Ivens, C. Newbold and A. Pain (2008) *Genome-wide discovery and verification of novel structured RNAs in Plasmodium falciparum*. *Genome Res.*, **18**, 281–292.
- [120] H. J. Bandelt und A. W. M. Dress (1992) *A Canonical Decomposition Theory for Metrics on a finite Set*. *Adv. Math.*, **92**, 47.
- [121] D. Bryant und V. Moulton (2004) *Neighbor-Net: An Agglomerative Method for the Construction of Phylogenetic Networks*. *Mol. Biol. Evol.*, **21**, 255–265.
- [122] A. Smit, R. Hubley und P. Green (1996–2004) *RepeatMasker Open-3.0*. , , .
- [123] M. Marz, T. Kirsten und P. F. Stadler (Nov 2008) *Evolution of Spliceosomal snRNA Genes in Metazoan Animals*. *J Mol Evol.*
- [124] A. Rajkovic, R. E. Davis, J. N. Simonsen und F. M. Rottman (Nov 1990) *A spliced leader is present on a subset of mRNAs from the human parasite Schistosoma mansoni*. *Proc Natl Acad Sci U S A*, **87** (22), 8879–8883.
- [125] J. D. Thompson, T. J. Gibson, F. Plewniak, F. Jeanmougin und D. G. Higgins (1997) *The CLUSTAL\_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools*. *Nucleic Acids Res.*, **25**, 4876–4882.
- [126] S. H. Bernhart, H. Tafer, U. Mückstein, C. Flamm, P. F. Stadler und I. L. Hofacker (2006) *Partition Function and Base Pairing Probabilities of RNA Heterodimers*. *Algorithms Mol. Biol.*, **1**, 3 [epub].
- [127] N. Saitou und M. Nei (1987) *The neighbor-joining method: a new method for reconstructing phylogenetic trees*. *Mol Biol. Evol.*, **4**, 406–425.
- [128] T. W. Nilsen (2003) *The spliceosome: the most complex macromolecular machine in the cell?*. *Bioessays*, **25**, 1147–1149.
- [129] S. Valadkhan, A. Mohammadi, C. Wachtel und J. L. Manley (2007) *Protein-free spliceosomal snRNAs catalyze a reaction that resembles the first step of splicing*. *RNA*, **13**, 2300–2311.
- [130] J. Lykke-Andersen, C. Aagaard, M. Semionov and R. A. Garrett (1997) *Archaeal introns: splicing, intercellular mobility and evolution*. *Trends Biochem Sci.*, **22**, 326–331.
- [131] D. M. Haugen P, Simon und B. D. (2005) *The natural history of group I introns*. *Trends Genet.*, **21**, 111–119.
- [132] O. Fedorova und N. Zingler (2007) *Group II introns: structure, folding and splicing mechanism*. *Biol. Chem.*, **388**, 665–678.
- [133] K. Calvin und H. Li (2008) *RNA-splicing endonuclease structure and function*. *Cell Mol Life Sci.*, **65**, 1176–1185.
- [134] X. Chen, T. S. Rozhddestvensky, L. J. Collins, J. Schmitz und D. Penny (2007) *Combined experimental and computational approach to identify non-protein-coding RNAs in the deep-branching eukaryote Giardia intestinalis*. *Nucleic Acids Res.* Doi:10.1093/nar/gkm474.
- [135] L. Chen, D. J. Lullo, E. Ma, S. E. Celniker, D. C. Rio und J. A. Doudna (2005) *Identification and analysis of U5 snRNA variants in Drosophila*. *RNA*, **11**, 1473–1477.
- [136] A. A. Patel und J. A. Steitz (2003) *Splicing double: insights from the second spliceosome*. *Nat. Rev. Mol. Cell Biol.*, **4**, 960–970.
- [137] N. Sheth, X. Roca, M. L. Hastings, T. Roeder, A. R. Krainer und R. Sachidanandam (2006) *Comprehensive splice-site analysis using comparative genomics*. *Nucleic Acids Res.*, **34**, 3955–3967.
- [138] H. König, N. Matter, R. Bader, W. Thiele und F. Müller (2007) *Splicing segregation: the minor spliceosome acts outside the nucleus and controls cell proliferation*. *Cell*, **131**, 718–729.
- [139] C. L. Will und R. Lührmann (2005) *Splicing of a rare class of introns by the U12-dependent spliceosome*. *Biol. Chem.*, **386**, 713–724.
- [140] K. Y. Kwek, S. Murphy, A. Furger, B. Thomas, W. O’Gorman, H. Kimura, N. J. Proudfoot und A. Akoulitchev (2002) *U1 snRNA associates with TFIIF and regulates transcriptional initiation*. *Nat. Struct. Biol.*, **9**, 800–805.
- [141] L. Collins und D. Penny (2005) *Complex Spliceosomal Organization Ancestral to Extant Eukaryotes*. *Mol. Biol. Evol.*, **22**, 1053–1066.
- [142] Z. J. Lorković, R. Lehner, C. Forstner und A. Barta (2005) *Evolutionary conservation of minor U12-type spliceosome between plants and humans*. *RNA*, **11**, 1095–1107.
- [143] A. G. Russell, J. M. Charette, D. F. Spencer und M. W. Gray (2006) *An early evolutionary origin for the minor spliceosome*. *Nature*, **443**, 863–866.

- [144] M. Dávila López, M. Alm Rosenblad und T. Samuelsson (2008) *Computational screen for spliceosomal RNA genes aids in defining the phylogenetic distribution of major and minor spliceosomal components*. Nucleic Acids Res.
- [145] Z. Palfi, B. Schimanski, A. Günzl, S. Lücke und A. Bindereif (2005) *U1 small nuclear RNP from Trypanosoma brucei: a minimal U1 snRNA with unusual protein components*. Nucleic Acids Res, **33**, 2493–2503.
- [146] N. N. Pouchkina-Stantcheva und A. Tunncliffe (2005) *Spliced leader RNA-mediated trans-splicing in phylum Rotifera*. Mol Biol Evol, **22**, 1482–1489.
- [147] K. E. Hastings (2005) *SL trans-splicing: easy come or easy go?*. Trends Genet., **21**, 240–247.
- [148] K. B. Lidie und F. M. van Dolah (2007) *Spliced leader RNA-mediated trans-splicing in a dinoflagellate, Karenia brevis*. J Eukaryot Microbiol, **54**, 427–435.
- [149] J. Pettitt, B. MÅller, I. Stansfield und B. Connolly (2008) *Spliced leader trans-splicing in the nematode Trichinella spiralis uses highly polymorphic, noncanonical spliced leaders..* RNA, **14**, 760–770.
- [150] T. Blumenthal (1995) *Trans-splicing and polycistronic transcription in Caenorhabditis elegans*. Trends Genet., **11**, 132–136.
- [151] S. Valadkhan (2005) *snRNAs as the catalysts of pre-mRNA splicing*. Curr. Op. Chem. Biol., **9**, 603–608.
- [152] S. Valadkhan (2007) *The spliceosome: caught in a web of shifting interactions*. Curr. Op. Struct. Biol., **17**, 310–315.
- [153] G. C. Shukla und R. A. Padgett (1999) *Conservation of functional features of U6atac and U12 snRNAs between vertebrates and higher plants*. RNA, **5**, 525–538.
- [154] C. Schneider, C. L. Will, J. Brosius, M. Frilander und R. Lührmann (2004) *Identification of an evolutionarily divergent U11 small nuclear ribonucleoprotein particle in Drosophila*. Proc. Natl. Acad. Sci. USA, **101** (26), 9584–9589.
- [155] L. J. Collins, T. J. Macke und D. Penny (2004) *Searching for ncRNAs in eukaryotic genomes: maximizing biological input with RNA motifs*. J. Integ. Bioinf., **1**, 2004–08–04.
- [156] P. C. Lo und S. M. Mount (1990) *Drosophila melanogaster genes for U1 snRNA variants and their expression during development*. Nucleic Acids Res, **18**, 6971–6979.
- [157] B. Stefanovic, J. M. Li, S. Sakallah und W. F. Marzluff (1991) *Isolation and characterization of developmentally regulated sea urchin U2 snRNA genes*. Dev Biol., **148**, 284–294.
- [158] E. J. Sontheimer und J. A. Steitz (1992) *Three novel functional variants of human U5 small nuclear RNA*. Mol. Cell. Biol., **12**, 734–746.
- [159] J. Morales, M. Borrero, J. Sumerel und S. C. (1997) *Identification of developmentally regulated sea urchin U5 snRNA genes*. DNA Seq., **7**, 243–259.
- [160] S. Pereira-Simon, J. M. Sierra-Montes, K. Ayes, L. Martinez, A. Socorro und R. J. Herrera (2004) *Variants of U1 small nuclear RNA assemble into spliceosomal complexes*. Insect Molecular Biology, **13**, 189–194.
- [161] C. Kyriakopoulou, P. Larsson, L. Liu, J. Schuster, F. Söderbom, L. A. Kirsebom und A. Virtanen (2006) *U1-like snRNAs lacking complementarity to canonical 5' splice sites*. RNA, **12**, 1603–1611.
- [162] A. Hinas, P. Larsson, L. Avesson, L. A. Kirsebom, A. Virtanen und F. Söderbom (2006) *Identification of the Major Spliceosomal RNAs in Dictyostelium discoideum Reveals Developmentally Regulated U2 Variants and Polyadenylated snRNAs*. Eukaryotic Cell, **5**, 924–934.
- [163] S. S. Smail, K. Ayes, J. M. Sierra-Montes und R. J. Herrera (2006) *U6 snRNA variants isolated from the posterior silk gland of the silk moth Bombyx mori*. Insect Biochem Mol Biol., **36**, 454–465.
- [164] D. Liao und A. M. Weiner (1995) *Concerted evolution of the tandemly repeated genes encoding primate U2 small nuclear RNA (the RNU2 locus) does not prevent rapid diversification of the (CT)<sub>n</sub>.(GA)<sub>n</sub> microsatellite embedded within the U2 repeat unit*. Genomics, **30**, 583–593.
- [165] D. Liao (1999) *Concerted evolution: molecular mechanism and biological implications*. Am J Hum Genet, **64**, 24–30.
- [166] M. Nei und A. P. Rooney (2005) *Concerted and birth-and-death evolution of multigene families*. Annu. Rev. Genet., **39**, 121–152.
- [167] R. A. Denison, S. W. Van Arsdell, L. B. Bernstein und A. M. Weiner (1981) *Abundant Pseudogenes for Small Nuclear RNAs are Dispersed in the Human Genome*. Proc. Natl. Acad. Sci. USA, **78**, 810–814.
- [168] M. J. Telford und P. W. H. Holland (1997) *Evolution of 28S Ribosomal DNA in Chaetognaths: Duplicate Genes and Molecular Phylogeny*. J. Mol. Evol., **44**, 135–144.
- [169] D. Papillon, Y. Perez, X. Caubit und Y. Le Parco (2006) *Systematics of Chaetognatha under the light of molecular data, using duplicated ribosomal 18S DNA sequences*. Mol Phylogenet Evol., **38**, 621–634.

