Washington University in St. Louis
Washington University Open Scholarship

All Theses and Dissertations (ETDs)

1-1-2011

Systematic Identification of Independent Functional Non-coding RNA Genes in Oxytricha trifallax

Seolkyoung Jung
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Division of Biology and Biomedical Sciences

Computational Biology

Dissertation Examination Committee:
Sean R. Eddy, Co-Chair
Barak A. Cohen, Co-Chair
Douglas Chalker
Susan K. Dutcher
Justin C. Fay
Kathleen B. Hall
Gary D. Stormo

Systematic identification of independent functional non-coding RNA genes in *Oxytricha trifallax*

by

Seolkyoung Jung

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfilment of the
requirements for the degree
of Doctor of Philosophy

December 2011

Saint Louis, Missouri

ABSTRACT OF THE DISSERTATION

Systematic Identification of

Independent Functional Non-coding RNA Genes

in *Oxytricha trifallax*

by

Seolkyoung Jung

Doctor of Philosophy in Biology and Biomedical Sciences

(Computational Biology)

Washington University in St. Louis, 2011

Sean R. Eddy and Barak A. Cohen, Co-Chairmen

Functional noncoding RNAs participate in a variety of biological processes: for example, modulating translation, catalyzing biochemical reactions, sensing environments etc. Independent of conventional approaches such as transcriptomics and computational comparative analysis, we took advantage of the unusual genomic organization of the ciliated unicellular protozoan *Oxytricha trifallax* to screen for eukaryotic independent functional noncoding RNA genes. The *Oxytricha* macronuclear genome consists of thousands of gene-sized “nanochromosomes”, each of which usually contains only a single gene. Using a draft *Oxytricha* genome assembly and a custom-written noncoding nanochromosome classifier, we identified a subset of nanochromosomes that lack any detectable protein-coding gene, thereby strongly enriching for nanochromosomes that carry noncoding RNA genes. Surprisingly, we found only a small proportion of noncoding nanochromosomes, suggesting that *Oxytricha* has few independent functional noncoding RNA genes besides homologs of already known noncoding RNAs. Other than new members of known noncoding RNA classes including C/D and H/ACA box small nucleolar RNAs, our screen identified a single novel family of small RNA genes, named the Arisong RNAs, which share some of the features of small nuclear RNAs. The small number of novel independent functional noncoding RNA genes identified in this screen contrasts to numerous recent reports of a large number of noncoding RNAs in a variety of eukaryotes. We think the difficulty of distinguishing functional noncoding RNA genes from other sources of putative noncoding RNAs has been underestimated.

Acknowledgements

I gratefully acknowledge financial support from the Samsung Scholarship Foundation for first four years and the Howard Hughes Medical Institute (HHMI). Sean Eddy's lab in Washington University in St. Louis, where this work took place between 2004 to 2006, was supported by National Institutes of Health (grant number R01-HG01363) and by an endowment from Alvin Goldfarb.

I think it is so fortunate to meet Sean Eddy as a PI, who is nice, thoughtful and also smart which seldom coexist in a person. It was a privilege to work with him and to learn how to design, conduct and evaluate experiments from him. I had a joyful time with the past and present lab members. Feedbacks of my work, paper and presentation from them and discussions with them were very helpful for me to build the "sound" scientific mind. Specially, I'd like to express my thanks to the previous lab member Jenny who taught me bench-work from ABC-of-experiments. My wonderful and warm-heated coordinators Melanie, Mary and Margaret were so supportive for all kind of paper works. I also appreciate scientific computing core facility, specially Goran, and molecular biology core facility in Janelia Farm Research Campus (JFRC) for their technical support on my doctoral work.

Without the collaboration with Laura Landweber's group in Princeton University and Genome sequencing Center in Washington University in St. Louis, I couldn't accomplish my graduate work. Specially, Joey and Estienne in Laura Landweber's lab were very supportive for cooperational work.

Finally, I'd like to thank my parents for their endless and continuous support with love and pray to keep me in faith that I'm under the God's protection and he will lead my life,

so I could do my best during the doctoral training period.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

In a 2009 special review issue of *Cell* called **RNA**, Phillip Sharp stated the following at the end of his introductory essay, *The centrality of RNA*:

> “ *The most surprising aspect of all of this is how late in the study of cell biology the importance and ubiquitous nature of RNA in gene regulation became widely recognized.*” [1]

Even though the potential functionality of noncoding RNAs as regulators and operators in protein synthesis was first presumed in 1961 [2], their functional importance and abundance in various cellular processes had been underappreciated for a long time.

Since ribosomal RNA (rRNA) was identified as the large RNA component of ribosomes in 1955 and alanine transfer RNA (tRNA) was first characterized in 1965 [3], various functionally important noncoding RNAs (ncRNAs) have been studied in a wide range of organisms. For example, small nuclear RNAs (snRNAs) were isolated and first analyzed in 1968 as a single species of “U” RNA having a high content of uridylic acid [4]. Later, it was found that snRNAs transcribed by RNA polymerase II or RNA polymerase III are present in all vertebrates, and that among them, U1, U2 and U6 are highly conserved [5].

Unlike other spliceosomal snRNAs which function in intron splicing, the function of 7SK RNA, an abundant snRNA first discovered in 1976 [6], was relatively recently elucidated as a negative regulator of the transcription elongation factor P-TEFb [7, 8]. Nonetheless, until 2000-2001 or so, protein-coding gene-oriented viewpoints had been dominated in biological research elucidating essential cellular processes. Up to that time, the number of discovered and studied ncRNAs was much less than that of protein-coding genes in a genome. While 3042 "ncRNA" nucleotide sequences (369 human ncRNAs) were deposited between 1986/1/1 and 2000/12/31 (Entrez nucleotide database), 215186 "protein" nucleotide sequences (44634 human mRNAs) were deposited.

However, the realization of the abundance of ncRNAs in cell regulation was stimulated by the discovery of several hundred eukaryotic microRNA (miRNA) genes in various genomes, which began to gain fame together with RNA interference (RNAi) in 2000, together with reports of more than a hundred new small nucleolar RNA (snoRNA) genes [9–12] and riboswitches [13–17] in the same time period. The first miRNA, lin-4 in *Caenorhabditis elegans* was positionally cloned in 1987 after being identified in genetic screens for larval development-related regulatory genes [18]. Six years later, in 1993, the Ambros group characterized its product as a small 21 nt ncRNA that functions as a posttranscriptional regulator (renamed later as miRNA) [19]. Seven years later, another miRNA, let-7, was discovered which also encodes a small $\sim$ 21 nt RNA with partial complementarity to the $3'$ untranslated region of target mRNAs [20]. In the next year, its conservation was revealed in an astonishingly wide range of organisms including fly, fish, mouse and human (but not in bacteria, yeast, sponge or plant) [21]. This suggested that lin-4 and let-7 were not nematode-specific oddities, and led to finding hundreds of new instances of miRNA genes in various other genomes from plants to human [22, 23] and also triggered a race to find novel ncRNA genes by systematic approaches in a variety of genomes [24–28].

In 2002, a key paper from the FANTOM Consortium of the RIKEN Mouse Gene Encyclopaedia Project was published, in which the group claimed to have discovered over 10,000 novel ncRNA candidates by characterizing new cDNA clones with the previous FANTOM I clone dataset [29]. After clustering cDNA clones into transcription units, representative cDNA clones from each cluster were examined for protein-coding potential by matching to known mouse DNA and protein sequences. Among transcription units which were not assigned some functional information, clones with computational predicted coding sequences (CDSs) of less than 100 amino acids (aa) were annotated as noncoding messages. Coupled with dramatic advances in sequencing capacity and the advent of microarray technology, subsequent genomics and transcriptomics approaches have reported more than tens of thousands of ncRNAs in a wide variety of species [30–35]. These reports have led some to hypothesize that regulatory networks by ncRNAs might explain most of the complexity of higher eukaryotic organisms [36]. However, whether these newly emerging RNA species are functional and whether they are even truly noncoding remains controversial [37] as I will discuss shortly in more detail, so a careful clarification on ncRNAs and reexamination on noncoding transcripts are required to distinguish meaningful functional ncRNAs among the collection of noise-prone transcription events in a genome.

1.1 Noncoding RNAs

The term “noncoding” seems to have been first used for tRNA and rRNA genes to contrast them with the coding RNA components of the central dogma, messenger RNAs (mRNAs) which produce proteins [38]. The dictionary definition is “not specifying the genetic code”.

As a generic term, “noncoding” can be used even to describe introns and untranslated regions (UTRs) which are transcribed as parts of a protein-coding mRNA, and there is no need for all such ncRNA to be functional. However, when people refer to the term “noncoding RNA”, they generally mean an RNA transcript that has a specific biological role, other than coding for protein as an mRNA; in other words, we use ncRNA, which is a much broader concept, to indicate only functional ncRNA transcripts. However, to clarify the notion of ncRNA, it is necessary to distinguish functional ncRNAs from other sources of noncoding RNA in a transcriptome.

1.1.1 Functional ncRNAs

Functional ncRNAs can be roughly divided into two groups according to the origin of their functionality: structural ncRNAs and guide ncRNAs. Many well-known ncRNAs adopt a compact tertiary structure and exert their various functions much as proteins do, either by themselves or by interacting with other biomolecules including proteins, other ncRNAs, mRNAs or small molecules. Some structural ncRNAs are components in large ribonucleoproteins such as the signal recognition particle (SRP) RNA, which contributes in binding and releasing of the signal peptide [39, 40]. Other structural ncRNAs are some catalytic RNAs (ribozymes) such as RNase P RNA, which participates in tRNA precursor processing [41], and self-splicing group I introns [42], both of which contributed to the proposed “RNA world hypothesis” [43]. RNase P RNA makes a complex with from one polypeptide chain (bacteria) to up to ten proteins (eukaryotes). Bacterial RNase P RNA still has a catalytic activity without protein subunits, but isolated archaeal or eukaryotic RNase P RNAs do not retain their biochemical activity even though they are functionally essential in holoenzymes. A final example of structural ncRNAs are riboswitches, which are naturally occurring RNA aptamers that sense the concentration of diverse small molecule metabolites including coenzymes, nucleosides, amino acids and an aminosugar through atomic interactions with well-positioned residues in a RNA tertiary structure. This RNA aptamer communicates with an “expression platform” ,a cis-acting genetic control module, to regulate the expression of a target gene. Most riboswitches are widespread only in bacteria [14–16], but the TPP riboswitch has been discovered in plants and certain fungi and predicted in archaea [44].

snoRNAs are a broad class of guide ncRNAs that were first identified by their localization to the nucleolus, where ribosome assembly takes place. snoRNAs have two main classes that have different sequence features, secondary structures and detailed functions: C/D box snoRNAs and H/ACA box snoRNAs. snoRNAs guide site-specific chemical modification, such as methylation (for C/D box snoRNAs) and pseudouridylation (for H/ACA box snoRNAs), or in a few cases, processing (for both) of mainly rRNAs and other RNAs [45] by providing a guide sequence to find their target position by complementary base pairing. Other examples of guide ncRNAs are miRNAs and small Argonaute-bound RNAs such as siRNA (small-interfering RNA) and piRNA (piwi-interacting RNA) which are functionally similar to miRNA genes in that they silence gene expression either by directing mRNA destruction or by inhibiting their translation or both, called RNAi (RNA interference). Incorporated with a RNA-induced silencing complex (RISC), processed ~ 22 nt small RNA from a longer source RNA, for instance, pre-miRNA transcript for miRNA, find specific target RNAs by binding to complementary sequences. siRNAs, like miRNAs, are broadly distributed in both phylogenetic and physiological terms, and associate with the Ago clade protein of Argonaute superfamily for RNA silencing. piRNAs are primarily found in animals, and function most clearly in the germline, where they associate with Piwi clade proteins [46].

There also exist functional ncRNAs for which both secondary structure and specific complementary interactions with primary sequence are critical to their function. For example, one of many known bacterial small regulatory RNAs, the *Escherichia coli* MicF gene, which was recognized in 1984 [47] and whose function in osmoregulatory expression of the OmpF gene was elucidated in 1987 [48], recognizes its target mRNA through complementary sequences residing in the loop regions of a conserved secondary structure.

Not all functional ncRNAs are genes. A lexical definition of a gene is “the basic physical unit of heredity; a linear sequence of nucleotides along a segment of DNA that provides the coded instructions for synthesis of RNA, which, when translated into protein, leads to the expression of hereditary character” [1]. Like many other biological terms, the “gene” is a vague concept. For example, in the 1900s, two terms had been used to indicate an indivisible unit of heredity: the English word “gene” and Johannsen’s German word “gen” which originated from Darwin’s English word ”pangen” [49]. By the 1910s, the idea that a gene is invariant and indivisible like an atom or a simple chemical compound had been refined: “the gene is stable but changes similarly, by definite steps” [50]. In modern biology, the definition has been expanded to include genes encoding functional RNA molecules, regulating operation of other genes or repressing such operation, and to include genes in viruses with RNA genome. Thus we can define a functional ncRNA gene as a physical heredity unit that is transcribed into RNA but not translated into protein, and which has a biological functional role as an RNA – but the term nonetheless is still ambiguous. However, in many cases, we can clearly recognize functional ncRNA “genes” such as the genes encoding

[1] `http://dictionary.reference.com`

RNase P RNA, snoRNAs and miRNAs. Among the above mentioned functional ncRNAs, group I introns and riboswitches are nongenic functional ncRNAs. Both are found as parts of protein-coding genes; group I introns are a special intron that has a self-splicing capability, and riboswitches are special UTRs that have the ability of regulating the expression of a protein-coding gene in cis by detecting environmental changes. There are other instances of such nongenic functional ncRNAs, especially cis-regulatory RNA motifs. The iron response element (IRE), a short conserved stem-loop structural sequence first found in 1987 [51, 52], is bound by iron response proteins (IRPs). IRE functions to either repress downstream iron metabolism-related translation (for IRE in 5′ UTR) or increase upstream mRNA stability (for IRE in 3′ UTR) [53, 54]. Internal ribosome entry sites (IRESs) were first discovered in 1988 in RNA viruses [55, 56] and later found also in mammalian mRNAs [57]. An IRES is an RNA structural motif allowing cap-independent eukaryotic translational initiation, including in the middle of a polycistronic mRNA. Some siRNAs that are not associated with protein-coding genes are also nongenic functional ncRNAs that can be found in a cell. They can be generated from exogenous sources such as long hairpin RNAs or double stranded RNAs derived from the foreign DNAs or RNAs. There are also many siRNAs generated from endogenous sources, such as piRNA, small-scanRNA (scnRNA), trans-acting siRNA (tasiRNA) and repeat-associated siRNA (rasiRNA), where the question of whether these are “genes” or not becomes purely semantic.

Functional ncRNAs may be transcribed from independent loci (“genes”), embedded in other transcripts (cis-acting), or produced from the processing of other longer transcripts. A good example of an independently transcribed functional ncRNA among the above mentioned functional ncRNAs is RNase P RNA. The genomic locus of the yeast 369 nt long RNase P RNA has an independent transcription start site, promoters, and a terminator signal for RNA polymerase III [58]. Most nongenic functional ncRNAs such as group I introns and riboswitches are associated with protein-coding genes. siRNAs arise from processing of other RNA transcripts [59] for genome defense or genome organization unlike miRNAs which are processed from the pre-miRNA products of endogenous pri-miRNA genes. However, some ncRNA genes show more complicated genomic distribution patterns. Most miRNAs are located in intergenic regions often as a cluster, which contain their own promoter and regulatory sequences of RNA polymerase II [22, 60], but intronic miRNAs also have been reported [61–63]. Among functional ncRNA genes, the most well-known representative genes that are associated with protein-coding genes are snoRNAs. But, actually, their genomic locations vary among organisms. The majority of mammalian snoRNAs are located in introns, but many yeast snoRNAs are transcribed independently in monocistronic or polycistronic transcripts [64]. Some mammalian snoRNAs are located in an intronic region of a noncoding host gene that has no protein-coding potential and just has a role as a carrier of snoRNA genes [65, 66]. In archaea, some snoRNAs are located in the $3'$ UTR of protein-coding genes [67]. Furthermore, intronic snoRNAs can mature differently. They are usually processed from excised introns by digesting undesired sequences, but a few snoRNAs are not dependent on splicing events and are endonucleolytically excised from introns [68].

1.1.2 Nonfunctional ncRNAs

It is not clear that every RNA transcript that a cell makes necessarily has a meaningful biological function. Nonfunctional ncRNAs can be transcribed from all other genomic regions aside from coding sequences and functional noncoding sequences. These nonfunctional noncoding sequences include introns, UTRs, pseudogenes, repeat sequences, transposons and integrated viral elements excluding those introns and UTRs that contain functional ncRNAs such as snoRNAs, miRNAs and cis-acting RNA motifs as we described above. These sequences comprise the majority of the genome, especially in mammals; for example, approximately 24% of the human genome is intronic and over 40% is derived from transposons, whereas the exonic coding region of the human genome is less than 2% [69, 70]. Therefore, it is possible that random transcription in these regions generate nonfunctional ncRNAs as "noisy" products. Alternatively, these genomic regions may contain as-yet undiscovered functional elements and the resulting transcripts may be functional ncRNAs.

Nonfunctional ncRNAs can also be generated as side-products of other functional transcriptional events. For example, transient ncRNA transcripts divergent from the adjacent genes can be generated due to the intrinsic bidirectional nature of some (and possibly most) eukaryotic promoters [71].

1.2 Noncoding transcripts

One of the methods used to search for new functional ncRNAs is to identify new apparently noncoding transcripts. However, it is neither easy to segregate noncoding transcripts from coding transcripts, nor to distinguish functional transcripts from nonfunctional transcripts and false positive signals.

1.2.1 Experimental noncoding transcripts detection methods

Experimental approaches for ncRNA gene identification were initiated by isolating highly abundant ncRNA species by size-separation in denaturing gels more than 45 years ago [72]. RNA populations can now be systematically enumerated by high-throughput sequencing or microarray methods, applied to various specialized cDNA libraries. Various approaches have been used to try to enrich cDNA libraries for novel ncRNAs by depleting mRNA and abundant rRNA and tRNA. For example, one approach is size selection. To remove mRNAs typically longer than ncRNA genes' transcripts, small size-selected cDNA libraries have been constructed and sequenced beginning with mouse [25] through other eukaryotes [73–76] and archaeal species [28, 77], and these studies found hundreds of novel ncRNA gene candidates including novel snoRNAs. As an example, lengths of the annotated protein and ncRNAs transcripts in human genome are listed in Table 1.1. The size-selected cDNA library construction approach was further improved by subtraction of rRNA fragments and other unwanted species using magnetic bead-attached complementary oligos [78]. Size selection has also been used in a narrow size range to identify members of specific subclasses of ncRNA genes such as miRNAs [22, 79–82]. Another ncRNA enrichment approach is to use immunoprecipitation of a RNA-binding protein to enrich specific classes of transcript. For example, several novel C/D and H/ACA snoRNA genes have been identified by co-immunoprecipitation with a snoRNA-binding protein such as fibrillarin in *Trypanosoma brucei* [9] or human [83, 84].

One drawback of cDNA sequencing is a non-uniform and biased cloning efficiency across the entire target population of ncRNAs due to their structure, chemical modification, variable abundance and/or tissue- or developmental stage-specificity. Thanks to advances

Table 1.1: Length (nt) of annotated protein and ncRNAs transcripts in the human genome, according to GENCODE v4.

# Gene class		min.	max.	avg.	num.
protein		30	21,723	2,414	20,631
tRNA		36	105	68	730
rRNA		35	161	113.1	531
snRNA		39	230	108.2	1,944
snoRNA		34	420	110.5	1,521
miRNA		47	195	92.5	1,756
lincRNA		232	9,047	2,293	1,451
misc_RNA	all	62	518	154.1	1,187
	telomerase RNA (TERC)	.	.	438	1
	RNaseP RNA	293	333	316	3
	vault RNA	89	103	97.6	9

According to GENCODE v4 transcript annotation available at `http://www.gencodegenes.org/`, each gene_type was retrieved. For tRNA gene_type, only mitochondrial tRNAs, mitochondrial tRNA pseudogenes and tRNA pseudogenes are annotated. For other gene_type, pseudogenes are removed, but some pseudogenes are accidently included because they are annotated as pseudogenes in this Gencode annotation version. For miRNA gene_type, the annotated feature length seems to be not the length of mature miRNA transcripts but that of pre-miRNA transcripts. A majority of gene_type "misc_RNA" is recently described "novel" ncRNA and the next most popular genes are Y RNA and 7SKRNA. The first two columns show the names of gene or gene_type (gene family). Other columns represent the smallest length, largest length, and average length among genes. Last column shows the total number of genes in each gene_type.

in high-throughput next-generation sequencing technologies, recently developed RNA-Seq deep-sequencing technology mitigates some of these problems by increasing sequence coverage. The sensitivity of RNA-seq raises other issues such as reduced specificity by amplifying biological noise, such as partially processed mRNAs, degradation products, or even random transcriptional products by RNA polymerases, and experimental noise, for example, occurring in the manipulation of fragile RNAs [85–88]. Several approaches to reduce such false positives have been attempted, such as RNPomics [89, 90] and dRNA-seq [91] which are RNA-Seq versions of the specialized cDNA library sequencing approach. RN-

Pomics expands target RNA of sequencing from specific subclasses of RNA genes which bind a specific RNA-binding protein through immunoprecipitation into all protein-bound RNAs by size-fractionation, and dRNA-seq selects a specific population of ncRNA transcripts which have an unprocessed 5′ tri-phosphate end on their primary transcripts. The improved specificity of these approaches comes at the expense of reduced sensitivity because only a subpopulation of ncRNAs are sampled. In addition, a more fundamental problem in cDNA sequencing, RNA-Seq and variants of RNA-seq is that the isolated RNA sequences themselves are not informative about their function: sequencing alone cannot determine if a transcript is noncoding or coding (mRNA), nor whether it is functional (as RNA) or nonfunctional transcriptional noise which is not biologically meaningful.

Another branch of experimental methods for transcriptomics uses microarrays. Microarrays, also known as DNA chips or expression arrays, were mostly used for mRNA expression profiling by hybridizing labeled samples to 25-70 nt oligonucleotide probes. First in bacteria *E. coli*, commercially available microarrays that had previously been limited to coding regions were expanded into the intergenic regions to create a technology called “tiled microarrays”, enabling the discovery of novel transcribed regions including ncRNAs [24, 92, 93]. The advent of customized tiling arrays with a few nucleotides resolution made it possible to annotate novel transcripts from the entire genomes of higher eukaryotes, ranging from worms to human [94–96]. Despite the merits of microarrays such as inexpensive cost and nonredundant readout of transcription level, microarray experiments share many of the same drawbacks with cDNA sequencing or RNA-Seq as mentioned above; in particular, detection of a transcript does not resolve whether it is coding or not, or functional or not.

1.2.2 Possible sources of false positives

Recent detailed studies have highlighted the issue of false positives resulting from technical and biological drawbacks found in transcriptomic analysis [97–99], challenging transcriptomic results that have reported large numbers of ncRNAs and pervasive noncoding transcription [100, 101]. The source of false positives is various: biological noise, technical artifacts and coding transcript classification error.

All biological systems produce varying levels of errors due to biochemical limitation of their components. For essential processes that require a high fidelity, the cell devises several proof-reading and correction mechanisms. However, it is impossible to make error-free cellular machinery. Transcriptional machinery can produce partially processed transcripts or generate random transcripts due to proximity to other promoters or cryptic promoters which are not genuine promoters. Splicing machinery could result in partially or totally unspliced transcripts. Degradation machinery could miss some introns that are spliced out from coding transcripts or generate partially degraded transcripts. Struhl extrapolated the fidelity of yeast RNA polymerase II (pol II) through chromatin immunoprecipitation experiments to measure the pol II and TATA-binding protein occupancy *in vivo*: by Struhl's calculation, around 90% of pol II transcribed loci are expected to be nonfunctional "noise" in transcription and an $\sim 10^4$-fold pol II initiation difference between an optimal site and an average noisy site is similar to the specificity of sequence-specific DNA-binding proteins and other biological processes [102]. Also, through a careful characterization of RNA-Seq results, Bakel and Hughes reported that most reported "ncRNA" transcripts are intronic transcripts that might be fragments of mRNAs or intergenic transcripts which are located near known genes, that is, byproducts of adjacent transcriptional machinery, and that the remaining singletons have characteristics of random sampling from a low-level background [103]. Furthermore, in the analysis of human chromosome 21 and/or 22 transcription, it was shown that only $\sim$7-20% of the novel transcribed regions are conserved in the mouse genome whereas $\sim$44% of the transcribed regions overlapping known genes are conserved [104, 105].

Several technical/experimental artifacts affect various steps of transcriptomics. RNA samples have low-level genomic DNA contamination, even after DNase treatment. Due to the fragile nature of RNA samples, fragments of coding transcripts could be generated. In the step of first-strand cDNA synthesis, wrongly primed products could be generated. For example, about 47% of FANTOM III “noncoding” transcripts, many of which are intronic transcripts, seem to be internally primed from genomically-encoded poly-A stretches in longer coding pre-mRNA transcripts by the oligo dT primer [98]. In a study by Kampa et al. to compare different sample preparation methods by using different oligo probes and hybridization materials (RNA vs. DNA) on the same microarray platform, only $\sim$35% of the positive probes overlapped with each other [105], which indicates a high false positive rate, probably resulting from both biological noise and technical artifacts. In microarray experiments, both sequence-specific and non-specific cross-hybridization are an important potential source of error. Attempting to increase sensitivity to detect low-copy RNA transcripts exacerbates the problem of distinguishing true signals from background cross-hybridization. A careful reanalysis on microarray data by Bakel and Hughes showed that a small increase in sensitivity can cause a dramatic loss in specificity for detection of exons over a broad range of parameter settings and that the estimated proportion of transcription events in microarray experiments is consistently higher than that found in RNA-seq experiments [103].

Coding region classification errors can result from the use of overly simple criteria to distinguish noncoding RNA from coding mRNAs; for example, as mentioned above, the FANTOM II project designated transcripts that have a less than 100 aa long open reading frame (ORF) as noncoding transcripts after eliminating coding transcripts detected by homology in DNA or protein level [29] even though many real proteins are smaller than 100 aa. Surveying several methods for discriminating protein-coding and noncoding, Dinger et al. systematically documented the existence of coding mRNAs that escape detection by simple criteria of ORF length or by ORF conservation constraint [106]. It was reported that in mammalian proteomes, the ORF length of $\sim$ 3700 protein genes is smaller than 100 aa [107], and that many yeast "orphan" ORFs, which have no known homologs, have detectable transcripts and/or translated products [108]. Some examples of small proteins are 11 aa long TAL protein which has a role in fruit fly development [109], a less than 33 aa long Cg-1 protein controlling tomato-nematode interaction [110], and 75-140 aa long CLE family proteins involved in *Arabidopsis* meristem development [111].

1.3 Newly emerging ncRNAs: Results of pervasive transcription

Transcriptomic analyses have been accumulating many instances of two new (but very general and crude) kinds of ncRNA populations, long ncRNAs (lncRNAs) and small ncRNAs (sRNAs), together with novel protein genes, new alternatively splicing exons, and antisense transcripts, implicating that almost whole genome is pervasively transcribed [112, 113]. Many of those ncRNAs are partially or entirely overlapped by another transcript in sense or antisense direction or located close to a neighboring gene in antisense direction. Some lncRNAs reside within an intron of another transcript or in the intergenic region (lincRNA) [114]. According to Bakel and Hughes's study on pervasive transcription [103], only 2.2-2.5% of reads of RNA-Seq transcriptome data are mapped into non-protein related genomic locations and most intergenic transcripts are adjacently located to annotated protein-coding genes either as extended transcripts or separate noncoding transcripts. Whether these RNAs have functions is still controversial: those could be confused with byproducts of natural transcripts such as cis-natural antisense transcripts or mRNA fragments as we discussed above [115, 116].

A large number of sRNAs are associated with protein-coding genes in a variety of ways: promoter-associated sRNAs (PASRs)[112], transcription start site (TSS) antisense RNAs (TSSa-RNAs) [117], nuclear run-on assay derived RNAs (NRO-RNAs) mapping 20-50 bp downstream of TSS [118], tiny transcription initiation RNAs (tiRNAs) mapping 20 bp downstream of TSS [87], promoter upstream transcripts (PROMPTs) mapping 0.5-2 kb upstream of TSS [119] and termini-associated sRNAs (TASRs) [112]. Dissection of the function of sRNAs might be highly challenging technically because biological effects of individual sRNA species may not be substantial enough to be detected by current experimental approaches [37]. Therefore, it is difficult to distinguish genuine functional ncRNAs from these ncRNAs even though a few of them are functionally elucidated.

Among recently reported ncRNAs, lincRNAs which are easily separable from adjacent coding loci are relatively more probable to be novel independently transcribed functional ncRNA genes compared to other small RNA species in a transcriptome. Several previously recognized large ncRNA genes that regulate protein-coding genes epigenetically or at the transcriptional level also might be considered as lincRNA. For example, HSR1 (heat-shock RNA-1) forms a complex with translation elongation factor eEF1A and stimulates trimerization of heat-shock transcription factor 1 (HSF1) to induce the transcription of heat-shock-induced genes indirectly [120]. The Xist gene and Tsix gene (antisense to Xist) are critical for dosage compensation in eutherian mammals as a component of the Xic (X-chromosome inactivation center) [121, 122]. However, although a few lincRNAs defined by the transcriptomic analysis such as a HOTAIR RNA have been functionally elucidated [123, 124], for others we generally only know a tissue-specific expression profile [125] and most remain poorly characterized.

1.4 Transcriptome-independent ncRNA finding: Computational analysis

Besides transcriptomic approaches, which have generated an astonishingly large number of controversial ncRNAs, an alternative approach to find ncRNAs is computational prediction of putative structural ncRNAs and cis-regulatory protein-binding structural motifs in mRNAs by identifying conserved patterns or stability of predicted RNA secondary structure [126–128]. These computational approaches have also generated large numbers of ncRNA candidates, relatively few of which have been experimentally validated, as described below in more detail.

1.4.1 Computational search by comparative analysis

One way to computationally find ncRNA genes is a homology search using the evolutionary constraint information imposed by the secondary structure and/or primary sequence of known ncRNA gene instances in a particular gene class or family. In addition to general homology search programs based on RNA family-specific models, such as the Infernal package [129], or other ncRNA homology search programs taking a single RNA sequence with its secondary structure as a query to search homologous sequences in database [130], several family-specific ncRNA genefinders [131–138] have been developed for higher sensitivity and specificity by adopting a probabilistic or heuristic model of family-specific features. Even though these ncRNA genefinders are specially trained and tuned homology detectors, some are still limited by a high false positive rate. Of course, homology detection by similarity search is not suitable for *de novo* novel ncRNA genefinding.

De novo ncRNA genefinding is a more difficult problem than that of protein-coding gene finding [139]. Whereas protein-coding genes have lots of known signals on the gene structure such as start/stop codons and splicing sites, ncRNA genes do not have such common primary sequence features. Moreover, poor primary sequence conservation of ncRNA genes across species makes the problem harder for comparative approaches as well. So, current *de novo* ncRNA genefinders fundamentally rely on phylogenetic secondary structural conservation information with or without thermodynamic structural stability information [126–128, 140] to find structural ncRNAs and cis-regulatory protein-binding structural motifs in mRNAs. For example, QRNA [126] assigns a class to a homologous sequence alignment by comparing scores of three probabilistic models: a pair-SCFG (stochastic context-free grammar) “RNA” model capturing co-evolutionary patterns of secondary structure, a pair-HMM (hidden Markov model) “protein” model representing triplet codon preservation, and a null “other” model emitting pair sequences independently from patterns. One problem of these phylogenetic approaches is the difficulty of detecting ncRNAs that are too conserved to show structure-induced conservation or too diverged to be accurately aligned, or that are unstructured or less structured, such as the C/D box snoRNA family. Another more serious problem is high false positive prediction rates of the current programs. For example, Babak *et al.*'s tests on several ncRNA search tools with shuffled alignments while preserving dinucleotide frequency showed high false positive rates; the score distributions on tiling windows of human-mouse alignments of chromosome 19 and the shuffled alignments are not distinguishable [141].