- [170] G. Giribet, G. D. Edgecombe und W. C. Wheeler (2001) *Arthropod phylogeny based on eight molecular loci and morphology*. Nature, **413**, 157–161.
- [171] E. Myslinksi, A. Krol und P. Carbon (2004) *Characterization of snRNA and snRNA-type genes in the pufferfish *Fugu rubripes**. Gene, **330**, 149–158.
- [172] J. W. Tichelaar, E. D. Wieben, R. Reddy, A. Vrabel und P. Camacho (1998) *In vivo expression of a variant human U6 RNA from a unique, internal promoter*. Biochemistry, **37**, 12943–12951.
- [173] The Chimpanzee Sequencing and Analysis Consortium (2005) *Initial sequence of the chimpanzee genome and comparison with the human genome*. Nature, **437**, 69–87.
- [174] International Chicken Genome Sequencing Consortium (2004) *Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution*. Nature, **432**, 695–716.
- [175] G. C. Shukla und R. A. Padgett (2004) *U4 small nuclear RNA can function in both the major and minor spliceosomes*. Proc. Natl. Acad. Sci. USA, **101**, 93–98.
- [176] D. Liao, T. Pavelitz, J. R. Kidd, K. K. Kidd und A. M. Weiner (1997) *Concerted evolution of the tandemly repeated genes encoding human U2 snRNA (the RNU2 locus) involves rapid intrachromosomal homogenization and rare interchromosomal gene conversion*. EMBO J., **16**, 588–598.
- [177] T. Pavelitz, D. Liao und A. M. Weiner (1999) *Concerted evolution of the tandem array encoding primate U2 snRNA (the RNU2 locus) is accompanied by dramatic remodeling of the junctions with flanking chromosomal sequences*. EMBO J., **18**, 3783–3792.
- [178] C. Ebel, C. Frantz, F. Paulus und P. Imbault (1999) *Trans-splicing and cis-splicing in the colourless *Euglenoid*, *Entosiphon sulcatum**. Curr Genet, **35**, 542–550.
- [179] F. Pelliccia, R. Barzotti, E. Bucciarelli und A. Rocchi (2001) *5S ribosomal and U1 small nuclear RNA genes: a new linkage type in the genome of a crustacean that has three different tandemly repeated units containing 5S ribosomal DNA sequences*. Genome, **44**, 331–335.
- [180] I. Cross und L. Rebordinos (2005) *5S rDNA and U2 snRNA are linked in the genome of *Crassostrea angulata* and *Crassostrea gigas* oysters: does the (CT)<sub>n</sub>·(GA)<sub>n</sub> microsatellite stabilize this novel linkage of large tandem arrays?*. Genome, **48**, 1116–1119.
- [181] M. Machado, E. Zuasti, I. Cross, A. Merlo, C. Infante und L. Rebordinos (2006) *Molecular characterization and chromosomal mapping of the 5S rRNA gene in *Solea senegalensis*: a new linkage to the U1, U2, and U5 small nuclear RNA genes*. Genome, **49**, 79–86.
- [182] D. M. Hillis und M. T. Dixon (1991) *Ribosomal DNA: molecular evolution and phylogenetic inference*. Q. Rev. Biol., **66**, 411–453.
- [183] C. Schlötterer und D. Tautz (1994) *Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution*. Curr. Biol., **4**, 777–783.
- [184] I. L. Gonzalez und J. E. Sylvester (2001) *Human rDNA: Evolutionary Patterns within the Genes and Tandem Arrays Derived from Multiple Chromosomes*. Genomics, **73**, 255–263.
- [185] J. E. Dahlberg und E. Lund (1988) *The genes and transcription of the major small nuclear RNAs*. M. L. Birnstiel, Herausgeber, *Structure and Function of Major and Minor Small Nuclear Ribonucleoprotein Particles*, 38–70. Springer-Verlag, Berlin.
- [186] Drosophila 12 Genomes Consortium (2007) *Evolution of genes and genomes on the *Drosophila* phylogeny*. Nature, **450**, 203–218.
- [187] S. M. Mount und J. A. Steitz (1981) *Sequence of U1 RNA from *Drosophila melanogaster*: implications for U1 secondary structure and possible involvement in splicing*. Nucleic Acids Res, **9**, 6351–6368.
- [188] T. P. Hausner, L. M. Giglio und A. M. Weiner (1990) *Evidence for base-pairing between mammalian U2 and U6 small nuclear ribonucleoprotein particles*. Genes Dev, **4**, 2146–2156.
- [189] A. Krol, C. Branlant, E. Lazar, H. Gallinaro und M. Jacob (1981) *Primary and secondary structures of chicken, rat and man nuclear U4 RNAs. Homologies with U1 and U5 RNAs*. Nucleic Acids Res, **9**, 2699–2716.
- [190] C. Branlant, A. Krol, E. Lazar, B. Haendler, M. Jacob, L. Galego-Dias und C. Pousada (1983) *High evolutionary conservation of the secondary structure and of certain nucleotide sequences of U5 RNA*. Nucleic Acids Res, **11**, 8359–8367.
- [191] K. A. Montzka und J. A. Steitz (1988) *Additional low-abundance human small nuclear ribonucleoproteins: U11, U12, etc.* Proc Natl Acad Sci U S A, **85**, 8885–8889.
- [192] G. C. Shukla, A. J. Cole, R. C. Dietrich und R. A. Padgett (2002) *Domains of human U4atac snRNA required for U12-dependent splicing in vivo*. Nucleic Acids Res, **30**, 4650–4657.
- [193] D. J. Forbes, M. W. Kirschner, D. Caput, J. E. Dahlberg und E. Lund (1984) *Differential expression of multiple U1 small nuclear RNAs in oocytes and embryos of *Xenopus laevis**. Cell, **38**, 681–689.

- [194] I. W. Mattaj und R. Zeller (1983) *Xenopus laevis U2 snRNA genes: tandemly repeated transcription units sharing 5' and 3' flanking homology with other RNA polymerase II transcribed genes*. EMBO J, **2**, 1883–1891.
- [195] E. Myslinski, C. Branlant, E. D. Wieben und T. Pederson (1984) *The small nuclear RNAs of Drosophila*. J. Mol. Biol., **180**, 927–945.
- [196] L. R. Otake, P. Scamborova, C. Hashimoto und J. A. Steitz (2002) *The divergent U12-type spliceosome is required for pre-mRNA splicing and is essential for development in Drosophila*. Mol Cell, **9**, 439–446.
- [197] J. M. Sierra-Montes, S. Pereira-Simon, A. V. Freund, L. M. Ruiz, M. N. Szmulewicz und R. J. Herrera (2003) *A diversity of U1 small nuclear RNAs in the silk moth Bombyx mori*. Insect Biochem Mol Biol, **33**, 29–39.
- [198] J. M. Sierra-Montes, A. V. Freund, L. M. Ruiz, M. N. Szmulewicz, D. J. Rowold und R. J. Herrera (2002) *Multiple forms of U2 snRNA coexist in the silk moth Bombyx mori*. Insect Mol Biol, **11**, 105–114.
- [199] R. Barzotti, F. Pelliccia und A. Rocchi (2003) *Identification and characterization of U1 small nuclear RNA genes from two crustacean isopod species*. Chromosome Res, **11**, 365–373.
- [200] J. D. Shambaugh, G. E. Hannon und T. W. Nilsen (1994) *The spliceosomal U small nuclear RNAs of Ascaris lumbricoides*. Mol Biochem Parasitol, **64**, 349–352.
- [201] Q. M. Mitrovich und C. Guthrie (2007) *Evolution of small nuclear RNAs in S. cerevisiae, C. albicans, and other hemiascomycetous yeasts*. RNA, **13**, 2066–2080.
- [202] N. H. Putnam, T. Butts, D. E. K. Ferrier, R. F. Furlong, U. K. Hellsten, Takeshi, M. Robinson-Rechavi, E. Shoguchi, A. Terry, J.-K. Yu, È. Benito-Gutiérrez, I. Dubchak, J. Garcia-Fernández, J. J. Gibson-Brown, I. V. Grigoriev, A. C. Horton, P. J. de Jong, J. Jurka, V. V. Kapitonov, Y. Kohara, Y. Kuroki, E. Lindquist, S. Lucas, K. Osoegawa, L. A. Pennacchio, A. A. Salamov, Y. Satou, T. Sauka-Spengler, J. Schmutz, T. Shin-I, A. Toyoda, M. Bronner-Fraser, A. Fujiyama, L. Z. Holland, P. W. H. Holland, N. Satoh und D. S. Rokhsar (2008) *The amphioxus genome and the evolution of the chordate karyotype*. Nature, **453**, 1064–1071.
- [203] M. J. Weber (2006) *Mammalian small nucleolar RNAs are mobile genetic elements*. PLoS Genet., **2**, e205.
- [204] J. Schmitz, A. Zemann, G. Churakov, H. Kuhl, F. Grützner, R. Reinhardt und J. Brosius (2008) *Retroposed SNOfall-A mammalian-wide comparison of platypus snoRNAs*. Genome Res., **18**, 1005–1010.
- [205] Z. J. Lu, D. H. Turner und D. H. Mathews (2006) *A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation*. Nucleic Acids Res., **34**, 4912–4924.
- [206] M. Zuker (Jul 2003) *Mfold web server for nucleic acid folding and hybridization prediction*. Nucleic Acids Res, **31** (13), 3406–3415.
- [207] S. Bonhoeffer, J. S. McCaskill, P. F. Stadler und P. Schuster (1993) *RNA multi-structure landscapes. A study based on temperature dependent partition functions*. Eur Biophys J, **22** (1), 13–24.
- [208] W. J. Murphy, K. P. Watkins und N. Agabian (Nov 1986) *Identification of a novel Y branch structure as an intermediate in trypanosome mRNA processing: evidence for trans splicing*. Cell, **47** (4), 517–525.
- [209] T. W. Nilsen (Dec 2001) *Evolutionary origin of SL-addition trans-splicing: still an enigma*. Trends Genet, **17** (12), 678–680.
- [210] M. Milhausen, R. G. Nelson, S. Sather, M. Selkirk und N. Agabian (Oct 1984) *Identification of a small RNA containing the trypanosome spliced leader: a donor of shared 5' sequences of trypanosomatid mRNAs?*. Cell, **38** (3), 721–729.
- [211] T. De Lange, T. M. Berkvens, H. J. Veerman, A. C. Frasch, J. D. Barry und P. Borst (Jun 1984) *Comparison of the genes coding for the common 5' terminal sequence of messenger RNAs in three trypanosome species*. Nucleic Acids Res, **12** (11), 4431–4443.
- [212] L. H. Tessier, M. Keller, R. L. Chan, R. Fournier, J. H. Weil und P. Imbault (Sep 1991) *Short leader sequences may be transferred from small RNAs to pre-mature mRNAs by trans-splicing in Euglena*. EMBO J, **10** (9), 2621–2625.
- [213] M. Krause und D. Hirsh (Jun 1987) *A trans-spliced leader sequence on actin mRNA in C. elegans*. Cell, **49** (6), 753–761.
- [214] N. A. Stover und R. E. Steele (May 2001) *Trans-spliced leader addition to mRNAs in a cnidarian*. Proc Natl Acad Sci U S A, **98** (10), 5693–5698.
- [215] A. E. Vandenberghe, T. H. Meedel und K. E. Hastings (Feb 2001) *mRNA 5'-leader trans-splicing in the chordates*. Genes Dev, **15** (3), 294–303.
- [216] P. Ganot, T. Kalløe, R. Reinhardt, D. Chourrout und E. M. Thompson (Sep 2004) *Spliced-leader RNA trans splicing in a chordate, Oikopleura dioica, with a compact genome*. Mol Cell Biol, **24** (17), 7795–7805.
- [217] N. N. Pouchkina-Stantcheva und A. Tunnacliffe (Jun 2005) *Spliced leader RNA-mediated trans-splicing in the lumen Rotifera*. Mol Biol Evol, **22** (6), 1482–1489.

- [218] H. Zhang, Y. Hou, L. Miranda, D. A. Campbell, N. R. Sturm, T. Gaasterland und S. Lin (Mar 2007) *Spliced leader RNA trans-splicing in dinoflagellates*. Proc Natl Acad Sci U S A, **104** (11), 4618–4623.
- [219] M. Blaxter und L. Liu (Oct 1996) *Nematode spliced leaders—ubiquity, evolution and utility*. Int J Parasitol, **26** (10), 1025–1033.
- [220] D. Evans und T. Blumenthal (Sep 2000) *trans splicing of polycistronic Caenorhabditis elegans pre-mRNAs: analysis of the SL2 RNA*. Mol Cell Biol, **20** (18), 6659–6667.
- [221] C. Frantz, C. Ebel, F. Paulus und P. Imbault (Jun 2000) *Characterization of trans-splicing in Euglenoids*. Curr Genet, **37** (6), 349–355.
- [222] C. MCDANIEL (2005) *MICROORGANISM COATING COMPONENTS, COATINGS, AND COATED SURFACES*. WIPO.
- [223] E. Alverca, S. Franca und S. M. Díaz de la Espina (Dec 2006) *Topology of splicing and snRNP biogenesis in dinoflagellate nuclei*. Biol Cell, **98** (12), 709–720.
- [224] D. B. Guiliano und M. L. Blaxter (Nov 2006) *Operon conservation and the evolution of trans-splicing in the phylum Nematoda*. PLoS Genet, **2** (11).
- [225] T. L. Bailey und C. Elkan (1994) *Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28–36. AAAI Press, Menlo Park, CA.
- [226] M. Jiang, J. Anderson, J. Gillespie und M. Mayne (2008) *uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts*. BMC Bioinformatics, **9**, 192–192.
- [227] H. Zhang, D. A. Campbell, N. R. Sturm und S. Lin (2009) *Dinoflagellate spliced leader RNA genes display a variety of sequences and genomic arrangements*. Mol Biol Evol.
- [228] H. Luo, G. Gilinger, D. Mukherjee und V. Bellofatto (Nov 1999) *Transcription initiation at the TATA-less spliced leader RNA gene promoter requires at least two DNA-binding proteins and a tripartite architecture that includes an initiator element*. J Biol Chem, **274** (45), 31947–31954.
- [229] G. B. Simpson, Alastair, E. K. MacQuarrie und A. J. Roger (2002) *Eukaryotic evolution: Early origin of canonical introns*. Nature, **419**, 270.
- [230] X. S. Chen, W. T. White, L. J. Collins und D. Penny (2008) *Computational identification of four spliceosomal snRNAs from the deep-branching eukaryote Giardia intestinalis*. PLoS ONE, **3**, e3106.
- [231] A. Simoes-Barbosa, D. Meloni, J. A. Wohlschlegel, M. M. Konarska und P. J. Johnson (2008) *Spliceosomal snRNAs in the unicellular eukaryote Trichomonas vaginalis are structurally conserved but lack a 5'-cap structure*. RNA, **14**, 1617–1631.
- [232] E. Bon, S. Casaregola, G. Blandin, B. Llorente, C. Neuvéglise, M. Munsterkotter, U. Guldener, H. W. Mewes, J. Van Helden, B. Dujon und C. Gaillardin (2003) *Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns*. Nucleic Acids Res., **31**, 1121–1135.
- [233] P. A. Maroney, G. J. Hannon, J. D. Shambaugh und T. W. Nilsen (Dec 1991) *Intramolecular base pairing between the nematode spliced leader and its 5' splice site is not essential for trans-splicing in vitro*. EMBO J, **10** (12), 3869–3875.
- [234] X. Chen, A. M. Quinn und S. L. Wolin (2000) *Ro ribonucleoproteins contribute to the resistance of Deinococcus radiodurans to ultraviolet resistance*. Genes Dev., **14**, 777–782.
- [235] M. Xie, A. Mosig, X. Qi, Y. Li, P. F. Stadler und J. J.-L. Chen (2008) *Size Variation and Structural Conservation of Vertebrate Telomerase RNA*. J. Biol. Chem., **283**, 2049–2059.
- [236] A. R. Gruber, D. Koper-Emde, M. Marz, H. Tafer, S. Bernhart, G. Obernosterer, A. Mosig, I. L. Hofacker, P. F. Stadler und B.-J. Benecke (2008) *Invertebrate 7SK snRNAs*. J. Mol. Evol., **107–115**, 66.
- [237] A. Gruber, C. Kilgus, A. Mosig, I. L. Hofacker, W. Hennig und P. F. Stadler (2008) *Arthropod 7SK RNA*. Mol. Biol. Evol., **1923–1930**, 25.
- [238] P. A. Maroney, Y. T. Yu, M. Jankowska und T. W. Nilsen (1996) *Direct Analysis of Nematode Cis and Trans-Spliceosomes: a Functional Role for U5 snRNA in Spliced Leader Addition Trans-Splicing and the Identification of Novel Sm snRNPs*. RNA, **2**, 735–745.
- [239] M. MacMorris, M. Kumar, E. Lasda, A. Larsen, B. Kraemer und T. Blumenthal (2007) *A Novel Family of C. elegans snRNPs Contains Proteins Associated With Trans-Splicing*. RNA, **13**, 511–520.
- [240] W. Deng, X. Zhu, G. Skogerboe, 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 und R. Chen (2006) *Organization of the Caenorhabditis elegans Small Non-Coding Transcriptome: Genomic Features, Biogenesis, and Expression*. Genome Res., **16**, 20–29.
- [241] A. Zemann, A. op de Bekke, M. Kiefmann, J. Brosius und J. Schmitz (2006) *Evolution of Small Nucleolar RNAs in Nematodes*. Nucleic Acids Research, **34**, 2676–2685.