1.4.2 Computational search by gene composition

Some special organisms, specifically hyperthermophiles, have allowed some unusual approaches independent of conventional ncRNA discovery methodology. A simple screen for GC-rich regions in the AT-rich *Methanococcus jannaschii* and *Pyrococcus furiosus* genomes provided ncRNA gene candidates, due to a strong DNA compositional bias toward G/C residues in structured ncRNA genes of some hyperthermophile genomes [26] to make more stable RNA structures in a high temperature environment. This screen found approximately five novel ncRNA genes in each organism. The small number of novel ncRNA genes is a stark contrast to ncRNA discovery efforts in other organisms with conventional transcriptomic or computational approaches, but it may be due to the characteristics of these hyperthermophiles: such as, selection against the use of ncRNA genes due to the constraints given by a high temperature environment and/or a relatively small genome in which the predicted protein-coding gene count is about half of that of *Escherichia coli*. Alternatively, a small number of ncRNAs might be predicted because of limitations in the ncRNA genefinders. This computational search using the difference in gene composition also has been applied into other AT-rich genomes, and similar to these hyperthermophiles, only a small number of novel ncRNA genes has been discovered in each genome [142, 143].

1.5 Conclusion

Even though computational ncRNA predictions by using base composition difference in hyperthermophiles or other AT-rich genomes provided a few novel ncRNA candidates, both transcriptomic approaches and other computational predictions for ncRNA detection have resulted in controversial reports of surprisingly large numbers of ncRNAs in a wide variety of species. These ncRNAs are a mixture of functional ncRNAs, biological background noises, technical artifacts, novel coding mRNAs, and/or computational false positives. Although some putative ncRNA candidates show cell-type specific expression, developmentally regulated expression and/or subcellular localization, these correlations cannot necessarily imply biological functionalities. Without a detailed careful examination of the identities and functions of these putative ncRNA candidates, it is neither possible to accurately estimate the functional ncRNA content in a genome, nor to conclude whether a genome is pervasively transcribed or not.

Chapter 2

Our approach

Independent of conventional transcriptomics and computational ncRNA prediction approaches that generates an overwhelming number of ncRNA candidates, including many false positives, a different systematic ncRNA identification approach might help to address a currently unsolved issue, the number of independent functional ncRNA genes in a genome. Transcriptomics does not distinguish genic from nongenic transcripts, noncoding from coding, functional from nonfunctional. Though these are many things that RNA could be doing, it would be nice to at least be confident of ncRNA genes - this is what a *Oxytricha* screen looks for. Although several studies show specific expression profiles on some of ncRNAs resulted from the transcriptomics [100, 144, 145], these patterns itself could not provide information about their functions. They provide hypotheses about the possible functions solely based on the correlation. Also, other ncRNAs from the transcriptomics might be byproducts of noisy eukaryotic transcriptional events that have no function as ncRNAs although transcriptional event itself may have a biological functionality, which is difficult to be distinguished from independent ncRNA genes of which transcribed products itself have a function on the cellular process. Moreover, the challenge is to figure out how these ncRNAs are generated exactly. For example, there appear to be ncRNAs produced from enhancers of coding genes [146], but because enhancers can be distant from their gene, it is difficult to distinguish enhancer-associated ncRNA transcription from an independent ncRNA gene. Distinguishing independent functional ncRNA genes from other sources of putative ncRNAs would be a step towards focusing effort on specific classes of ncRNAs and RNA function, rather than treating all "ncRNAs" and "ncRNA genes" as a homogeneous class. Therefore, an important question that has not been addressed well by current approaches, "How many independent functional ncRNA genes exist in the genome?", could be answered by *O. trifallax* at least in part.

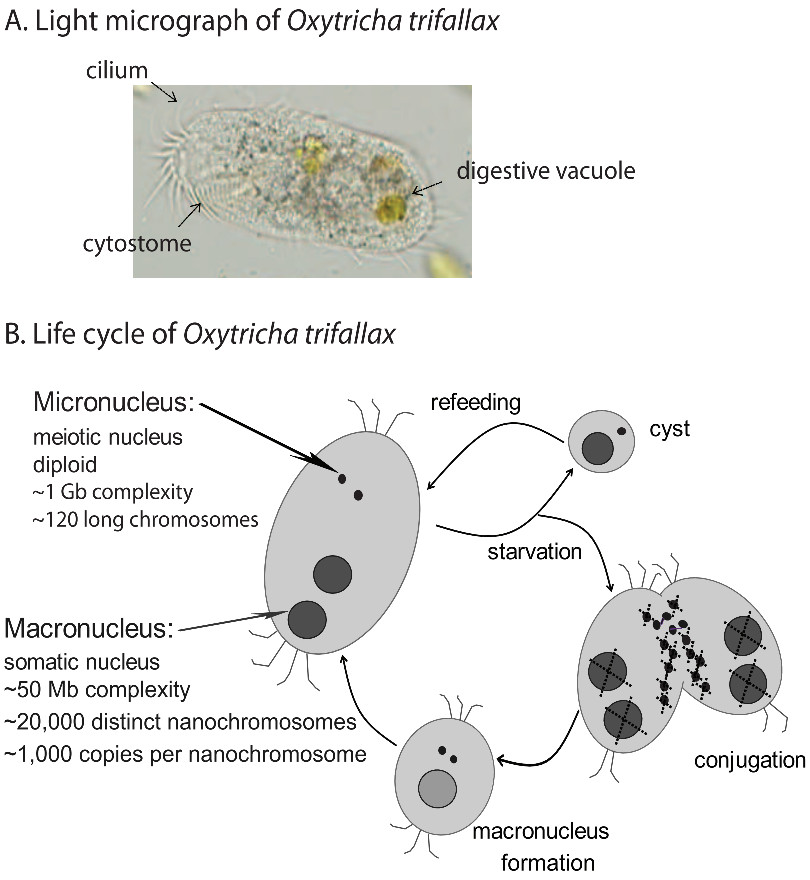
2.1 *Oxytricha trifallax*

Oxytricha trifallax (also known as *Sterkiella histriomuscorum* [147]) is a unicellular ciliated protozoan in class *Spirotrichea*, one of the extensively studied classes among 10 ciliate classes (Figure 2.1.A). Ciliates, known as a birthplace of telomere biochemistry [148] and self-splicing Group I intron RNA study [42], diverged from other microbial eukaryotes, so they are a phylogenetic outgroup of the crown eukaryotes, including metazoans, plants, and fungi. *Oxytricha* is also quite diverged from two other sequenced oligohymenophoran ciliates, *Tetrahymena thermophila* [149] and *Paramecium tetraurelia* [150].

Ciliates (phylum Ciliophora) have a nuclear dimorphism: a diploid meiotic germ-line nucleus (micronucleus) and a somatic nucleus (macronucleus). The number of each nuclear type per cell varies among ciliates: *O. trifallax* has two micronuclei and two macronuclei. The micronucleus mainly serves as the template material during conjugation and is almost transcriptionally silent. Conversely, the macronucleus is a highly specialized expression organelle providing all genes for normal cell function during vegetative growth and asexual reproduction [151, 152]. After several cycles of asexual reproduction, some ciliates entirely lacking micronuclei occur in the wild [153]. The life cycle of ciliates is simple: in the absence of food, it forms a cyst, a biologically inert form of the ciliate that retains only one macronucleus and one micronucleus, or undergoes cell mating, and otherwise, it proliferates continuously (Figure 2.1.B) [154].

2.1.1 The genomic characteristics of *O. trifallax*

The micronuclear genome consists of several large chromosomes similar to typical eukaryotic chromosomes. Genes are scattered along the chromosome and are separated by large stretches of “spacer DNA”, which seems to provide a safe place for the invasion of foreign DNA sequences, as a defense mechanism. The micronuclear genes are enigmatically interrupted by multiple A/T-rich noncoding sequences called internal eliminated segments (IESs), which will be spliced out during macronucleus development [155] (Figure 2.2.A). Most IESs are intact $\sim$ 4-5 kb long transposons, and short IESs (less than 0.5 kb) seem to be degenerated non-autonomous transposons that retain the cis-acting sequences required for precise excision. Most IESs in hypotrichs are less than 100 bp long. It is estimated that there are 100,000 to 200,000 IESs per haploid genome [154]. The telomeres of micronuclear chromosomes are made up of hundreds of duplex repeats of the sequence $5'$-C_4A_4-$3'$/$3'$-G_4T_4-$5'$ and terminate with a “t-loop” that provides a general mechanism for chromosomal end protection and telomere replication [156, 157]. The t-loop is stably formed by a foldback of a single-stranded $3'$ tail into the downstream double-stranded telomere repeat



A. Light micrograph of *Oxytricha trifallax*

B. Life cycle of *Oxytricha trifallax*

Figure 2.1: Life cycle of *O. trifallax*

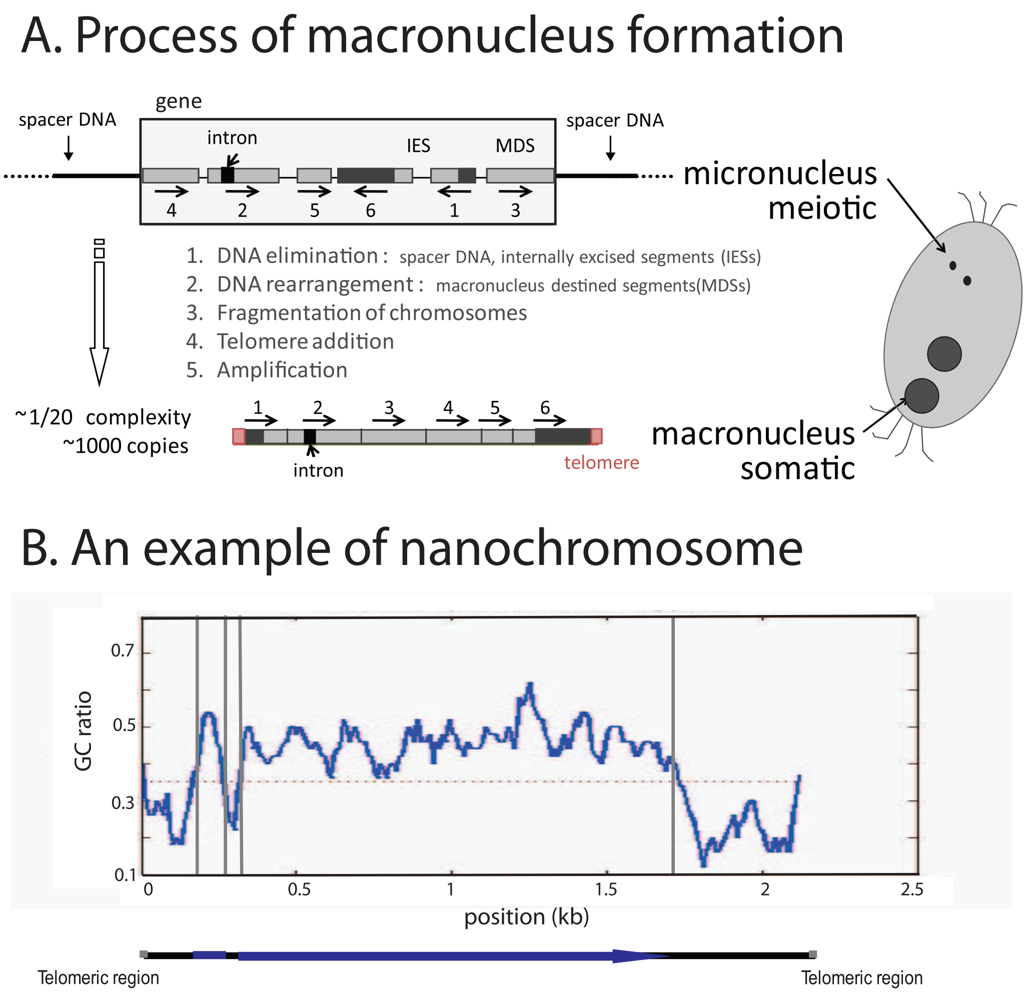
A. A light micrograph of the stretched *Oxytricha trifallax* due to Protoslo (protozoa quieting solution) slowing the movement of cells to keep them in focus and in the field of view while preserving characteristic motion of cells. Several structural features are detectable under a light microscope without staining. **B.** The macronuclei are represented as big circles and the micronuclei are represented as tiny circles to express their cytological size although the haploid complexity of the macronuclear genome sequence is much less than that of the micronuclear genome sequence.

region. The micronucleus undergoes meiosis during cell mating. Two haploid micronuclei exchanged between two cells in a mating pair are fused to form a diploid zygotic nucleus in each cell. After separated from a mating pair, unused haploid micronuclei and the old macronuclei are destroyed, and at the same time, a new macronucleus develops from a mitotic copy of the newly formed diploid micronucleus (Figure 2.1.B) [154].

The macronuclear genome consists of many small, linear, acentric chromosomes which are produced from the micronuclear genome by a baroque ncRNA-dependent process of splicing out the micronucleus limited sequences during sexual conjugation. This process includes not only genome fragmentation and spacer DNA elimination, but also rearrangement and unscrambling of the macronucleus destined sequences (MDSs) that are separated by an IES, in some ciliates including *Oxytricha* [153, 158–161] (Figure 2.2.A). These DNA processing events apparently depend both on the pairs of repeats that flank IESs for recombination and on nongenic transcription of long RNAs [160, 162], even though the detailed mechanism in each gene in each ciliate might be different. At least in *Tetrahymena*, these events involve large numbers of Argonaute-bound small RNAs [163–166]. The degree of genome fragmentation varies among ciliates. It reaches an extreme in the spirotrich ciliates including *Oxytricha*, *Stylonychia*, and *Euplotes*, where the macronuclear genome is composed of many thousands of gene-sized nanochromosomes [167, 168]. In *Oxytricha trifallax*, the micronuclear haploid DNA content of $\sim$1 Gb is reduced by 95% to $\sim$50-55 Mb of sequence complexity in the macronucleus. The macronucleus is thought to contain $\sim$17,000-25,000 different nanochromosomes almost entirely in the range of 1-8 kb, with a mean of 2.2-2.5 kb [153, 169, 170]. Each nanochromosome is amplified to an average copy number of $\sim$ 1000. Remarkably, each nanochromosome usually contains only a single gene, which usually has just a few small introns with an average 118 nt, and also has very short 5′ and 3′ UTRs with a median length of about 130 nt and short subtelomeric non-coding sequences and telomeres [171–174]. A typical example of a nanochromosome that represents well these characteristics is shown in Figure 2.2.B. Some macronuclear chromosomes are generated by alternative fragmentation of the polytene chromatids during differentiation, reproducibly [175, 176]. One alternative fragmentation mechanism seems to be correlated with a variant form of telomere addition within $\sim$ 100bp subtelomeric regions.

2.2 Our approach

If eukaryotes generally have a large proportion of independent ncRNA genes, then the *Oxytricha* macronucleus should have a large proportion of noncoding nanochromosomes. In effect, in these ciliates with gene-sized nanochromosomes, the organism itself has solved the hard eukaryotic genefinding problem for us. Most genes and their cis-regulatory signals have been isolated on individual chromosomes, their locations demarcated by telomere addition, and most of their nonessential noncoding DNA has been eliminated [173]. Given the assumption that most nanochromosomes contain a single gene in it, we can identify and discard nanochromosomes carrying protein-coding genes among the macronuclear genome sequences, because identifying coding genes computationally is far easier than identifying ncRNA genes. Coding gene identification in *Oxytricha* is even easier than in many eukaryotes, because its protein-coding gene structures are simple, with few introns, and those introns that do occur are small, with a mean length of 118 nt [168, 174]. The resulting subset of apparently noncoding nanochromosomes should be enriched for nanochromosomes carrying independently transcribed ncRNA genes. We took advantage of the availability of a draft macronuclear *O. trifallax* genome sequence assembly [168] to conduct such a



```
ccccaaaaccccaaaaccccaaagatatgtggctggattttaaaatatgactaaagataattgagattctcgtttttttattggtagtcgttagatttacctct
atataatccaaattaattccctccattttataaattaatttaacaattcttaaatcaaatattacaatgtcatcagctgctaaaaaccaaagatcaacttcaa
gagtcacaaagaagaagaccacaggtcccaaggacgctgcaccaaagaaagccgaaaagggaagcaaagtaagcaaataa…………
……………………………………………………………………………………………….... gacaccagaat
tgtctactgaacagacttgtgaattcaaatggaagggagagcctagttctctttttctataataaatctctaacagttttgataaccctaacaggttaagtaa
tgatttttaaaattttataatttcttcctaaaaattaaaaacatactgtaactctttaagtttatatattaacacataatttatataactattaagctgcacataattt
aatccaatacaccttattatgttttatgaattgtaaaatgtttcttctcacatactaaatagtctgaagtctttcaatgttcctttaataatgcaaatgtatgaattt
ttagaaatttgtcatacattcataaatttaaattttacaaattggtaaacagctaatactttaatatatctctatttatttgattcattatttatattctcaattcattag
tctggggttttggggttttgggg
```

Figure 2.2: The micronuclear and macronuclear genome of *O. trifallax*

A. Organization of micronuclear chromosomes and genes. The micronuclear genome is processed to generate the macronuclear genome. Arrows and numbers below or above the macronucleus destined segments (MDSs) indicate the relative direction and order of segments in the macronuclear genome to explain the rearrangement process of the scrambled MDSs. Dark gray regions in the MDSs represent the subtelomeric regions in the macronuclear nanochromosome. **B.** An example of a macronuclear nanochromosome with GC ratio over the chromosome, gene structure, and genome sequence. The average GC ratio of the *O. trifallax* draft genome assembly is 0.34 which is indicated as a dotted line in graph. GC ratio is calculated by sliding 50 nt segment window with 10 nt step size.

screen.

To winnow out nanochromosomes containing protein-coding genes, we developed a nanochromosome classifier “nanoclassifier” based on a Hidden Markov Model (HMM) that is easier to train and adjust for a desired sensitivity/specificity tradeoff, than conventional custom-trainable genefinders. We conducted comparative analyses on the related ciliate *Stylonychia lemnae* to remove false positives from the nanoclassifier, to find functional conserved ncRNA genes, and to refine the possible genic regions within a noncoding nanochromosome. For the final ncRNA candidate gene sets, we experimentally validated their *in vivo* transcripts with Northern and RACE-PCR and manually analyzed the consensus secondary structures and regulating elements if possible.

2.3 Outline of this work

Chapter 3 surveys non-redundant full-length nanochromosomes from the draft genome assembly of *O. trifallax* and describes how the known ncRNA genes are distributed on the nanochromosomes and the experiment used to characterize how complete and how biased our sample of nanochromosomes is.

Chapter 4 details the computational ncRNA screens we designed and executed. Technical specifications of nano-classifier and nano-genefinder we built are described and how the comparative analysis was conducted with *S. lemnae* genome is illustrated.

Chapter 5 describes how the ncRNA candidates in the computationally-identified final data set were verified with Northern and RACE-PCR experiments and details the characteristics of ncRNA candidates.

Chapter 6 describes the conservation and transcriptional features of a novel ncRNA family, which we called the Arisong RNA, in four ciliate species and speculates about the possible function of this family.

Chapter 7 investigates other ncRNA genes in *O. trifallax* which are not detected in this screen to examine the soundness or weakness of this screen.

Chapter 8 offers a brief description of the result of this screen and concluding thoughts on the number of independently-transcribed ncRNA genes on *Oxytricha* and other eukaryotic genomes.

Appendix A describes tRNA gene analysis on the total *O. trifallax* dataset; Appendix B lists coordinates of telomere endpoints of a subset of full-length *Oxytricha* nanochromosomes (WGS2.1.1 dataset among stage 3 dataset); Appendix C catalogs all the *Oxytricha* genes which were mentioned in this screen; Appendix D displays Northern blot of some of known ncRNA genes in *Oxytricha* and lists all Northern blot oligonucleotide probes; Appendix E lists RACE-PCR gene-specific probes; Appendix F displays the results of comparative analysis on the final candidate data set (stage 5 dataset); Appendix G shows sequence alignments of regulatory motifs in *O. trifallax* and *S. lemnae*; and Appendix H mentions programs and databases we used, and data availability.

Chapters 2 to 8 and the appendices are derived from a published paper:

S. Jung, E. C. Swart, P. J. Minx, V. Magrini, E. R. Mardis, L. F. Landweber, and S. R. Eddy. Exploiting *Oxytricha trifallax* nanochromosomes to screen for non-coding RNA genes. *Nucl. Acids Res.*, in press, first published online June 28, 2011

Chapter 3

O. trifallax genome sequence

O. trifallax macronuclear genome sequencing is an ongoing project through collaboration between the Genome Center at Washington University and the Landweber laboratory at Princeton [`http://genome.wustl.edu/genomes/view/oxytricha_trifallax/`]. We utilized the draft genome assembly data and conducted several analyses on this incomplete dataset as a computational screen for independent functional ncRNA genes.

3.1 *O. trifallax* draft genome assembly

We obtained two draft datasets for *O. trifallax* genome sequence: a "WGS" dataset and a "pilot" dataset.

The WGS dataset is a prepublication whole genome shotgun draft assembly version 2.1.1 (June 2007), comprising 54982 contig sequences (79.2 Mb) averaging 1.44 kb in length. Whole cell DNA were prepared from vegetatively growing *O. trifallax* strain JRB310 [173] and <7kb nanochromosomes were selected by gel purification to avoid the abundant rDNA nanochromosome. Nonetheless, this size fractionation captures the great majority of the macronuclear genome, which is primarily pieces in 1-8 kb [153, 169, 170]. After excluding singletons, PCAP[177] assembled 728,035 ABI 3730 shotgun reads (583.7 Mb) into contigs. Overall assembly contiguity is less than that expected given the 7.4X mean shotgun coverage in part because macronuclear nanochromosomes have variable copy numbers and coverage per nanochromosome is non-uniformly distributed. The assembly also appears to be contaminated with a second *Oxytricha* strain, 510, and with bacterial DNA from food in the culture. Surprisingly, a substantial fraction of contigs retains vector sequences at both or either ends of contigs of various length. For instance, among contigs that have detectable telomeres at the both ends, 848 contigs contain $\geq$ 100nt vector sequence on one or both ends.

The "pilot" dataset is a collection of pilot sequencing data comprising 1976 complete nanochromosome sequences (1.96 Mb) averaging 0.9 kb in length. It consists of 254 complete nanochromosome sequences from a Princeton/Utah pilot genome project [168, 173, 174], 1707 nanochromosomes generated by paired-end sequencing of full-length plasmid inserts cloned from a size-selected $<$1kb nanochromosome fraction, the 7.6 kb ribosome DNA nanochromosome, and 14 additional full-length nanochromosome sequences.

Overall, the combination of the WGS and pilot datasets consists of 56,958 sequences (total 81,114,275 nt), with contigs ranging from 42 to 13,846 nt and averaging 1.42 kb in length.

3.2 Non-redundant full-length nanochromosomes: stage 1 dataset

Our screening strategy involves classification of full-length nanochromosomes as coding or noncoding, so we should begin the screen with full-length nanochromosomes. The genome assembly is somewhat crude, with a large amount of untrimmed vector sequence, many incomplete contigs, and some bacterial contamination. From our WGS and pilot genome datasets, we extracted a nonredundant, merged set of presumptive full-length *Oxytricha* nanochromosomes (the "stage 1" dataset). All 1976 contigs in the pilot dataset were assumed to be full length. In the WGS 2.1.1 assembly, we searched the terminal 400 nt of each contig end for matches to partial telomere consensus sequences ($[\text{CCCCAAAA}]_3$ at each contig's $5'$ end and $[\text{GGGGTTTT}]_3$ at the $3'$ end) after removing any flanking x's by requiring a local Smith/Waterman alignment score of ≥ 80 using gapcost = -3, match = 5, mismatch = -4. If a telomere was identified internal to the contig, we required that the extra flanking sequence matched the known cloning vector with at least 80% identity through a glocal (global with respect to the vector, local with respect to the nanochromosome) alignment using gapcost = -2, match = 5, mismatch = 1. This defined the minimal telomere endpoint coordinate. A small number of nanochromosomes were additionally defined as "full length" after further inspection of borderline results. We identified 8565 complete nanochromosomes in the WGS 2.1.1 assembly by this procedure. The telomere endpoints coordinates of a subset of them (stage 3 dataset) are listed in Appendix B.

To remove nanochromosomes that appear redundantly in both the pilot and WGS datasets, we used WU-BLASTN with default parameters to identify near-identical pairs that satisfied $E \leq 10^{-100}$ and % identity $\geq 98\%$ and which differ in length by $\leq 10\%$ of the longer sequence. We chose one sequence of such pairs at random, thereby removing 894 redundant sequences.

The stage 1 dataset consists of 9647 full-length nanochromosome sequences of average length 1.9kb. Four typical examples of *Oxytricha* full-length nanochromosome organization are shown in Figure 3.1, including annotations by methods we describe below.

3.3 Quasialleles in the draft assembly and stage 1 dataset

There are usually several identical or near-identical copies of each locus in the assembly. Highly identical contigs were removed from the stage 1 dataset. However, even after this step, the stage 1 dataset still includes some very similar copies of each locus. The full-length chromosomes in the stage 1 dataset can be grouped in up to seven nanochromosomes with approximately 3.4% mean sequence difference. There could be several explanations for this.

Oxytricha is a diploid. The sequenced *Oxytricha* culture was an inadvertent mixture of two mating types, 310 and 510 (Laura Landweber; personal communication). The study on IESs and introns of *Oxytricha* 81 locus by *Seegmiller et al.* estimated the divergence of these two strains at about 0.1 changes/site, but did not address the allelic difference within each strain [179]. However, this divergence rate was calculated from non-coding regions of DNA, so it is not directly applicable to distinguish alleles of ncRNA genes. There also appears to be a substantial fraction of alternatively processed nanochromosomes with different sizes and breakpoints. Without a micronuclear genome sequence and a more complete assembly, we cannot distinguish alleles, products of alternative DNA processing, and highly identical paralogs.

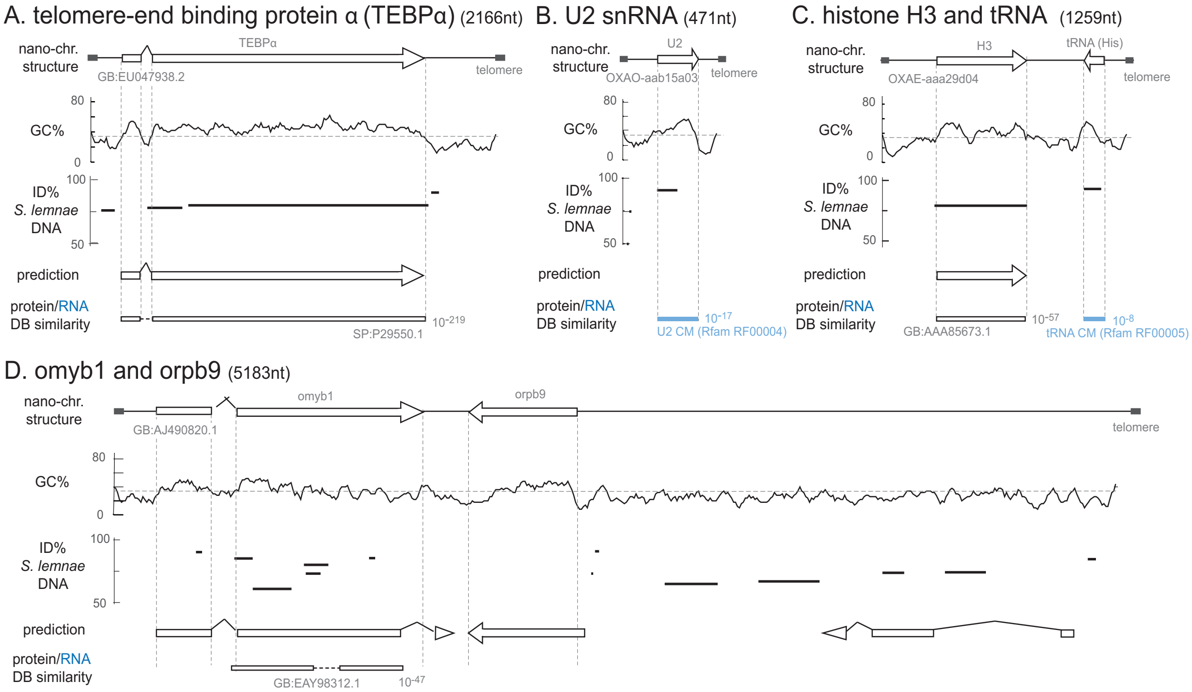


Figure 3.1: Examples of *O. trifallax* nanochromosomes.

A. A typical nanochromosome containing a single protein-coding gene (telomere-end binding protein α); **B.** A typical nanochromosome containing a single ncRNA gene (U2 snRNA); **C.** A nanochromosome containing both a protein-coding gene (histone H3) and an ncRNA gene (a tRNA-His); **D.** A nanochromosome containing two protein-coding genes (omyb1 and orpb9). Data tracks below each nanochromosome show some of the features we used in suggesting regions of coding potential, conservation, and/or functionality, as follows. **nano-chr. structure**: gene structures as annotated in GenBank (A,D) or predicted by us by similarity (B,C). **GC%**: calculated GC% in sliding 50nt windows with 10nt step size (the average GC% of *O. trifallax* is 34%, and a higher GC ratio tends to correlate with genic regions); **ID% *S. lemnae* DNA**: best WU-BLASTN matches to *Stylonychia lemnae* shotgun sequence data (see Methods); **prediction**: coding gene prediction from our *Oxytricha* genefinding program (nanogenefinder); **protein/RNA DB similarity**: best significant WU-BLASTX matches to NCBI NR protein database excluding *O. trifallax* proteins (black) or Infernal cmsearch [129] matches to the Rfam RNA database [178] (blue).

Operationally, we manually grouped highly identical loci (roughly $>$85% identical in DNA sequence flanking each locus) into what we call “quasiallele” groups. In this threshold, we generally identify up to four apparent “alleles” of any given sequence, which can be reasonably interpreted to include two alleles for each strain. For each quasiallele group, we assign a representative locus. In subsequent sections, we refer to numbers of “distinct” (representative) loci versus total numbers of sequences including “quasialleles”. We named and numbered each distinct locus “Onc1”, “Onc2”, etc. (for “*Oxytricha* noncoding candidate”), and numbered each additional quasiallele “Onc1.2”, “Onc1.3”, etc. Names, coordinates, and other information for all examined loci, including candidate loci described in the screen below, are listed in Appendix C.

3.4 The “known” ncRNA gene distribution in the stage 1 dataset

Previous studies indicate that *Oxytricha* nanochromosomes usually contain just a single gene [168, 172, 174, 180] with a few exceptions [172, 179, 181, 182], but these studies were largely focused on coding genes and were based on small numbers of nanochromosomes. Our screening strategy depends crucially on an assumption that ncRNA genes usually occur alone on their own nanochromosome, with no coding gene on the same nanochromosome. To test this assumption, we first investigated the 24 publicly available *O. trifallax* nanochromosomes for their potential to contain ncRNA genes on the same chromosome. Among 24 NCBI-retrieved *O. trifallax* nanochromosomes, only two nanochromosomes encode two protein-coding genes. We cannot detect any prominent known ncRNA homologs by using the cmsearch program of Infernal 1.0.2 [129] against 1372 ncRNA models in the Rfam 9.1 database [178].