- [242] S. F. Altschul, T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller und D. J. Lipman (1997) *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs*. *Nucleic Acids Res.*, **25**, 3389–3402.
- [243] E. P. Nawrocki und S. R. Eddy (2007) *Query-Dependent Banding (QDB) for Faster RNA Similarity Searches*. *PLoS Comput. Biol.*, **3**, e56.
- [244] 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 und A. Bateman (Jan 2009) *Rfam: updates to the RNA families database*. *Nucleic Acids Res.*, **37** (Database issue), 136–140.
- [245] W. Sudhaus und K. Kiontke (2007) *Comparison of the Cryptic Nematode Species Caenorhabditis brenneri sp. n. and C. remanei (Nematoda: Rhabditidae) With the Stem Species Pattern of the Caenorhabditis Elegans Group*. *Zootaxa*, **1456**, 45–62.
- [246] M. L. Blaxter, P. D. Ley, J. R. Garey, L. X. Liu, P. Scheldeman, A. Vierstraete, J. R. Vanfleteren, L. Y. Mackey, M. Dorris, L. M. Frisse, J. T. Vida und W. K. Thomas (1998) *A Molecular Evolutionary Framework for the Phylum Nematoda*. *Nature*, **392**, 71–75.
- [247] M. Mitreva, M. L. Blaxter, D. M. Bird und J. P. McCarter (2005) *Comparative Genomics of Nematodes*. *Trends Genet.*, **21**, 573–581.
- [248] International Human Genome Sequencing Consortium (2001) *Initial Sequencing and Analysis of the Human Genome*. *Nature*, **409**, 860–921.
- [249] B. J. Haas, M. Berriman, H. Hirai, G. G. Cerqueira, P. T. LoVerde und N. M. El-Sayed (2007) *Schistosoma mansoni Genome: Closing in on a Final Gene set*. *Exp. Parasitol.*, **117**, 225–228.
- [250] P. Dehal, Y. Satou, R. K. Campbell, J. Chapman, B. Degnan, A. D. Tomaso, B. Davidson, A. D. Gregorio, M. Gelpke, D. M. Goodstein, N. Harafuji, K. E. Hastings, I. Ho, K. Hotta, W. Huang, T. Kawashima, P. Lemaire, D. Martinez, I. A. Meinertzhagen, S. Nacula, M. Nonaka, N. Putnam, S. Rash, H. Saiga, M. Satake, A. Terry, L. Yamada, H. G. Wang, S. Awazu, K. Azumi, J. Boore, M. Branno, S. Chin-Bow, R. DeSantis, S. Doyle, P. Francino, D. N. Keys, S. Haga, H. Hayashi, K. Hino, K. Imai, K. Inaba, S. Kano, K. Kobayashi, M. Kobayashi, B. I. Lee, K. W. Makabe, C. Manohar, G. Matassi, M. Medina, Y. Mochizuki, S. Mount, T. Morishita, S. Miura, A. Nakayama, S. Nishizaka, H. Nomoto, F. Ohta, K. Oishi, I. Rigoutsos, M. Sano, A. Sasaki, Y. Sasakura, E. Shoguchi, T. Shin-i, A. Spagnuolo, D. Stainier, M. M. Suzuki, O. Tassy, N. Takatori, M. Tokuoka, K. Yagi, F. Yoshizaki, S. Wada, C. Zhang, P. D. Hyatt, F. Larimer, C. Detter, N. Doggett, T. Glavina, T. Hawkins, P. Richardson, S. Lucas, Y. Kohara, M. Levine, N. Satoh und D. S. Rokhsar (2002) *The Draft Genome of Ciona intestinalis: Insights Into Chordate and Vertebrate Origins*. *Science*, **298**, 2157–2167.
- [251] T. Blumenthal (2005) *Trans-Splicing and Operons*. The *C. elegans* Research Community, Herausgeber, *Worm-Book*. doi/10.1895/wormbook.1.5.1, <http://www.wormbook.org>.
- [252] S. I. Lee und J. A. Steitz (1990) *Herpesvirus saimiri U RNAs Are Expressed and Assembled Into Ribonucleoprotein Particles in the Absence of Other Viral Genes*. *J. Virol.*, **64**, 3905–3915.
- [253] K. Mowry und J. A. Steitz (1987) *Identification of the human U7 snRNP as one of several factors involved in the 3' end maturation of histone premessenger RNAs*. *Science*, **238**, 1682–1687.
- [254] S. Fabry, K. Müller, A. Lindauer, P. B. Park, T. Cornelius und R. Schmitt (1995) *The organization structure and regulatory elements of Chlamydomonas histone genes reveal features linking plant and animal genes*. *Curr. Genet.*, **28**, 333–345.
- [255] W. D. Townley-Tilson, S. A. Pendergrass, W. F. Marzluff und M. L. Whitfield (2006) *Genome-wide analysis of mRNAs bound to the histone stem-loop binding protein RNA*. *RNA*, **12**, 1853–1867.
- [256] T. J. Golembe, J. Yong und G. Dreyfuss (2005) *Specific sequence features, recognized by the SMN complex, identify snRNAs and determine their fate as snRNPs*. *Mol. Cell Biol.*, **25**, 10989–11004.
- [257] T. N. Azzouz und D. Schümperli (2003) *Evolutionary conservation of the U7 small nuclear ribonucleoprotein in Drosophila melanogaster*. *RNA*, **9**, 1532–1541.
- [258] R. S. Pillai, M. Grimmer, G. Meister, C. L. Will, R. Lührmann, U. Fischer und D. Schümperli (2003) *Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing*. *Genes. Dev.*, **17**, 2321–2333.
- [259] D. Schümperli und R. S. Pillai (2004) *The special Sm core structure of the U7 snRNP: far-reaching significance of a small nuclear ribonucleoprotein*. *Cell. Mol. Life Sci.*, **61**, 2560–2570.
- [260] N. G. Kolev und J. A. Steitz (2006) *In vivo assembly of functional U7 snRNP requires RNA backbone flexibility within the Sm-binding site*. *Nat. Struct. Mol. Biol.*, **13**, 347–353.
- [261] S. Jaeger, F. Martin, J. Rudinger-Thirion, R. Giegé und G. Eriani (2006) *Binding of human SLBP on the 3'-UTR of histone precursor H4-12 mRNA induces structural rearrangements that enable U7 snRNA anchoring*. *Nucleic Acids Res.*, **34**, 4987–4995.
- [262] C. Brun, D. Suter, C. Pauli, P. Dunant, H. Lochmüller, B. J.-M., D. Schümperli und J. Weis (2003) *U7 snRNAs induce correction of mutated dystrophin pre-mRNA by exon skipping*. *Cell. Mol. Life Sci.*, **60**, 557–566.

- [263] A. Goyenvalle, A. Vulin, F. Fougerousse, F. Leturcq, J.-C. Kaplan, L. Garcia und O. Danos (2004) *Rescue of dystrophic muscle through U7 snRNA-mediated exon skipping*. *Science*, **306**, 1796–1799.
- [264] D. Soldati und D. Schümperli (1988) *Structural and functional characterization of mouse U7 small nuclear RNA active in 3' processing of histone pre-mRNA*. *Mol. Cell Biol.*, **8**, 1518–1524.
- [265] A. Gruber, D. Soldati, M. Burri und D. Schümperli (1991) *Isolation of an active gene and of two pseudogenes for mouse U7 small nuclear RNA*. *Biochim. Biophys. Acta*, **1088**, 151–154.
- [266] S. C. Phillips und P. C. Turner (1992) *A transcriptional analysis of the gene encoding mouse U7 small nuclear RNA*. *Gene*, **116**, 181–186.
- [267] S. C. Phillips und P. C. Turner (1991) *Sequence and expression of a mouse U7 snRNA type II pseudogene*. *DNA Seq.*, **1**, 401–404.
- [268] Y.-T. Yu, W.-Y. Tarn, T. A. Yario und J. A. Steitz (1996) *More Sm snRNAs from Vertebrate Cells*. *Exp. Cell Res.*, **229**, 276–281.
- [269] K. Strub, G. Galli, M. Busslinger und M. L. Birnstiel (1984) *The cDNA sequences of the sea urchin U7 small nuclear RNA suggest specific contacts between histone mRNA precursor and U7 RNA during RNA processing*. *EMBO J.*, **3**, 2801–2807.
- [270] M. De Lorenzi, U. Rohrer und M. L. Birnstiel (1986) *Analysis of a sea urchin gene cluster coding for the small nuclear U7 RNA, a rare RNA species implicated in the 3' editing of histone precursor mRNAs*. *Proc. Natl. Acad. Sci. USA*, **83**, 3243–3247.
- [271] G. M. Gilmartin, F. Schaufele, G. Schaffner und M. L. Birnstiel (1988) *Functional analysis of the sea urchin U7 small nuclear RNA*. *Mol. Cell Biol.*, **8**, 1076–1084.
- [272] C. Southgate und M. Busslinger (1989) *In vivo and in vitro expression of U7 snRNA genes: cis- and trans-acting elements required for RNA polymerase II-directed transcription*. *EMBO J.*, **8**, 539–549.
- [273] S. C. Phillips und M. L. Birnstiel (1992) *Analysis of a gene cluster coding for the Xenopus laevis U7 snRNA*. *Biochim. Biophys. Acta*, **1131**, 95–98.
- [274] N. J. Watkins, S. C. Phillips und P. C. Turner (1992) *The U7 small nuclear RNA genes of Xenopus borealis*. *Biochem. Soc. Trans.*, **20**, 301S.
- [275] C.-H. H. Wu und J. G. Gall (1993) *U7 small nuclear RNA in C snurposomes of the Xenopus germinal vesicle*. *Proc. Natl. Acad. Sci. USA*, **90**, 6257–6259.
- [276] Z. Dominski, X.-c. Yang, M. Purdy und W. F. Marzluff (2003) *Cloning and characterization of the Drosophila U7 small nuclear RNA*. *Proc. Natl. Acad. Sci. USA*, **100**, 9422–9427.
- [277] R. Lück, S. Gräf und G. Steger (1999) *ConStruct: A tool for thermodynamic controlled prediction of conserved secondary structure*. *Nucl. Acids Res.*, **27**, 4208–4217.
- [278] G. Hernandez Jr., F. Valafar und W. E. Stumph (2007) *Insect small nuclear RNA gene promoters evolve rapidly yet retain conserved features involved in determining promoter activity and RNA polymerase specificity*. *Nucleic Acids Res.*, **35**, 21–34.
- [279] D. Soldati und D. Schümperli (1990) *Structures of four human pseudogenes for U7 small nuclear RNA*. *1990*, **95**, 305–306.
- [280] Z. Dominski, X.-C. Yang, M. Purdy und W. Marzluff (2005) *Differences and similarities between Drosophila and mammalian 3' end processing of histone pre-mRNAs*. *RNA*, **11**, 1835–1847.
- [281] E. P. Nawrocki und S. R. Eddy (2007) *Query-Dependent Banding for Faster RNA Similarity Searches*. *PLoS Comp. Biol.*, **3**, e56. Doi:10.1371/journal.pcbi.0030056.
- [282] Mouse Genome Sequencing Consortium (2002) *Initial sequencing and comparative analysis of the mouse genome*. *Nature*, **420**, 520–562.
- [283] T. B. Nesterova, S. Y. Slobodyanyuk, E. A. Elisaphenko, A. I. Shevchenko, C. Johnston, M. E. Pavlova, I. B. Rogozin, N. N. Kolesnikov, N. Brockdorff und S. M. Zakian (2001) *Characterization of the genomic Xist locus in rodents reveals conservation of overall gene structure and tandem repeats but rapid evolution of unique sequence*. *Genome Res.*, **11**, 833–849.
- [284] J. Ponjavic, C. P. Ponting und G. Lunter (2007) *Functionality or transcriptional noise? Evidence for selection within long noncoding RNAs*. *Genome Res.*, **17**, 556–565.
- [285] A. Stark, M. F. Lin, P. Kheradpour, J. S. Pedersen, L. Parts, J. W. Carlson, M. A. Crosby, M. D. Rasmussen, S. Roy, A. N. Deoras, J. G. Ruby, J. Brennecke, Harvard FlyBase curators, Berkeley Drosophila Genome Project, E. Hodges, A. S. Hinrichs, A. Caspi, B. Paten, S. W. Park, M. V. Han, M. L. Maeder, B. J. Polansky, B. E. Robson, S. Aerts, J. van Helden, B. Hassan, D. G. Gilbert, D. A. Eastman, M. Rice, M. Weir, M. W. Hahn, Y. Park, C. N. Dewey, L. Pachter, W. J. Kent, D. Haussler, E. C. Lai, D. P. Bartel, G. J. Hannon, T. C. Kaufman, M. B. Eisen, A. G. Clark, D. Smith, S. E. Celniker, W. M. Gelbart und M. Kellis (2007) *Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures*. *Nature*, **450**, 219–232.

- [286] Drosophila 12 Genomes Consortium (2007) *Evolution of genes and genomes on the Drosophila phylogeny*. Nature, **450**, 203–218.
- [287] A. Stark, P. Kheradpour, L. Parts, J. Brennecke, E. Hodges, G. J. Hannon und M. Kellis (2007) *Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes*. Genome Res, **17** (12), 1865–1879.
- [288] P. Kheradpour, A. Stark, S. Roy und M. Kellis (2007) *Reliable prediction of regulator targets using 12 Drosophila genomes*. Genome Res, **17** (12), 1919–1931.
- [289] L. P. Lim und C. B. Burge (2001) *A computational analysis of sequence features involved in recognition of short introns*. Proc Natl Acad Sci USA, **98**, 11193–11198.
- [290] S. M. Mount, C. Burks, G. Hertz, G. D. Stormo, O. White und C. Fields (Aug 1992) *Splicing signals in Drosophila: intron size, information content, and consensus sequences*. Nucleic Acids Res, **20** (16), 4255–4262.
- [291] M. Deutsch und M. Long (1999) *Intron-exon structures of eukaryotic model organisms*. Nucleic Acids Res., **27**, 3219–3228.
- [292] C. N. Dewey, I. B. Rogozin und E. V. Koonin (2006) *Compensatory relationship between splice sites and exonic splicing signals depending on the length of vertebrate introns*. BMC Genomics, **7**, 311.
- [293] D. L. Halligan und P. D. Keightley (Jul 2006) *Ubiquitous selective constraints in the Drosophila genome revealed by a genome-wide interspecies comparison*. Genome Res, **16** (7), 875–884.
- [294] J. Parsch (Dec 2003) *Selective constraints on intron evolution in Drosophila*. Genetics, **165** (4), 1843–1851.
- [295] M. Hiller, S. Findeiß, S. Lein, M. Marz, C. Nickel, D. Rose, C. Schulz, R. Backofen, S. J. Prohaska, G. Reuter und P. F. Stadler (May 2009) *Conserved introns reveal novel transcripts in Drosophila melanogaster*. Genome Res.
- [296] K. Okamura, W.-J. Chung, J. G. Ruby, H. Guo, D. P. Bartel und E. C. Lai (2008) *The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs*. Nature, **453**, 803–806.
- [297] D. Karolchik, R. M. Kuhn, R. Baertsch, G. P. Barber, H. Clawson, M. Diekhans, B. Giardine, R. A. Harte, A. S. Hinrichs, F. Hsu, K. M. Kober, W. Miller, J. S. Pedersen, A. Pohl, B. J. Raney, B. Rhead, K. R. Rosenbloom, K. E. Smith, M. Stanke, A. Thakkapallayil, H. Trumbower, T. Wang, A. S. Zweig, D. Haussler und W. J. Kent (2008) *The UCSC Genome Browser Database: 2008 update*. Nucleic Acids Res., **36**, D773–D779.
- [298] K. Tamura, S. Subramanian und S. Kumar (2004) *Temporal patterns of fruit fly (Drosophila) evolution revealed by mutation clocks*. Mol Biol Evol, **21** (1), 36–44.
- [299] J. L. Tupy, A. M. Bailey, G. Dailey, M. Evans-Holm, C. W. Siebel, S. Misra, S. E. Celniker und G. M. Rubin (2005) *Identification of putative noncoding polyadenylated transcripts in Drosophila melanogaster*. Proc Natl Acad Sci U S A, **102** (15), 5495–5500.
- [300] S. Misra, M. A. Crosby, C. J. Mungall, B. B. Matthews, K. S. Campbell, P. Hradecky, Y. Huang, J. S. Kaminker, G. H. Millburn, S. E. Prochnik, C. D. Smith, J. L. Tupy, E. J. Whitfield, L. Bayraktaroglu, B. P. Berman, B. R. Bettencourt, S. E. Celniker, A. D. N. J. de Grey, R. A. Drysdale, N. L. Harris, J. Richter, S. Russo, A. J. Schroeder, S. Q. Shu, M. Stapleton, C. Yamada, M. Ashburner, W. M. Gelbart, G. M. Rubin und S. E. Lewis (2002) *Annotation of the Drosophila melanogaster euchromatic genome: a systematic review*. Genome Biol, **3** (12), RESEARCH0083.
- [301] K. C. Panga, M. C. Fritha und J. S. Mattick (2006) *Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function*. Trends Genet, **22** (1), 1–5.
- [302] R. L. Kelley und M. I. Kuroda (2000) *Noncoding RNA genes in dosage compensation and imprinting*. Cell, **103** (1), 9–12.
- [303] K. Okamura, J. W. Hagen, H. Duan, D. M. Tyler und E. C. Lai (2007) *The mirtron pathway generates microRNA-class regulatory RNAs in Drosophila*. Cell, **130** (1), 89–100.
- [304] E. Berezikov, W.-J. Chung, J. Willis, E. Cuppen und E. C. Lai (2007) *Mammalian mirtron genes*. Mol Cell, **28** (2), 328–336.
- [305] M. Hiller, K. Huse, M. Platzer und R. Backofen (2005) *Non-EST based prediction of exon skipping and intron retention events using Pfam information*. Nucleic Acids Res, **33** (17), 5611–5621.
- [306] V. Atzorn, P. Fragapane und T. Kiss (2004) *U17/snR30 Is a Ubiquitous snoRNA with Two Conserved Sequence Motifs Essential for 18S rRNA Production*. Mol Cell Biol., **24**, 1769–1778.
- [307] K. Tyc und J. A. Steitz (Oct 1989) *U3, U8 and U13 comprise a new class of mammalian snRNPs localized in the cell nucleolus*. EMBO J, **8** (10), 3113–3119.
- [308] D. Jia, L. Cai, H. He, G. Skogerbo, T. Li, M. N. Aftab und R. Chen (2007) *Systematic identification of non-coding RNA 2,2,7-trimethylguanosine cap structures in Caenorhabditis elegans*. BMC Mol Biol, **8**, 86–86.



- [309] S. Nabavi und R. N. Nazar (Oct 2008) *U3 snoRNA promoter reflects the RNA's function in ribosome biogenesis*. *Curr Genet*, **54** (4), 175–184.
- [310] C. Verheggen, D. L. Lafontaine, D. Samarsky, J. Mouaikel, J. M. Blanchard, R. Bordonné und E. Bertrand (Jun 2002) *Mammalian and yeast U3 snoRNPs are matured in specific and related nuclear compartments*. *EMBO J*, **21** (11), 2736–2745.
- [311] S. Boulon, C. Verheggen, B. E. Jady, C. Girard, C. Pescia, C. Paul, J. K. Ospina, T. Kiss, A. G. Matera, R. Bordonné und E. Bertrand (Dec 2004) *PHAX and CRMI are required sequentially to transport U3 snoRNA to nucleoli*. *Mol Cell*, **16** (5), 777–787.
- [312] A. V. Borovjagin und S. A. Gerbi (Jun 2004) *Xenopus U3 snoRNA docks on pre-rRNA through a novel base-pairing interaction*. *RNA*, **10** (6), 942–953.
- [313] Y. Sasano, Y. Hokii, K. Inoue, H. Sakamoto, C. Ushida und T. Fujiwara (Jun 2008) *Distribution of U3 small nucleolar RNA and fibrillarin during early embryogenesis in Caenorhabditis elegans*. *Biochimie*, **90** (6), 898–907.
- [314] M. Antal, A. Mougín, M. Kis, E. Boros, G. Steger, G. Jakab, F. Solymosy und C. Branlant (Aug 2000) *Molecular characterization at the RNA and gene levels of U3 snoRNA from a unicellular green alga, Chlamydomonas reinhardtii*. *Nucleic Acids Res*, **28** (15), 2959–2968.
- [315] C. Marshallsay, S. Connelly und W. Filipowicz (Sep 1992) *Characterization of the U3 and U6 snRNA genes from wheat: U3 snRNA genes in monocot plants are transcribed by RNA polymerase III*. *Plant Mol Biol*, **19** (6), 973–983.
- [316] J. C. Mottram, S. D. Bell, R. G. Nelson und J. D. Barry (Sep 1991) *tRNAs of Trypanosoma brucei. Unusual gene organization and mitochondrial importation*. *J Biol Chem*, **266** (27), 18313–18317.
- [317] J. A. Wise und A. M. Weiner (Nov 1980) *Dictyostelium small nuclear RNA D2 is homologous to rat nucleolar RNA U3 and is encoded by a dispersed multigene family*. *Cell*, **22** (1 Pt 1), 109–118.
- [318] M. P. Skupski, D. A. Jackson und D. O. Natvig (Feb 1997) *Phylogenetic Analysis of Heterothallic Neurospora Species*. *Fungal Genet Biol*, **21** (1), 153–162.
- [319] C. K. Tsui, H. M. Daniel, V. Robert und W. Meyer (Jun 2008) *Re-examining the phylogeny of clinically relevant Candida species and allied genera based on multigene analyses*. *FEMS Yeast Res*, **8** (4), 651–659.
- [320] D. S. Hibbett, M. Binder, J. F. Bischoff, M. Blackwell, P. F. Cannon, O. E. Eriksson, S. Huhndorf, T. James, P. M. Kirk, R. Lücking, H. Thorsten Lumbsch, F. Lutzoni, P. B. Matheny, D. J. McLaughlin, M. J. Powell, S. Redhead, C. L. Schoch, J. W. Spatafora, J. A. Stalpers, R. Vilgalys, M. C. Aime, A. Aptroot, R. Bauer, D. Begerow, G. L. Benny, L. A. Castlebury, P. W. Crous, Y. C. Dai, W. Gams, D. M. Geiser, G. W. Griffith, C. Gueidan, D. L. Hawksworth, G. Hestmark, K. Hosaka, R. A. Humber, K. D. Hyde, J. E. Ironside, U. Köljalg, C. P. Kurtzman, K. H. Larsson, R. Lichtwardt, J. Longcore, J. Miadlikowska, A. Miller, J. M. Moncalvo, S. Mozley-Standridge, F. Oberwinkler, E. Parmasto, V. Reeb, J. D. Rogers, C. Roux, L. Ryvarden, J. P. Sampaio, A. Schüssler, J. Sugiyama, R. G. Thorn, L. Tibell, W. A. Untereiner, C. Walker, Z. Wang, A. Weir, M. Weiss, M. M. White, K. Winka, Y. J. Yao und N. Zhang (May 2007) *A higher-level phylogenetic classification of the Fungi*. *Mycol Res*, **111** (Pt 5), 509–547.
- [321] R. Fournier, F. Brulé, V. Ségault, A. Mougín und C. Branlant (Mar 1998) *U3 snoRNA genes with and without intron in the Kluyveromyces genus: yeasts can accommodate great variations of the U3 snoRNA 3'-terminal domain*. *RNA*, **4** (3), 285–302.
- [322] D. K. Willkomm und R. K. Hartmann (2007) *An important piece of the RNase P jigsaw solved*. *Trends Biochem Sci*, **32**, 247–250.
- [323] S. C. Walker und D. R. Engelke (2006) *Ribonuclease P: the evolution of an ancient RNA enzyme*. *Crit Rev Biochem Mol Biol*, **41**, 77–102.
- [324] P. Piccinelli, M. A. Rosenblad und T. Samuelsson (2005) *Identification and analysis of ribonuclease P and MRP RNA in a broad range of eukaryotes*. *Nucleic Acids Res*, **33**, 4485–4495.
- [325] M. D. Woodhams, P. F. Stadler, D. Penny und L. J. Collins (2007) *RNase MRP and the RNA Processing Cascade in the Eukaryotic Ancestor*. *BMC Evol. Biol*, **7**, S13.
- [326] L. J. Collins, V. Moulton und D. Penny (2000) *Use of RNA secondary structure for studying the evolution of RNase P and RNase MRP*. *J Mol Evol*, **51**, 194–204.
- [327] T. V. Aspinall, J. M. B. Gordon, H. J. Bennet, P. Karahalios, J.-P. Bukowski, S. C. Walker, D. R. Engelke und J. M. Avis (2007) *Interactions between subunits of Saccharomyces cerevisiae RNase MRP support a conserved eukaryotic RNase P/MRP architecture*. *Nucleic Acids Res*, **35**, 6439–6450.
- [328] R. Kachouri, V. Stribinskis, Y. Zhu, K. S. Ramos, E. Westhof und Y. Li (2005) *A surprisingly large RNase P RNA in Candida glabrata*. *RNA*, **11**, 1064–1072.
- [329] J. de la Cruz und A. Vioque (2003) *A structural and functional study of plastid RNAs homologous to catalytic bacterial RNase P RNA*. *Gene*, **321**, 47–56.