Next, we identified homologs of known ncRNA genes in the stage 1 dataset and examined those ncRNA-containing nanochromosomes for protein-coding potential. The cmsearch program was also used to search 9647 stage 1 nanochromosomes at an E≤ 0.001 threshold per query model and 461 hits met this threshold. We manually removed 324 hits that we judged to be either redundant (different Rfam models for the same family: snoU18 and SNORD18) or false positives, including 318 weak miRNA similarities (most of which fell in telomeric repeats, and all of which appear to be false positives). Remaining were 135 ncRNA homologs from 11 Rfam families on 134 different nanochromosomes, including 106 transfer RNA (tRNA) genes (Table 3.1). In all but one case that has homologs of two known ncRNA genes, RNase MRP and snoZ 196, we found a single ncRNA homolog per nanochromosome.

To estimate how many of these 134 nanochromosomes contain coding genes in addition to an ncRNA gene, we masked the homologous ncRNA regions plus an extra 20nt on each side of the identified Infernal alignment, by converting the sequence to N's and any vector sequence was removed using telomere endpoint coordinates described above. One nanochromosome carrying the ribosomal RNA genes, which were identified by the presence of 5.8S rRNA, was manually masked for SSU rRNA and LSU rRNA by comparing with *Tetrahymena* ribosomal RNA gene sequence because Rfam does not include complete models of the large SSU and LSU rRNAs.

Three different methods were used to look for possible coding genes: (1) BLASTX for the identification of significantly similar regions to the annotated proteins in the database. (2) BLASTN for the examination of significant sequence conservation with *Stylonychia*

Table 3.1: Coding potential of ncRNA gene-containing nanochromosomes.

ncRNA	**Rfam accession**	**# nanos**		**X/NR**		**N/Sty**		**nanocl**		**any**		**all**	
tRNA	RF00005	106	**51**	35	**19**	41	**22**	66	**34**	68	**35**	35	**19**
5S rRNA	RF00001	13	**1**	.	**.**	.	**.**	.	**.**	.	**.**	.	**.**
5.8S rRNA	RF00002	1	**1**	.	**.**	.	**.**	.	**.**	.	**.**	.	**.**
U2	RF00004	4	**1**	.	**.**	.	**.**	.	**.**	.	**.**	.	**.**
U6atac	RF00619	2	**1**	.	**.**	2	**1**	2	**1**	2	**1**	.	**.**
SRP	RF00017	1	**1**	1	**1**	1	**1**	1	**1**	1	**1**	1	**1**
snoU18	RF01159	3	**1**	.	**.**	.	**.**	1	**1**	1	**1**	.	**.**
RNase_MRP,snoZ196	RF00030,RF00134	1	**1**	.	**.**	.	**.**	1	**1**	1	**1**	.	**.**
snoR38	RF00213	1	**1**	1	**1**	1	**1**	1	**1**	1	**1**	1	**1**
snoMe28S_Cm2645	RF00530	2	**1**	.	**.**	.	**.**	2	**1**	2	**1**	.	**.**
Total		134	**60**	37	**21**	45	**25**	74	**40**	76	**41**	37	**21**

The first two columns show the names of known ncRNAs and their accession numbers in the Rfam database [178]; the third column, “# nanos” is the number of nanochromosomes found to contain homologs of these known ncRNAs; both the total number of loci including all quasialleles, followed (in bold) by the number of distinct loci. “X/NR”, “N/Sty”, and “nanocl” columns show the number of these nanochromosomes that have significant similarity to known proteins by BLASTX, the number with another region of significant DNA conservation with *Stylonychia* by BLASTN, and the number with coding genes called by our nanoclassifier. The final two columns show the number that are called coding by at least one of the three methods (any), and the number called coding by all three methods (all).

lemnae[1]. (3) Our nanoclassifier for the detection of protein coding gene potential. For detection of the similarity to known proteins, WU-BLASTX was used on the NCBI NR database with “filter=seg filter=xnu C=6” (C=6 is the ciliate genetic code) options and with a $E < 10^{-5}$ threshold. For investigating genome sequence conservation on the close ciliate genome, WU-BLASTN was used to our *Stylonychia* shotgun data with “filter=seg filter=dust” options and with an $E < 10^{-10}$ threshold, and additionally required $> 70\%$ sequence identity for the best alignment. For the nanoclassifier, we used a $P \leq 0.09$ threshold, based on the benchmark ROC curve which will be described in the next chapter in Figure 4.2.

[1] a description of this genome assembly will be discussed in the next chapter in detail

Results are summarized in Table 3.1. BLASTX detects 37/134 (28%) with significant similarity to protein sequences in the NCBI NR database. BLASTN detects 45/134 (34%) with additional DNA conservation to *Stylonychia*. The nanoclassifier calls 74/134 (55%) of these as containing coding sequence.

Each method for detecting coding genes has limitations, in terms of both sensitivity and specificity. In terms of sensitivity, rapidly evolving or "*Oxytricha*-specific" genes will not be detected by BLASTX to the protein database or even by BLASTN to *Stylonychia*. Some will not show BLASTN hits to *Stylonychia* due to the partial coverage of our *Stylonychia* shotgun data. Our nanoclassifier has an estimated coding sensitivity of about 94% (Chapter 4). Analysis of a randomly chosen set of 200 stage 1 nanochromosomes showed 130/200 (65%) with BLASTX hits to NR; 148/200 (74%) with *Stylonychia* BLASTN hits; and 189/200 (94%) called coding by the nanoclassifier. If almost all *Oxytricha* nanochromosomes carry at least one coding gene, these numbers would approximate the sensitivity of each method. In terms of specificity for coding regions, some ncRNAs show BLASTX hits to the "protein" databases because some noncoding RNA genes have been erroneously translated and deposited in the databases. BLASTN conservation to *Stylonychia* can have multiple interpretations besides a conserved coding region, including an ncRNA or a large regulatory DNA sequence. Finally, we determined that our nanoclassifier has about a 17% false positive rate (Figure 4.2).

Using these expected false negative and false positive rates, we can extrapolate a corrected rough estimate of the total number of coding regions in these ncRNA-containing nanochromosomes. Correcting the BLASTX results for a 65% sensitivity (and assuming that essentially 100% of BLASTX conservation is truly due to coding regions) gives 0.28/0.65 = 43% of ncRNA-carrying nanochromosomes estimated to also carry one or more coding genes. Correcting the BLASTN results for 74% sensitivity (and ignoring possible false positives from noncoding conservation) gives 0.34/0.74 = 46%. Correcting the nanoclassifier results for 94% sensitivity and 17% false positives gives (0.55 - 0.17) / (0.94 - 0.17) = 49%. Therefore we conclude that about 50-60% of ncRNA-containing *Oxytricha* nanochromosomes carry no coding gene, at least for the known types of ncRNAs we can identify by homology searches.

3.5 The completeness and bias of the stage 1 dataset

The stage 1 dataset is an incomplete sample of the macronuclear genome. It was not feasible to obtain a complete assembly. One difficulty is that the unusual properties of *Oxytricha* nanochromosomes tend to violate assumptions made by standard production-scale genome sequencing methods. Improving the quality of the assembly likely will require a nonstandard assembly effort beyond the scope of this work. However, because our main question is about the relative *proportion* of independent functional ncRNA genes versus coding genes, not absolute numbers, a statistical sample of the genome will suffice, provided it is sufficiently unbiased. We therefore sought to characterize the completeness and the two most important sources of potential bias in the stage 1 dataset, as follows.

We estimate that the dataset includes 40-65% of the macronuclear genome, based on two different estimates. First, by dividing the kinetic complexity of the macronucleus (50-55 Mb) by the average nanochromosome size (2.2-2.5kb) [153, 169, 170], *Oxytricha* is thought to contain about 20,000-25,000 different nanochromosomes; 9,647 would represent around 40-50% coverage of the genome. Second, we measured coverage of a set of conserved core single-copy eukaryotic protein genes [183].

Parra *et al.* described a method to estimate the completeness of a eukaryotic genome assembly by assessing the presence of 248 "core eukaryotic genes" (CEGs), chosen for their wide orthologous conservation but low frequency of paralogous duplication [183]. We modified and simplified their method described in [184] for use on a low-pass, low-contiguity shotgun assembly without full-length gene predictions. We searched each CEG with TBLASTX against each of our ciliate datasets, collected all hits of $E < 10^{-10}$, calculated what fraction of each CEG sequence was covered by these alignments. We considered the CEG "present" if this fraction was >70%. By this definition, 215 (∼87% coverage) in the combined *Oxytricha* WGS+pilot dataset, and 162 (∼65% coverage) in the *Oxytricha* stage 1 dataset.

Similarly, completeness can be estimated by counting the conserved single-copy coding genes (CSCCGs) using the *Tetrahymena* genome sequences as a outgroup reference genome among the available sequenced ciliate genomes. Note that the *Paramecium* genome is problematic due to whole genome duplications. Using the simple TBLASTN approach to detect conserved coding genes from *Oxytricha* and *Stylonychia* to *Tetrahymena*, each CSCCG can be used to estimate the coverage of each other genome. However, this gives much higher coverage estimation than expected (∼89% coverage in *Stylonychia* and ∼97% coverage in *Oxytricha*). Also, when using the resulting CSCCGs sets as a reciprocal data set to reestimate coverage in each other genome, it gives very different coverage estimation (∼47% coverage in *Stylonychia* and ∼84% coverage in *Oxytricha*). So this approach using simple schematics was not deemed reliable.

Even though the dataset of full length nanochromosomes is estimated to be only 40-65% complete, in principle even just a small sample of about a hundred would suffice to investigate the proportion of noncoding nanochromosomes, so long as that sample was random and unbiased. However, there are two important sources of bias to consider in the stage 1 dataset as follows.

We expect a bias towards shorter nanochromosomes. Shorter nanochromosomes are easier to assemble, and the WGS part of the assembly is from a size-selected $<$7 kb fraction of macronuclear DNA. We compared the length distribution of the stage 1 dataset to two different estimates of the actual length distribution of the overall macronuclear genome (Figure 3.2). The actual distribution has been characterized previously by measuring the contour lengths of $\sim$1000 individual nanochromosomes in electron micrographs [170]. We extracted the EM contour length histogram from reference [170]. The actual distribution also can be obtained through digitizing an ethidium-stained agarose electropherogram used generally for nanochromosome detection. We extracted pixel intensities from a digital image of an ethidium-stained agarose electropherogram of *Oxytricha* DNA and averaged over sections of 0.1 kb as measured from adjacent size standards, assuming a logarithmic relationship between gel migration distance and DNA length in nucleotides. Intensity values were assumed to be proportional to DNA mass because ethidium is an intercalating dye and converted to relative molar nanochromosome abundance by dividing by DNA length [185].

Both methods produced similar overall length histograms. Overall, nanochromosomes have a mean length of 2.2-2.5 kb, ranging up to 10-20 kb, whereas the stage 1 data have a somewhat smaller mean length of 1.9 kb, ranging up to 13.8 kb. About 2% of nanochromosomes run out on a gel are larger than 7 kb, whereas only nine contigs in the stage 1 data are longer than 7 kb (0.1%). This indicates substantial (20x) undersampling of the 2% tail of longest nanochromosomes. About 15% on a gel are 4-7kb, where we have 202 contigs in the stage 1 data (2%), indicating moderate (7-8x) undersampling in this length range.

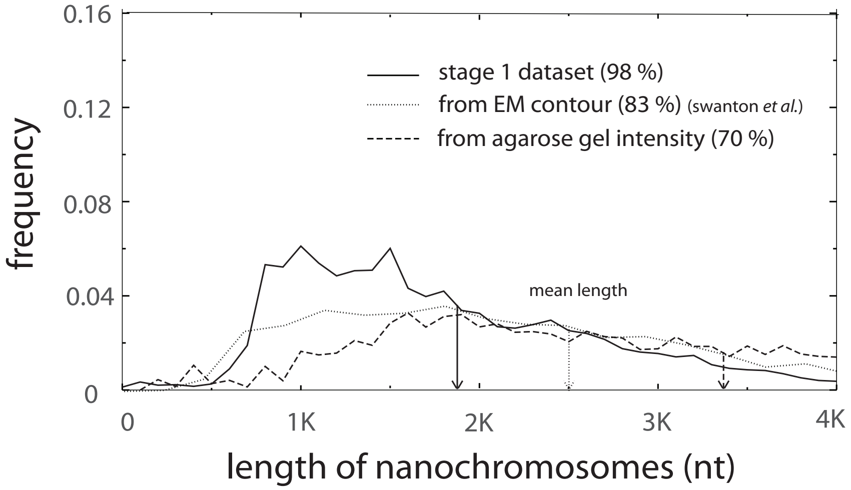


Figure 3.2: Length distribution of full-length nanochromosomes
The length distribution of the stage 1 dataset (solid line) and other reference length distributions; The dashed line shows the actual nanochromosome length distribution as estimated from an agarose gel electropherogram, and the dotted line shows the actual nanochromosome length distribution as estimated by Swanton et al. [170] from contour length in electron microscope images.

For the 80% of nanochromosomes that are <4kb, there is only a modest sampling bias.

The principal concern with a bias towards shorter nanochromosomes is that we could overlook ncRNA genes like the recently described mammalian long intergenic noncoding RNAs (lincRNAs) [88, 186]. However, lincRNAs are only “long” relative to other previously well-studied ncRNAs, which are often 100-400 nt. Mammalian lincRNAs seem to be about the same length distribution as coding mRNAs. According to GENCODE v4 transcript annotation[2], human lincRNAs and protein-coding mRNA transcripts show a similar length distribution, with mean lengths of 2.4 kb versus 2.3 kb, respectively (Figure 3.3). The length distribution of the stage 1 dataset covers the great majority of coding nanochromosomes, so it is also expected to cover lincRNA-like ncRNA genes. Second, and more

[2] `http://www.gencodegenes.org/`

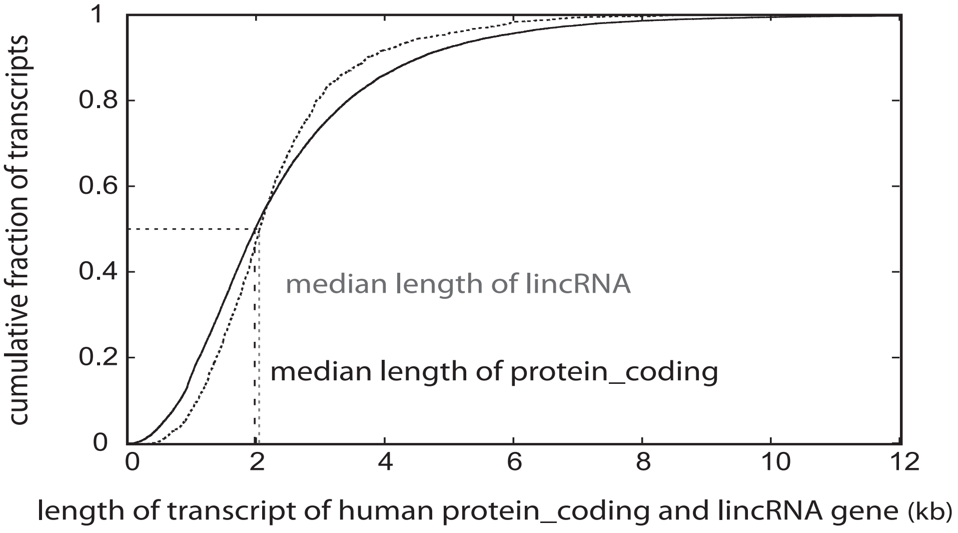


Figure 3.3: Cumulative length distribution of human genes
The solid line represents the cumulative length distribution of the transcripts of human protein-coding gene and the dotted line shows the cumulative length distribution of the transcripts of human lincRNA genes. Two vertical lines represent the median length of the each distribution.

generally, if a class of large ncRNA-containing nanochromosomes were present even at a few percent, we would have expected to sample some noncoding nanochromosomes among the 211 nanochromosomes longer than 4 kb in the stage 1 dataset.

We also expect a bias towards assembling more abundant (high-copy number) nanochromosomes, which get higher sequence coverage. Each *Oxytricha* nanochromosome occurs with a mean of $\sim$1000 copies per macronucleus [153, 170], but some nanochromosomes are known to be maintained at different copy numbers. The most extreme case is the rDNA nanochromosome, found to be present at about 100,000 copies. The rDNA appears as a prominent 7.6kb band on agarose gels of macronuclear DNA [153], where a distinctive species-specific pattern of 100-200 overrepresented bands is also seen [187]. Several examples of about six-fold copy number differences have been observed when the copy number of individual non-rDNA nanochromosomes has been measured [188, 189], and a few cases of extreme overamplifications have been observed during prolonged vegetative growth [190]. However, reassociation kinetics experiments have shown that bulk macronuclear DNA reanneals as if the great majority of sequences occur in roughly equal numbers [151, 169, 191]. In order to gauge the extent and impact of copy number control, we examined the distribution of sequencing coverage of individual assembled nanochromosomes in the WGS subset of the stage 1 data. We found a right-skewed distribution ranging from 1.1- to 87.4-fold coverage, with mean 10.4, median 7.3, and a mode of about 5 (data not shown). As expected from previous published results, this coverage distribution is consistent with nonuniform copy number varying over perhaps an order of magnitude, and it appears we have likely sampled the bulk of that distribution. Combined with the estimate of 40-65% completeness of the stage 1 dataset, it seems unlikely that a population of low-copy ncRNA-carrying nanochromosomes exists that has been entirely missed, as opposed to somewhat undersampled.

Chapter 4

Computational screens for ncRNA genes

Our scheme relies on being able to sensitively identify protein-coding regions, in order to screen out as many nanochromosomes containing protein-coding genes as possible. Homology searches are one way to identify probable coding regions, but while homology searching is specific, it is not very sensitive. Many proteins may have no detectable homologs, either because they are clade-specific or rapidly evolving. Therefore we aimed to use computational protein "genefinding" to sensitively identify protein-coding regions by their statistical signals, and later homology search was done to remove the undetected by genefinder but conserved protein coding genes. To find more probably functional ncRNA candidate and define the ncRNA region within the non-coding nanochromosome, comparative analysis with *Stylonychia* was conducted. All the following processes to detect ncRNA genes are summarized in Figure 4.1.

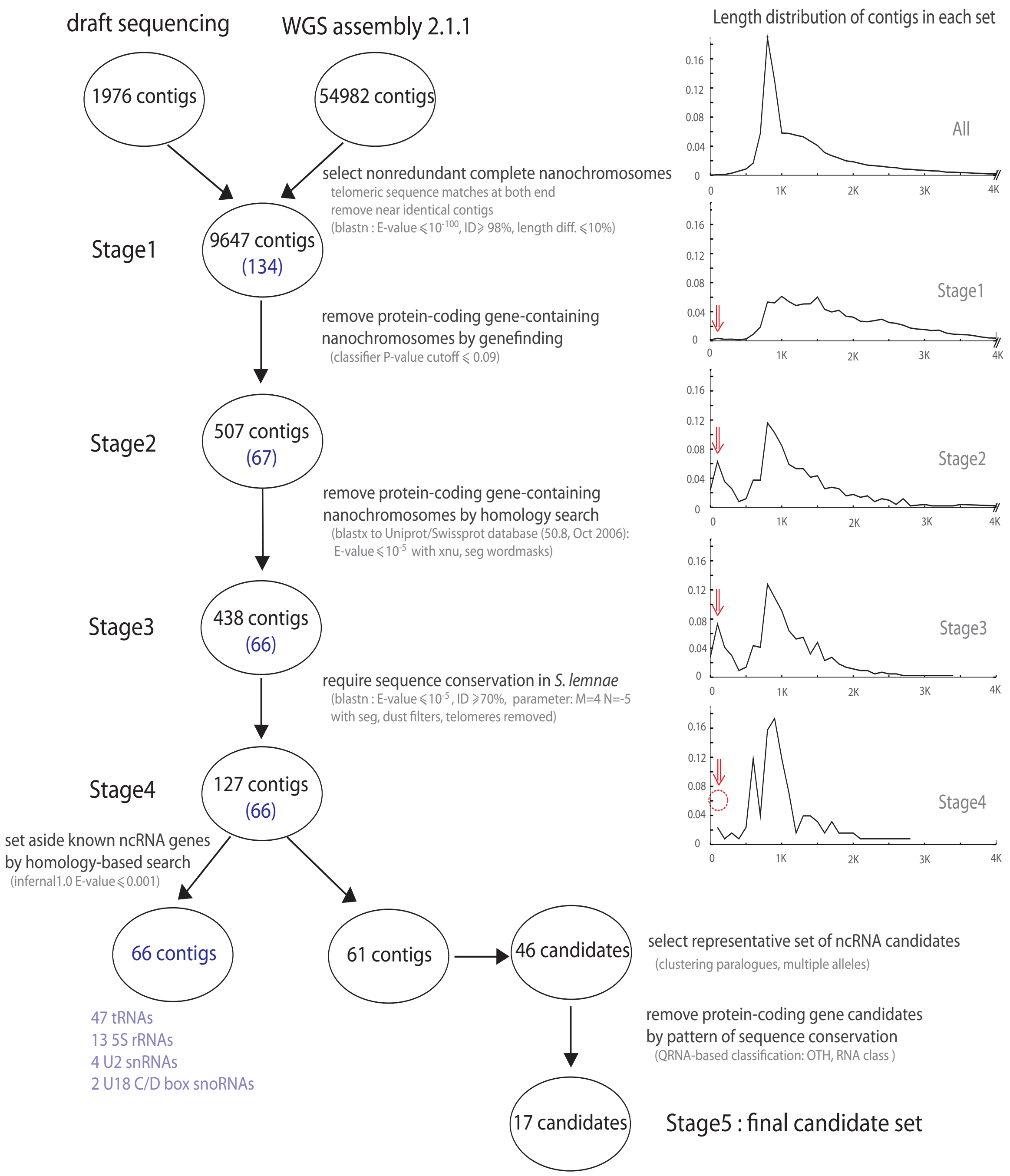

draft sequencing
WGS assembly 2.1.1
1976 contigs
54982 contigs
select nonredundant complete nanochromosomes
telomeric sequence matches at both end
remove near identical contigs
(blastn : E-value ⩽10^-100, ID⩾ 98%, length diff. ⩽10%)
Stage1
9647 contigs
(134)
remove protein-coding gene-containing nanochromosomes by genefinding
(classifier P-value cutoff ⩽ 0.09)
Stage2
507 contigs
(67)
remove protein-coding gene-containing nanochromosomes by homology search
(blastx to Uniprot/Swissprot database (50.8, Oct 2006): E-value ⩽10^-5 with xnu, seg wordmasks)
Stage3
438 contigs
(66)
require sequence conservation in S. lemnae
(blastn : E-value ⩽10^-5 , ID ⩾70%, parameter: M=4 N=-5 with seg, dust filters, telomeres removed)
Stage4
127 contigs
(66)
set aside known ncRNA genes by homology-based search
(infernal1.0 E-value ⩽ 0.001)
66 contigs
47 tRNAs
13 5S rRNAs
4 U2 snRNAs
2 U18 C/D box snoRNAs
61 contigs
46 candidates
select representative set of ncRNA candidates
(clustering paralogues, multiple alleles)
remove protein-coding gene candidates by pattern of sequence conservation
(QRNA-based classification: OTH, RNA class)
17 candidates
Stage5 : final candidate set
Length distribution of contigs in each set
0.16
0.12
0.08
0.04
0
0
1K
2K
3K
4K
All
Stage1
Stage2
Stage3
Stage4

Figure 4.1: Flowchart of the screen for noncoding nanochromosomes.

The graphs to the right show the length distribution of the dataset at each stage of the screen. Red arrows indicate a peak of small (presumably artifactual) noncoding contigs that is initially enriched, then removed when a requirement for DNA sequence conservation to *Stylonychia* is imposed.

4.1 Non-coding nanochromosome classification

To winnow the protein-coding gene-containing nanochromosomes in respect of ncRNA gene screen so as to make our screen effective, we need a coding genefinder to have high sensitivity. We are less concerned with the comprehensiveness of our screen's ability to detect ncRNA genes – the stage 1 dataset is already only a sample – so we can tolerate somewhat low specificity, which means relatively higher rate of miscalling a noncoding nanochromosome as coding so to throw it away because signal/noise ratio in terms of ncRNA genes is affected largely by the predominant protein-coding genes in gene population. We strived to develop a coding gene classifier with about 95% sensitivity and at most a 20% false positive rate, based on the following "back of the envelope" argument. Suppose there were 100 ncRNA-only nanochromosomes in the stage 1 dataset, with the balance (9547) containing one or more coding genes. At 95% sensitivity, about 475 (5%) of nanochromosomes carrying coding genes would be misclassified as noncoding. At a 20% false positive rate, 20 ncRNA nanochromosomes would be mistakenly discarded because we falsely predict a coding region on them. Thus we would find about 555 "noncoding" candidate nanochromosomes, only 80 of which contain true ncRNA genes (15%). This would be a barely tolerable signal/noise level in a candidate set that we could sort out using further computational and experimental analysis, while maintaining a reasonable sample of ncRNA genes for novel ncRNA gene discovery. We are not concerned with the detailed exon/intron accuracy of a genefinding prediction for this problem, only with the sensitivity and specificity of classification of an entire nanochromosome.

4.1.1 Test and training dataset

To evaluate how accurately these programs could distinguish coding nanochromosomes from noncoding random sequences of the same size and composition, we constructed the following training and test datasets. Here, we oriented into the protein gene detection problem, so positive data contains protein coding genes and noncoding gene containing nanochromosome is defined as a negative. For positive training and test data, we identified a dataset of nanochromosomes which contain teh conserved region to known protein. A set of 2520 nanochromosomes were identified in the stage 1 dataset as follows. First, 6702 (69%) stage 1 nanochromosomes had BLASTX hits of E $\leq 10^{-5}$ to proteins in the NR database and were considered likely to contain coding genes. To reduce redundancy at the protein similarity level, these 6702 nanochromosomes were compared all-vs-all by TBLASTX and single linkage clustered at an E-value threshold of 10^{-20}, and one nanochromosome was randomly selected from each of the 2520 clusters. Each was randomly assigned to one of ten jackknife datasets of 252 sequences each. To train *Oxytricha*-relevant model parameters, we had to partially annotate coding exon/intron structure in the positive data. The top scoring homologous protein sequence was aligned to the nanochromosome using the protein2genome program in Exonerate 1.2.0 [192]. To get fully annotated exon/intron structure, we also identified an additional training data set of all 26 annotated *Oxytricha* genes of 24 nanochromosomes in Genbank, and 33 genes manually annotated using expressed sequence tag (EST) coverage[1].

For negative test data, we generated 2500 random nanochromosome-sized sequences flanked by simulated telomere repeats. We used an HMM model consisting of three states:

[1] EST data was generated by the pilot project of protist EST program (PEP). By clustering and filtering from 3066 EST reads, 1225 sequences were obtained. Among them, only 33 sequences can be annotated to include the whole protein gene structure.

$5'$-telomere, sequence, and $3'$-telomere. Two $5'$ and $3'$ telomere states use explicit length distributions derived from the draft sequencing data set and emit a complete telomere subsequence. The sequence state emits one nucleotide at a time using 2^{nd} order Markov statistics trained on the entire stage 1 dataset.

We jackknifed the positive and negative datasets to construct ten different test sets of 252 positives and 250 negatives, leaving 90% of the positive data for training on Exonerate-annotated partial gene structures.

4.1.2 Available eukaryotic end-user trainable genefinders

To develop a high-sensitivity classifier, we first searched coding genefinding programs that are already available. Eukaryotic protein genefinders depend on species-specific statistical signals such as codon or hexamer bias, splice site signals, and intron length. *Oxytricha* genefinding also presents a special problem because it uses a variant genetic code, reading UAG and UAA codons as glutamine and only using UGA as a stop codon [193]. We surveyed available eukaryotic genefinding programs to identify programs that could deal with the ciliate genetic code, that we could easily retrain ourselves for *Oxytricha*'s statistical features, and that (ideally) we could train on limited datasets consisting of incomplete gene structures, because we have few cDNA-validated gene structures for *Oxytricha*. We chose Genezilla [194], Unveil[195], GeneID[196] and Augustus[197] for evaluation. Genezilla was the program used for genefinding by the *Tetrahymena thermophila* genome project [149], and GeneID was used by the *Paramecium tetraurelia* genome project [150]. Among those available genefinders, only GeneID program can be trained with partial gene structures owing to the simple structure and detailed description of parameter files enabling to generate it by ourseleves.

We trained GeneID ten times on a combination of the 57 human-annotated sequences with a different jackknifed training set of 90% of the positive data (2268+57=2325 sequences total). Genezilla, Unveil, and Augustus require complete gene structures for training, so we could not use the partially annotated positives for these programs. Instead we only trained these three programs once, relying exclusively on the very limited set of 57 full-length Genbank+EST annotated genes. We expected these limited training data to put these three programs at a significant disadvantage. Each genefinder was then tested ten times on jackknifed sets of 252 positives and 250 negatives for its ability to discriminate coding nanochromosomes from synthetic noncoding nanochromosome-like sequences.

Figure 4.2 shows the benchmarking results as a ROC (receiver operator characteristic) plot. None of the genefinders we tested reached our desired level of sensitivity and specificity. We suspect it is due to the dearth of well-annotated *Oxytricha* gene structures for training data. It may be possible to improve the performance of any of these genefinders on this unorthodox application, if we had expert inside knowledge of their implementation. However, we turned instead to developing our own specialized computational *Oxytricha* “nanoclassifier” algorithm and software implementation.

4.1.3 The *Oxytricha* nanoclassifier

We used hidden Markov model methodology [198] to specify a probabilistic model of *Oxytricha* nanochromosomes containing coding genes. Figure 4.2.B shows a schematic of our model. It includes standard statistical features for eukaryotic genefinding [199], such as 5th-order Markov (hexamer) statistics for residues in coding exons and an intron model

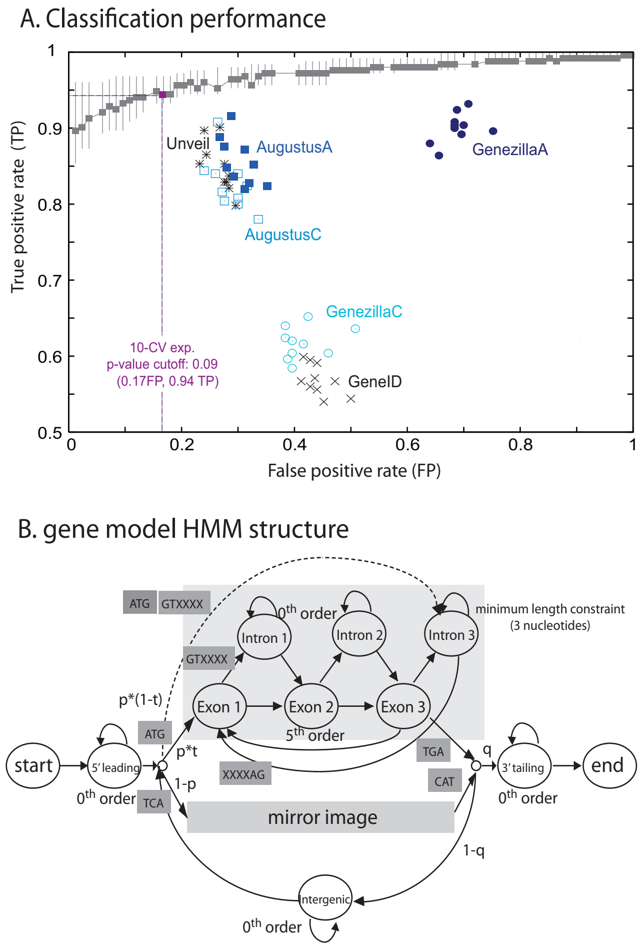


Figure 4.2: A hidden Markov model based coding nanoclassifier.

A. ROC curve of classification performance. **10CV exp** - results of ten-fold cross-validation on jackknifed test sets of 252 positive sequences and 250 negative sequences, showing the average (grey box) and range of the ten test results as P-value threshold is varied. **GeneID** - results of ten-fold cross-validation of the GeneID program, where a GeneID annotation of a complete coding gene structure is counted as a positive coding classification. **other genefinders** - each point represents a result on one jackknifed test dataset, but each of these genefinders was only trained once on a set of complete gene structures, not on the partial gene structures of the jackknifed positive training data. Unveil, AugustusC, GeneZillaC points call a complete coding gene structure annotation as a positive classification. AugustusA, GenezillaA points call a partial or a complete gene structure annotation as a positive classification. **B.** Schematic of the HMM state architecture of nanoclassifier gene model.

consisting of hexamer 5′ and 3′ splice site consensus, a frame, a minimum length, and a geometric length distribution tailing off from the minimum length. The model consists of six exon states, six intron states, 5′ and 3′ flanking sequence states and one intergenic state to allow more than one protein gene per nanochromosome in the same or opposite orientation. A start state emits an ATG (exactly), and a stop state emits a TGA codon (exactly). For intron signals, hexamer nucleotide frequencies including exact GT or AGs are estimated from the training set. We included minimum length constraints on the intron state. The overall model includes a mirror image of the coding gene model for the reverse strand, allowing more than one coding region to occur per nanochromosome on either strand. Additional states in the model generate noncoding extragenic and intragenic DNA segments, so the overall model is that of a complete full-length nanochromosome containing one or more coding genes. The background (null hypothesis) model has the same HMM state-structure as the gene model, but the emission statistics of all states are changed to background: 5^{th} order Markov (hexamer) background statistics in the exon states (estimated from the entire stage 1 dataset), and 0^{th} order background nucleotide frequencies in all other states and also for start and stop codons and GT/AG splice sites. This preserves the same length distribution for both coding and null models. If we used a different model structure for the null hypothesis, it would be hard to match overall length distributions implied by the two models, and sequences could get classified spuriously by length rather than statistical coding signals.

One advantage of this model to us is that we fully control its parameterization, and could tailor it for *Oxytricha* and for the types of partial data we had available for training. Another is that we have full control over thresholding model scores, so we can trade off sensitivity against specificity as needed.

Here we are interested in nanochromosome classification rather than genefinding which will be described in the later section. To solve classification problem more effectively, we could improve its implementation rather than simply using genefinder methodology which predicts the precise gene structure. First, it is advantageous to set up a hypothesis test with two specified models, that is, having an explicit model of the "Null hypothesis": a coding gene model and a background model generating non-coding nanochromosomes as described above. Second, it is advantageous to use the Forward algorithm rather than the Viterbi algorithm which calculates the most probable path to fit the sequence into this generative gene model to produce the parsing of the gene structure of one or more genes in a nanochromosome. The Forward algorithm calculates the probability of all probable paths to be fitted into the model, so it admits the possibility of the existence of uncertainty in parsing as opposed to one exact parsing, which is appropriate for classification problem to figure out whether protein coding genes exist or not.

Given a full-length nanochromosome sequence, we calculate a log likelihood for both models by using the HMM Forward algorithm, and report the log-odds likelihood ratio in units of nats (natural logs). A positive log-odds score indicates stronger evidence for the coding model than the null hypothesis, and the higher the score, the more evidence for coding potential. In principle, we could threshold the log-odds likelihood scores to distinguish coding from noncoding nanochromosomes, but length and residue composition effects in a given individual nanochromosome introduce biases into log-odds scores toward the coding model. To mitigate these effects, we calculate a P-value statistic for each log-odds score by order statistics (i.e. by brute force simulation), by shuffling the sequence 30,000 times by 3-mers to roughly preserve 2^{nd} order statistics, calculating a score for each shuffle, and reporting where the score of the real sequence falls in that simulated null distribution. A low P-value means higher confidence that a nanochromosome contains one or more coding regions. Classification is based on thresholding the P-value.

We tested the classification performance of our nanoclassifier using the same jackknifed training/test data used for GeneID. Figure 4.2.A shows the results for varying choices of P-value threshold. At a P-value threshold of 0.09, the average of the 10 jackknifed experiments is 94% sensitivity and 17% false positive rate. This estimated performance was acceptable for our screening strategy. We then retrained the classifier on the entire positive dataset (not just a jackknifed subset) for subsequent use.

4.1.4 The nanochromosome classification screen

The results above establish the basis for the idea that we should be able to systematically identify ncRNA genes in *Oxytricha* by computationally identifying *coding* genes in full-length nanochromosomes, and subtracting these coding nanochromosomes to leave a subset of apparently noncoding nanochromosomes for further analysis. We applied our nanoclassifier to each of the 9647 presumptive full-length nanochromosomes in the stage 1 dataset (Figure 4.1). Unexpectedly, this identified a stage 2 dataset of only 507 noncoding contigs (5.3%).

This small number is consistent with the expected false negative rate of the nanoclassifier, so many of these contigs are still likely to contain coding regions. Given the estimated sensitivity of 94% for our nanoclassifier, if all 9647 contigs were coding, we expect about 580 (6%) to pass.

4.1.5 Exclusion of “known” protein coding genes

To further increase the stringency of the screen, we used BLASTX to identify nanochromosomes with significant similarity to known proteins with “wordmask=seg, wordmask=xnu” options and with “E = 10^{-5}” threshold on UniProt/Swissprot database which contains more experimentally validated proteins in spite of its smaller size than the NR database. This process removed 69 more contigs, leaving a stage 3 dataset of 438 noncoding contigs.

This small number is surprising, and a main result of the work. If *Oxytricha* contained large numbers of ncRNA *genes*, we would expect to find large numbers of noncoding *nanochromosomes* at this stage of the screen, but we do not. Indeed, the actual number of noncoding nanochromosomes is even smaller. The 438 stage 3 nanochromosomes still include undetected coding genes and assembly artifacts, as described below. We established that ncRNA genes occur alone on single-gene nanochromosomes sufficiently often, that our nanoclassifier is sufficiently accurate, and that the stage 1 sample of full-length nanochromosomes is sufficiently representative, that this result is expected to be robust. In what follows, we exploit comparative analysis against the *Stylonychia* draft genome sequence to look deeper at this set of 438 nanochromosomes to see whether we have nonetheless sampled some interesting new ncRNA genes, and to further study possible sample biases.

4.1.6 A lightweight nanogenefinder

As mentioned in the nanoclassifier section, *O. trifallax* has different content and signal statistics for gene structures compared to other eukaryotes, so pre-trained gene-finding programs for other eukaryotes may work very poorly. For example, the Genscan [200] program, one of the conventional *ab initio* HMM gene finders, with the human parameter files performed at 0.325 sensitivity and 0.996 specificity at the nucleotide level and 0.133 sensitivity and 0.182 specificity at the "exact" exon level when tested on the *Oxytricha* 26 protein genes deposited in GenBank among the additional training dataset of nanoclassifier. A more difficult problem is the very small size of training dataset that annotated the whole gene structure. Therefore, to make our own genefinder which can be tuned on *Oxytricha* even with partial data might be advantageous.

Our genefinder, named "nanogenefinder" to reflect both the characteristics of its structural simplicity and the fact that it was trained specially for *Oxytricha* nanochromosome, is based on the same HMM gene model in nanoclassifier. The differences are an unnecessity of the background model, and using the Viterbi algorithm to calculate the probabilities from the model as described in the nanoclassifier section. The nanogenefinder, a lightweight program having simple structure and relatively small number of parameters, can be trained on a small amount of data.

The performance test result on nanogenefinder and other custom-trainable genefinders which are the same programs which were used for comparing the performance of nanoclassifier due to the same reasons as mentioned above is shown at the Table 4.1. Two sets of data both of which were used for training the genefinders as additional data in the above classification problem were also used to train and test exchangeably: NCBI and EST. NCBI dataset consists of all 26 genes on 24 nanochromosomes annotated in Genbank. It includes total 16 single exons, 10 initial/final exons, and 10 internal exons. EST dataset consists of 33 manually fully annotated genes on 33 nanochromosomes using EST data. It includes total 14 single exons, 19 initial/final exons, and 7 internal exons. As mentioned in the nanoclassifier section, because the GeneID package does not provide a training program, we generated parameter files for *Oxytricha* by hand. Rather than using the same parameter complexity as the default, we reduced the size of signal information for start and stop codon, donor and acceptor sites and changed transition probability matrix for trimer, not for hexamer, to get a better performance on this small data set.

The nanogenefinder and Augustus show relatively better performance than other programs on this small size of training and test data although their performances are not satisfying compared to that of conventional genefinders trained on other eukaryotic genome. For example, genscan reported 0.78/0.81 sensitivity/specificity at exon level and 0.93/0.93 at the nucleotide level on the 570 vertebrate gene sets [201]. The performance of Augustus trained on NCBI set and tested on EST set is similar to this level. If Augustus and nanogenefinder are trained and tested both on NCBI and EST, their performances can be achieved in the acceptable level although additional information from the partial data which have no full gene-structure annotation and are the computationally predicted is supplemented for nanogenefinder training.

Two anecdotal examples shown at Figure 4.3 illustrate the performance of genefinders. GeneZilla and GeneID have a tendency to over-predict genes and under-predict exons on the nanochromosome. The fact that a lightweight nanogenefinder shows almost equivalent performance to the more sophisticated genefinder Augustus might indicate some possible need for custom-trainable and partial training data-acceptable lightweight genefinder in the paucity of training data when a new genome is sequenced.

Table 4.1: Performance test on the genefinders

Program	Training/Test set	Sensitivity				Specficicity			
		Seq	Gene	Exon	Nuc	Seq	Gene	Exon	Nuc
GeneZilla	NCBI/EST	0.03	0.03	0.034	0.032	1.0	1.0	1.0	1.0
	EST/NCBI	0.25	0.231	0.13	0.326	0.462	0.429	0.375	0.992
Unveil	NCBI/EST	0.394	0.394	0.288	0.766	0.448	0.351	0.459	0.911
	EST/NCBI	0.269	0.269	0.174	0.504	0.368	0.259	0.286	0.986
GeneID	NCBI/EST	0.03	0.03	0.034	0.402	0.032	0.03	0.061	0.645
	EST/NCBI	0.042	0.077	0.043	0.512	0.042	0.054	0.054	0.844
Augustus	NCBI/EST	0.727	0.727	0.763	0.973	0.750	0.727	0.804	0.954
	EST/NCBI	0.333	0.308	0.5	0.762	0.348	0.333	0.242	0.991
	All/All	0.895	0.881	0.914	0.957	0.927	0.912	0.914	0.991
nanogenefinder	NCBI/EST	0.485	0.485	0.458	0.801	0.533	0.533	0.711	0.956
	EST/NCBI	0.5	0.462	0.478	0.636	0.667	0.667	0.711	1.0
	All/All	0.789	0.797	0.895	0.983	0.789	0.746	0.855	0.976

The sensitivities and specificities of each genefinder are shown in several levels: Nuc (in nucleotide level), Exon (in exon level), Gene (in whole gene level) and Seq (in whole nanochromosome level). For instance, Nuc sensitivity is the ratio of the number of correctly predicted nucleotides to the number of all annotated nucleotides, and Nuc specificity is the ratio of the number of correctly predicted nucleotides to the number of all predicted nucleotides.

4.2 Comparative analysis of the candidate nanochromosomes

We wanted to utilize comparative sequence analysis to identify conserved sequences likely to encode functional ncRNA genes, to distinguish such conserved RNA sequences from the distinctive codon-dependent conservation pattern of coding regions, to confine the possible ncRNA regions within the nanochromosome, and to assist in secondary structure prediction of any structural RNAs found. Therefore we sought the macronuclear genome sequence of another ciliate at a suitable evolutionary distance for comparative sequence analysis of *Oxytricha*. Ten stichotrich ciliate isolates were surveyed by our collaborators in Princeton University through PCR and sequencing of four conserved protein-coding genes: telomere-end binding proteins α and β, HSP70, and DNA polymerase α. The number of

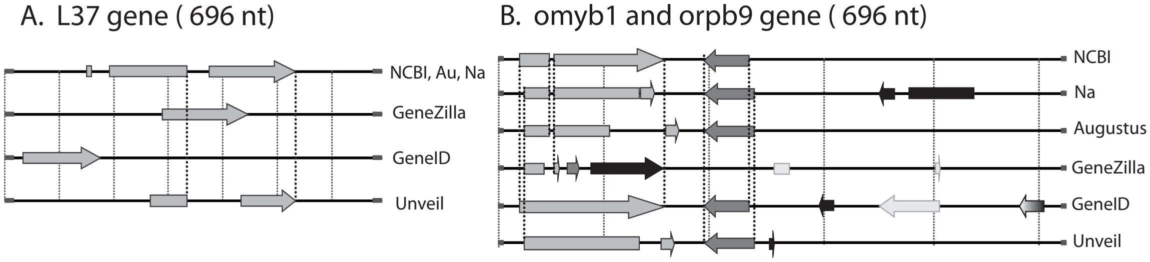


Figure 4.3: Examples of the results from genefinders

Examples of genefinder results. NCBI represents the annotation from GenBank, Au represents Augustus program, and Na represents our nanogenefinder. **A.** A small nanochromosome known to encode one gene. **B.** A long nanochromosome known to encode two genes.

substitutions observed in synonymous four-box codons in alignments to homologous *O. trifallax* sequence was used as a proxy of neutral evolutionary distance. We aimed to identify a species at about 0.4 neutral substitutions/site [202]. Two isolates (*Oxytricha fallax* and *Oxytricha* “Bath”) were too closely related, but eight isolates (“*Sterkiella histriomuscorum*”, *Oxytricha nova*, *Oxytricha* Maryland, *Stylonychia lemnae*, *Stylonychia mytilus*, *Laurentellia* sp., *Paraurostyla* sp., and *Urostyla* sp.) were all suitable, ranging from 0.3 to 0.6 substitutions/4box-site. We chose the stichotrich *Stylonychia lemnae* because it appears to have a neutral evolutionary distance of approximately 0.4 substitutions per site to *Oxytricha*, roughly comparable to mouse/human sequence comparison, a distance well suited both for detection of conserved coding exons and comparative analysis of conserved RNA structure in pairwise alignments [202].

Like *Oxytricha*, *Stylonychia* is a stichotrich, and its biology is thus comparable to that of *Oxytricha*. Because it is physically larger, it was used extensively in early cytogenetic studies of macronuclear genome development [203–205].

4.2.1 *Stylonychia lemnae* genome sequencing

We sequenced whole cell DNA obtained from *Stylonychia lemnae* strain 2x8/2, which was kindly provided by Francziska Jönsson and Hans Lipps (University of Witten, Germany). At the Genome Sequencing Center (Washington University in St. Louis), a sample was sequenced in one 454FLX run without purification of macronuclear DNA away from micronuclear DNA because macronuclear DNA is in vast excess. This produced 568,094 reads (146 Mb), about 3x average shotgun coverage of the presumed ∼50Mb macronuclear genome with reads averaging 260 nt in length. The Newbler program, which can be downloadable from [`https://valicertext.roche.com/`], assembled these reads into 53,806 contigs (27.3 Mb) ranging in size from 95 to 9947 nt.

This *Stylonychia* assembly contain 131(∼53% coverage) CEGs, so we estimate this assembly covers about 50% of the *Stylonychia* genome. We therefore expect to be able to detect approximately 50% of single-copy evolutionary conserved regions in *Oxytricha* by comparison with the *Stylonychia* dataset.

4.2.2 Sequence alignment to *Stylonychia lemnae*

We expect the steps to this point have also enriched for artifactual "noncoding" contigs that arise either from sequence assembly errors or DNA processing errors in *Oxytricha*. Figure 4.1 shows length distributions for the contigs in each dataset, which show the progressive enrichment of a peak of small (∼100 nt) contigs in the stage 3 data that are likely assembly artifacts due to false overlaps in low-complexity subtelomeric sequence. To enrich for nanochromosomes containing functional genes, we screened for contigs with significant DNA similarity to our *Stylonychia* shotgun data. We produced pairwise

Oxytricha/*Stylonychia* local alignments using WU-BLASTN with options “filter=seg filter=dust maskextra=10 M=4 N=-5” to detect relatively short alignments, which partially resulted from low-continuity shotgun assembly of *Stylonychia* genome sequence and relatively small size of ncRNA genes, and selected alignments of $\geq$ 70% identity and E $\leq 10^{-5}$. This identified a dataset of 127 conserved noncoding nanochromosomes (stage 4 dataset; Figure 4.1). The peak of small contigs disappears in this stage.

The stage 4 dataset is highly enriched for nanochromosomes carrying known ncRNA genes (66/127, 52%). Although it is possible that these 66 nanochromosomes contain additional novel ncRNA genes, we excluded them from further analysis, leaving a set of 61 conserved noncoding nanochromosomes with no significant similarity to known ncRNA genes. These are candidates for harboring novel ncRNAs.

Several of these appeared to be quasialleles or paralogs of each other. We clustered the 61 nanochromosomes by sequence similarity and chose a representative set of 46 distinct loci. This clustering included both identifying “quasialleles” (11 contigs were considered to be quasialleles of others), and also clustering obvious paralogs together. In particular, 9/61 of the nanochromosomes at this stage represent one family of ncRNAs which will be described in the next chapter. Five of them are distinct loci (Onc91, Onc92, Onc94, Onc95, Onc96) after clustering quasialleles. After clustering paralogs by sequence similarity, two of these nanochromosomes were chosen as representative (Onc91 represents a cluster including Onc92; Onc94 represents Onc95 and Onc96).

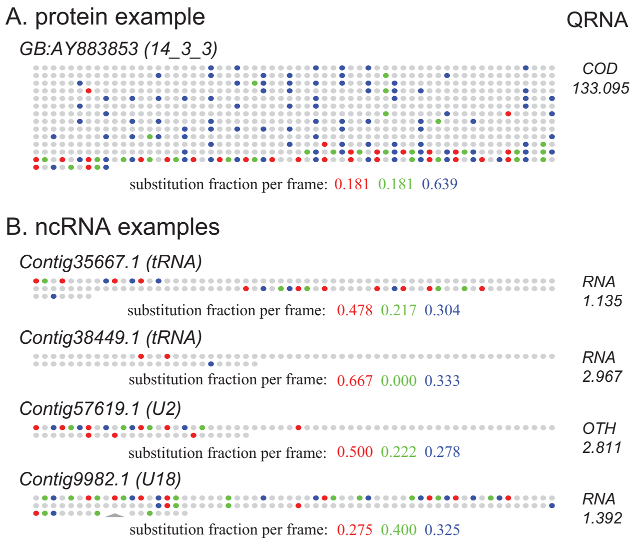


Figure 4.4: Comparative sequence analysis of sequence regions conserved with *Stylonychia*,

Examples of one known protein and four ncRNA genes were shown. Starting with a BLASTN alignment of homologous *Oxytricha* and *Stylonychia* sequences, each residue represent each colum of the alignment, and residue substitution events in three frames are colored into red, green, and blue. A whole sequence alignment is splitted into several lines to display 60 characters per line. QRNA[126] classification results for each alignment are shown in the right column, showing the best scoring class ("COD" for coding, "RNA" for structural RNA, and "OTH" for other) and a QRNA classification log-odds score in bits. Coding regions generally stand out both visually and by QRNA because of the periodicity of three in their substitution events (i.e. the predominance of one color in a large region of the protein example).

4.2.3 Analysis of alignment patterns

Despite all the steps taken so far, we still expect that more than half of these 46 contigs will carry coding genes that we have failed to recognize. Given that the BLASTX step at stage 3 removed 69 contigs, and we expect from the previous section that about 65% of *Oxytricha* proteins have significant similarity to known proteins, then we expect approximately 37 coding regions to pass into stage 4. About 70% (26) of these would pass the stage 4 conservation screen against the incomplete *Stylonychia* dataset. Therefore, as a final step to remove coding nanochromosomes, we used the pattern of residue substitution observed in the region of DNA sequence conservation with *Stylonychia*.

Because we selected *Stylonychia* to be at a neutral distance of about 0.4 substitutions per site, we expect substitutions in many near-neutral codon third positions, and thus a distinctive periodicity of three is seen in the pattern of observed substitutions in conserved coding regions. Figure 4.4 shows examples of this periodic pattern in a known coding region as opposed to known ncRNA genes. This pattern can be evaluated by brute force comparing of numbers of the substitution in each codon position or by the RFC (reading frame conservation) test [206] that measures the portion of nucleotides in a fully aligned interval for which the reading-frame has been locally conserved. The RFC score calculation was implemented by a Perl program according to the published description [206]. As shown on Figure 4.5, it is not easy to discriminate the protein-coding gene like alignments in *Oxytricha* and *Stylonychia* alignment, using RFC.

Instead, we scored this pattern by running “eqrna -a” to the pairwise BLASTN alignments processed by “blastn2qrnadepth.pl” in the QRNA [126] package, which has explicit probabilistic protein-coding gene (codon position-dependent alignment), non-coding gene (ncRNA structure-constrained alignment) and other class (position-independent alignment) models for the pairwise alignment. Although QRNA was originally designed to identify structural ncRNAs by comparative analysis, which is a task that remains difficult with high false positive rate, its statistical model for discriminating conserved coding regions from other types of sequence conservation is effective. Performance benchmarking of QRNA’s ability to detect coding nanochromosomes was done using the same 2520 presumptive coding sequences in the positive test data for the nanoclassifier 10CV data set. A total of 1517 pairwise alignments passed the above criteria. After splitting long alignments into alignments of a maximum length of 1000 columns, 5033 pairwise alignments were used as a positive data set. For negative data, we produced the simulated noncoding

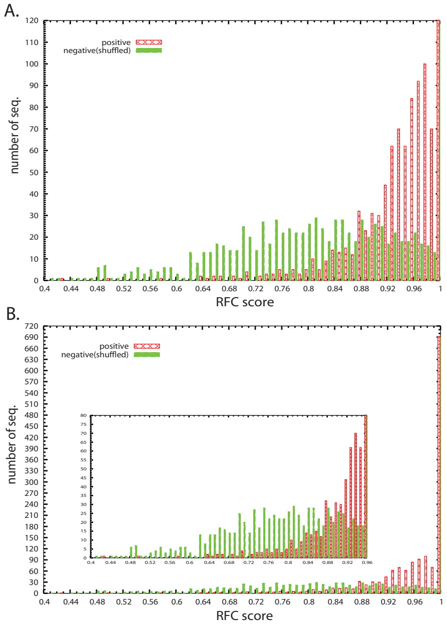


Figure 4.5: RFC score distribution

The graphs show the distribution of RFC scores in the positive data set and the negative dataset (which consists of the shuffled alignments of the positive data set). The positive dataset is a set of BLASTN alignments to *Stylonychia* of the same dataset used at 10-fold CV performance test of nanoclassifier. **A.** The average RFC score of the multiple alignment segments of a sequence is used as a representative. **B.** The best RFC score among the multiple alignment segments of a sequence is selected as a representative.

Oxytricha/Stylonychia alignments by shuffling the pairwise alignments by columns thus preserving mean base composition and percent identity. For longer alignment, it is split into several pieces to match up the limitation of alignment length in QRNA program. On this dataset, we estimated that QRNA has a true positive rate of 95% and a false positive rate of 3% for distinguishing conserved coding regions.

QRNA classified the conserved regions in 29 of the 46 contigs (63%) as probable coding regions, consistent with our statistical expectation. The final candidate set (stage 5) consists of 17 representative, distinct, conserved, full-length, apparently noncoding nanochromosomes.

Chapter 5

Analysis of the final candidate nanochromosomes

5.1 Experimental validation of ncRNA gene predictions

To test whether our 17 candidate nanochromosomes express RNA transcripts from the identified regions of sequence conservation, we performed Northern blots. As positive controls, we also performed Northerns for 13 homologs of known ncRNAs (8 C/D snoRNAs, 4 tRNAs, and one U2 RNA locus) and identified small RNA transcripts of the expected size for all 13. Images of these Northerns (Figure D.1) and probe sequences (Table D.1) are provided in Appendix D. For positive candidates on Northerns, we performed 5′ and 3′ RACE-PCRs (Rapid Amplification of cDNA Ends) and sequenced multiple clones from each in order to define complete transcript sequences.

5.1.1 *O. trifallax* culture and RNA extraction

O. trifallax strain JRB310 [207] was cultured to ≥ 5000 cells/ml density in 8x12 inch Pyrex dishes with 300 ml ciliate medium. They were fed a mixture of an alga (*C. elongatum*, University of Texas) grown in 500 ml flasks under light in Euglena medium, and bacteria (*K. pneumoniae*). *O. triallax* cultures was split into two cultures with fresh vegetatively grown *Oxytricha* cells were collected on several layers of gauze to exclude clumps of algae, then filtered on a Nitex nylon membrane to get rid of bacteria and culture medium. For long term storage, inactive cyst forms of *Oxytricha* cells that were generated by starvation of the collected cells were stored at $-70\,^{\circ}\mathrm{C}$ after mixing with equal volume of 20% DMSO. To hatch up *Oxytricha* cells from cysts, cysts were diluted in an excess of ciliate medium several times and fed. For RNA extraction, the medium of the collected cells is brought to 0.05M EDTA to immobilize the motile cells and to reduce RNase activity, and cells were collected by centrifugation at $4\,^{\circ}\mathrm{C}$. Total RNA was extracted by a standard Trizol (Invitrogen) protocol that uses chloroform for phase separation and propanol for RNA precipitation, and stored in 1mM EDTA at $-20\,^{\circ}\mathrm{C}$.

5.1.2 Testing the existence of RNA transcripts: Northern blot

Northern blot was initiated with runing $2 - 10\ \mu\mathrm{g}$ of total *Oxytricha* RNA on 4% acrylamide gels. Then, the gel was electroblotted using a semi-dry electrophoretic transfer unit (BioRad), and UV crosslinked to a ZetaProbe charged membrane (BioRad). DNA oligonucleotide probes (38-44nt) were end-labeled with γP^{32}-ATP using T4 polynucleotide kinase and hybridized to the Northern blots in UltraHyb Oligo solution (Ambion) at $42\,^{\circ}\mathrm{C}$ for overnight at least 15 hours. Blots were washed twice in a solution of 2X saline-sodium citrate (SSC) buffer and 0.1% sodium dodecyl sulfate (SDS) solution at $55\,^{\circ}\mathrm{C}$ for 5 minutes and 15 minutes, then again twice in 0.1X SSC and 0.1% SDS solution at $55\,^{\circ}\mathrm{C}$ for 5 minutes and 15 minutes. Blots were either visualized by phosphorimager (Amersham Biosciences), or exposed at least one day on X-ray film (FujiFilm) at $-80\,^{\circ}\mathrm{C}$. For some probes with lower calculated melting temperatures, an additional Northern blot was performed using less stringent hybridization and washing temperatures: $37\,^{\circ}\mathrm{C}$ for hybridization, $42\,^{\circ}\mathrm{C}$ for washing. A ^{32}P-labeled 50bp dsDNA ladder (New England BioLabs) was used for molecular weight standards.

Each candidate nanochromosome was tested with one or several single-stranded oligonucleotide probes directed against the conserved and/or relatively high GC-content regions of the nanochromosome. We probed each strand separately to decipher the strand specificity, using total RNA extracted in vegetative growth condition as described above. For 7 of the 17 candidates we detected small RNA transcripts (Figure 5.1). All probe sequences are listed in Appendix D.

5.1.3 Verifying the ends of transcripts: RACE-PCR

To conduct RACE experiments on the Northern postive candidates, poly-A tails were added to total RNA by terminal deoxynucleotidyl transferase to sample all RNA species in the following RACE protocols, which provide the anchor for transcript polymerization. We used two different commercial RACE protocols: SMART-RACE (Clontech) and GeneRacer (Invitrogen). The SMART-RACE 5′ RACE protocol relies on the addition of 3-5 untemplated C residues at the 3′ end of the first-strand cDNA synthesized by reverse transcriptase. This protocol is less efficient in our hands, but may be relatively insensitive to unusual 5′RNA

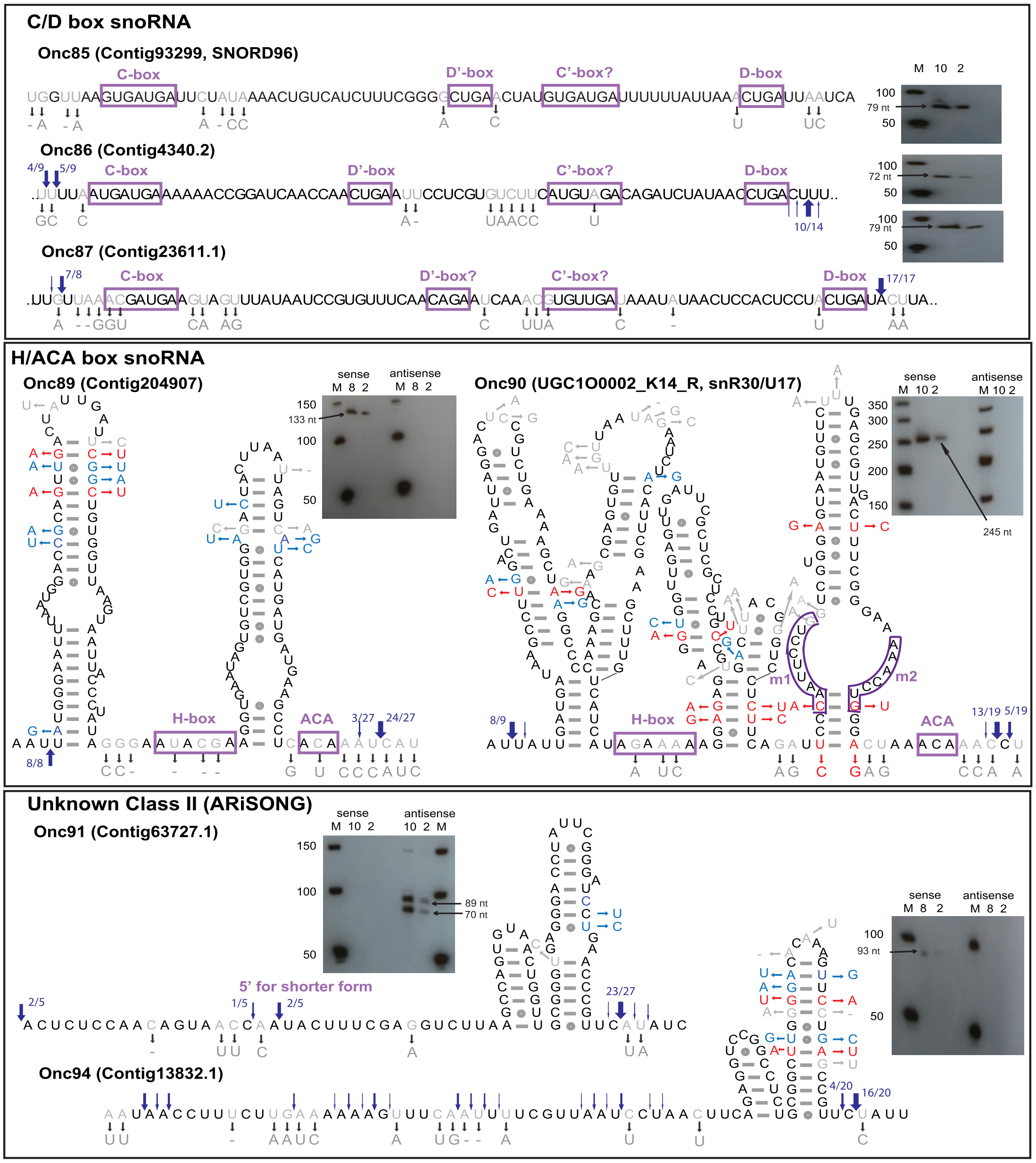


Figure 5.1: Experimental confirmation of candidate ncRNA transcripts.