- [330] E. R. Seif, L. Forget, N. C. Martin und B. F. Lang (2003) *Mitochondrial RNase P RNAs in ascomycete fungi: lineage-specific variations in RNA secondary structure*. RNA, **9**, 1073–1083.
- [331] L. Randau, I. Schröder und D. Söll (2008) *Life without RNase P*. Nature, **453**, 120–123.
- [332] S. M. Marquez, J. K. Harris, S. T. Kelley, J. W. Brown, S. C. Dawson, E. C. Roberts und N. R. Pace (2005) *Structural implications of novel diversity in eucaryal RNase P RNA*. RNA, **11**, 739–751.
- [333] Y. Zhu, D. K. Pulukkunat und Y. Li (2007) *Deciphering RNA structural diversity and systematic phylogeny from microbial metagenomes*. Nucleic Acids Res., **35**, 2283–2294.
- [334] S. Barth, B. Shalem, A. Hury, I. D. Tkacz, X. H. Liang, S. Uziel, I. Myslyuk, T. Doniger, M. Salmon-Divon, R. Unger und S. Michaeli (Jan 2008) *Elucidating the role of C/D snoRNA in rRNA processing and modification in Trypanosoma brucei*. Eukaryot Cell, **7** (1), 86–101.
- [335] W. Krüger und B. J. Benecke (1987) *Structural and functional analysis of a human 7 S K RNA gene*. J. Mol. Biol., **195**, 31–41.
- [336] S. Murphy, C. Di Liegro und M. Melli (1987) *The in vitro transcription of the 7SK RNA gene by RNA polymerase III is dependent only on the presence of an upstream promoter*. Cell, **51**, 81–87.
- [337] H.-C. Gürsoy, D. Koper und B.-J. Benecke (2000) *The vertebrate 7S K RNA separates hagfish (Myxine glutinosa) and lamprey (Lampetra fluviatilis)*. J. Mol. Evol., **50**, 456–464.
- [338] S. Egloff, E. Van Herreweghe und T. Kiss (2006) *Regulation of polymerase II transcription by 7SK snRNA: two distinct RNA elements direct P-TEFb and HEXIM1 binding*. Mol. Cell. Biol., **26**, 630–642.
- [339] A. A. Michels, Q. Fraldi, A. Li, T. E. Adamson, F. Bonnet, V. T. Nguyen, S. C. Sedore, J. P. Price, D. H. Price, L. Lania, und O. Bensaude (2004) *Binding of the 7SK snRNA turns the HEXIM1 protein into a P-TEFb (CDK9/cyclin T) inhibitor*. EMBO J., **23**, 2608–2619.
- [340] D. Blazek, M. Barboric, J. Kohoutek, I. Oven und B. M. Peterlin (2005) *Oligomerization of HEXIM1 via 7SK snRNA and coiled-coil region directs the inhibition of P-TEFb*. Nucleic Acids Res., **33**, 7000–7010.
- [341] B. M. Peterlin und D. H. Price (297-305) *Controlling the elongation phase of transcription with P-TEFb*. Mol. Cell., **2006**, 23.
- [342] W. J. He, R. Chen, Z. Yang und Q. Zhou (2006) *Regulation of two key nuclear enzymatic activities by the 7SK small nuclear RNA*. Cold Spring Harb Symp Quant Biol., **71**, 301–311.
- [343] J. H. Yik, R. Chen, R. Nishimura, J. L. Jennings, A. J. Link und Q. Zhou (Oct 2003) *Inhibition of P-TEFb (CDK9/Cyclin T) kinase and RNA polymerase II transcription by the coordinated actions of HEXIM1 and 7SK snRNA*. Mol Cell, **12** (4), 971–982.
- [344] A. A. Michels, V. T. Nguyen, A. Fraldi, V. Labas, M. Edwards, F. Bonnet, L. Lania und O. Bensaude (2003) *MAQ1 and 7SK RNA interact with CDK9/cyclin T complexes in a transcription-dependent manner*. Mol Cell Biol, **23** (14), 4859–4869.
- [345] A. Markert, M. Grimm, J. Martinez, J. Wiesner, A. Meyerhans, O. Meyuhas, A. Sickmann und U. Fischer (Jun 2008) *The La-related protein LARP7 is a component of the 7SK ribonucleoprotein and affects transcription of cellular and viral polymerase II genes*. EMBO Rep., **9**, 569–575.
- [346] D. Sürig, S. Bredow und B. J. Benecke (1993) *The seemingly identical 7SK and U6 core promoters depend on different transcription factor complexes*. Gene Expr., **3**, 175–185.
- [347] D. A. Wassarman und J. A. Steitz (1991) *Structural analyses of the 7SK ribonucleoprotein (RNP), the most abundant human small RNP of unknown function*. Mol. Cell. Biol., **11**, 3432–3445.
- [348] S. A. Byers, J. P. Price, J. J. Cooper, Q. Li und D. H. Price (2005) *HEXIM2, a HEXIM1-related protein, regulates positive transcription elongation factor b through association with 7SK*. J Biol Chem., **280**, 16360–16377.
- [349] H. Nishihara, S. Maruyama und N. Okada (2009) *Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals*. Proc Natl Acad Sci USA, **106**, 5235–5240.
- [350] A. Mosig, J. L. Chen und P. F. Stadler (2007) *Homology Search with Fragmented Nucleic Acid Sequence Patterns*. R. Giancarlo und S. Hannenhalli, Herausgeber, Algorithms in Bioinformatics (WABI 2007), Band 4645 von Lecture Notes in Computer Science, 335–345. Springer Verlag, Berlin, Heidelberg.
- [351] C. Jeronimo, D. Forget, A. Bouchard, Q. Li, G. Chua, C. Poitras, C. Thérien, D. Bergeron, S. Bourassa, J. Greenblatt, B. Chabot, G. G. Poirier, T. R. Hughes, M. Blanchette, D. H. Price und B. Coulombe (Jul 2007) *Systematic analysis of the protein interaction network for the human transcription machinery reveals the identity of the 7SK capping enzyme*. Mol Cell, **27** (2), 262–274.
- [352] J. A. Box, J. T. Bunch, W. Tang und P. Baumann (Dec 2008) *Spliceosomal cleavage generates the 3' end of telomerase RNA*. Nature.
- [353] J. D. Podlevsky, C. J. Bley, R. V. Omana, X. Qi und J. J. Chen (2008) *The telomerase database*. Nucleic Acids Res., **36**, D339–D343.

- [354] Y. Tzfati, Z. Knight, J. Roy und E. H. Blackburn (2003) *A novel pseudoknot element is essential for the action of a yeast telomerase*. *Genes & Dev.*, **17**, 1779–1788.
- [355] J. Leonardi, J. A. Box, J. T. Bunch und P. Baumann (2008) *TER1, the RNA subunit of fission yeast telomerase*. *Nat Struct Mol Biol*, **15**, 26–33.
- [356] C. J. Webb und V. A. Zakian (2008) *Identification and characterization of the Schizosaccharomyces pombe TER1 telomerase RNA*. *Nat Struct Mol Biol*, **15**, 34–42.
- [357] J. Franke, J. Gehlen und A. E. Ehrenhofer-Murray (Nov 2008) *Hypermethylation of yeast telomerase RNA by the snRNA and snoRNA methyltransferase Tgs1*. *J Cell Sci*, **121** (Pt 21), 3553–3560.
- [358] N. B. Ulyanov, K. Shefer, T. L. James und Y. Tzfati (2007) *Pseudoknot structures with conserved base triples in telomerase RNAs of ciliates*. *Nucleic Acids Res.*, **35**, 6150–6160.
- [359] K. Chakrabarti, M. Pearson, L. Grate, T. Sterne-Weiler, J. Deans und M. Donohue, J P amd Ares Jr. (2007) *Structural RNAs of known and unknown function identified in malaria parasites by comparative genomics and RNA analysis*. *RNA*, **13**, 1923–1939.
- [360] J.-L. Chen und C. W. Greider (2004) *An emerging consensus for telomerase RNA structure*. *Proc. Natl. Acad. Sci. USA*, **101**, 14683–14684.
- [361] A. T. Dandjinou, N. Lévesque, S. Larose, J.-F. Lucier, S. A. Elela und R. J. Wellinger (2004) *A phylogenetically based secondary structure for the yeast telomerase RNA*. *Curr. Biol.*, **14**, 1148–1158.
- [362] D. C. Zappulla und T. R. Cech (2004) *Yeast telomerase RNA: a flexible scaffold for protein subunits*. *Proc. Natl. Acad. Sci. USA*, **101**, 10024–10029.
- [363] Y. Brown, M. Abraham, S. Pearl, M. M. Kabaha, E. Elboher und Y. Tzfati (2007) *A critical three-way junction is conserved in budding yeast and vertebrate telomerase RNAs*. *Nucleic Acids Res.*, **35**, 6280–6289.
- [364] J. R. Mitchell, J. Cheng und C. K. (1999) *A box H/ACA small nucleolar RNA-like domain at the human telomerase 3' end*. *Mol. Cell Biol.*, **19**, 567–576.
- [365] B. E. Jády, E. Bertrand und T. Kiss (2004) *Human telomerase RNA and box H/ACA scaRNAs share a common Cajal body specific localization signal*. *J. Cell Biol.*, **164**, 647–652.
- [366] O. Voigt, A. G. Collins, V. B. Pearse, J. S. Pearse, A. Ender, H. Hadrys und B. Schierwater (2004) *Placozoa – no longer a phylum of one*. *Curr. Biol.*, **14**, R944–R945.
- [367] T. Syed und B. Schierwater (2002) *Trichoplax adhaerens: Discovered as a missing link, forgotten as a hydrozoan, re-discovered as a key to metazoan evolution*. *Vie et Milieu*, **52**, 177–187.
- [368] B. Schierwater, M. Eitel, W. Jakob, H. J. Osigus, H. Hadrys, S. L. Dellaporta, S. O. Kolokotronis und R. Desalle (Jan 2009) *Concatenated analysis sheds light on early metazoan evolution and fuels a modern "urmetazoon" hypothesis*. *PLoS Biol*, **7** (1).
- [369] W. Jakob, S. Sagasser, S. Dellaporta, P. Holland, K. Kuhn und B. Schierwater (2004) *The Trox-2 Hox/ParaHox gene of Trichoplax (Placozoa) marks an epithelial boundary*. *Dev Genes Evol.*, **214**, 170–175.
- [370] M. S. Srivastava, E. Begovic, J. Chapman, N. H. Putnam, U. Hellsten, T. Kawashima, A. Kuo, T. Mitros, M. L. Carpenter, A. Y. Signorovitch, M. A. Moreno, K. Kamm, H. Shapiro, I. V. Grigoriev, L. W. Buss, B. Schierwater, S. L. Dellaporta und D. S. Rokhsar (2008) *The Trichoplax Genome and the Nature of Placozoans*. *Nature*, **454**, 955–960.
- [371] R. N. Nazar (2004) *Ribosomal RNA processing and ribosome biogenesis in eukaryotes*. *IUBMB Life*, **56**, 457–465.
- [372] P. O. Wainright, G. Hinkle, M. L. Sogin und S. K. Stickel (1993) *The monophyletic origins of the metazoa; an unexpected evolutionary link with fungi*. *Science*, **260**, 340–342.
- [373] D. M. Odorico und D. J. Miller (1997) *Internal and external relationships of the Cnidaria: implications of primary and predicted secondary structure of the 5'-end of the 23S-like rDNA*. *Proc. R. Soc. Lond., B, Biol. Sci.*, **264**, 77–82.
- [374] F. Britto da Silva, V. Muschner und S. L. Bonatto (2007) *Phylogenetic position of Placozoa based on large subunit (LSU) and small subunit (SSU) rRNA genes*. *Genetics Mol. Biol.*, **30**, 127–132.
- [375] K. M. Val'ekho-Roman, V. K. Bobrova, A. V. Troitskiĭ, A. B. Tsetlin und I. L. Okshteĭn (1990) *[New data on Trichoplax: the nucleotide sequence of 5S rRNA]*. *Dokl Akad Nauk SSSR*, **311**, 500–503.
- [376] K. Nagai, C. Oubridge, A. Kuglstatter, E. Menichelli, C. Isel und L. Jovine (2003) *Structure, function and evolution of the signal recognition particle*. *EMBO J.*, **22**, 3479–3485.
- [377] M. Alm Rosenblad, G. J., B. Knudsen, C. Zwieb und T. Samuelsson (2003) *SRPDB (Signal Recognition Particle Database)*. *Nucleic Acids Res.*, **31**, D363–364.
- [378] J.-P. Bachelierie, J. Cavallé und A. Hüttenhofer (2002) *The expanding snoRNA world*. *Biochimie*, **84**, 775–790.

- [379] C. A. Enright, E. S. Maxwell, G. L. Eliceiri und B. Sollner-Webb (1996) *5'ETS rRNA processing facilitated by four small RNAs: U14, E3, U17, and U3*. RNA, **2**, 1094–1099.
- [380] S. Ro, C. Park, J. Jin, K. M. Sanders und W. Yan (2006) *A PCR-based method for detection and quantification of small RNAs*. Biochem Biophys Res Comm, **351**, 756–763.
- [381] S. L. Stricklin, S. Griffiths-Jones und S. R. Eddy (2005) *C. elegans noncoding RNA genes*. WormBook, doi/10.1895/wormbook.1.7.1. [http://www.wormbook.org/chapters/www\\_noncodingRNA/noncodingRNA.html](http://www.wormbook.org/chapters/www_noncodingRNA/noncodingRNA.html).
- [382] L. F. Sempere, C. N. Cole, M. A. McPeck und K. J. Peterson (2006) *The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint*. J Exp Zool B Mol Dev Evol., **306B**, 575–588.
- [383] S. E. Prochnik, D. S. Rokhsar und A. A. Aboobaker (2007) *Evidence for a microRNA expansion in the bilaterian ancestor*. Dev Genes Evol., **217**, 73–77.
- [384] A. Grimson, M. Srivastava, B. Fahey, B. J. Woodcroft, H. R. Chiang, N. King, A. M. Degnan, D. S. Rokhsar und D. P. Bartel (2008) *Early origins and evolution of miRNAs and Piwi-interacting RNAs in animals*. Nature.
- [385] J. Hertel, M. Lindemeyer, K. Missal, C. Fried, A. Tanzer, C. Flamm, I. L. Hofacker, P. F. Stadler und The Students of Bioinformatics Computer Labs 2004 and 2005 (2006) *The Expansion of the Metazoan MicroRNA Repertoire*. BMC Genomics, **7**, 15 [epub].
- [386] R. Niwa und F. J. Slack (2007) *The evolution of animal microRNA function*. Curr. Op. Gen. Devel., **17**, 145–150.
- [387] C. T. Lee, T. Risom und W. M. Strauss (2007) *Evolutionary conservation of microRNA regulatory circuits: an examination of microRNA gene complexity and conserved microRNA-target interactions through metazoan phylogeny*. DNA Cell Biol., **26**, 209–218.
- [388] A. Hinas, J. Reimegård, E. G. Wagner, W. Nellen, V. Ambros und F. Söderbom (2007) *The small RNA repertoire of Dictyostellium discoideum and its regulation by components of the RNAi pathway*. Nucleic Acids Res., **6714-6726**, 35.
- [389] T. Zhao, G. Li, S. Mi, S. Li, G. J. Hannon, X. J. Wang und Y. Qi (2007) *A complex system of small RNAs in the unicellular green alga Chlamydomonas reinhardtii*. Genes Dev., **21**, 1190–1203.
- [390] A. Molnár, F. Schwach, D. Studholme, E. C. Thuenemann und D. C. Baulcombe (2007) *miRNAs control gene expression in the single-cell alga Chlamydomonas reinhardtii*. Nature, **447**, 1126–1129.
- [391] B. Zhang, X. Pan, C. H. Cannon, G. P. Cobb und T. A. Anderson (2006) *Conservation and divergence of plant microRNA genes*. Plant J., **46**, 243–259.
- [392] M. J. Axtell, J. A. Snyder und D. P. Bartel (2007) *Common functions for diverse small RNAs of land plants*. Plant Cell, **19**, 1750–1769.
- [393] R. Sunkar und G. Jagadeeswaran (2008) *In silico identification of conserved microRNAs in large number of diverse plant species*. BMC Plant Biol., **8**, 37.
- [394] C. W. Dunn, A. Hejnol, D. Q. Matus, K. Pang, W. E. Browne, S. A. Smith, E. Seaver, G. W. Rouse, M. Obst, G. D. Edgecombe, M. V. Sorensen, S. H. Haddock, A. Schmidt-Rhaesa, A. Okusu, R. M. Kristensen, W. C. Wheeler, M. Q. Martindale und G. Giribet (Apr 2008) *Broad phylogenomic sampling improves resolution of the animal tree of life*. Nature, **452** (7188), 745–749.
- [395] M. Marz, A. Mosig, B. M. R. Stadler und P. F. Stadler (2007) *U7 snRNAs: A Computational Survey*. Geno. Prot. Bioinf., **5**, 187–195.
- [396] M. Dávila López und T. Samuelsson (2008) *Early evolution of histone mRNA 3' end processing*. RNA, **14**, 1–10.
- [397] A. Mosig, M. Guofeng, B. M. R. Stadler und P. F. Stadler (2007) *Evolution of the Vertebrate Y RNA Cluster*. Th. Biosci., **126**, 9–14.
- [398] J. Perreault, J.-P. Perreault und G. Boire (2007) *The Ro associated Y RNAs in metazoans: evolution and diversification*. Mol. Biol. Evol., **24**, 1678–1689.
- [399] P. F. Stadler, J. J.-L. Chen, J. Hackermüller, S. Hoffmann, F. Horn, P. Khaitovich, A. K. Kretschmar, A. Mosig, S. J. Prohaska, X. Qi, K. Schutt und K. Ullmann (2008) *Evolution of Vault RNAs*. Submitted.
- [400] J. L. Chen, M. A. Blasco und C. W. Greider (2000) *Secondary structure of vertebrate telomerase RNA*. Cell, **100**, 503–514.
- [401] D. Blair, G. M. Davis und B. Wu (2001) *Evolutionary relationships between trematodes and snails emphasizing schistosomes and paragonimids.. Parasitology*, **123 Suppl**, S229–S243.
- [402] S. V. Brant und E. S. Loker (Nov 2005) *Can specialized pathogens colonize distantly related hosts? Schistosome evolution as a case study.. PLoS Pathog*, **1** (3), 167–169.

- [403] B. L. Webster, V. R. Southgate und D. T. J. Littlewood (Jul 2006) *A revision of the interrelationships of Schistosoma including the recently described Schistosoma guineensis*. Int J Parasitol, **36** (8), 947–955.
- [404] E. Jiménez-Guri, H. Philippe, B. Okamura und P. W. H. Holland (2007) *Buddenbrockia is a cnidarian worm*. Science, **317**, 116–118.
- [405] R. A. Wilson, P. D. Ashton, S. Braschi, G. P. Dillon, M. Berriman und A. Ivens (2007) *'Oming in on schistosomes: prospects and limitations for post-genomics*. Trends Parasitol, **23**, 14–20.
- [406] H. Hirai, T. Taguchi, Y. Saitoh, M. Kawanaka, H. Sugiyama, S. Habe, M. Okamoto, M. Hirata, M. Shimada, W. U. Tiu, K. Lai, E. S. Upatham und T. Agatsuma (2000) *Chromosomal differentiation of the Schistosoma japonicum complex*. Int J Parasitol, **30**, 441–452.
- [407] W. Hu, Q. Yan, D. K. Shen, F. Liu, Z. D. Zhu, H. D. Song, X. R. Xu, Z. J. Wang, Y. P. Rong, L. C. Zeng, J. Wu, X. Zhang, J. J. Wang, X. N. Xu, S. Y. Wang, G. Fu, X. L. Zhang, Z. Q. Wang, P. J. Brindley, D. P. McManus, C. L. Xue, Z. Feng, Z. Chen und Z. G. Han (2003) *Evolutionary and biomedical implications of a Schistosoma japonicum complementary DNA resource*. Nat Genet., **35**, 139–147.
- [408] S. Verjovski-Almeida, D. R. E. A. Martins, P. E. Guimarães, E. P. Ojopi, A. C. Paquola, J. P. Piazza, M. Y. Nishiyama Jr, J. P. Kitajima, R. E. Adamson, P. D. Ashton, M. F. Bonaldo, P. S. Coulson, G. P. Dillon, L. P. Farias, S. P. Gregorio, P. L. Ho, R. A. Leite, L. C. Malaquias, R. C. Marques, P. A. Miyasato, A. L. Nascimento, F. P. Ohlweiler, E. M. Reis, M. A. Ribeiro, R. G. Sá, G. C. Stukart, M. B. Soares, C. Gargioni, T. Kawano, V. Rodrigues, A. M. Madeira, R. A. Wilson, C. F. Menck, J. C. Setubal, L. C. Leite und E. Dias-Neto (2003) *Transcriptome analysis of the acoelomate human parasite Schistosoma mansoni*. Nat. Genet., **35**, 148–157.
- [409] A. Schulmeister, O. Heyers, M. E. Morales, P. J. Brindley, R. Lucius, G. Meusel und B. H. Kalinna (2005) *Organization and functional analysis of the Schistosoma mansoni cathepsin D-like aspartic protease gene promoter*. Biochim Biophys Acta, **1727**, 27–34.
- [410] C. S. Copeland, V. H. Mann und P. J. Brindley (2007) *Both sense and antisense strands of the LTR of the Schistosoma mansoni Pao-like retrotransposon Sinbad drive luciferase expression*. Mol. Genet. Genomics, **277**, 161–170.
- [411] B. Brejová, T. Vinař, Y. Chen, S. Wang, G. Zhou, D. G. Brown, M. Li und Y. Zhou (2009) *Finding genes in Schistosoma japonicum: annotating novel genomes with help of extrinsic evidence*. Nucleic Acids Res., **37**, e52.
- [412] G. Ferbeyre, J. M. Smith und R. Cedergren (1998) *Schistosome satellite DNA encodes active hammerhead ribozymes*. Mol Cell Biol, **18**, 3880–3888.
- [413] T. Laha, D. P. McManus, A. Loukas und P. J. Brindley (2000) *Sja elements, short interspersed element-like retrotransposons bearing a hammerhead ribozyme motif from the genome of the oriental blood fluke Schistosoma japonicum*. Biochim Biophys Acta, **1492**, 477–482.
- [414] C. S. Copeland, O. Heyers, B. H. Kalinna, A. Bachmair, P. F. Stadler, I. L. Hofacker und P. J. Brindley (2004) *Structural and evolutionary analysis of the transcribed sequence of Boudicca, a Schistosoma mansoni retrotransposon*. Gene, **329**, 103–114.
- [415] D. Rollinson, A. Kaukas, D. A. Johnston, A. J. Simpson und M. Tanaka (1997) *Some molecular insights into schistosome evolution*. Int J Parasitol, **27**, 11–28.
- [416] D. T. Littlewood, A. E. Lockyer, B. L. Webster, D. A. Johnston und T. H. Le (2006) *The complete mitochondrial genomes of Schistosoma haematobium and Schistosoma spindale and the evolutionary history of mitochondrial genome changes among parasitic flatworms*. Mol Phylogenet Evol, **39**, 452–467.
- [417] E. Kim, T. A. Day, J. L. Bennett und R. A. Pax (1995) *Cloning and functional expression of a Shaker-related voltage-gated potassium channel gene from Schistosoma mansoni (Trematoda: Digenea)*. Parasitology, **110**, 171–180.
- [418] A. J. Simpson, J. B. Dame, F. A. Lewis und T. F. McCutchan (1984) *The arrangement of ribosomal RNA genes in Schistosoma mansoni. Identification of polymorphic structural variants*. Eur J Biochem, **139**, 41–45.
- [419] K. Sheppard, P. M. Akochy und D. Söll (2008) *Assays for transfer RNA-dependent amino acid biosynthesis*. Methods, **44**, 139–145.
- [420] A. Ambrogelly, S. Palioura und D. Söll (2007) *Natural expansion of the genetic code*. Nat Chem Biol, **3**, 29–35.
- [421] N. Hubert, R. Walczak, C. Sturchler, E. Myslinski, C. Schuster, E. Westhof, P. Carbon und A. Krol (1996) *RNAs mediating cotranslational insertion of selenocysteine in eukaryotic selenoproteins*. Biochimie, **78**, 590–596.
- [422] H. van Keulen, P. T. Loverde, L. A. Bobek und D. M. Rekosch (1985) *Organization of the ribosomal RNA genes in Schistosoma mansoni*. Mol Biochem Parasitol, **15**, 215–230.
- [423] K. Scheibye-Alsing, S. Hoffmann, A. M. Frankel, P. Jensen, P. F. Stadler, Y. Mang, N. Tommerup, M. J. Gilchrist, A.-B. N. Hillig, S. Cirera, C. B. Jørgensen, M. Fredholm und J. Gorodkin (2009) *Sequence Assembly*. Comp. Biol. Chem., **33**, 121–136.
- [424] J. P. Staley und J. L. Woolford Jr. (2009) *Assembly of ribosomes and spliceosomes: complex ribonucleoprotein machines*. Curr Opin Cell Biol., **21**, 109–118.