Sequences, predicted secondary structures, and Northern and RACE data for seven candidates with detected transcripts. Genomic sequences are shown for each locus, with 5′ and 3′ ends of transcripts determined by RACE indicated by dark blue arrows. Arrows pointing between nucleotides indicate an unambiguously determined end; arrows pointing at an 3′ end nucleotide A indicate an ambiguity where we cannot distinguish an nucleotide A in the native transcript from the artificially appended poly-A tail. For Northern blots, 10/2 or 8/2 lanes indicate the amount of total RNA loaded in each lane (in μg). M indicates a radiolabeled 50bp DNA ladder. "Sense/antisense" refers to the orientation of probes on the reference genome sequence, not the transcipt. For C/D box snoRNAs, only one probe was tested because we predicted the correct strand by sequence analysis. Secondary structures of transcript were initially predicted by RNAalifold [208] then manually modified based on comparative sequence analysis and other features (such as the predicted target sites for the two H/ACA box snoRNAs). Conservation of sequence and structure in *Stylonychia* alignments is annotated using a color scheme, with red indicating a compensatory base pair substitution that supports the structure prediction, blue indicating a wobble base pair substitution consistent with the structure prediction, and gray indicating all other substitutions, including those in single-stranded regions and those that are inconsistent with the structure prediction.

structure. The GeneRacer 5′ RACE protocol ligates an RNA oligo to the 5′ phosphate end of an RNA transcript and uses that oligo as the amplification annealing site. Because this protocol is optimized for capped mRNA transcripts, we skipped the first phosphatase step and added a kinase step to assure that ncRNAs with a variety of possible capped and uncapped 5′ ends could be amplified. Both kits use the same approach for 3′ RACE, using one oligo-dT primer against the added poly-A tail, and one gene-specific primer internal in the transcript. RACE-PCR products were cloned and sequenced by standard methods.

We conducted RACE experiments on 6 of the 7 candidates detected by Northern, and one C/D snoRNA, Onc85, was not examined because we had it classified as a "known RNA" by its weak SNORD96 homology at the time we designed the RACE experiments. In each case, except for the indeterminate 5′ ends of two loci described in the next chapter, transcript sequence(s) implied by RACE-PCR sequencing were consistent with the band(s) observed by Northern (Figure 5.1). All gene-specific probe sequences of the tested candidate genes are listed in Appendix E.

5.2 Manual examination

We analyzed each of the 17 candidate loci in detail, taking particular advantage of the pattern of *Stylonychia* conservation including multiple alignments where possible. *Stylonychia* conservation patterns of the final candidate dataset are shown in Figure F in Appendix F. For ncRNA loci that appear to conserve an intramolecular RNA secondary structure, we used manual comparative analysis to infer the structure based on the RNAalifold [208] prediction .

Of the ten candidates for which we detected no small RNA expression, upon detailed examination, five contain small fragments of coding genes found on other nanochromosomes. These nanochromosomes possibly arose as assembly errors or errors in macronuclear DNA processing. Two more appear to be fragments of nanochromosomes containing pieces of conserved promoter sequence. Another has only a small patch of conservation. Finally, 2 of these 10 candidates (Onc98, Onc106) have well-conserved regions that appear to be plausible ncRNA genes, but because we did not observe any expression from these loci, we cannot be sure of the bounds (or mature RNA sequence) of any transcript. We do not consider them to be confirmed ncRNA loci.

Of the seven candidates for which we did detect small RNA expression, five are snoRNAs: three C/D snoRNAs (Onc85, Onc86, Onc87) and two H/ACA snoRNAs (Onc89, Onc90) (Figure 5.1). The C/D snoRNAs and the Onc89 H/ACA snoRNA have typical structures for these classes of eukaryotic snoRNAs. Those snoRNAs are also detectable by snoscan and snoGPS programs [134, 209] although the prediction score of Onc89 H/ACA snoRNA is too low to be discernible from a noise. The predicted rRNA methylation sites of C/D snoRNA Onc85, Onc86 and Onc87 are LSU832, SSU8 and/or LSU1242, and LSU678, respectively. The Onc89 H/ACA snoRNA's predicted target site is LSU2517; however, its homologous position is a C residue in human and not pseudouridylated in yeast. A 3′ box regulatory motif which will be described in detail in the next chapter is detected in the downstream of Onc89 H/ACA snoRNA. The Onc90 H/ACA snoRNA has an unusual and distinctive structure, with large helices inserted in positions that H/ACA snoRNAs are known to tolerate additional helices (known as the IH1 and IH2 locations; [210]). Based on this unusual structure and the conservation of distinctive sequence elements (m1 and m2) in a bulge in the 3′-most stem, Onc90 is likely to be the *Oxytricha* homolog of the "ubiquitous" eukaryotic U17/snR30 H/ACA snoRNA. This is the only H/ACA snoRNA that does not function as a pseudouridylation guide, but instead is involved in rRNA processing via a presumed interaction with SSU rRNA [211, 212]. The proposed interaction for yeast snR30 and human U17 with their cognate SSU rRNAs is conserved for Onc90 with *Oxytricha* SSU rRNA (not shown) [213].

The remaining two candidates that show small RNA expression (Onc91 and Onc94) share a well-conserved predicted RNA structure (Figure 5.1), but are not detectably homologous to any well-known eukaryotic small RNA families. A detailed description of these ncRNAs is given in the next chapter.

Chapter 6

Arisong ncRNA gene class

This computational screen found two related novel ncRNA genes: Onc91 and Onc94. Their RNA transcripts exhibit well-defined bands on Northern blots, and their 3′ ends were readily mapped by RACE-PCR. However, we had difficulty obtaining 5′ RACE-PCR products for them, and the products we did obtain mapped diffusely and failed to define consistent 5′ ends. We are unsure whether this represents mere technical failure although we had much less difficulty with other RNAs, or if it reflects a peculiarity of the structure of these RNAs that might interfere with a 5′ RACE protocol, such as a lariat structure or an unusual 5′ cap although we used two different 5′ RACE protocols, one of which should be insensitive to unusual 5′ end structure. Onc91 showed two distinctive Northern bands of approximately equal intensity, while Northern of Onc94 revealed a single apparent band. Neither are consistent with the 5′ end variance detected by RACE experiments.

In primary sequence level, Onc91 and Onc94 are 49% identical [1] over the region pre-

[1] There is no consensus on exactly how to calculate the percent identity between two sequences. Here, we showed the ClustalW2 [214] score in the primary sequence alignment by neglecting the secondary structure conservation; explicitly, this score (in DNA) is similar to the percentage of the exactly matched columns in the whole alignment.

sented in Figure 6.1 in which most of variable 5′ end sequences are removed, and neither can detect the other in a genomic BLASTN search with “W=5 E=10” option. However, they share a common secondary structure, and in both cases, determining the mature 5′ end has been problematic. Therefore, novel ncRNA Onc91 and Onc94 genes seem to belong to the same small RNA family despite the primary sequences dissimilarity. We named these the “Arisong” family of ncRNAs. “Arisong” is named after a Korean verb “arisong hada” that means to be in an unsure or confusing status. This Arisong ncRNA family is expanded into related ciliate genomes by an iterative search of homologs and investigated their conserved secondary structure and regulating elements. To gain some understanding of their function, we attempted to disrupt one or several ncRNAs in this family through RNAi knock-down. In the absence of clear results, we speculated about the possible functions of Arisong RNA genes based on the shared features.

6.1 Homologous gene search

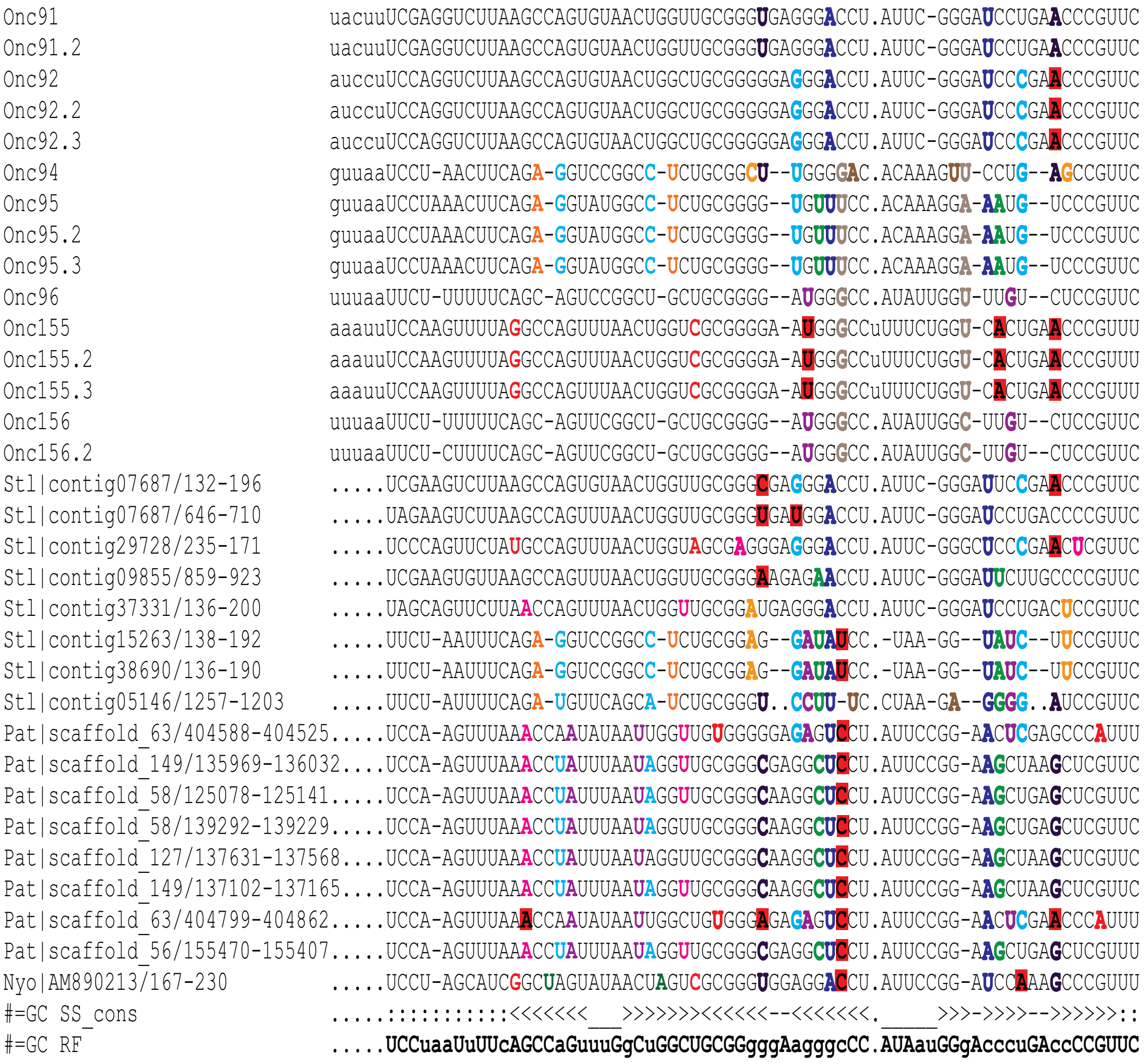
We used Infernal [129] and BLASTN to iteratively search for additional Arisong RNA homologs in the *Oxytricha* genome, our *Stylonychia* genome, and other available ciliate genome sequences: *Tetrahymena thermophila* (Nov06 version) [149], *Paramecium tetraurelia* (Dec06(v1) version) [150, 215] and *Nyctotherus ovalis* [216]. We started with a modified RNAalifold [208] RNA secondary structure alignment prediction of Onc91, Onc94 and their quasialleles to build a covariance model for the conserved secondary structure. As indicated above, we have not been able to determine a clear and consistent 5′ end by RACE experiments. Further, the 5′ end sequence and also its length do not appear to be conserved among family members. Thus, we chose to model only the conserved two consecutive stem-loop structure regions in the form of profile seed alignments by using "cm-build" program in the Infernal package. As new homologs were identified, we refined our covariance model of the consensus secondary structure, and tried to various models with different combination of Arisong RNA sequences to search for new homologs. We found a total of 15 Arisong loci in the entire *Oxytricha* dataset that include partially assembled nanochromosomes. These loci can be clustered into 7 distinct loci (Onc91, Onc92, Onc94, Onc95, Onc96, Onc155, Onc156) that appear to be paralogous rather than allelic. The peculiar nature of the spirotrich macronuclear nanochromosomes make it difficult to make a definitive identification of alleles versus paralogs. However, subtelomeric sequence flanking each locus supports an organization of paralogous genes (data not shown). We found 8 loci in *Stylonychia*, 8 in *Paramecium*, and 1 in *Nyctotherus*. Six of the 8 *Paramecium* loci have been previously predicted to be small RNA genes called PM01_1-6 by computationally identifying the RNA polymerase II signature such as a USE (upstream sequence element) and TATA box of conserved sequences [217]. Oddly, we could not detect any Arisong RNA homologs in another ciliate, *Tetrahymena*, even though the evolutionary distance between *Oxytricha* and *Tetrahymena* is similar to that of *Oxytricha* and *Paramecium*. A total of 32 Arisong RNA genes was retrieved.

6.2 Consensus structure and regulatory elements

All of these loci were predicted to share a consensus secondary structure consisting of two coaxially-stacked stems, with a highly conserved and well defined 3′ GUUC tail, and a highly variable 5′ end (Figure 6.2). The length of the first stem is more conserved than that of the second stem; it consists of six or seven base pairs. The second stem has bulges and an internal loop of variable size. But the base region of this stem, which is part of the stacked junction, is relatively conserved in the length as well as the primary sequence. The high primary sequence conservation of the stacked junction region leads us to conjecture that this coaxially-stacked dumbbell structure could be a functionally important region and might indicate a protein binding site. This proposed structure is well-supported by a number of compensatory changes in base pairs observed in the multiple alignment (Figure 6.1).

All seven distinct *Oxytricha* Arisong loci are flanked on their nanochromosome by two conserved motifs. A 17 nt motif TgACCCATnAAnnnTTA occurs about 50-60 nt upstream of the putative 5′ end of Arisong genes. This motif is likely to be the Proximal Sequence Element (PSE) found upstream of spliceosomal RNA genes in many organisms [219, 220] including other ciliates [221]. A 19-20 nt motif AAAnGAAAnnGTTTGATTAg occurs 5-10 nt downstream of the putative 3′ end of Arisong RNA. This motif is likely the functional analog (if not the homolog) of the 3′ box motif which is responsible for 3′ end processing in snRNAs and other small RNAs in many organisms [222]. Other ncRNA genes such as RNase P, telomerase RNA, and SRP RNA have upstream PSE elements. Moreover, most splicesomal snRNAs and U3 snoRNA are flanked by both of PSE and 3′ box elements, except for U6 and U6atac which show the hallmark T$_n$ terminator of RNA polymerase III-transcribed small RNAs instead of a 3′ box element. This suggests that the Arisong loci are transcribed and processed similarly to spliceosomal RNAs, probably by RNA polymerase II.

This consensus PSE element is much shorter than that of *C. elegans* [219]. However, the core conserved “ACCC” in our Arisong PSEs (Figure 6.2) is relatively well conserved also in *C. elegans*. Additionally, their relative distance from the start site of transcript (60-70 nt upstream) is similar. As noted by Chen and colleagues [217], Arisong RNAs in *Parame-*



```
Onc91                         uacuuUCGAGGUCUUAAGCCAGUGUAACUGGUUGCGGGUGAGGGACCU.AUUC-GGGAUCCUGAACCCGUUC
Onc91.2                       uacuuUCGAGGUCUUAAGCCAGUGUAACUGGUUGCGGGUGAGGGACCU.AUUC-GGGAUCCUGAACCCGUUC
Onc92                         auccuUCCAGGUCUUAAGCCAGUGUAACUGGCUGCGGGGGAGGGACCU.AUUC-GGGAUCCCGAACCCGUUC
Onc92.2                       auccuUCCAGGUCUUAAGCCAGUGUAACUGGCUGCGGGGGAGGGACCU.AUUC-GGGAUCCCGAACCCGUUC
Onc92.3                       auccuUCCAGGUCUUAAGCCAGUGUAACUGGCUGCGGGGGAGGGACCU.AUUC-GGGAUCCCGAACCCGUUC
Onc94                         guuaaUCCU-AACUUCAGA-GGUCCGGCC-UCUGCGGCU--UGGGGAC.ACAAAGUU-CCUG--AGCCGUUC
Onc95                         guuaaUCCUAAACUUCAGA-GGUAUGGCC-UCUGCGGGG--UGUUUCC.ACAAAGGA-AAUG--UCCCGUUC
Onc95.2                       guuaaUCCUAAACUUCAGA-GGUAUGGCC-UCUGCGGGG--UGUUUCC.ACAAAGGA-AAUG--UCCCGUUC
Onc95.3                       guuaaUCCUAAACUUCAGA-GGUAUGGCC-UCUGCGGGG--UGUUUCC.ACAAAGGA-AAUG--UCCCGUUC
Onc96                         uuuaaUUCU-UUUUUCAGC-AGUCCGGCU-GCUGCGGGG--AUGGGCC.AUAUUGGU-UUGU--CUCCGUUC
Onc155                        aaauuUCCAAGUUUUAGGCCAGUUUAACUGGUCGCGGGGA-AUGGGCCuUUUCUGGU-CACUGAACCCGUUU
Onc155.2                      aaauuUCCAAGUUUUAGGCCAGUUUAACUGGUCGCGGGGA-AUGGGCCuUUUCUGGU-CACUGAACCCGUUU
Onc155.3                      aaauuUCCAAGUUUUAGGCCAGUUUAACUGGUCGCGGGGA-AUGGGCCuUUUCUGGU-CACUGAACCCGUUU
Onc156                        uuuaaUUCU-UUUUUCAGC-AGUUCGGCU-GCUGCGGGG--AUGGGCC.AUAUUGGC-UUGU--CUCCGUUC
Onc156.2                      uuuaaUUCU-CUUUUCAGC-AGUUCGGCU-GCUGCGGGG--AUGGGCC.AUAUUGGC-UUGU--CUCCGUUC
Stl|contig07687/132-196       .....UCGAAGUCUUAAGCCAGUGUAACUGGUUGCGGGCGAGGGACCU.AUUC-GGGAUUCCGAACCCGUUC
Stl|contig07687/646-710       .....UAGAAGUCUUAAGCCAGUUUAACUGGUUGCGGGUGAUGGACCU.AUUC-GGGAUCCUGACCCCGUUC
Stl|contig29728/235-171       .....UCCCAGUUCUAUGCCAGUUUAACUGGUAGCGAGGGAGGGACCU.AUUC-GGGCUCCCGAACUCGUUC
Stl|contig09855/859-923       .....UCGAAGUGUUAAGCCAGUUUAACUGGUUGCGGGAAGAGAACCU.AUUC-GGGAUUCUUGCCCCGUUC
Stl|contig37331/136-200       .....UAGCAGUUCUUAACCAGUUUAACUGGUUGCGGAUGAGGGACCU.AUUC-GGGAUCCUGACUCCGUUC
Stl|contig15263/138-192       .....UUCU-AAUUUCAGA-GGUCCGGCC-UCUGCGGAG--GAUAUCC.-UAA-GG--UAUC--UUCCGUUC
Stl|contig38690/136-190       .....UUCU-AAUUUCAGA-GGUCCGGCC-UCUGCGGAG--GAUAUCC.-UAA-GG--UAUC--UUCCGUUC
Stl|contig05146/1257-1203     .....UUCU-AUUUUCAGA-UGUUCAGCA-UCUGCGGGU..CCUU-UC.CUAA-GA--GGGG..AUCCGUUC
Pat|scaffold_63/404588-404525.....UCCA-AGUUUAAACCAAUAUAAUUGGUUGUGGGGGAGAGUCCU.AUUCCGG-AACUCGAGCCCAUUU
Pat|scaffold_149/135969-136032....UCCA-AGUUUAAACCUAUUUAAUAGGUUGCGGGCGAGGCUCCU.AUUCCGG-AAGCUAAGCUCGUUC
Pat|scaffold_58/125078-125141.....UCCA-AGUUUAAACCUAUUUAAUAGGUUGCGGGCAAGGCUCCU.AUUCCGG-AAGCUGAGCUCGUUC
Pat|scaffold_58/139292-139229.....UCCA-AGUUUAAACCUAUUUAAUAGGUUGCGGGCAAGGCUCCU.AUUCCGG-AAGCUGAGCUCGUUC
Pat|scaffold_127/137631-137568....UCCA-AGUUUAAACCUAUUUAAUAGGUUGCGGGCAAGGCUCCU.AUUCCGG-AAGCUAAGCUCGUUC
Pat|scaffold_149/137102-137165....UCCA-AGUUUAAACCUAUUUAAUAGGUUGCGGGCAAGGCUCCU.AUUCCGG-AAGCUAAGCUCGUUC
Pat|scaffold_63/404799-404862.....UCCA-AGUUUAAACCAAUAUAAUUGGCUGUGGGAGAGAGUCCU.AUUCCGG-AACUCGAACCCAUUU
Pat|scaffold_56/155470-155407.....UCCA-AGUUUAAACCUAUUUAAUAGGUUGCGGGCGAGGCUCCU.AUUCCGG-AAGCUGAGCUCGUUU
Nyo|AM890213/167-230          .....UCCU-AGCAUCGGCUAGUAUAACUAGUCGCGGGUGGAGGACCU.AUUCCGG-AUCCAAAGCCCGUUU
#=GC SS_cons                  .....:::::::::::<<<<<<<___>>>>>>><<<<<<--<<<<<<<._____>>>->>>>-->>>>>>::
#=GC RF                       .....UCCuaaUuUUcAGCCaGuuuGgCuGGCUGCGGgggAagggcCC.AUAauGGgAcccuGAccCCGUUC
```

Figure 6.1: Sequence alignment of the Arisong RNA genes

The sequence portions which correspond to the conserved secondary structure (two consecutive stems) of all ciliate Arisong RNA genes were aligned in a Stockholm format. Compensatory changes supporting the secondary structure are indicated by color in pair, and inconsistent changes in the conserved common secondary structure are indicated by red filled boxes.

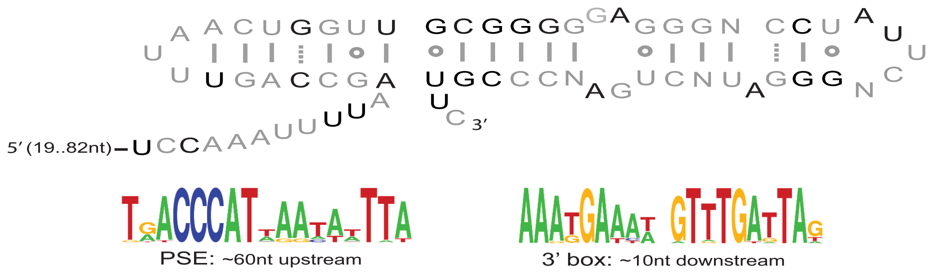


Figure 6.2: Consensus secondary structure of the Arisong RNAs and their flanking regulatory elements.
The structure shown is extrapolated from the individual structures of Arisong RNAs (examples are shown in Figure 5.1). Each sequence is the majority-rule sequence consensus of a multiple alignment of 32 Arisong RNAs. Highly conserved residues (identical in $\geq 80\%$ of aligned sequences) are shown in black; variable residues (identical in $< 50\%$) are shown as “N”; weakly conserved residues are in grey. Dotted lines for base pairs indicate that not all sequences conserve those base pairs at that position. Consensus motifs for the PSE and $3'$ box regulatory elements were generated from multiple alignments using the WebLogo program [218], after removing columns containing $> 50\%$ gaps.

cium were flanked by typical transcription and processing signals of RNA polymerase II-transcribed small RNAs, a USE, TATA box and a downstream distal element (DE). The *Paramecium* Arisong RNA predictions and snRNAs have identical USE sequences and have a TATA box at around 18 nt downstream of a USE. However, the PSE of *Oxytricha* Arisong RNA has sequence variability and does not include an obvious downstream TATA box. A PSE sequence very similar to *Oxytricha* also can be detected at 60 nt upstream of Arisong genes in *Stylonychia* and *Nyctotherus* for cases where sufficient upstream sequence is available. *Paramecium* Arisong RNAs also have an identical downstream DE at 4-5nt downstream from the conserved “GTTY” $3'$ end, which partially overlaps with the *Oxytricha* $3'$ box element, thus suggesting a different DE consensus.

6.3 Possible functions of Arisong RNA genes

The Arisong RNA has no detectable homology to any “known” ncRNA gene families (in Rfam). We attempted to determine the functions of Arisong RNA genes experimentally and conceptually.

6.3.1 Gene knock-down - RNAi

As we began to investigate the functionality of Arisong RNA, we attempted to knock down Arisong RNA expression by RNAi. This approaches has been adapted first for *Paramecium* [223, 224] and *Tetrahymena* [225], and recently for several protein-coding genes in spirotrichous ciliates [226, 227] as well as specific long RNA species transcribed from the macronucleus during conjugation in *Oxytricha trifallax*[161], though not for small ncRNA genes in vegetative growth condition (L. Landweber, personal communication).

Using five Arisong RNA genes in the stage 4 dataset (Onc91, Onc91.2, Onc94, Onc95, Onc96), we designed seven knock-down experiments that include knocking down each Arisong RNA genes individually, all five Arisong RNA genes in combination, and a negative control. We selected a 500 nt partial sequence of bacteriophage λ DNA that has no significant similarity to *Oxytricha* genome as a negative control. PCR fragments of almost full-length regions of individual Arisong RNA gene were inserted into L4440 plasmid to transform HT115 *E. coli*, and constructs were confirmed by PCR (data not shown). After induction of RNA transcription in the transformed *E.coli* using IPTG, we fed these bacteria to *Oxytricha* along with a small algal aliquot only for the first feeding [2] . Knock-

[2] The reason to supply the algal culture in the starting day of feeding the bacteria with RNAi construct is that *Oxytricha* does not grow up well without them. (This was learned from Landweber’s lab in Princeton University.)

ing down any individual Arisong RNA gene did not result in any detectable phenotypes. However, knocking down all five Arisong RNA genes caused sudden death after two and half days feeding. Due to the sudden death, we could not isolate RNA to confirm in vivo RNA knock-down. So, we replicated our experiments several times with total RNA extractions from samples of *Oxytricha* culture in time series, but these experiments gave us inconsistent results: partial death (reduced size of *Oxytricha* population) or no observable phenotype. Northern blots with extracted RNAs failed to show any detectable knock down of the targeted Arisong RNA species (data not shown).

6.3.2 Speculation on the functions of Arisong RNA genes

From the features of conserved secondary structure, transcript, and regulatory elements, we can speculate about the function of Arisong RNA genes. Several lines of weak evidence suggest that Arisong RNAs may have a function related to spliceosomal RNAs. The Arisong RNAs have the same conserved flanking PSE and 3′ box motifs as spliceosomal snRNAs. We identified U4atac and U6atac snRNAs in *Oxytricha*, which demonstrates the presence of a minor spliceosome, but the U11 and U12 homologs remain unidentified. The consensus structure of the Arisong RNAs does not appear to resemble U11 or U12, and the draft genome assembly is not complete, so it is possible not to sample the minor snRNA gene containing nanochromosomes partially or entirely. Furthermore, both flanking elements are not unique to splicesomal RNAs; for example, the U3 snoRNA gene has nothing to do with the spliceosome, yet is also flanked by these elements.

The 5′ sequence variability of Arisong loci, and the two different sizes of the Onc91 Arisong RNA are somewhat evocative of splice leaders (SLs), a special class of spliceosomal snRNAs found in organisms that perform trans-splicing. A SL provides a splice donor site in trans; thus, the 5′ end of the SL is spliced to a mRNA. The 3′ region of an SL gene has snRNA-like features, including an Sm-binding site [228–230]. However, the structure of the Arisong RNAs does not resemble other known splice leader RNAs, nor do we see a convincing conserved Sm binding site although we do identify conserved putative Sm binding sites in *Oxytricha* traditional snRNAs. We do not see the 5′ sequence of Arisong RNAs on *Oxytricha* ESTs or cDNAs, suggesting that it is not an SL gene though trans-splicing is found to be quite rare in some organisms, for instance, *Schistosoma mansoni* [231].

Chapter 7

Other known ncRNA genes in *O. trifallax*

In the course of this screen, we identified a total of 150 known ncRNA genes: 106 tRNAs (51 distinctive loci), 13 5S rRNAs (1 loci), a rRNA, 6 snRNAs (2 loci), a SRP RNA, a RNase_MRP and 12 snoRNAs (9 loci). Through homology-based search by Infernal/Rfam on incomplete nanochromosomes, other essential ncRNA genes were also detected: 3 RNase P (1 loci) and a telomerase RNA. Considering the 40-65% genome completeness of the stage 1 data set, it is possible to miss other ncRNA genes such as miRNA genes. However, spliceosomal snRNA and snoRNA genes that are among the largest known ncRNA gene families appear to still be underrepresented. So, we sought to study in more detail how the spliceosomal small nuclear RNA (snRNA) gene family and small nucleolar RNA (snoRNA) gene family behaved in the screen, in order to confirm that we were sampling them at the expected frequency, and to look for any unexpected reasons why we might miss ncRNA genes.

7.1 Spliceosomal small nuclear RNA (snRNA) genes

Only the U2 and U6atac snRNAs were identified in the stage 1 dataset. If *Oxytricha* has U2, it should have all RNA components of the major U1/U2 spliceosome. If it has U6atac, it should also have all RNA components of the minor U11/U12 spliceosome [232]. This was a surprising shortfall of knwon ncRNAs, and wanted to see if they could be found in the non-full length nanochromosome data. We analyzed the incomplete contigs of the total dataset using Rfam/Infernal homology searches and identified two additional distinct U2 snRNAs and one distinct locus each for U1, U4, U5, U6, and U4atac snRNA genes, essentially as expected. Additional sequence analysis including *Stylonychia* conservation and/or upstream PSE elments supported these loci. The presence of both U4atac and U6atac strongly suggests that *Oxytricha* possesses a minor spliceosome, although we were unable to identify homologs of U11 or U12 snRNAs.

Thus only two of nine different distinct snRNA loci are contained in the stage 1 dataset. We expected about half of them, given our coverage estimate of 40-65%. This finding could indicate that the stage 1 data may contain a smaller fraction of the *Oxytricha* gene set than we estimated earlier, but these numbers are small. Note that these five additional snRNA genes were not missed by our screen. They were never included in the stage 1 data set because they were not on fully sequenced nanochromosomes, and incomplete nanochromosomes interfered on this screen as it relied on identifying protein coding genes.

7.2 Small nucleolar RNA (snoRNA) genes

In Eukarya and Archaea, two large families of snoRNAs direct site-specific nucleotide modifications of rRNA and other target RNAs: C/D snoRNAs directing 2'-O-methylations, and H/ACA snoRNAs directing pseudouridylations. We expect that like other eukaryotes, *Oxytricha* has tens to hundreds of snoRNAs [12]. *Oxytricha* clearly has both snoRNA-dependent modification systems, because we detect homologs of the conserved catalytic protein components of the yeast C/D and H/ACA snoRNPs (Nop1/fibrillarin and Cbf5/dyskerin) and other C/D and H/ACA snoRNP core proteins [233] in *Oxytricha* by TBLASTN. However, the similarity search analysis in Table 3.1 only identified four distinctive C/D snoRNAs and no H/ACA snoRNAs, which was also a concern.

However, in contrast to the highly conserved spliceosomal snRNAs, it is not surprising that we would have difficulty identifying snoRNAs by homology searches. snoRNAs evolve rapidly and are difficult to detect reliably and systematically by computational analysis alone [234]. We put additional effort into identifying a set of probable *Oxytricha* snoRNAs in the WGS+pilot dataset, to see how snoRNA loci behaved in our screen. We used a pair of gene class-specific genefinding programs: snoscan and snoGPS [134, 209]. These programs suffer from a high false positive rate. To reduce this, we require *Stylonychia* sequence conservation. In addtion, we performed low-stringency Rfam/Infernal homology searches and searches for conserved regions flanked by the PSE motif identified above as well as manual sequence analysis.