- [425] M. Marz, N. Vanzo und P. F. Stadler (2009) *Carnival of SL RNAs: Structural variants and the possibility of a common origin*. RNA. Submitted.
- [426] A. McNair, K. Zemzoumi, H. Lütke, C. Guillermin, A. Boitelle, A. Capron und C. Dissous (1995) *Cloning of a signal-recognition-particle subunit of Schistosoma mansoni*. Parasitol Res, **81**, 175–177.
- [427] L. A. Kirsebom (2007) *RNase P RNA mediated cleavage: substrate recognition and catalysis*. Biochimie, **89**, 1183–1194.
- [428] E. Kikowska, S. G. Svård und L. A. Kirsebom (2007) *Eukaryotic RNase P RNA mediates cleavage in the absence of protein*. Proc. Natl. Acad. Sci. USA, **104**, 2062–2067.
- [429] G. Krautz-Peterson und P. J. Skelly (2008) *Schistosoma mansoni: the dicer gene and its expression*. Exp. Parasitol., **118**, 122–128.
- [430] M. S. Gomes, F. J. Cabral, L. K. Jannotti-Passos, O. Carvalho, V. Rodrigues, E. H. Baba und R. G. Sá (2009) *Preliminary analysis of miRNA pathway in Schistosoma mansoni*. Parasitol Int., **58**, 61–68.
- [431] F. Liu, J. Lu, W. Hu, S. Y. Wang, S. J. Cui, M. Chi, Q. Yan, X. R. Wang, H. D. Song, X. N. Xu, J. J. Wang, X. L. Zhang, X. Zhang, Z. Q. Wang, C. L. Xue, P. J. Brindley, D. P. McManus, P. Y. Yang, Z. Feng, Z. Chen und Z. G. Han (2006) *New perspectives on host-parasite interplay by comparative transcriptomic and proteomic analyses of Schistosoma japonicum*. PLoS Pathog, **2**, e29.
- [432] X. Xue, J. Sun, Q. Zhang, Z. Wang, Y. Huang und W. Pan (2008) *Identification and characterization of novel microRNAs from Schistosoma japonicum*. PLoS ONE, **3**, e4034.
- [433] A. G. Matera, R. Terns und Terns (2007) *Non-coding RNAs: lessons from the small nuclear and small nucleolar RNAs*. Nat. Rev. Mol. Cell Biol., **8**, 209–220.
- [434] G. Dieci, M. Preti und B. Montanini (2009) *Eukaryotic snoRNAs: A paradigm for gene expression flexibility*. Genomics.
- [435] H. Hirai und P. T. LoVerde (1996) *Identification of the telomeres on Schistosoma mansoni chromosomes by FISH*. J. Parasitol., **82**, 511–512.
- [436] C. A. Theimer und J. Feigon (2006) *Structure and function of telomerase RNA*. Curr Opin Struct Biol, **16**, 307–318.
- [437] C. Copeland, M. Marz, D. Rose, J. Hertel, P. Brindley, C. Santana, C. Attolini und S. PF (2009) *Non-coding RNA Annotation of the Schistosoma mansoni Genome*.
- [438] N. Lartillot, H. Brinkmann und H. Philippe (2007) *Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model*. BMC Evolutionary Biology, **7**, S4.
- [439] B. A. Gaëta, S. J. Sharp und T. S. Stewart (Mar 1990) *Saturation mutagenesis of the Drosophila tRNA(Arg) gene B-Box intragenic promoter element: requirements for transcription activation and stable complex formation*. Nucleic Acids Res, **18** (6), 1541–1548.
- [440] K. Rutherford, J. Parkhill, J. Crook, T. Horsnell, P. Rice, M. A. Rajandream und B. Barrell (Oct 2000) *Artemis: sequence visualization and annotation*. Bioinformatics, **16** (10), 944–945.
- [441] R. M. Kuhn, D. Karolchik, A. S. Zweig, T. Wang, K. E. Smith, K. R. Rosenbloom, B. Rhead, B. J. Raney, A. Pohl, M. Pheasant, L. Meyer, F. Hsu, A. S. Hinrichs, R. A. Harte, B. Giardine, P. Fujita, M. Diekhans, T. Dreszer, H. Clawson, G. P. Barber, D. Haussler und W. J. Kent (Jan 2009) *The UCSC Genome Browser Database: update 2009*. Nucleic Acids Res, **37** (Database issue), 755–761.
- [442] S. Ohno (Aug 1993) *A song in praise of peptide palindromes*. Leukemia, **7 Suppl 2**, 157–159.
- [443] S. Ohno (1987) *Repetition as the essence of life on this earth: music and genes*. Haematol Blood Transfus, **31**, 511–518.
- [444] S. Ohno und M. Ohno (1986) *The all pervasive principle of repetitious recurrence governs not only coding sequence construction but also human endeavor in musical composition*. Immunogenetics, **24** (2), 71–78.
- [445] M. D. Hansen, E. Chapp, S. Lodha, D. Meads und A. Pang (1999) *PROMUSE: a system for multi-media data presentation of protein structural alignments*. Pac Symp Biocomput, 368–379.
- [446] T. Hermann, P. Meinicke und H. Ritter (2000) *Principal Curve Sonification*. in *Proceedings of the Int. Conf. on Auditory Display*, 81–86.
- [447] R. Takahashi und J. H. Miller (2007) *Conversion of amino-acid sequence in proteins to classical music: search for auditory patterns*. Genome Biol, **8** (5), 405–405.
- [448] S. K. Lodha, D. Whitmore, M. Hansen und E. Chapp (2000) *Analysis and user evaluation of a musical-visual system: Does music make any difference*. in *Proceedings of the Int. Conf. on Auditory Displays*, 167–172.
- [449] K. Hayashi und N. Munakata (Jul 1984) *Basically musical*. Nature, **310** (5973), 96–96.

- [450] P. Gena and C. Strom (1995) *Musical synthesis of DNA sequences*. in *XI Colloquio di Informatica Musicale*, 203–204. Bologna, I.
- [451] P. Gena and C. Strom (2001) *A physiological approach to DNA music*. in *Proceedings of CADE 2001*, 81–86. Glasgow School of Art Press, Glasgow, UK.
- [452] J. Dunn and M. A. Clark (1999) *Life Music: The Sonification of Proteins*. Leonardo, **32** (1), 25–32.
- [453] T. Hermann und H. Ritter (10 2005) *Crystallization sonification of high-dimensional datasets*. ACM Trans. Applied Perception, **2** (4), 550–558.
- [454] H. Taube. *Common Music Website* - <http://commonmusic.sourceforge.net/doc/cm.html>.
- [455] S. Kawai. *Gauche Scheme* - <http://practical-scheme.net/gauche/index.html>.
- [456] HASKELL Community. *Common Music Website* - <http://www.haskell.org/>.
- [457] D. . G. Consortium (Nov 2007) *Evolution of genes and genomes on the Drosophila phylogeny*. Nature, **450** (7167), 203–218.
- [458] P. A. Maroney, J. A. Denker, E. Darzynkiewicz, R. Laneve und T. W. Nilsen (Sep 1995) *Most mRNAs in the nematode Ascaris lumbricoides are trans-spliced: a role for spliced leader addition in translational efficiency*. RNA, **1** (7), 714–723.
- [459] R. S. Dassanayake, N. V. Chandrasekharan und E. H. Karunanayake (May 2001) *Trans-spliced leader RNA, 5S-rRNA genes and novel variant orphan spliced-leader of the lymphatic filarial nematode Wuchereria bancrofti, and a sensitive polymerase chain reaction based detection assay*. Gene, **269** (1-2), 185–193.
- [460] D. L. Redmond und D. P. Knox (Sep 2001) *Haemonchus contortus SL2 trans-spliced RNA leader sequence*. Mol Biochem Parasitol, **117** (1), 107–110.
- [461] K. Z. Lee und R. J. Sommer (Dec 2003) *Operon structure and trans-splicing in the nematode Pristionchus pacificus*. Mol Biol Evol, **20** (12), 2097–2103.
- [462] R. E. Davis, H. Singh, C. Botka, C. Hardwick, M. Ashraf el Meanawy und J. Villanueva (Aug 1994) *RNA trans-splicing in Fasciola hepatica. Identification of a spliced leader (SL) RNA and SL sequences on mRNAs*. J Biol Chem, **269** (31), 20026–20030.
- [463] K. Brehm, K. Jensen und M. Frosch (Dec 2000) *mRNA trans-splicing in the human parasitic cestode Echinococcus multilocularis*. J Biol Chem, **275** (49), 38311–38318.
- [464] R. M. Zayas, T. D. Bold und P. A. Newmark (Oct 2005) *Spliced-leader trans-splicing in freshwater planarians*. Mol Biol Evol, **22** (10), 2048–2054.
- [465] S. I. Miller, S. M. Landfear und D. F. Wirth (Sep 1986) *Cloning and characterization of a Leishmania gene encoding a RNA spliced leader sequence*. Nucleic Acids Res, **14** (18), 7341–7360.
- [466] M. L. Muhich, D. E. Hughes, A. M. Simpson und L. Simpson (Apr 1987) *The monogenetic kinetoplastid protozoan, Crithidia fasciculata, contains a transcriptionally active, multicopy mini-exon sequence*. Nucleic Acids Res, **15** (7), 3141–3153.
- [467] D. A. Campbell (Feb 1992) *Bodo caudatus medRNA and 5S rRNA genes: tandem arrangement and phylogenetic analyses*. Biochem Biophys Res Commun, **182** (3), 1053–1058.
- [468] G. Zieve und S. Penman (1976) *Small RNA species of the HeLa cell: metabolism and subcellular localization*. Cell, **8**, 19–31.
- [469] S. Murphy, F. Altruda, E. Ullu, M. Tripodi, L. Silengo und M. Melli (1984) *DNA sequences complementary to human 7 SK RNA show structural similarities to the short mobile elements of the mammalian genome*. J. Mol. Biol., **177**, 575–590.
- [470] I. S. Moon und M. O. Krause (1991) *Common RNA polymerase I, II, and III upstream elements in mouse 7SK gene locus revealed by the inverse polymerase chain reaction*. DNA Cell Biol., **10**, 23–32.
- [471] R. Reddy, D. Henning, C. S. Subrahmanyam und H. Busch (1984) *Primary and secondary structure of 7-3 (K) RNA of Novikoff hepatoma*. J. Biol. Chem., **259**, 12265–12270.
- [472] J. Cook (1966) *Adaptations to temperature in two closely related strains of Euglena gracilis*. Biological Bulletin.
- [473] W. Kusber (1998) *A study on Phacus smulkowskianus (Euglenophyceae) - a rarely reported taxon found in waters of the Botanic Garden Berlin-Dahlem*. Wlldenowia, **28**, 239–247.
- [474] T. L. Webb und D. Francis (Nov 1969) *Mating types in Stentor coeruleus*. J Protozool, **16** (4), 758–763.
- [475] J. Kielhorn und G. Rosner (1996) *Morpholine. First draft*. INTERNATIONAL PROGRAMME ON CHEMICAL SAFETY.
- [476] J. P. Bruzik, K. Van Doren, D. Hirsh und J. A. Steitz (Oct 1988) *Trans splicing involves a novel form of small nuclear ribonucleoprotein particles*. Nature, **335** (6190), 559–562.

- [477] A. Manaia, M. de Souza, E. Lustosa und R. I (1981) *Leptomonas lactosovorans* n.sp., a *Lactose-Utilizing Trypanosomatid: Description and Nutritional Requirements*. J.Protozool, **28** (1), 124–126.
- [478] J. da Silva und R. I (1982) *Effect of Temperature and Osmolarity on Growth of Crithidia fasciculata, C. hutneri, C. luciliae thermophila, and Herpetomonas samuelpeessoai*. J.Protozool, **29** (2), 269–272.
- [479] J. D. Berman und F. A. Neva (Mar 1981) *Effect of temperature on multiplication of Leishmania amastigotes within human monocyte-derived macrophages in vitro*. Am J Trop Med Hyg, **30** (2), 318–321.
- [480] D. L. LEHMANN (Aug 1962) *Culture Forms of Trypanosoma ranarum (Lankester, 1871). II. Effect of Temperature upon Reproduction and Cyclic Development*. J Protozool, **9**, 325–326.
- [481] M. Gray, B. Wawrik, J. Paul und E. Casper (Sep 2003) *Molecular detection and quantitation of the red tide dinoflagellate Karenia brevis in the marine environment*. Appl Environ Microbiol, **69** (9), 5726–5730.
- [482] H. D. Park, N. E. Sharpless und A. B. Ortmeyer (Dec 1965) *Growth and differentiation in Hydra. I. The effect of temperature on sexual differentiation in Hydra littoralis*. J Exp Zool, **160** (3), 247–254.
- [483] C. Ricci, M. Caprioli und D. Fontaneto (2007) *Stress and fitness in parthenogens: is dormancy a key feature for bdelloid rotifers?*. BMC Evol Biol, **7 Suppl 2**.
- [484] L. Lebedeva und T. Gerasimova (1985) *Peculiarities of Philodina roseola growth and reproduction under various temperature conditions*. Int. Revue ges. Hydrobiol., **70(4)**, 509–525.
- [485] W. A. Van Voorhies und S. Ward (Sep 1999) *Genetic and environmental conditions that increase longevity in Caenorhabditis elegans decrease metabolic rate*. Proc Natl Acad Sci U S A, **96** (20), 11399–11403.
- [486] J. Martinez, J. Perez-Serrano, W. E. Bernadina und F. Rodriguez-Caabeiro (Dec 2002) *Oxidative, heat and anthelmintic stress responses in four species of Trichinella: comparative study*. J Exp Zool, **293** (7), 664–674.
- [487] A. Pires da Silva (2005) *Pristionchus pacificus genetic protocols*. WormBook, 1–8.
- [488] L. F. Lejambre und J. H. Whitlock (May 1973) *Optimum temperature for egg development of phenotypes in Haemonchus contortus cayugensis as determined by Arrhenius diagrams and Sacher's entropy function*. Int J Parasitol, **3** (3), 299–310.
- [489] N. J. Lwambo, E. S. Upatham, M. Kruatrachue und V. Viyanant (Jun 1987) *The host-parasite relationship between the Saudi Arabian Schistosoma mansoni and its intermediate and definitive hosts. 2. Effects of temperature, salinity and pH on the infection of mice by S. mansoni cercariae*. Southeast Asian J Trop Med Public Health, **18** (2), 166–170.
- [490] J. R. Claxton, J. Sutherst, P. Ortiz und M. J. Clarkson (Mar 1999) *The effect of cyclic temperatures on the growth of Fasciola hepatica and Lymnaea viatrix*. Vet J, **157** (2), 166–171.
- [491] D. Palakodeti, M. Smielewska und B. R. Graveley (Sep 2006) *MicroRNAs from the Planarian Schmidtea mediterranea: a model system for stem cell biology*. RNA, **12** (9), 1640–1649.
- [492] M. Novak (Apr 1983) *Growth of Echinococcus multilocularis in gerbils exposed to different environmental temperature*. Experientia, **39** (4), 414–414.
- [493] J. Petersen und H. Riisgard (1992) *Filtration capacity of the ascidian Ciona intestinalis and its grazing impact in a shallow fjord*. Mar.Ecol.Prog.Ser, **88**, 9–17.
- [494] F. Broms und P. Tiselius (2003) *Effects of temperature and body size on the clearance rate of Oikopleura dioica*. J.Plan.Res., **25(5)**, 573–577.