To search for the PSE promoter motif, we built an HMM model from upstream 28 nt sequences of 12 *Oxytricha* and *Stylonychia* "known" ncRNAs such as snRNAs, SRP RNA, RNase P, telomerase RNA by using hmmbuild and hmmcalibrate in HMMER2.3.2

package (Figure G.A in Appendix G). By running hmmsearch on all sequences and their reverse complementary sequences in the WGS+pilot (total) dataset, 476 sequences were found to contain one or more PSEs. The PSE has insufficient information content on its own for distinguishing false positives particularly in an AT-rich genome. Thus, we required downstream conservation between *Oxytricha* and *Stylonychia*. We used BLASTN with the "wordmask=dust wordmask=seg extramask=16 M=4 N=-5 E=10^{-5}" option. A total of 56 sequences have putative PSEs located within the BLASTN alignment area with additional 10nt flanking sequences at both ends and in non-protein region predicted by nanogenefinder. Among them, 13 PSE predictions are "known sites" which are located at upstream of "known" ncRNAs and arisong RNAs. All these PSE motif contain exact "ACCCAT" subsequence in their match, and an additional PSE prediction contains it. However, this unknown PSE prediction is located at the intronic region by nanogenefinder but protein region by BLASTX match. So solely relying on the PSE motif screen we could not discriminate the false positives to find other snoRNA genes.

For the 3' box motif search, we built an HMM model from downstream 21 nt sequences of *Oxytricha* arisong RNAs, H/ACA snoRNAs (Onc89 and Onc90), snRNAs and some *Stylonychia* snRNAs by using hmmbuild and hmmcalibrate in HMMER1.8.5 package (Figure G.C in Appendix G). After running hmmsearch on the total dataset, we manually investigated top 80 hits of which the lowest bit score is 24.22, wheares the highest bit score acquired from the shuffled sequences (negative control) is 24.96. Among the examined top 80 hits, 31 predictions are "known sites" which are located at downstream of known ncRNAs and alleles of known ncRNAs. In the remained hits, 26 redundant predictions were not examined because they are located at downstream of alleles of other loci, and 11 predictions of which upstream regions are not conserved in *Stylonychia* seem to be false positives.

Only 4 new snoRNAs (3 C/D box snoRNAs and 1 H/ACA box snoRNA) which have conserved snoRNA features and structures in *Stylonychia* could be detected confidently from the 3' box predictions; however, 3' box motif is conserved only in one C/D snoRNA. Remained predictions are located at downstream of suspicious ncRNA candidates which are not conserved in *Stylonychia*.

For computational detection of apparently missing C/D snoRNAs, snoscan0.9b [134] was run on the stage 1 data set, using *Oxytricha* LSU, SSU, 5.8S and 5S rRNAs as target sequences. To reduce false positives, BLASTN was used for conservation detection in *Stylonychia* with a 10^{-10} E-value cutoff. The snoRNA gene candidiate containing nanochromosomes were examined for the possible existence of protein-conding genes and their structures using our nanogenefinder and QRNA.

For computational detection of H/ACA snoRNAs, we ran snoGPS0.2 [209] on the stage 1 dataset. This software requires candidate pseudouridylation target sites. We provided 31 target sites in LSU and SSU rRNA that are conserved pseudouridylation target sites and also conserved in *Oxytricha*[1]. For estimating the false positive rate, we ran the program on 1000 randomly generated sequences, the same dataset used for the performance test of nanoclassifier, using all uridine ribonucleotide sequences as target sites. The haca2stemv7.table and haca2stemv4a.desc files in the package were used as the score table and descriptor. To select candidates for further analysis, we used 38.5 score cutoff of which the estimated false positive rate from random sequences is 8.9% in sequence level. H/ACA snoRNA genes also have weak signals in primary sequence features and this software has a high false positive rate as determined by random sequence controls. So we applied stringent conservation

[1] The conserved target pseudouridylation sites are infered by snoRNAs correspondence between human and yeast which are listed in snoRNABase database website `http://www-snorna.biotoul.fr/human_yeast/`

requirements to *Stylonychia* by using BLASTN with the “wordmask=dust wordmask=seg extramask=16 M=4 N=-5 E=10^{-5}” option.

All told, after analyzing the entire assembly (not only stage 1 data), we predicted 35 distinct snoRNA loci. This includes 29 distinct methylation guide C/D snoRNAs, five distinct H/ACA snoRNAs, and one distinct U3 snoRNA locus, on 20 different contigs. Only four of the 20 contigs (20%) are incomplete and fail to reach the stage 1 dataset, somewhat fewer than expected from 40-65% coverage. Nine of 16 (56%) nanochromosomes in the stage 1 dataset are are classified as coding, about what is expected from our observation that 50% ncRNAs occur on noncoding nanochromosomes. Of the 7 lacking a clear protein gene, one carries a known snoRNA (U18), and another has no sequence coverage in *Stylonychia*. Five nanochromosomes, each apparently carrying a single snoRNA gene, pass the entire screen and are included in our stage 5 final candidate pool.

The majority of the identified C/D snoRNAs are in two large arrays on incomplete contigs: a 3.5 kb contig that contains 12 C/D snoRNAs and a 1.5 kb contig that contains 4 C/D snoRNAs. snoRNAs are known to occur in clusters in many other organisms [235, 236], sometimes because an entire cluster is carried on one long precursor noncoding RNA that is processed to release multiple snoRNAs [65, 237]. In both identified arrays, the C/D snoRNAs appear to be intronic within a noncoding RNA carrier transcript, as indicated by the presence of strongly conserved 5′ splice sites and lack of other conservation or significant protein coding potential in the contigs. 3′ splice sites are less conserved and more difficult to identify in AT-rich *Oxytricha* sequence. In addition to these snoRNAs in two arrays, other four C/D snoRNAs were intronic in coding genes in the stage 1 nanochromosomes, and one (a U18 homolog) was flanked by a strong conserved consensus 5′ splice site and is probably intronic as well. It therefore appears likely that many, perhaps most snoRNAs in

Oxytricha are intron-encoded in a combination of coding genes and noncoding transcripts. Large arrays may be on large contigs that are less likely to be fully assembled in the current dataset: it was only back luck in assembly that these arryas did not appear in our stage 1 dataset. Intron-encoded ncRNAs in coding genes will consistently be screened out by the coding gene classifier step; however, it is a "feature" of our screen to find independently-transcribed functional ncRNA genes, which misses ncRNA genes processed out of introns from protein coding genes. Although the screen successfully detects both C/D and H/ACA snoRNAs, they are likely underrepresented for these reasons.

Chapter 8

Conclusions

Our screen identified a novel family of at least 32 small Arisong RNA genes that are flanked by detectable upstream and downstream regulatory elements in four ciliate species, which encompass a previous prediction of six small RNA loci in *Paramecium tetraurelia* called PM01_1-6 [217]. We confirm not only Chen et al.'s computational prediction of six novel small RNAs, but also extend it by: a) expanding the Arisong RNA family by including other paralogs in *Paramecium* and homologs in other three ciliate species; b) confirming the expression and defining $3'$ boundaries of transcripts of some representatives of this family in *Oxytricha* by Northern and RACE-PCR; and c) recognizing that all members of the family share a distinctive consensus secondary structure.

If we assume that there is no systematic bias in independent functional ncRNA gene distribution across nanochromosomes, we can estimate that the probability of sampling any given ncRNA gene in the complete screen is roughly 10%. This estimate is calculated by multiplying $\sim$50% completeness of the genome, $\sim$50% of known ncRNA genes found on noncoding nanochromosomes (per the 134 known ncRNAs in Table 3.1), $\sim$80% specificity of the computational nanoclassifier (derived from the observed 0.17 FP rates for coding predictions), and $\sim$50% *Stylonychia* coverage for detecting conserved regions (0.5 * 0.5 * 0.8 * 0.5 = 10%). Our estimated 10% overall sampling rate is roughly consistent with the rate at which the major and minor spliceosomal RNAs made it to stage 4 in our screen (2 of 9). Including all full-length nanochromosome candidates which are not obviously coding fragments, a total of 88 “quasialleles” of independent functional ncRNA genes passed through to stage 4 in our screen. Our screen identified no long mRNA-like ncRNA genes, other than probable noncoding host transcripts for arrays of intronic snoRNAs. This result suggests that *O. trifallax* has roughly a thousand alleles of independent functional ncRNA genes. Note that “88” is not the number of genes, but rather the number of alleles - these may include paralogs or orthologs from the other strain 510. The estimated number should be lowered by two- to four fold to obtain the number of genes. In terms of ncRNA gene families, the number should be reduced further. Otherwise, the slight sampling bias in the stage 1 dataset as discussed in Chapter 3 might not cause an order of magnitude difference in total ncRNA gene numbers. Therefore, if *Oxytricha* has large numbers of undiscovered ncRNAs in its macronuclear genome, they are unlikely to be from independently transcribed ncRNA genes. Clearly some ncRNA genes are located within introns of protein genes, which are invisible to our screen. Additional ncRNAs may come from nongenic transcription or from processes associated with transcription of coding mRNAs in cis, including cis-antisense RNA and other cis-transcribed ncRNAs (overlapping coding regions or regulatory regions for coding genes) such as RNAs involved in chromatin modification or transcriptional interference. Our conclusion is consistent with recent arguments that most of the “ncRNA” that has been observed in mammalian systems is a mixture of technical artifact, introns, alternatively processed polyA sites/promoters/exons and RNAs arising from cis-acting processes associated with transcription of nearby coding genes [103].

If novel independently transcribed ncRNA genes were numerous in eukaryotes, we would have expected to see many noncoding nanochromosomes carrying single ncRNA genes. Instead, our coding nanoclassifier immediately classified 95% of nanochromosomes as protein-coding – which is essentially all of them, because the nanoclassifier has an estimated 6% false negative rate of misclassifying coding nanochromosomes. This conclusion in *Oxytricha*, the relative paucity of independent functional ncRNA genes, might be expanded into other eukaryotes although there are the following limitations. One limitation is that our approach does not look for the possibility of ncRNA genes in the micronucleus. Although the micronucleus is generally transcriptionally silent and not considered to harbor active genes, it becomes briefly transcriptionally active after conjugation, during the process of forming a new macronucleus. Among the micronuclear RNAs expressed at this time are transcripts of a major transposon family (TBE1) [161]. To propagate in a normally silent germ line, micronuclear-limited transposon genes presumably need special adaptations. Another limitation is that the unicellular ciliates are evolutionarily distant from the most commonly studied lineages of plants and animals. Ciliates clearly utilize functional but nongenic ncRNA transcripts extensively in DNA elimination and rearrangements [160, 162–166]. Nonetheless ciliates might systematically lack ncRNA-dependent regulatory systems that are important in other lineages. A screen in a unicellular ciliate therefore does not bear directly on the question of whether there are large numbers of ncRNA genes specific to “complex” multicellular organisms [30, 238]. It does, however, have a bearing on the question of whether there are large numbers of undiscovered ncRNA genes in eukaryotes in general.

Our study also illustrates some of the difficulties in distinguishing noncoding RNA genes from other RNA products, such as mRNAs for small, unusual, or rapidly evolving coding genes. At each step of our screen – probabilistic genefinding, similarity to known proteins, and using the evolutionary pattern of coding gene evolution in *Oxytricha*/*Stylonychia* sequence alignments – we detected and removed a large proportion of apparent coding genes. Even so, after all these steps, in the final set of 17 apparently noncoding conserved nanochromosomes, 5/17 appear to us upon manual analysis to contain fragments of conserved coding sequence. Although some features are *Oxytricha*-specific due to the extraordinary genome processing, this multistep analysis may be contrasted to studies that have identified large numbers of putative "noncoding" RNAs using overly simple definitions such as the lack of an ORF $>$ 100 aa [29, 144], or finding cDNA transcripts that do not overlap with Ensembl gene predictions [239]. We believe one reason that we find so few ncRNA genes, whereas other studies find so many, results from different standards in computational analysis of coding genes. In light of our results here, we believe that "noncoding" RNA loci in other organisms merit careful reexamination, as others have argued [97, 99, 141].

Appendix A

Appendix: tRNA gene analysis on the total dataset

tRNAscan-SE program [131] found total 243 tRNA genes in the total dataset: 40 tRNA genes in the pilot dataset and 203 tRNA genes in WSG dataset. By removing redundancy, we found 143 unique loci for tRNA genes which encode 78 unique tRNA genes; that is, some of tRNA genes in different loci encode identical tRNA gene sequences. They represent all 20 amino acid tRNA types which include 1 pseudo-gene and 4 undefined tRNA types. They also represent 48 anticodons. When considering wobble rules in the third position in codon, they still miss 2 anticodon types, wheares 3 anticodon types which don't follow wobble rules are included.

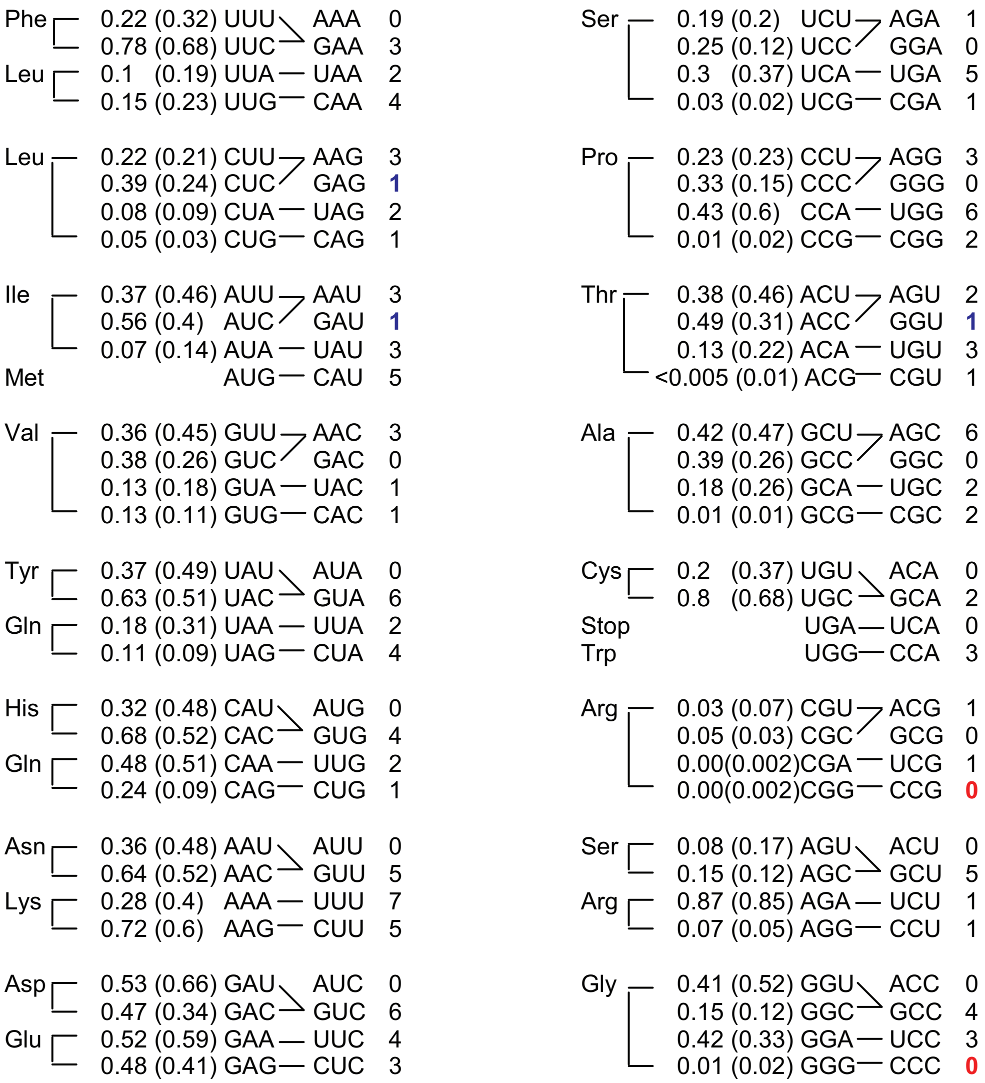


Figure A.1: *Oxytricha* genetic code and associated tRNA genes

For each of the 64 codons, this figure shows following things: the corresponding amino acid, the observed frequency of *Oxytricha* codon from the Prescott paper [171], the observed frequency of *O. trifallax* codon from publicly available 26 genes in NCBI, predicted wobble pairing to a tRNA anticodon, an unmodified tRNA anticodon sequence, and the number of tRNA genes found with the corresponding anticodon. Red numbers indicate missed tRNA genes which have the corresponding anticodons. Blue numbers indicate tRNA genes having the corresponding anticodons which don't follow wobble rules.

Appendix B

Appendix: Telomere endpoints coordinates of contigs in the stage 3 dataset

Coordinates of telomere endpoints of contigs in WGS2.1.1 dataset among stage 3 dataset (full-length nanochromosomes) are listed because all full-length nanochromosomes are too much to be listed. These telomeres are detected by Smith/Waterman alignment with minimal sequences of 5′ and 3′ telomeres, so their end points might not be the genuine end points of telomeres. In the list, “start” means the starting position of minimal 5′ telomere and “end” means the ending position of minimal 3′ telomere, i.e. starting and endind point of nanochromosomes inclusive of those telomeres.

Table B.1: Coordinates of telomere ends

Contig name	start	end	Contig name	start	end	Contig name	start	end
Contig72558.2	1	1665	Contig71216.1	17	814	Contig44272.1	16	913
Contig65279.1	98	738	Contig12754.1	19	1201	Contig64166.1	182	791
Contig41939.1	14	1611	Contig57493.1	12	993	Contig90597	1	2508
Contig58322.1	15	746	Contig75252.1	1	2510	Contig47631.1	3	1016
Contig63601.1	6	670	Contig36794.2	459	1065	Contig72592.3	160	1258
Contig81663.1	21	1748	Contig56717.1	17	515	Contig44542.1	63	711
Contig54011.1	9	802	Contig40627.1	13	1442	Contig56466.1	24	637
Contig32977.1	1	1274	Contig63692.1	5	1433	Contig56085.1	12	1868
Contig63165.1	23	2037	Contig201758	18	1062	Contig350.1	24	640
Contig69833.1	16	1792	Contig36467.1	1	2551	Contig62738.1	1	1440
Contig42360.1	6	1093	Contig65036.1	167	773	Contig52618.1	1	2081
Contig45979.1	2	1417	Contig70704.1	12	898	Contig36544.2	23	1605
Contig64625.1	58	2002	Contig53544.1	9	1768	Contig33474.1	135	874
Contig46462.1	59	1336	Contig40060.1	13	1855	Contig42218.1	89	1466
Contig64682.1	4	1639	Contig20293.1	57	1800	Contig71362.1	12	1314
Contig79287.2	367	836	Contig77349.1	201	1024	Contig60577.1	12	2134
Contig37941.1	65	3047	Contig68529.1	58	1950	Contig66981.1	12	1773
Contig80041.1	12	1182	Contig30968.1	1	1076	Contig38743.1	58	1481
Contig41267.1	57	1674	Contig54158.1	1	789	Contig20377.1	13	1181
Contig63886.1	1	1285	Contig65897.1	57	1486	Contig70376.1	21	1398
Contig73674.2	263	879	Contig46616.1	14	630	Contig63474.1	12	1522
Contig45079.1	57	1822	Contig70731.1	12	803	Contig71432.1	8	597
Contig80821.1	366	1286	Contig56341.1	59	1644	Contig54157.1	13	1118
Contig34132.1	1	1160	Contig19117.1	166	841	Contig37944.1	59	1655

Continued on next page

Table B.1 – continued from previous page

Contig name	start	end	Contig name	start	end	Contig name	start	end
Contig56473.2	246	859	Contig67238.1	13	871	Contig50047.1	8	1307
Contig21168.1	2	773	Contig90796	58	1452	Contig57364.1	13	1819
Contig71359.1	24	845	Contig27888.1	424	1152	Contig62704.1	9	1384
Contig42299.1	15	916	Contig55772.1	56	1035	Contig61428.1	18	934
Contig79483.1	16	838	Contig71781.1	9	1150	Contig15953.1	57	1587
Contig28252.1	12	1023	Contig27958.1	20	1110	Contig50209.1	209	842
Contig57707.1	15	1150	Contig64767.1	9	963	Contig65115.1	1	1306
Contig77913.1	62	1892	Contig38310.2	58	2110	Contig42625.1	56	1465
Contig38449.1	428	1007	Contig47514.1	390	1106	Contig37041.1	57	1602
Contig82949.1	1	3491	Contig37987.1	56	1731	Contig7084.1	184	830
Contig47030.2	8	1174	Contig57787.1	5	1198	Contig73911.1	62	1530
Contig40198.1	128	732	Contig49984.1	17	1067	Contig47714.1	212	848
Contig44659.1	8	890	Contig64181.1	10	1057	Contig23611.1	15	936
Contig53962.1	229	790	Contig53992.1	178	820	Contig49976.1	6	455
Contig27153.1	20	1532	Contig63683.1	41	844	Contig52483.1	57	1635
Contig41931.1	182	1528	Contig44069.1	16	913	Contig62533.1	14	601
Contig64310.1	8	559	Contig36242.1	147	854	Contig63998.1	350	1224
Contig55752.1	57	1445	Contig57521.1	69	1929	Contig52959.1	1	1405
Contig76009.1	1	1553	Contig42303.1	8	1062	Contig69957.1	6	672
Contig41708.1	56	1464	Contig42246.1	60	933	Contig200270	20	914
Contig15155.2	449	1896	Contig71999.1	9	1502	Contig46853.1	4	1454
Contig47794.1	56	2351	Contig71398.1	530	999	Contig82141.1	5	746
Contig74167.1	15	1664	Contig50244.2	172	821	Contig63694.1	14	754
Contig46532.1	16	1050	Contig52250.1	52	2178	Contig11669.1	20	1182
Contig70680.1	3	1237	Contig50051.1	127	801	Contig22505.1	1	1609

Continued on next page

Table B.1 – continued from previous page

Contig name	start	end	Contig name	start	end	Contig name	start	end
Contig71072.1	8	867	Contig83246.1	59	2391	Contig71315.1	2	904
Contig71215.1	233	720	Contig64079.1	13	1046	Contig62742.1	11	1346
Contig28854.1	224	761	Contig64780.1	281	1581	Contig38453.1	348	1057
Contig55875.1	11	1756	Contig63032.1	1	597	Contig51710.1	9	1535
Contig63348.1	14	614	Contig50237.1	13	662	Contig63260.1	5	1403
Contig85069.1	59	1452	Contig27013.1	9	737	Contig11508.1	3	1484
Contig53197.1	316	814	Contig8162.1	55	2725	Contig44402.1	250	863
Contig58094.1	1	1460	Contig20965.1	8	753	Contig13832.1	254	887
Contig56080.1	1	1130	Contig63545.1	13	797	Contig63671.1	14	1173
Contig84680.1	57	2366	Contig60049.1	252	1681	Contig39846.1	1	1335
Contig42494.1	1	1458	Contig57944.1	12	757	Contig53310.1	17	1096
Contig52417.1	58	1585	Contig37908.2	256	872	Contig64187.1	89	862
Contig45510.1	181	861	Contig49716.1	17	827	Contig62743.1	10	1130
Contig21532.1	170	805	Contig20923.1	16	909	Contig58516.1	65	782
Contig4340.2	58	1334	Contig72585.2	227	859	Contig54150.1	8	1436
Contig53762.1	9	1009	Contig71212.1	12	1325	Contig57129.1	26	1867
Contig72791.1	367	838	Contig49753.1	7	908	Contig44757.1	1	1019
Contig49185.1	8	741	Contig49074.1	1	1392	Contig79964.2	4	767
Contig40129.1	17	1110	Contig201267	322	881	Contig78823.1	3	1254
Contig70148.1	20	609	Contig57150.1	7	1227	Contig202913	9	1100
Contig42445.1	54	2400	Contig9982.1	270	1319	Contig200640	9	1125
Contig47479.1	15	934	Contig52992.1	502	832	Contig56296.1	58	2605
Contig46225.1	58	1633	Contig72727.1	376	1538	Contig63727.1	15	912
Contig47353.1	3	1630	Contig42454.1	17	918	Contig16863.1	10	1068
Contig34069.1	8	836	Contig57675.1	119	840	Contig24276.1	56	943

Continued on next page

Table B.1 – continued from previous page

Contig name	start	end	Contig name	start	end	Contig name	start	end
Contig84022.1	486	2774	Contig76779.1	223	802	Contig20685.1	93	881
Contig45856.1	60	1013	Contig74674.2	113	788	Contig204907	282	788
Contig42078.1	552	1778	Contig63555.1	14	932	Contig71254.1	8	1154
Contig45583.1	15	628	Contig63306.1	3	1169	Contig53857.1	9	1042
Contig16891.1	19	1121	Contig25788.1	4	1091	Contig44349.1	11	952
Contig64742.1	16	1604	Contig79253.2	246	862	Contig93299	58	2197
Contig57619.1	312	783	Contig44235.1	10	812	Contig66475.1	9	842
Contig14700.1	11	2135	Contig42186.1	2	1500	Contig54139.1	9	637
Contig64473.1	54	1722	Contig52231.1	55	1897	Contig57109.2	461	2342
Contig27559.1	13	928	Contig72503.1	13	1170	Contig5306.1	70	1217
Contig202880	233	812	Contig204375	10	1509	Contig63304.1	14	1384
Contig80981.1	16	1494	Contig35667.1	204	834	Contig36615.1	7	2208
Contig44498.1	13	1345	Contig41841.1	55	1696	Contig50763.1	2	1960
Contig90468	13	1035	Contig47258.1	9	832	Contig47315.1	214	726
Contig45590.2	256	869	Contig54172.1	1	853	Contig53584.1	277	1277
Contig82969.1	292	875	Contig202273	22	1537	Contig72578.1	244	875
Contig58176.1	8	1414	Contig35598.1	12	2255	Contig60019.1	12	1144
Contig69506.1	29	1528	Contig91146	272	749	Contig54004.1	14	810
Contig24059.1	52	1336	Contig83373.1	295	921	Contig57769.1	21	868
Contig27097.1	14	836						

Appendix C

Appendix: *Oxytricha* gene list

Oxytricha trifallax gene list contains the following fields:

1. Locus name. ”Distinct” loci are called “OncXX”; other loci in the same “quasiallelic” group are called “OncXX.y”.

2. coordinates. start..end If start is greater than end, locus is on Crick strand.

3. Source contig. A sequence name in the WGS+pilot dataset.

4. Stage. Highest stage the contig reached in the screen: 0-5. 0=incomplete contigs in WGS+pilot; 1-5 = stage 1..stage 5 datasets.

5. Locus type. Semi-controlled classification vocabulary: U2_snRNA or C/D_snoRNA, for example.

6. Length. Length of the locus in nt.

7. Experiments. N = size, expression confirmed on Northern; R = size, expression, ends confirmed by overlapping 5′, 3′ RACE-PCRs. N,R=both .=neither. (The lowercase letters indicate that the experiments were not done directly at these loci but results were infered from a representative locus, because transcripts of them are same.)

8. Etc. Homologous Rfam family name detected by cmsearch and its evalue or other detection method

Although the OncXX numbering reaches Onc154, the numbers are not all used. For example,there are not 154 distinct loci in this file. The assignment of “distinct loci” is subjective, and we rearranged it late in our analysis when we realized there’s a pretty clear pattern of up to four ”alleles” per quasiallelic group - suggesting that the macronuclear genome is tetraploid. This converted many loci from ”distinct” to ”quasialleles”. Also, the OncXX numbers are not consecutive in this file. OncXX numbers were assigned as we discovered new loci, whereas the list is arranged into sensible sections: tRNAs, rRNAs, miscRNAs, snRNAs, snoRNAs, and finally the “novel” candidates identified by the nanogenefinder screen.

Table C.1: Summary of all *Oxytricha trifallax* loci

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc1	377..449	Contig19117.1	4	tRNA(Asn,GTT)	73nt	.	Rfam:RF00005:1e-13
Onc1.2	324..396	Contig74674.2	4	tRNA(Asn,GTT)	73nt	.	Rfam:RF00005:1e-13
Onc2	701..629	Contig76935.2	1	tRNA(Asn,GTT)	73nt	.	Rfam:RF00005:1e-13

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc3	259..188	Contig200270	4	tRNA(Lys,TTT)	72nt	N	Rfam:RF00005:4e-14
Onc3.2	265..194	Contig40129.1	4	tRNA(Lys,TTT)	72nt	n	Rfam:RF00005:4e-14
Onc3.3	257..186	Contig53310.1	4	tRNA(Lys,TTT)	72nt	n	Rfam:RF00005:4e-14
Onc53	176..248	OXAO-aab15f07	4	tRNA(Lys,CTT)	73nt	N	Rfam:RF00005:5e-13
Onc53.2	229..301	Contig35865.1	1	tRNA(Lys,CTT)	73nt	n	Rfam:RF00005:5e-13
Onc53.3	1184..1112	OXAO-aab17e12	1	tRNA(Lys,CTT)	73nt	n	Rfam:RF00005:5e-13
Onc7	381..453	Contig38385.1	1	tRNA(Lys,CTT)	73nt	n	Rfam:RF00005:5e-13
Onc7.2	188..260	Contig91659	1	tRNA(Lys,CTT)	73nt	n	Rfam:RF00005:5e-13
Onc8	607..535	Contig42095.1	1	tRNA(Lys,CTT)	73nt	n	Rfam:RF00005:5e-13
Onc8.2	197..269	UGC1O0003_C05_F	1	tRNA(Lys,CTT)	73nt	n	Rfam:RF00005:5e-13
Onc9	451..524	Contig201758	4	tRNA(Val,AAC)	74nt	.	Rfam:RF00005:8e-12
Onc9.2	619..546	OXAC-aaa03e08	4	tRNA(Val,AAC)	74nt	.	Rfam:RF00005:8e-12
Onc11	266..194	Contig82806.1	1	tRNA(Val,TAC)	73nt	.	Rfam:RF00005:9e-14
Onc12	2257..2185	Contig75612.1	1	tRNA(Val,CAC)	73nt	.	Rfam:RF00005:7e-11
Onc12.2	2259..2187	Contig81063.2	1	tRNA(Val,CAC)	73nt	.	Rfam:RF00005:7e-11
Onc13	433..363	Contig202880	4	tRNA(Gln,CTG)	71nt	.	Rfam:RF00005:3e-13
Onc13.2	423..353	Contig76779.1	4	tRNA(Gln,CTG)	71nt	.	Rfam:RF00005:3e-13

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc13.3	492..422	Contig82969.1	4	tRNA(Gln,CTG)	71nt	.	Rfam:RF00005:3e-13
Onc14	248..177	Contig78714.1	1	tRNA(Gln,TTG)	72nt	.	Rfam:RF00005:3e-14
Onc14.2	302..231	Contig79809.1	1	tRNA(Gln,TTG)	72nt	.	Rfam:RF00005:3e-14
Onc15	1064..1134	Contig42065.1	1	tRNA(Gln,CTA)	71nt	.	Rfam:RF00005:3e-12
Onc15.2	1064..1134	Contig50398.1	1	tRNA(Gln,CTA)	71nt	.	Rfam:RF00005:5e-12
Onc39	243..172	OXAD-aaa02e11	4	tRNA(Gln,TTA)	72nt	.	Rfam:RF00005:3e-14
Onc39.2	243..172	OXAD-aaa04h06	4	tRNA(Gln,TTA)	72nt	.	Rfam:RF00005:3e-14
Onc16	419..490	Contig35667.1	4	tRNA(Gly,TCC)	72nt	n	Rfam:RF00005:2e-13
Onc16.2	232..303	OXAE-aad39c12	4	tRNA(Gly,TCC)	72nt	n	Rfam:RF00005:2e-13
Onc16.3	224..295	UGC1O0005_L02_F	4	tRNA(Gly,TCC)	72nt	n	Rfam:RF00005:2e-13
Onc17	458..529	Contig72578.1	4	tRNA(Gly,TCC)	72nt	N	Rfam:RF00005:2e-13
Onc17.2	442..513	Contig72585.2	4	tRNA(Gly,TCC)	72nt	n	Rfam:RF00005:2e-13
Onc17.3	375..446	Contig72592.3	4	tRNA(Gly,TCC)	72nt	n	Rfam:RF00005:2e-13
Onc18	617..547	Contig38449.1	4	tRNA(Gly,GCC)	71nt	.	Rfam:RF00005:4e-12
Onc18.2	388..458	OXAO-aaa61c10	4	tRNA(Gly,GCC)	71nt	.	Rfam:RF00005:4e-12
Onc18.3	196..126	Contig93074	1	tRNA(Gly,GCC)	71nt	.	Rfam:RF00005:4e-12
Onc20	367..439	Contig41931.1	4	tRNA(Thr,AGT)	73nt	.	Rfam:RF00005:1e-11

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc21	206..135	OXAB-aaa03d02	1	tRNA(Thr,TGT)	72nt	.	Rfam:RF00005:5e-14
Onc21.2	989..1060	Contig36821.1	1	tRNA(Thr,TGT)	72nt	.	Rfam:RF00005:5e-14
Onc23	230..159	Contig42299.1	4	tRNA(Met,CAT)	72nt	.	Rfam:RF00005:5e-10
Onc23.2	216..145	OXAO-aab16f12	4	tRNA(Met,CAT)	72nt	.	Rfam:RF00005:5e-10
Onc23.3	233..162	Contig42454.1	4	tRNA(Met,CAT)	72nt	.	Rfam:RF00005:5e-10
Onc24	221..293	Contig78307.1	1	tRNA(Met,CAT)	73nt	.	Rfam:RF00005:4e-12
Onc25	192..120	Contig44069.1	4	tRNA(Ala,AGC)	73nt	.	Rfam:RF00005:1e-09
Onc25.2	737..809	Contig44272.1	4	tRNA(Ala,AGC)	73nt	.	Rfam:RF00005:1e-09
Onc25.3	176..104	OXAE-aae01d07	4	tRNA(Ala,AGC)	73nt	.	Rfam:RF00005:1e-09
Onc25.4	724..796	OXAC-aaa08h01	4	tRNA(Ala,AGC)	73nt	.	Rfam:RF00005:1e-09
Onc26	185..113	Contig202071	1	tRNA(Ala,AGC)	73nt	.	Rfam:RF00005:1e-09
Onc26.2	2253..2181	Contig7063.1	1	tRNA(Ala,AGC)	73nt	.	Rfam:RF00005:1e-09
Onc28	210..281	Contig83003.1	1	tRNA(Ala,TGC)	72nt	.	Rfam:RF00005:5e-13
Onc28.2	252..323	Contig83010.2	1	tRNA(Ala,TGC)	72nt	.	Rfam:RF00005:5e-13
Onc29	1032..961	Contig56335.1	1	tRNA(Ala,CGC)	72nt	.	Rfam:RF00005:8e-13
Onc29.2	193..264	Contig57947.1	1	tRNA(Ala,CGC)	72nt	.	Rfam:RF00005:8e-13
Onc31	655..727	Contig50209.1	4	tRNA(Pro,TGG)	73nt	.	Rfam:RF00005:2e-09

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc31.2	183..111	OXAO-aab17f04	4	tRNA(Pro,TGG)	73nt	.	Rfam:RF00005:2e-09
Onc31.3	183..111	OXAD-aaa08e07	4	tRNA(Pro,TGG)	73nt	.	Rfam:RF00005:2e-09
Onc34	1363..1434	Contig63474.1	4	tRNA(Pro,AGG)	72nt	.	Rfam:RF00005:2e-09
Onc35	3309..3238	Contig53709.1	1	tRNA(Pro,CGG)	72nt	.	Rfam:RF00005:4e-11
Onc36	662..590	Contig67238.1	4	tRNA(Phe,GAA)	73nt	.	Rfam:RF00005:6e-12
Onc37	234..162	Contig71359.1	4	tRNA(Arg,TCT)	73nt	N	Rfam:RF00005:2e-12
Onc37.2	226..154	Contig79483.1	4	tRNA(Arg,TCT)	73nt	n	Rfam:RF00005:2e-12
Onc37.3	815..887	Contig77349.1	4	tRNA(Arg,TCT)	73nt	n	Rfam:RF00005:2e-12
Onc37.4	211..139	OXAD-aaa04e08	4	tRNA(Arg,TCT)	73nt	n	Rfam:RF00005:2e-12
Onc38	6108..6180	Contig45919.1	1	tRNA(Arg,CCT)	73nt	.	Rfam:RF00005:6e-13
Onc40	284..213	Contig50373.1	1	tRNA(Cys,GCA)	72nt	.	Rfam:RF00005:1e-13
Onc40.2	292..221	Contig79368.1	1	tRNA(Cys,GCA)	72nt	.	Rfam:RF00005:1e-13
Onc41	204..132	OXAE-aaa18e02	4	tRNA(Ile,AAT)	73nt	.	Rfam:RF00005:2e-13
Onc41.2	616..688	OXAD-aaa01f07	1	tRNA(Ile,AAT)	73nt	.	Rfam:RF00005:2e-13
Onc41.3	186..114	OXAD-aaa07f12	1	tRNA(Ile,AAT)	73nt	.	Rfam:RF00005:2e-13
Onc41.4	228..156	UGC1O0003_P04_R	4	tRNA(Ile,AAT)	73nt	.	Rfam:RF00005:2e-13
Onc43	3333..3405	Contig2009.1	1	tRNA(Ile,TAT)	73nt	.	Rfam:RF00005:6e-14

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc44	246..167	OXAE-aad49e12	4	tRNA(Leu,AAG)	80nt	.	Rfam:RF00005:4e-11
Onc44.2	255..176	UGC1O0002_B18_R	4	tRNA(Leu,AAG)	80nt	.	Rfam:RF00005:4e-11
Onc46	231..148	OXAE-aad58g07	1	tRNA(Leu,CAA)	84nt	.	Rfam:RF00005:5e-11
Onc46.2	921..1004	OXAD-aaa01e03	1	tRNA(Leu,CAA)	84nt	.	Rfam:RF00005:5e-11
Onc48	2804..2725	Contig81525.1	1	tRNA(Leu,CAG)	80nt	.	Rfam:RF00005:1e-11
Onc49	2298..2215	Contig41162.1	1	tRNA(Leu,TAA)	84nt	.	Rfam:RF00005:4e-10
Onc49.2	1911..1828	Contig60633.1	1	tRNA(Leu,TAA)	84nt	.	Rfam:RF00005:4e-10
Onc50	207..122	OXAO-aaa58g12	4	tRNA(Tyr,GTA)	86nt	.	Rfam:RF00005:1e-09
Onc50.2	201..116	UGC1O0001_O16_R	4	tRNA(Tyr,GTA)	86nt	.	Rfam:RF00005:1e-09
Onc50.3	246..161	Contig66860.1	1	tRNA(Tyr,GTA)	86nt	.	Rfam:RF00005:1e-09
Onc56	180..252	OXAO-aab16e01	4	tRNA(Glu,TTC)	73nt	.	Rfam:RF00005:3e-10
Onc56.2	806..733	OXAO-aab16g01	1	tRNA(Glu,TTC)	74nt	.	Rfam:RF00005:4e-10
Onc57	307..379	Contig42902.1	1	tRNA(Glu,TTC)	73nt	.	Rfam:RF00005:3e-10
Onc57.2	207..279	Contig47833.1	1	tRNA(Glu,TTC)	73nt	.	Rfam:RF00005:3e-10
Onc57.3	256..328	Contig61332.1	1	tRNA(Glu,TTC)	73nt	.	Rfam:RF00005:3e-10
Onc58	204..132	OXAE-aaa06f08	1	tRNA(Glu,CTC)	73nt	.	Rfam:RF00005:5e-12
Onc58.2	211..139	UGC1O0002_B22_R	1	tRNA(Glu,CTC)	73nt	.	Rfam:RF00005:5e-12

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc60	174..246	Contig44957.1	1	tRNA(Glu,CTC)	73nt	.	Rfam:RF00005:5e-12
Onc61	199..128	OXAE-aaf79d05	4	tRNA(Asp,GTC)	72nt	.	Rfam:RF00005:2e-12
Onc61.2	499..570	OXAO-aab15d01	4	tRNA(Asp,GTC)	72nt	.	Rfam:RF00005:2e-12
Onc63	276..205	Contig52579.1	1	tRNA(Asp,GTC)	72nt	.	Rfam:RF00005:2e-12
Onc63.2	220..149	Contig74115.1	1	tRNA(Asp,GTC)	72nt	.	Rfam:RF00005:2e-12
Onc65	1082..1011	OXAE-aaa29d04	1	tRNA(His,GTG)	72nt	.	Rfam:RF00005:3e-11
Onc65.2	2271..2200	Contig51714.1	1	tRNA(His,GTG)	72nt	.	Rfam:RF00005:3e-11
Onc65.3	2253..2182	Contig73221.2	1	tRNA(His,GTG)	72nt	.	Rfam:RF00005:3e-11
Onc67	1214..1143	Contig37344.1	1	tRNA(His,GTG)	72nt	.	Rfam:RF00005:3e-11
Onc68	2162..2082	Contig47137.1	1	tRNA(Ser,AGA)	81nt	.	Rfam:RF00005:5e-12
Onc68.2	2155..2075	Contig76987.1	1	tRNA(Ser,AGA)	81nt	.	Rfam:RF00005:5e-12
Onc69	2304..2386	Contig74257.1	1	tRNA(Ser,GCT)	83nt	.	Rfam:RF00005:3e-11
Onc69.2	2293..2375	Contig84423.1	1	tRNA(Ser,GCT)	83nt	.	Rfam:RF00005:3e-11
Onc70	312..393	Contig35394.1	1	tRNA(Ser,TGA)	82nt	.	Rfam:RF00005:2e-10
Onc70.2	1744..1663	Contig37182.1	1	tRNA(Ser,TGA)	82nt	.	Rfam:RF00005:2e-10
Onc72	244..316	Contig41301.1	1	tRNA(Trp,CCA)	73nt	.	Rfam:RF00005:8e-12
Onc72.2	240..312	Contig85126.1	1	tRNA(Trp,CCA)	73nt	.	Rfam:RF00005:8e-12

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc74	192..262	Contig72432.1	1	tRNA(Undet)	71nt	.	Rfam:RF00005:1e-06
Onc75	441..323	Contig350.1	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.2	462..580	Contig73674.2	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.3	427..309	UGC1O0003_F04_R	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.4	429..311	Contig45583.1	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.5	670..552	Contig45590.2	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.6	438..320	Contig56466.1	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.7	660..542	Contig56473.2	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.8	458..576	Contig37908.2	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.9	216..334	Contig46616.1	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.10	660..542	Contig79253.2	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.11	449..567	Contig44402.1	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.12	199..317	OXAO-aab14b11	4	5S_rRNA	119nt	.	Rfam:RF00001:3e-23
Onc75.13	411..293	OXAO-aaa59b10	4	5S_rRNA	119nt	.	Rfam:RF00001:1e-22
Onc113	3652..3804	rDNA	2	5.8S_rRNA	153nt	.	Rfam:RF00002:6e-40
Onc152	1750..3521	rDNA	2	SSU_rRNA	1772nt	.	gb:FJ545743.1

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc153	4008..7392	rDNA	2	LSU_rRNA	3385nt	.	gb:FJ545743.1
Onc81	2567..2850	Contig82592.1	1	SRP_euk_arch	284nt	.	Rfam:RF00017:1e-36
Onc81.2	557..274	Contig56789.1	0	SRP_euk_arch	284nt	.	Rfam:RF00017:1e-36
Onc82	1118..790	OXAO-aaa59f01	1	RNase_MRP	329nt	.	Rfam:RF00030:4e-32
Onc82.2	1141..812	Contig77949.2	0	RNase_MRP	330nt	.	Rfam:RF00030:4e-32
Onc82.3	1181..853	Contig77942.1	0	RNase_MRP	329nt	.	Rfam:RF00030:4e-32
Onc82.4	143..471	Contig84242.1	0	RNase_MRP	329nt	.	Rfam:RF00030:4e-32
Onc149	573..246	Contig62031.2	0	RNaseP_nuc	328nt	.	Rfam:RF00009:3e-27
Onc149.2	270..597	Contig39152.2	0	RNaseP_nuc	328nt	.	Rfam:RF00009:2e-26
Onc149.3	582..255	Contig62024.1	0	RNaseP_nuc	328nt	.	Rfam:RF00009:3e-19
Onc150	419..234	Contig41445.1	0	Telomerase_cil	186nt	.	Rfam:RF00025:3e-14
Onc120	1860..1698	Contig51351.1	0	U1_snRNA	163nt	.	Rfam:RF00003:7e-24
Onc120.2	149..311	Contig75046.1	0	U1_snRNA	163nt	.	Rfam:RF00003:7e-24
Onc120.3	223..386	Contig206433	0	U1_snRNA	164nt	.	Rfam:RF00003:7e-24
Onc121	979..790	Contig36667.2	0	U2_snRNA	190nt	.	Rfam:RF00004:4e-42
Onc121.2	627..438	Contig58942.2	0	U2_snRNA	190nt	.	Rfam:RF00004:4e-42
Onc77	465..654	Contig57619.1	4	U2_snRNA	190nt	N	Rfam:RF00004:2e-37

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc77.2	558..369	Contig71215.1	4	U2_snRNA	190nt	n	Rfam:RF00004:2e-37
Onc77.3	675..486	Contig72791.1	4	U2_snRNA	190nt	n	Rfam:RF00004:2e-37
Onc77.4	163..352	OXAO-aab15a03	4	U2_snRNA	190nt	n	Rfam:RF00004:2e-37
Onc123	527..721	Contig201159	0	U2_snRNA	195nt	.	Rfam:RF00004:2e-34
Onc123.2	195..389	Contig201160	0	U2_snRNA	195nt	.	Rfam:RF00004:2e-34
Onc125	393..265	Contig51899.1	0	U4_snRNA	129nt	.	Rfam:RF00015:1e-20
Onc125.2	3763..3891	Contig71125.2	0	U4_snRNA	129nt	.	Rfam:RF00015:1e-20
Onc125.3	530..658	Contig54249.2	0	U4_snRNA	129nt	.	Rfam:RF00015:1e-20
Onc127	649..537	Contig44339.2	0	U5_snRNA	113nt	.	Rfam:RF00020:3e-12
Onc128	485..382	Contig70576.1	0	U6_snRNA	104nt	.	Rfam:RF00026:5e-29
Onc151	1312..1247	Contig200992	0	U4atac_snRNA	66nt	.	Rfam:RF00618:3e-3
Onc151.2	764..829	Contig42569.2	0	U4atac_snRNA	66nt	.	Rfam:RF00618:2e-3
Onc151.3	1676..1741	Contig76036.2	0	U4atac_snRNA	66nt	.	Rfam:RF00618:3e-3
Onc151.4	630..565	Contig71676.2	0	U4atac_snRNA	66nt	.	Rfam:RF00618:3e-3
Onc114	171..273	Contig73994.1	1	U6atac_snRNA	103nt	.	Rfam:RF00619:8e-06
Onc114.2	2568..2466	Contig74999.1	1	U6atac_snRNA	103nt	.	Rfam:RF00619:8e-06

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc114.3	168..270	Contig60127.1	0	U6atac_snRNA	103nt	.	Rfam:RF00619:8e-06
Onc124	2618..2387	Contig66111.1	0	U3_snoRNA	232nt	.	Rfam:RF00012:1e-18
Onc124.2	3175..2944	Contig68732.1	0	U3_snoRNA	232nt	.	Rfam:RF00012:1e-18
Onc78	1134..1206	Contig9982.1	4	C/D_snoRNA: U18	73nt	N	Rfam:RF01159:2e-05
Onc78.2	144..216	OXAE-aae58h12	4	C/D_snoRNA: U18	73nt	n	Rfam:RF01159:2e-05
Onc78.3	181..109	OXAO-aab15a01	1	C/D_snoRNA: U18	73nt	n	Rfam:RF01159:2e-05
Onc78.4	209..137	Contig36677.1	0	C/D_snoRNA: U18	73nt	n	Rfam:RF01159:2e-05
Onc78.5	242..170	Contig36684.2	0	C/D_snoRNA: U18	73nt	n	Rfam:RF01159:2e-05
Onc78.6	197..125	UGC1O0005_I02_R	0	C/D_snoRNA: U18	73nt	n	Rfam:RF01159:2e-05
Onc83	325..240	OXAO-aaa59f01	1	C/D_snoRNA: snoZ196	86nt	N	Rfam:RF00134:3e-06
Onc83.2	381..296	Contig77942.1	0	C/D_snoRNA: snoZ196	86nt	n	Rfam:RF00134:3e-06
Onc83.3	340..255	Contig77949.2	0	C/D_snoRNA: snoZ196	86nt	n	Rfam:RF00134:3e-06
Onc83.4	943..1028	Contig84242.1	0	C/D_snoRNA: snoZ196	86nt	n	Rfam:RF00134:3e-06
Onc84	1953..2040	Contig147.1	1	C/D_snoRNA: snoR38	88nt	N	Rfam:RF00213:3e-05
Onc84.2	1948..2035	Contig90550	0	C/D_snoRNA: snoR38	88nt	n	Rfam:RF00213:5e-05
Onc84.3	599..686	Contig90551	0	C/D_snoRNA: snoR38	88nt	n	Rfam:RF00213:5e-05
Onc84.4	1915..2002	Contig81691.1	0	C/D_snoRNA: snoR38	88nt	n	Rfam:RF00213:5e-05

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc84.5	523..610	Contig81698.2	0	C/D_snoRNA: snoR38	88nt	n	Rfam:RF00213:5e-05
Onc116	665..587	Contig60671.1	1	C/D_snoRNA: snoMe28S-Cm2645	79nt	.	Rfam:RF00530:2e-3
Onc116.2	1056..1134	Contig74810.1	1	C/D_snoRNA: snoMe28S-Cm2645	79nt	.	Rfam:RF00530:2e-3
Onc116.3	698..620	Contig74184.1	0	C/D_snoRNA: snoMe28S-Cm2645	79nt	.	Rfam:RF00530:2e-3
Onc108	688..609	Contig83501.3	0	C/D_snoRNA: SNORD36	80nt	N	Rfam:RF00049:4e-3
Onc109	1777..1691	Contig201200	0	C/D_snoRNA: SNORD24	87nt	N	Rfam:RF00069:2e-04
Onc110	1389..1318	Contig76679.1	1	C/D_snoRNA	72nt	N	snoscan
Onc110.2	1054..1123	Contig46873.1	1	C/D_snoRNA	70nt	n	snoscan
Onc110.3	1658..1589	Contig77473.1	0	C/D_snoRNA	70nt	n	snoscan
Onc111	4173..4099	Contig70178.1	1	C/D_snoRNA	75nt	N	snoscan
Onc111.2	116..190	Contig546.2	0	C/D_snoRNA	75nt	n	snoscan
Onc112	1164..1066	Contig80897.1	1	C/D_snoRNA	99nt	N	snoscan
Onc112.2	2005..2103	Contig51398.1	0	C/D_snoRNA	99nt	n	snoscan
Onc112.3	2403..2501	Contig66429.1	0	C/D_snoRNA	99nt	n	snoscan
Onc130	3208..3126	Contig201200	0	C/D_snoRNA: snR77	83nt	.	array:Onc109
Onc131	2866..2794	Contig201200	0	C/D_snoRNA	73nt	.	array:Onc109
Onc132	2494..2410	Contig201200	0	C/D_snoRNA	85nt	.	array:Onc109

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc133	2249..2124	Contig201200	0	C/D_snoRNA	126nt	.	array:Onc109
Onc134	1930..1871	Contig201200	0	C/D_snoRNA	60nt	.	array:Onc109
Onc135	1541..1474	Contig201200	0	C/D_snoRNA	68nt	.	array:Onc109
Onc136	1362..1301	Contig201200	0	C/D_snoRNA	62nt	.	array:Onc109
Onc137	1225..1136	Contig201200	0	C/D_snoRNA	90nt	.	array:Onc109
Onc138	962..878	Contig201200	0	C/D_snoRNA	85nt	.	array:Onc109
Onc139	689..615	Contig201200	0	C/D_snoRNA	75nt	.	array:Onc109
Onc140	416..332	Contig201200	0	C/D_snoRNA	85nt	.	array:Onc109
Onc141	1263..1180	Contig83501.3	0	C/D_snoRNA	84nt	.	array:Onc108
Onc142	923..830	Contig83501.3	0	C/D_snoRNA	94nt	.	array:Onc108
Onc143	342..428	Contig83501.3	0	C/D_snoRNA	87nt	.	array:Onc108
Onc144	754..836	Contig4340.2	5	H/ACA_snoRNA	83nt	.	array:Onc86
Onc130.2	1885..1803	Contig1162.2	0	C/D_snoRNA: snR77	83nt	.	array
Onc131.2	1547..1475	Contig1162.2	0	C/D_snoRNA	73nt	.	array
Onc132.2	1176..1092	Contig1162.2	0	C/D_snoRNA	85nt	.	array
Onc133.2	931..806	Contig1162.2	0	C/D_snoRNA	126nt	.	array
Onc134.2	612..553	Contig1162.2	0	C/D_snoRNA	60nt	.	array

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc109.2	459..373	Contig1162.2	0	C/D_snoRNA	87nt	.	array
Onc135.2	224..158	Contig1162.2	0	C/D_snoRNA	67nt	.	array
Onc130.3	224..306	Contig19537.1	0	C/D_snoRNA: snR77	83nt	.	array
Onc131.3	562..634	Contig19537.1	0	C/D_snoRNA	73nt	.	array
Onc132.3	933..1017	Contig19537.1	0	C/D_snoRNA	85nt	.	array
Onc133.3	1178..1303	Contig19537.1	0	C/D_snoRNA	126nt	.	array
Onc130.4	442..360	Contig201201	0	C/D_snoRNA: snR77	83nt	.	array
Onc131.4	103..31	Contig201201	0	C/D_snoRNA	73nt	.	array
Onc137.2	1209..1129	Contig1162.1	0	C/D_snoRNA	81nt	.	array
Onc138.2	954..871	Contig1162.1	0	C/D_snoRNA	84nt	.	array
Onc139.2	682..608	Contig1162.1	0	C/D_snoRNA	75nt	.	array
Onc140.2	409..325	Contig1162.1	0	C/D_snoRNA	85nt	.	array
Onc137.3	1218..1130	Contig46059.1	0	C/D_snoRNA	89nt	.	array
Onc138.3	954..871	Contig46059.1	0	C/D_snoRNA	84nt	.	array
Onc139.3	682..608	Contig46059.1	0	C/D_snoRNA	75nt	.	array
Onc140.3	409..325	Contig46059.1	0	C/D_snoRNA	85nt	.	array
Onc145	492..562	Contig6909.1	1	C/D_snoRNA	71nt	-	3box_screen

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc146	1972..1887	Contig47518.1	1	C/D_snoRNA	86nt	.	3box_screen
Onc146.2	1037..1122	Contig92891	0	C/D_snoRNA	86nt	.	3box_screen
Onc146.3	1965..1880	Contig42170.1	0	C/D_snoRNA	86nt	.	3box_screen
Onc146.4	167..252	Contig92893	0	C/D_snoRNA	86nt	.	3box_screen
Onc147	3094..2955	Contig48640.1	0	H/ACA_snoRNA	140nt	.	3box_screen
Onc147.2	709..570	Contig61584.1	0	H/ACA_snoRNA	140nt	.	3box_screen
Onc147.3	999..1138	Contig8120.2	0	H/ACA_snoRNA	140nt	.	3box_screen
Onc147.4	1003..1142	Contig74386.1	0	H/ACA_snoRNA	140nt	.	3box_screen
Onc148	2280..2223	Contig79173.1	1	C/D_snoRNA	58nt	.	3box_screen
Onc148.2	163..220	Contig36772.1	0	C/D_snoRNA	58nt	.	3box_screen
Onc154	674..545	Contig50244.2	3	H/ACA_snoRNA	130nt	.	3box_screen
Onc154.2	706..577	Contig47714.1	3	H/ACA_snoRNA	130nt	.	3box_screen
Onc154.3	330..459	Contig7084.1	3	H/ACA_snoRNA	130nt	.	3box_screen
Onc154.4	515..386	Contig50237.1	3	H/ACA_snoRNA	130nt	.	3box_screen
Onc85	395..317	Contig93299	5	C/D_snoRNA: SNORD96	79nt	N	nano_screen
Onc85.2	393..315	Contig76610.1	0	C/D_snoRNA: SNORD96	79nt	n	nano_screen
Onc86	360..431	Contig4340.2	5	C/D_snoRNA	72nt	N/R	nano_screen

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc87	418..496	Contig23611.1	5	C/D_snoRNA	79nt	N/R	nano_screen
Onc87.2	772..850	Contig80821.1	4	C/D_snoRNA	79nt	n/r	nano_screen
Onc87.3	407..485	OXAC-aaa05b11	4	C/D_snoRNA	79nt	n/r	nano_screen
Onc87.4	515..437	OXAE-aaa21b05	4	C/D_snoRNA	79nt	n/r	nano_screen
Onc87.5	419..497	Contig47479.1	4	C/D_snoRNA	79nt	n/r	nano_screen
Onc87.6	530..452	Contig63555.1	4	C/D_snoRNA	79nt	n/r	nano_screen
Onc89	550..418	Contig204907	5	H/ACA_snoRNA	133nt	N/R	nano_screen
Onc89.2	280..148	Contig63528.1	0	H/ACA_snoRNA	133nt	n/r	nano_screen
Onc89.3	301..169	Contig204908	0	H/ACA_snoRNA	133nt	n/r	nano_screen
Onc89.4	281..149	Contig60002.1	0	H/ACA_snoRNA	133nt	n/r	nano_screen
Onc89.5	550..682	Contig42459.2	0	H/ACA_snoRNA	133nt	n/r	nano_screen
Onc90	485..241	UGC1O0002_K14_R	5	H/ACA_snoRNA: U17/snR30	245nt	N/R	nano_screen
Onc90.2	527..283	Contig49311.1	0	H/ACA_snoRNA	245nt	n/r	nano_screen
Onc90.3	607..363	Contig68403.1	0	H/ACA_snoRNA	245nt	n/r	nano_screen
Onc90.4	3492..3736	Contig53912.1	0	H/ACA_snoRNA	245nt	n/r	nano_screen
Onc91	165..253	Contig63727.1	5	arisong	89nt	N/R	nano_screen
Onc91.2	752..664	Contig71315.1	4	arisong	89nt	n/r	nano_screen

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc92	693..610	Contig36242.1	4	arisong	84nt	.	nano_screen
Onc92.2	534..451	OXAO-aab16b04	4	arisong	84nt	.	nano_screen
Onc92.3	543..460	Contig65300.1	0	arisong	84nt	.	nano_screen
Onc94	761..670	Contig13832.1	5	arisong	92nt	N/R	nano_screen
Onc95	580..496	Contig44542.1	4	arisong	85nt	R	nano_screen
Onc95.2	132..216	OXAD-aaa04a12	4	arisong	85nt	r	nano_screen
Onc95.3	531..447	Contig57964.1	0	arisong	85nt	r	nano_screen
Onc96	285..146	OXAO-aab17f07	4	arisong	140nt	R	nano_screen
Onc155	343..428	Contig48963.1	1	arisong	86nt	.	nano_screen
Onc155.2	255..340	Contig65636.1	0	arisong	86nt	.	nano_screen
Onc155.3	209..294	Contig38772.1	0	arisong	86nt	.	nano_screen
Onc156	1587..1719	Contig203665	1	arisong	133nt	.	nano_screen
Onc156.2	628..760	Contig203666	0	arisong	133nt	.	nano_screen
Onc97	371..499	Contig91146	5	?	129nt	-	nano_screen
Onc98	444..554	Contig63260.1	5	?	111nt	-	nano_screen
Onc99	18..123	OXAE-aae57g05	5	?	106nt	-	nano_screen
Onc100	281..396	Contig40627.1	5	?	116nt	-	nano_screen

Continued on next page

Table C.1 – continued from previous page

Locus name	Coordinate	Contig name	Stage	Locus type	Length	Experiments	Etc
Onc100.2	325..440	Contig65897.1	4	?	116nt	-	nano_screen
Onc102	21..94	OXAE-aae64g09	5	?	74nt	-	nano_screen
Onc103	657..717	Contig47258.1	5	?	61nt	-	nano_screen
Onc104	1447..1675	Contig64625.1	5	?	229nt	-	nano_screen
Onc105	489..553	Contig63601.1	5	?	65nt	-	nano_screen
Onc105.2	241..279	Contig69957.1	4	?	39nt	-	nano_screen
Onc106	480..554	Contig54011.1	5	?	75nt	-	nano_screen
Onc107	40..130	OXAE-aaa57c10	5	?	91nt	-	nano_screen

Appendix D

Appendix: Northern blot experiments

D.1 Northern blot results for the known genes

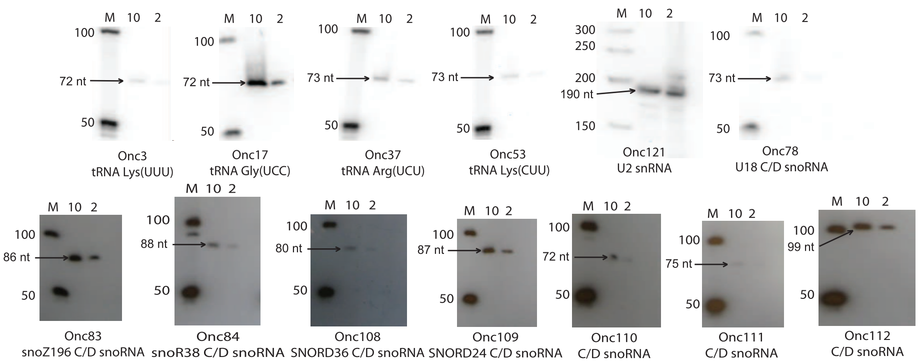


Figure D.1: Experimental confirmation of known RNA gene transcripts.

10/2 lanes indicate the amount of total RNA loaded in each lane (in μg). M indicates a radiolabeled 50bp DNA ladder. The arrow indicates the transcript band in blot.

D.2 List of probe sequences for Northern blots.

“AS” in the “direction” column in tables indicates that the probe sequence is reverse-complementary to the reference genome sequence.