## Selbstständigkeitserklärung

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

.....  
Manuela Marz

Leipzig, den 6. Juli 2009



## Curriculum Vitae

### Personal Details

---

Name	Manuela Marz (b. Lindemeyer)
Snail	Elsterblick 20, 04159 Leipzig
Email	manja@bioinf.uni-leipzig.de
Birthday and place	06.05.1981, Leipzig, Germany
Nationality	German
Family status	Married with Dr. Michael Marz One son: Ferdinand Marz, born 24.06.2006

### Education

---

1987-1992	10-klassige Polytechnische Oberschule
1992-1999	Geschwister-Scholl-Gymnasium Taucha
07/1999	Abitur at Geschwister-Scholl-Gymnasium

### Scientific Education

---

1999-2005	9 terms Study of Biology at the University of Leipzig
2001-2006	7 terms Study of Computer Science at the University of Leipzig (Specialism: Bioinformatics)
2002-2003	Study of Biology at University of Edinburgh
2002-2003	Study of Computer Science at University of Edinburgh
2004	Study of Biology at TU Darmstadt
02/2005	Certificate: Diploma Biology (Grade: 1.7) Diploma Thesis: Arbeiten zur evolutiven Optimierung des HI-Virus: Erzeugung, funktionelle Bewertung und Sequenzierung von Enzymvarianten (Grade: 2.3)
02/2006	Certificate: Diploma Computer Science (Grade: 1.2) Diploma thesis: Evolution of Spliceosomal RNAs in Metazoan Animals (Grade: 1.0)
12/2005-01/2006	Research Assistant at the University of Leipzig
02/2006-12/2008	PhD Scholarship from "Graduiertenkolleg Wissensrepräsentation" at Universität of Leipzig in cooperation with "Deutsche Forschungs-Gemeinschaft".
01/2009-03/2009 since 04/2009	Research Assistant at the University of Leipzig PhD Scholarship from "Landestipendium Sachsen" at Universität of Leipzig.

## Practical Work

---

08/2003 - 09/2003	Student Assistant at Computer Science Department, University of Leipzig
10/2002 - 12/2002	Student Assistant at Biology Department, University of Leipzig
05/2002 - 06/2002	Student Assistant at Biology Department, University of Leipzig
08/2001 - 09/2001	Student Assistant at Computer Science Department, University of Leipzig
01/2001 - 03/2001	Student Assistant at Biology Department, University of Leipzig
10/2000 - 12/2000	Student Assistant at Biology Department, University of Leipzig

## Publications Manuela Marz

---

(Joint first authorships are marked by asterisks.)

**Marz M**, Donath A, Stadler PF, Bensaude O, *Evolution of 7SK RNA and its Protein Partners in Metazoa*, submitted

Ingalls T, Martius G, Hellmuth M, **Marz M**, Prohaska SJ *Converting DNA to Music: ComposAlign*, accepted for GCB 2009

**Marz M**, Stadler PF, *Phylogentic range of U3 snoRNA in eukaryots*, submitted

Copeland CS\*, **Marz M\***, Rose D\*, Hertel J\*, Brindley PJ, Santana CB, Kehr S, Attolini CSO, Stadler PF, *Non-coding RNA Annotation of the Schistosoma mansoni Genome*, submitted

**Marz M**, Vanzo N, Stadler PF, *Temperature-Dependent Structural Variability of RNAs: Spliced Leader RNAs and their Evolutionary History*, resubmitted

Hiller M, Findeiss S, Lein S, **Marz M**, Nickel C, Rose D, Schulz C, Backofen R, Prohaska SJ, Reuter G, Stadler PF, *Conserved introns reveal novel transcripts in Drosophila melanogaster*, Genome Res (2009), , **19**, 1290–1300; DOI 10.1101/gr.090050.108

Hertel J, de Jong D, **Marz M**, Rose D, Tafer H, Tanzer A, Schierwater B, Stadler PF, *Non-Coding RNA Annotation of the Genome of Trichoplax adhaerens*, Nucleic Acids Res (2009), **37**, 1602–1615, DOI 10.1093/nar/gkn1084

Jones TA\*, Otto W\*, **Marz M\***, Eddy SR, and Stadler PF, *A Survey of Nematode SmY RNAs*, RNA Biology (2009),**6**, 5–8

**Marz M**, Kirsten T, Stadler PF, *Evolution of Spliceosomal snRNA Genes in Metazoan Animals*, J.Mol.Evol. (2008),**67**, 594–607, DOI 10.1007/s00239-008-9149-6

Donath A, Findeiß S, Hertel J, **Marz M**, Otto W, Schulz C, Stadler PF, Wirth S, *Non-Coding RNAs*, Evolutionary Genomics, Caetano-Anolles, Gustavo, Wiley, 2008, in press

**Marz M**, Mosig A, Stadler BM, Stadler PF., *U7 snRNAs: a computational survey.*, J Mol Evol. (2008), **66**:107–115, DOI 0.1007/s00239-007-9052-6

Gruber AR\*, Koper-Emde D\*, **Marz M\***, Tafer H\*, Bernhart S, Obernosterer G, Mosig A, Hofacker IL, Stadler PF, Benecke BJ., *Invertebrate 7SK snRNAs.*, J Mol Evol. (2008), **66**, 107–115, DOI 10.1007/s00239-007-9052-6

ENCODE Project Consortium, *Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project.*, Nature (2007), **447**, 799–816, doi:10.1038/nature05874

Washietl S, Pedersen JS, Korb J, Stocsits C, Gruber AR, Hackermüller J, Hertel J, **Lindemeyer M**, Reiche K, Tanzer A, Ucla C, Wyss C, Antonarakis SE, Denoeud F, Lagarde J, Drenkow J, Kapranov P, Gingeras TR, Guigo R, Snyder M, Gerstein MB, Reymond A, Hofacker IL, Stadler PF., *Structured RNAs in the ENCODE selected regions of the human genome.*, Genome Res. (2007), **17**, 852–864, DOI 10.1101/gr.5650707

Hertel J, **Lindemeyer M**, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF, and the students of the bioinformatics computer labs 2004 and 2005, *The Expansion of the Metazoan microRNA repertoire*, BMC Genomics. (2006), **15**, 25, 10.1186/1471-2164-7-25

**Lindemeyer M**, *Evolution of snRNAs*, diploma thesis in computer science (2006), University of Leipzig

**Lindemeyer M**, *Arbeiten zur evolutiven Optimierung des HI-Virus*, German, diploma thesis in biology (2005), University of Leipzig

The authors of **The ENCODE Consortium** are:

Birney, E. and Stamatoyannopoulos, J. A. and Dutta, A. and Guigo, R. and Gingeras, T. R. and Mar-

gules, E. H. and Weng, Z. and Snyder, M. and Dermitzakis, E. T. and Thurman, R. E. and Kuehn, M. S. and Taylor, C. M. and Neph, S. and Koch, C. M. and Asthana, S. and Malhotra, A. and Adzhubei, I. and Greenbaum, J. A. and Andrews, R. M. and Flicek, P. and Boyle, P. J. and Cao, H. and Carter, N. P. and Clelland, G. K. and Davis, S. and Day, N. and Dhami, P. and Dillon, S. C. and Dorschner, M. O. and Fiegler, H. and Giresi, P. G. and Goldy, J. and Hawrylycz, M. and Haydock, A. and Humbert, R. and James, K. D. and Johnson, B. E. and Johnson, E. M. and Frum, T. T. and Rosenzweig, E. R. and Karnani, N. and Lee, K. and Lefebvre, G. C. and Navas, P. A. and Neri, F. and Parker, S. C. and Sabo, P. J. and Sandstrom, R. and Shafer, A. and Vetric, D. and Weaver, M. and Wilcox, S. and Yu, M. and Collins, F. S. and Dekker, J. and Lieb, J. D. and Tullius, T. D. and Crawford, G. E. and Sunyaev, S. and Noble, W. S. and Dunham, I. and Denoeud, F. and Reymond, A. and Kapranov, P. and Rozowsky, J. and Zheng, D. and Castelo, R. and Frankish, A. and Harrow, J. and Ghosh, S. and Sandelin, A. and Hofacker, I. L. and Baertsch, R. and Keefe, D. and Dike, S. and Cheng, J. and Hirsch, H. A. and Sekinger, E. A. and Lagarde, J. and Abril, J. F. and Shahab, A. and Flamm, C. and Fried, C. and Hackermuller, J. and Hertel, J. and **Lindemeyer, M.** and Missal, K. and Tanzer, A. and Washietl, S. and Korbel, J. and Emanuelsson, O. and Pedersen, J. S. and Holroyd, N. and Taylor, R. and Swarbreck, D. and Matthews, N. and Dickson, M. C. and Thomas, D. J. and Weirauch, M. T. and Gilbert, J. and Drenkow, J. and Bell, I. and Zhao, X. and Srinivasan, K. G. and Sung, W. K. and Ooi, H. S. and Chiu, K. P. and Foissac, S. and Alioto, T. and Brent, M. and Pachter, L. and Tress, M. L. and Valencia, A. and Choo, S. W. and Choo, C. Y. and Ucla, C. and Manzano, C. and Wyss, C. and Cheung, E. and Clark, T. G. and Brown, J. B. and Ganesh, M. and Patel, S. and Tammanna, H. and Chrast, J. and Henrichsen, C. N. and Kai, C. and Kawai, J. and Nagalakshmi, U. and Wu, J. and Lian, Z. and Lian, J. and Newburger, P. and Zhang, X. and Bickel, P. and Mattick, J. S. and Carninci, P. and Hayashizaki, Y. and Weissman, S. and Hubbard, T. and Myers, R. M. and Rogers, J. and Stadler, P. F. and Lowe, T. M. and Wei, C. L. and Ruan, Y. and Struhl, K. and Gerstein, M. and Antonarakis, S. E. and Fu, Y. and Green, E. D. and Karaoz, U. and Siepel, A. and Taylor, J. and Liefer, L. A. and Wetterstrand, K. A. and Good, P. J. and Feingold, E. A. and Guyer, M. S. and Cooper, G. M. and Asimenos, G. and Dewey, C. N. and Hou, M. and Nikolaev, S. and Montoya-Burgos, J. I. and Loytynoja, A. and Whelan, S. and Pardi, F. and Massingham, T. and Huang, H. and Zhang, N. R. and Holmes, I. and Mullikin, J. C. and Ureta-Vidal, A. and Paten, B. and Seringhaus, M. and Church, D. and Rosenbloom, K. and Kent, W. J. and Stone, E. A. and Batzoglu, S. and Goldman, N. and Hardison, R. C. and Haussler, D. and Miller, W. and Sidow, A. and Trinklein, N. D. and Zhang, Z. D. and Barrera, L. and Stuart, R. and King, D. C. and Ameur, A. and Enroth, S. and Bieda, M. C. and Kim, J. and Bhinge, A. A. and Jiang, N. and Liu, J. and Yao, F. and Vega, V. B. and Lee, C. W. and Ng, P. and Shahab, A. and Yang, A. and Moqtaderi, Z. and Zhu, Z. and Xu, X. and Squazzo, S. and Oberley, M. J. and Inman, D. and Singer, M. A. and Richmond, T. A. and Munn, K. J. and Rada-Iglesias, A. and Wallerman, O. and Komorowski, J. and Fowler, J. C. and Couttet, P. and Bruce, A. W. and Dovey, O. M. and Ellis, P. D. and Langford, C. F. and Nix, D. A. and Euskirchen, G. and Hartman, S. and Urban, A. E. and Kraus, P. and Van Calcar, S. and Heintzman, N. and Kim, T. H. and Wang, K. and Qu, C. and Hon, G. and Luna, R. and Glass, C. K. and Rosenfeld, M. G. and Aldred, S. F. and Cooper, S. J. and Halees, A. and Lin, J. M. and Shulha, H. P. and Zhang, X. and Xu, M. and Haidar, J. N. and Yu, Y. and Ruan, Y. and Iyer, V. R. and Green, R. D. and Wadelius, C. and Farnham, P. J. and Ren, B. and Harte, R. A. and Hinrichs, A. S. and Trumbower, H. and Clawson, H. and Hillman-Jackson, J. and Zweig, A. S. and Smith, K. and Thakkapallayil, A. and Barber, G. and Kuhn, R. M. and Karolchik, D. and Armengol, L. and Bird, C. P. and de Bakker, P. I. and Kern, A. D. and Lopez-Bigas, N. and Martin, J. D. and Stranger, B. E. and Woodroffe, A. and Davydov, E. and Dimas, A. and Eyraes, E. and Hallgrimsdottir, I. B. and Huppert, J. and Zody, M. C. and Abecasis, G. R. and Estivill, X. and Bouffard, G. G. and Guan, X. and Hansen, N. F. and Idol, J. R. and Maduro, V. V. and Maskeri, B. and McDowell, J. C. and Park, M. and Thomas, P. J. and Young, A. C. and Blakesley, R. W. and Muzny, D. M. and Sodergren, E. and Wheeler, D. A. and Worley, K. C. and Jiang, H. and Weinstock, G. M. and Gibbs, R. A. and Graves, T. and Fulton, R. and Mardis, E. R. and Wilson, R. K. and Clamp, M. and Cuff, J. and Gnerre, S. and Jaffe, D. B. and Chang, J. L. and Lindblad-Toh, K. and Lander, E. S. and Koriabine, M. and Nefedov, M. and Osoegawa, K. and Yoshinaga, Y. and Zhu, B. and de Jong, P. J.