D.2.1 Northern blot probes for the final candidate genes

Table D.1: Northern blot probes for the final candidate genes

Contig name	**Gene name**	**Direction**	**Probe Size**	**Sequence**
Contig93299	Onc85(SNORD96)	S	40 nt	CACATAGTTCAGCCCCGAAAGATGACAGTTTTATAGAATC
Contig4340.2	Onc86(C/D snoRNA)	AS	42 nt	AGACACGAGGAATTCAGTTGGTTGATCCGGTTTTTTCATCAT
Contig23611.1	Onc87(C/D snoRNA)	AS	44 nt	CAGTAGGAGTGGAGTTATATTTATCAACACGTTTGATTCTGTTG
Contig204907	Onc89(H/ACA snoRNA)	S	41 nt	CCACAGCCGAATCAATAGTCAACTGCGGTCCATTAAATTCC
UGC1O0002_K14_R	Onc90(snR30/U17)	S	41 nt	CACGGCAGGAGCGAGCGAATCAACTCAACCACCTCTCTCCT
Contig63727.1	Onc91(Arisong)	S	42 nt	GTCTTAAGCCAGTGTAACTGGTTGCGGGTGAGGGACCTATTC
Contig13832.1.1	Onc94(Arisong)	S	39 nt	CTCAGGAACTTTGTGTCCCCAAGCCGCAGAGGCCGGACC

D.2.2 Northern blot probes for known genes

Table D.2: Northern blot probes for known genes

Class	**Contig name**	**Gene name**	**Direction**	**Probe Size**	**Sequence**
tRNA	Contig200270	Onc3(tRNA/Lys)	S	39 nt	GATTAAAAGTCTGGCGCTCTACCACTGAGCTAGACGGGC
	Contig72578.1	Onc17(tRNA/Gly)	AS	42 nt	CCGGGTCAAGTGCTTGGAAGGCACCTATCCTAACCACTAGAC
	Contig71359.1	Onc37(tRNA/Arg)	S	40 nt	GATTAGAAGTCTGATGCGCTATCCATTGCGCCACGAAGAC
	OXAO-aab15f07	Onc53(tRNA/Lys)	AS	40 nt	GGTTAAGAGCCAAGCGCTCTACCGACTGAGCTAGACGGGC
C/D snoRNA	Contig9982.1	Onc78(U18)	AS	40 nt	TGAGTTAGAGTCAGACATTGGACAGGTTATCGTCAATCGA
	OXAO-aaa59f01	Onc83(snoZ196)	S	41 nt	GGTGTGTATGAGTTGTATCATCAATGAATGACTCAGTGTGG
	Contig147.1	Onc84(snoR38)	AS	38 nt	CTCATCAATGATCTTGTCTATGACAGGGATAACTGTTG
	Contig83501.3	Onc108(SNORD36)	S	43 nt	GTTCATCAAGAAAATTATGTCGTAAAATAACAAGTGTATCATC
	Contig201200	Onc109(SNORD24)	S	40 nt	GGCCCTTTCGAGTCATGATCAGAAGTAGCAATTATTTTTG
	Contig76679.1	Onc110(C/D snoRNA)	S	40 nt	GTCAGAATTGCAGAACCATATATCGTCAAATTGATTTCAG
	Contig70178.1	Onc111(C/D snoRNA)	S	38 nt	GTAAGAATCACAGGGATTGTCATAAAGAACGCAGCAAC
	Contig80897.1	Onc112(C/D snoRNA)	S	40 nt	CGCCAATGGGTTCATGTATCAGCGACAATAGCCAACCTTC
snRNA	Contig36667.2	Onc121(U2)	S	42 nt	AAAGTGTAGGTCCAAGGCGACTCTGTAAGAGTGATGCGCAAG

D.2.3 Probes sequences for negative Northern blots

For snoRNA candidate, Northern blot was done on the predicted region. For the final candidates in stage 5 dataset, Northern blot was done twice with sense and antisense probe sequence on the same candidate region of which G/C contents is relatively high.

Table D.3: Northern blot probes for the tested snoRNA candidates

Contig name	**Gene name**	**Direction**	**Probe Size**	**Sequence**
OXAO-aaa59f01	.	S	39 nt	CTTGGTTTCAATTCAGAAGAACGAAAGTAAATTAGCATC
Contig76351.1	.	AS	41 nt	GAGTGAGCCTGACTATAATAATGATCTATAAAATGAGAGCC
Contig6909.1	Onc145	AS	40 nt	CAGAGTAACTATGACGGCATCCATCTCATTTAGAGTCATG

Table D.4: Northern blot probes for the tested final candidates

Contig name	Gene name	Probe Size	Sequence
Contig91146	Onc97	42 nt	CTAATTAACACAGTCTTAATTAAAATATTAATATTCCCTCTC
Contig63260.1	Onc98	42 nt	CTCCAAAAACCTAGCCAACCTCACTTAAAATAAAGCAGATGG
OXAE-aae57g05	Onc99	46 nt	CCACATTTTTAGATTTAGTTTTTATATCTTTTTTATGGTTAATTTG
Contig40627.1	Onc100	39 nt	CTTGAGTGGCCCCCTGAAATGTGAAAGAGTCACAAAGCC
OXAE-aae64g09	Onc102	39 nt	CCTCGAAGACGAAGACAGCAGACAGAGAACTTTGAAGAC
		43 nt	GAACGGAAAGTACGAAGTTCCCTTAGGACTCAACCTCGAAGAC
Contig47258.1	Onc103	40 nt	GAAGCACAATGGATCTTATTTAGAGTAGAGAATGAAAATG
Contig64625.1	Onc104	41 nt	CCAGTACCGTGGAGTCTCAAAGAACGGGATTTAATGGCAGG
Contig63601.1	Onc105	40 nt	CAACTCATTACATGGACGAAGCTGATATTCTTGTTGAGAG
Contig54011.1	Onc106	38 nt	GTTGGAGTTTAAATGTTTGATTAAAGAAAATTTAGTAG
OXAE-aaa57c10	Onc107	39 nt	CATTAATAATTTGAAAATATAAAGTTCTTAATAACATCC

Appendix E

Appendix: RACE probes

List of gene specific probe (GSP) sequences of RACE experiments.

Table E.1: RACE-PCR GSPs for the final candidate genes

Contig name	Gene name	Probe	Probe Size	Sequence
Contig4340.2	Onc86 (C/D snoRNA)	5GSP	35	GTCAGGTTATAGATCTGTCTACATGAAGACACGAG
Contig4340.2	Onc86 (C/D snoRNA)	3GSP	30	CCGGATCAACCAACTGAATTCCTCGTGTCT
Contig23611.1	Onc87 (C/D snoRNA)	5GSP	35	CAGTAGGAGTGGAGTTATATTTATCAACACGTTTG
Contig23611.1	Onc87 (C/D snoRNA)	3GSP	33	ACGATGAAGTAGTTTATAATCCGTGTTTCAACA
Contig204907	Onc89 (H/ACA snoRNA)	5GSP	33	CCACAGCCGAATCAATAGTCAACTGCGGTCCAT
Contig204907	Onc89 (H/ACA snoRNA)	3GSP	29	GTTGACTATTGATTCGGCTGTGGTTAAGT
UGC1O0002K_14_R	Onc90 (H/ACA snoRNA)	5GSP	29	GAGGACCCGTAAGTCACGGCAGGAGCGAG
UGC1O0002K_14_R	Onc90 (H/ACA snoRNA)	3GSP	32	GGAGAGAGGTGGTTGAGTTGATTCGCTCGCTC
UGC1O0002K_14_R	Onc90 (H/ACA snoRNA)	3GSP2	25	GCCTTGGACTGATTAGGACTCCGTC
Contig63727.1	Onc91 (classII)	5GSP	31	CGGGTTCAGGATCCCGAATAGGTCCCTCACC
Contig63727.1	Onc91 (classII)	3GSP	31	GGTCTTAAGCCAGTGTAACTGGTTGCGGGTG
Contig63727.1	Onc91 (classII)	3GSP2	24	CAACAGTAACCAATACTTTCGAGG
Contig13832.1	Onc94 (classII)	5GSP	29	CAGGAACTTTGTGTCCCCAAGCCGCAGAG
Contig13832.1	Onc94 (classII)	3GSP	31	GGTCCGGCCTCTGCGGCTTGGGGACACAAAG
Contig44542.1	Onc95 (classII)	5GSP	31	CCTTTGTGGAAACACCCCGCAGAGGCCATAC
Contig44542.1	Onc95 (classII)	3GSP	31	GGTATGGCCTCTGCGGGGTGTTTCCACAAAG
OXAO_aab17f07	Onc96 (classII)	5SGP	29	ATATGGCCCATCCCCGCAGCAGCCGGACT
OXAO_aab17f07	Onc96 (classII)	3GSP	30	GTCCGGCTGCTGCGGGGATGGGCCATATTG

Appendix F

Appendix: Comparative analysis on the stage 5 dataset

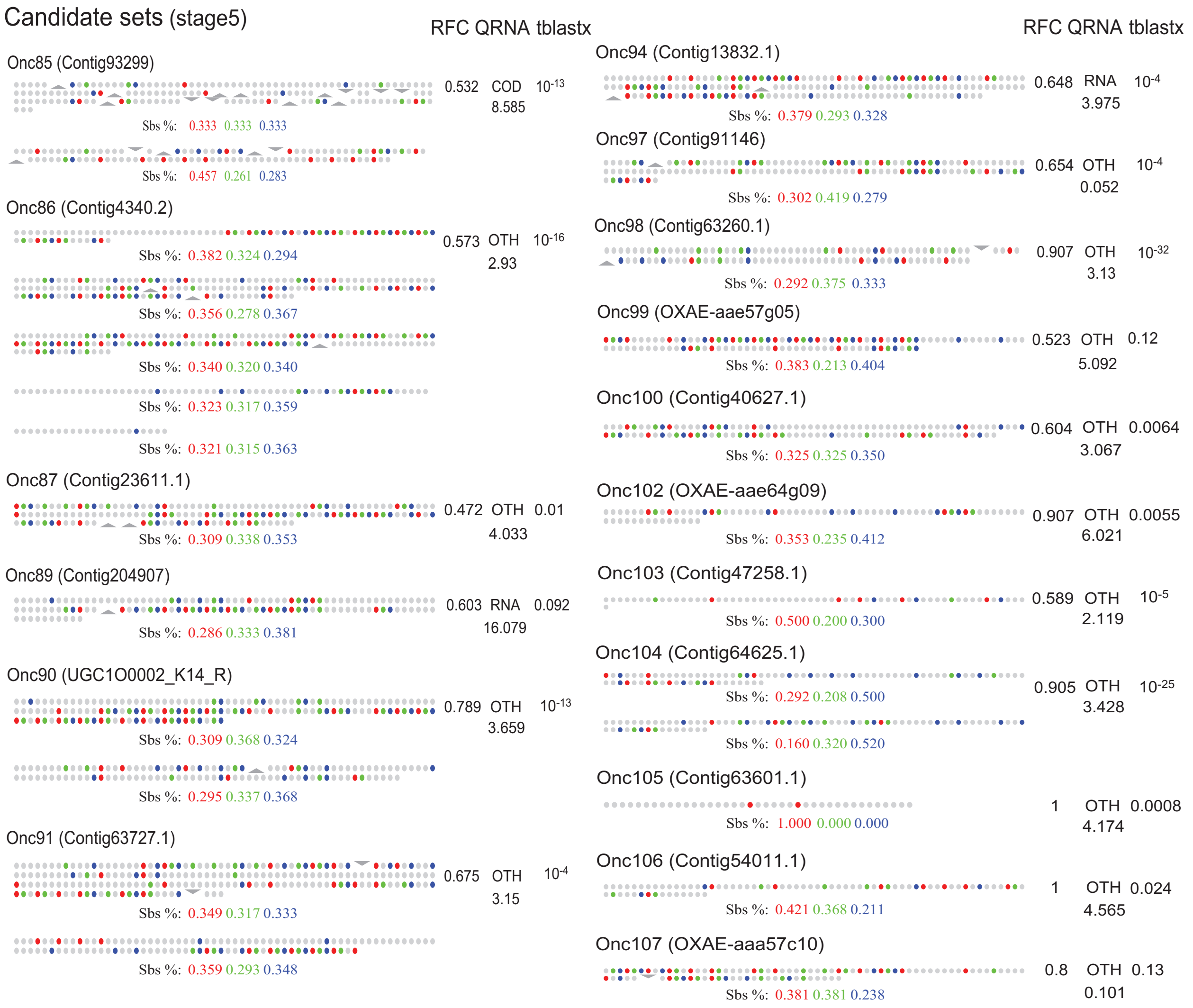


Figure F.1: *Stylonychia* conservation patterns of the final candidate nanochromosomes
The left column on the alignment shows the RFC score, QRNA's class and score and the best evalue of tblastx to the NR database

Appendix G

Appendix: Sequences of regulatory motifs

A. Sequences for PSE motif search

U1	TATTTATGACCCATAAATATTTAGGCCA
U2	AAAATATGACCCATTAATATTTAACGGA
U3	AATAGTTAACCCATTAATAATTTGGTAG
U4	TAAAATTAACCCATAAATATTTAAGTGG
U5	TAAATATGACCCATTAATATTTAAATCA
U6	ATAATTTAACCCATTAATATTTAAGGTG
SRP	TATAACTAACCCATAAACTTTTAATTAA
RNaseP	AATAAATGACCCATTAACTATTAATCTG
Telomerase	AAATATTGACCCATAAATATTTAAGCGG
styU6	TAATTCTGACCCATTAACAAATAGCGAG
StyRNaseP	AATAAGCAACCCATTAACTTTTAATTCT
StySRP	AAAATATGACCCATAAACTATTAGAATT

C. Sequences for 3’ box motif search

Onc120.1/U1	AAATGAAaa.GTTTGA.TTAG
Onc121/U2	AAAGGAtaatGTTTGA.TTAT
Onc123.1/U2	AAATGATaatGTTTGA.TTAG
Onc77.1/U2	AAAGGAtaatGTTTGA.TTAG
Onc125.1/U4	AAATGAAat.GTTTGA.TTAG
Onc127/U5	AAATGAAattGTTTGA.TTAG
StyOnc120/U1	AAATGAAtt.GTTTGAaTTAG
StyOnc127/U5	AAATGAAtt.GTTTGAaTTAA
Onc124.1/U3	AAAGGAAttaGTTTGA.TTAG
Onc91.1/ClassII	AAATGAACTTGTTTGA.TTAG
Onc92/ClassII	AAATGAACTCGTTTGA.TTAT
Onc94/ClassII	AAATGAAAT.GTTTGA.ATAA
Onc95/ClassII	AAATGAAATAGTTTGA.GTAG
Onc96/ClassII	AAATGAAAA.GTTTGATTTAG
Onc95.3/ClassII	AAATGAAAA.GTTTGA.TTAG
Onc89/HACA	AAGGGAAATTGTTTGA.TTAG
Onc90/HACA	AAATGAAAACGTGTGA.TTAG

B. Sequences of Oxytricha PSE

Onc120/U1	TGACCCATAAATATTTA
Onc121/U2	TGACCCATTAATATTTA
Onc123/U2	TGACCCATTAGTATTTA
Onc77/U2	TGACCCATTAGTATTTA
Onc124.1/U3	TAACCCATTAATAATTT
Onc125/U4	TGACCCATTAATATTTA
Onc127/U5	TGACCCATTAATATTTA
Onc128/U6	TAACCCATTAATATTTA
Onc151/U4atac	TAACCCATAGAAACTTA
Onc114/U6atac	TGACCCATAGAAAATTA
Onc81/SRP	TAACCCATAAACTTTTA
Onc149/RNaseP	TGACCCATTAACTATTA
Onc150/Telomerase	TGACCCATAAATATTTA
Onc91/classII	TGACCCATGAATTATTA
Onc92/classII	TAACCCATAAATAATTA
Onc94/classII	TTACCCATAAACAATTA
Onc95/classII	TGACCCATTAATATTTA
Onc96/classII	TGACCCATTAAAAGTTA
Onc155/classII	GATCCCATCAATTTTAT
Onc156/classII	TAACCCATTAATAATTA

D. Sequences of Oxytricha 3’ box

Onc120/U1	AAATGAAAA.GTTTGA..TTAG
Onc121/U2	AAAGGATAAtGTTTGA..TTAT
Onc123.1/U2	AAATGATAAtGTTTGA..TTAG
Onc77.1/U2	AAAGGATAAtGTTTGA..TTAG
Onc124/U3	AAAGGAATTaGTTTGA..TTAG
Onc125/U4	AAATGAAAT.GTTTGA..TTAG
Onc127/U5	AAATGAAATtGTTTGA..TTAG
Onc151/U4atac	AAATGAAAAtGTTTGTttTTAT
Onc91/ClassII	AAATGAACTaGTTTGA..TTAG
Onc92/ClassII	AAATGAACTcGTTTGA..TTAT
Onc94/ClassII	AAATGAAAT.GTTTGA..ATAA
Onc95/ClassII	AAATGAAATaGTTTGA..GTAG
Onc96/ClassII	AAATGAAAA.GTTTGAt.TTAG
Onc155/ClassII	AAATGAAAA.GTTTGA..TTAG
Onc156/classII	AAATGAAAT.GTATGA..GTAA
Onc89/HACA	AAGGGAAATtGTTTGA..TTAG
Onc90/HACA	AAATGAAAAcGTGTGA..TTAG
Onc145/CD	AAAGGAAATaGTTTGA..TTAG
Onc146/CD	AAATGAAATgGTTTGA..TTAG
Onc147/HACA	AAATGAAATgGTTTGA..GTAG
Onc148/CD	AAATGAAAAaATTTGA..TTAG
Onc154/HACA	AAAGGAATAtGTTTGA..TTAT

Figure G.1: Instances of PSE and 3’ box motif sequences

A and C are sequence alignments used to build HMM models for PSE and 3’ box screening, respectively. (”Sty” in front of gene name indicates the *Stylonychia* sequences.) B and D are sequence alignments of *O. trifallax* PSE and 3’ box motifs, respectively.

Appendix H

Appendix: Information for programs, databases and datasets

H.1 Programs and databases

Infernal 1.0.2 was used for RNA similarity searches [129]. Infernal models of known ncRNA families were from the Rfam 9.1 database [178]. For routine sequence manipulations we used a variety of miniapps provided by the Easel library package included in Infernal 1.0.2. All BLAST comparisons used Washington University BLAST (WU-BLAST) version 2.0MP-WashU [04-May-2006]. All comparisons to the NCBI NR protein database used a version of NR downloaded on 13 April 2009. In the screen, to remove nanochromosomes containing a detectable homolog of known protein, UniProt/Swissprot database version 50.8 downloaded on October 2006 was used. To evaluate the performance for nanochromosome classification, Genezilla [194], Unveil v1.0 [195], GeneID v1.2 [196] and Augustus 2.0 [197] were examined. To evaluate the performance for nanogenefinder, the same programs and additionally Genscan [200] were used. PSE and 3′ box consensus motif searches were done with HMMER 2.3.2. Multiple alignments were produced using MUSCLE [240] or CLUSTALW [214] and manually edited in Emacs using the RALEE alignment editing mode [241]. For a computational prediction of additional snoRNAs, snoGPS 0.2 [209] and snoscan 0.9b [134] were used. List of conserved pseudouridylation target sites in human and yeast is extracted from the SnoRNABase database version3 website `http://www-snorna.biotoul.fr`. Analysis of cDNA/genome alignments used Exonerate 1.0.2 [192], and unpublished cDNA/EST data. For comparative analysis of coding gene sequence conservation patterns, we used QRNA 2.0.3c [126]. Sequence logos were generated with WebLogo 2.8.2 [218].

H.2 Dataset availability.

A compressed tar archive containing the *Oxytricha* and *Stylonychia* sequence data, the nanoclassifier source code, training and test data, parsable tables of results, and other datasets described in the paper are available for download at `http://selab.janelia.org/publications.html/#JungEddy11`.

H.3 Accession number

A modified version of the *Stylonychia* data was deposited to DDBJ/EMBL/GenBank (accession ADNZ01000000) after trimming terminal Ns and removing 951 contigs deemed to be low-quality or foreign contamination.

Bibliography

[1] P. A. Sharp. The centrality of RNA. *Cell*, 136:577–580, 2009.

[2] F. Jacob and J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.*, 3:318–356, 1961.

[3] G. E. Palade. A small particulate component of the cytoplasm. *J. Biophys. Biochem. Cytol.*, 1:59–67, 1955.

[4] J. L. Hodnett and H. Busch. Isolation and characterization of uridylic acid-rich 7S ribonucleic acid of rat liver nuclei. *J. Biol. Chem.*, 243:6334–6342, 1968.

[5] H. Busch, R. Reddy, L. Rothblum, and Y. C. Choi. SnRNAs, SnRNPs, and RNA processing. *Annu. Rev. Biochem.*, 5:617–654, 1982.

[6] G. Zieve and S. Penman. Small RNA species of the HeLa cell: Metabolism and subcellular localization. *Cell*, 8:19–31, 1976.

[7] V. T. Nguyen, T. Kiss, A. A. Michels, and O. Bensaude. 7SK small nuclear RNA binds to and inhibits the activity of CDK9/cyclin T complexes. *Nature*, 414:322–325, 2001.

[8] Z. Yang, Q. Zhu, K. Luo, and Q. Zhou. The 7SK small nuclear RNA inhibits the CDK9/cyclin T1 kinase to control transcription. *Nature*, 414:317–322, 2001.

[9] D. A. Dunbar, S. Wormsley, T. M. Lowe, and S. J. Baserga. Fibrillarin-associated box C/D small nucleolar RNAs in *Trypanosoma brucei*. Sequence conservation and implications for 2'-O-ribose methylation of rRNA. *J. Biol. Chem.*, 275:14767–14776, 2000.

[10] L. H. Qu, Q. Meng, H. Zhou, and Y. Q. Chen. Identification of 10 novel snoRNA gene clusters from arabidopsis thaliana. *Nucl. Acids Res.*, 29:1623–1630, 2001.

[11] F. Barneche, C. Gaspin, R. Guyot, and M. Echeverria. Identification of 66 box C/D snoRNAs in *Arabidopsis thaliana*: Extensive gene duplications generated multiple isoforms predicting new ribosomal RNA 2'-O-methylation sites. *J. Mol. Biol.*, 311: 57–73, 2001.

[12] T. Kiss. Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, 109:145–148, 2002.

[13] J. Miranda-Ríos, M. Navarro, and M. Soberón. A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. USA*, 98:9736–9741, 2001.

[14] A. Nahvi, N. Sudarsan, M. S. Ebert, X. Zou, K. L. Brown, and R. R. Breaker. Genetic control by a metabolite binding mRNA. *Chem. Biol.*, 9:1043–1049, 2002.

[15] A. S. Mironov, I. Gusarov, R. Rafikov, L. E. Lopez, K. Shatalin, R. A. Kreneva, D. A.

Perumov, and E. Nudler. Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell*, 111:747–756, 2002.

[16] W. C. Winkler, S. Cohen-Chalamish, and R. R. Breaker. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. USA*, 99:15908–15913, 2002.

[17] A. G. Vitreschak, D. A. Rodionov, A. A. Mironov, and M. S. Gelfand. Regulation of the vitamin b12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA*, 9:1084–1097, 2003.

[18] V. Ambros and H. R. Horvitz. The lin-14 locus of *Caenorhabditis Elegans* controls the time of expression of specific postembryonic developmental events. *Genes Dev.*, 1:398–414, 1987.

[19] R. C. Lee, R. L. Feinbaum, and V. Ambros. The *C. elegans* heterochronic gene *lin-4* encodes small RNAs with antisense complementarity to *lin-14*. *Cell*, 75:843–854, 1993.

[20] B. J. Reinhart, F. J. Slack, M. Basson, A. E. Pasquinelli, J. C. Bettinger, A. E. Rougvie, H. R. Horvitz, and G. Ruvkun. The 21-nucleotide *let-7* RNA regulates developmental timing in *Caenorhabditis elegans*. *Nature*, 403:901–906, 2000.

[21] A. E. Pasquinelli, B. J. Reinhart, F. Slack, M. Q. Martindale, M. I. Kuroda, B. Maller, D. C. Hayward, E. E. Ball, B. Degnan, P. Müller, Jürg Spring, A. Srinivasan, M. Fishman, J. Finnerty, J. Corbo, M. Levine, P. Leahy, E. Davidson, and G. Ruvkun. Conservation of the sequence and temporal expression of *let-7* heterochronic regulatory RNA. *Nature*, 408:86–89, 2000.

[22] D. P. Bartel. MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, 116:281–297, 2004.

[23] L. He and G. J. Hannon. MicroRNAs: small RNAs with a big role in gene regulation. *Nat. Rev. Genet.*, 5:522–531, 2004.

[24] K. M. Wassarman, F. Repoila, C. Rosenow, G. Storz, and S. Gottesman. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.*, 15:1637–1651, 2001.

[25] A. Hüttenhofer, M. Kiefmann, S. Meier-Ewert, J. O'Brien, H. Lehrach, J. P. Bachellerie, and J. Brosius. RNomics: an experimental approach that identifies 201 candidates for novel, small, non-messenger RNAs in mouse. *EMBO J.*, 20:2943–2953, 2001.

[26] R. J. Klein, Z. Misulovin, and S. R. Eddy. Noncoding RNA genes identified in AT-rich hyperthermophiles. *Proc. Natl. Acad. Sci. USA*, 99:7542–7547, 2002.

[27] C. Llave, K. D. Kasschau, M. A. Rector, and J. C. Carrington. Endogenous and silencing-associated small RNAs in plants. *Plant Cell*, 14:1605–1619, 2002.

[28] T. H. Tang, J. P. Bachellerie, T. Rozhdestvensky, M. L. Bortolin, H. Huber, M. Drungowski, T. Elge, J. Brosius, and A. Hüttenhofer. Identification of 86 candidates for small non-messenger RNAs from the Archaeon *Archaeoglobus fulgidus*. *Proc. Natl. Acad. Sci. USA*, 99:7536–7541, 2002.

[29] 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. Schonbach, 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. Brusic, 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. Kawasawa, 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, J. Reid, 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, and Y. Hayashizaki. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, 420:563–573, 2002.

[30] J. S. Mattick. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, 2:986–991, 2001.

[31] J. S. Mattick. The hidden genetic program of complex organisms. *Sci. Am.*, 291: 60–67, 2004.

[32] T. W. Nilsen. RNA 1997–2007: A remarkable decade of discovery. *Mol. Cell*, 28: 715–720, 2008.

[33] J. R. Hogg and K. Collins. Structured non-coding RNAs and the RNP renaissance. *Curr. Opin. Chem. Biol.*, 12:684–689, 2008.

[34] J. E. Wilusz, H. Sunwoo, and D. L. Spector. Long noncoding RNAs: Functional surprises from the RNA world. *Genes Dev.*, 23:1494–1504, 2009.

[35] J. S. Mattick. The genetic signatures of noncoding RNAs. *PLoS Genetics*, 5: e1000459, 2009.

[36] J. S. Mattick and I. V. Makunin06. Non-coding RNA. *Hum. Mol. Genet.*, 15:R17–R29, 2006.

[37] P. Carninci. RNA dust: Where are the genes? *DNA Res.*, 17:51–59, 2010.

[38] F. Crick. Central dogma of molecular biology. *Nature*, 227:561–563, 1970.

[39] J. M. Bishop, W. E. Levinson, D. Sullivan, L. Fanshier, N. Quintrell, and J. Jackson. The low molecular weight RNAs of rous sarcoma virus. II. the 7 s RNA. *Virology*, 42:927–937, 1970.

[40] T. A. Walker, N. R. Pace, R. L. Erikson, E. Erikson, and F. Behr. The 7S RNA common to oncornaviruses and normal cells is associated with polyribosomes. *Proc. Natl. Acad. Sci. USA*, 71:3390–3394, 1974.

[41] C. Guerrier-Takada, K. Gardiner, T. Marsh, N. Pace, and S. Altman. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell*, 35:849–857, 1983.

[42] T. R. Cech. The chemistry of self-splicing RNA and RNA enzymes. *Science*, 236: 1532–1539, 1987.

[43] W. Gilbert. The RNA world. *Nature*, 319:618, 1986.

[44] N. Sudarsan, J. E. Barrick, and R. R. Breaker. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, 9:644–647, 2003.

[45] A. W. Prestayko, M. Tonato, B. C. Lewis, and H. Busch. Heterogeneity of nucleolar u3 ribonucleic acid of the *Novikoff Hepatoma*. *J. Biol. Chem.*, 246:182–187, 1971.

[46] R. W. Carthew and E. J. Sontheimer. Origins and mechanisms of miRNAs and siRNAs. *Cell*, 136:642–655, 2009.

[47] T. Mizuno, M.-Y. Chou, and M. Inouye. A unique mechanism regulating gene expression: Translational inhibition by a complementary RNA transcript (micRNA). *Proc. Natl. Acad. Sci. USA*, 81:1966–1970, 1984.

[48] H. Aiba, S. Matsuyama, T. Mizuno, and S. Mizushima. Function of MicF as an antisense RNA in osmoregulatory expression of the OmpF gene in *Escherichia coli*. *J. Bacteriol.*, 169:3007–3012, 1987.

[49] G. H. Shull. ”GENES” OR ”GENS”? *Science*, 35:819, 1912.

[50] W. E. Castle. Piebald rats and the theory of genes. *Proc. Natl. Acad. Sci. USA*, 5: 126–130, 1919.

[51] N. Aziz and H. N. Munro. Iron regulates ferritin mRNA translation through a segment of its 5’ untranslated region. *Proc. Natl. Acad. Sci. USA*, 84:8478–8482, 1987.

[52] M. W. Hentze, S. W. Caughman, T. A. Rouault, J. G. Barriocanal, A. Dancis, J. B. Harford, and R. D. Klausner. Identification of the iron-responsive element for the translational regulation of human ferritin mRNA. *Science*, 238:1570–1573, 1987.

[53] D. J. Haile, M. W. Hentze, T. A. Rouault, J. B. Harford, and R. D. Klausner. Regulation of interaction of the iron-responsive element binding protein with iron-responsive RNA elements. *Mol. Cell Biol.*, 9:5055–5061, 1989.

[54] M. W. Hentze and L. C. Kühn. Molecular control of vertebrate iron metabolism: mRNA-based regulatory circuits operated by iron, nitric oxide, and oxidative stress. *Proc. Natl. Acad. Sci. USA*, 93:8175–8182, 1996.

[55] J. Pelletier and N. Sonenberg. Internal initiation of translation of eukaryotic mRNA directed by a sequence derived from poliovirus RNA. *Nature*, 334:320–325, 1988.

[56] S. K. Jang, H. G. Krusslich, M. J. Nicklin, G. M. Duke, A. C. Palmenberg, and E. Wimmer. A segment of the 5’ nontranslated region of *Encephalomyocarditis* virus RNA directs internal entry of ribosomes during in vitro translation. *J. Virol.*, 62:2636–2643, 1988.

[57] J. Fernandez, I. Yaman, W. C. Merrick, A. Koromilas, R. C. Wek, R. Sood, J. Hensold, and M. Hatzoglou. Regulation of internal ribosome entry site-mediated translation by eukaryotic initiation factor-2alpha phosphorylation and translation of a small upstream open reading frame. *J. Biol. Chem.*, 277:2050–2058, 2002.

[58] J. Y. Lee, C. F. Evans, and D. R. Engelke. Expression of RNase P RNA in *Saccharomyces Cerevisiae* is controlled by an unusual RNA polymerase III promoter. *Proc. Natl. Acad. Sci. USA*, 88:6986–6990, 1991.

[59] T. A. Farazi, S. A. Juranek, and T. Tuschl. The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development*, 135: 1201–1214, 2008.

[60] Y. Lee, M. Kim, J. Han, K. H. Yeom, S. Lee, S.H. Baek, and V. N. Kim. MicroRNA genes are transcribed by RNA polymerase II. *EMBO J.*, 23:4051–4060, 2004.

[61] V. Ambros, R. C. Lee, A. Lavanway, P. T. Williams, and D. Jewell. MicroRNAs and other tiny endogenous RNAs in *C. Elegans*. *Curr. Biol.*, 13:807–818, 2003.

[62] S. Y. Ying and S. L. Lin. Intron-derived microRNAs–fine tuning of gene functions. *Gene*, 342:25–28, 2004.

[63] A. Rodriguez, S. Griffiths-Jones, J. L. Ashurst, and A. Bradley. Identification of mammalian microRNA host genes and transcription units. *Genome Res.*, 14:1902–1910, 2004.

[64] L. B. Weinstein and J. A. Steitz. Guided tours: From precursor snoRNA to functional snoRNP. *Curr. Opin. Cell Biol.*, 11:378–384, 1999.

[65] K. T. Tycowski, M. D. Shu, and J. A. Steitz. A mammalian gene with introns instead of exons generating stable RNA products. *Nature*, 379:464–466, 1996.

[66] P. Pelczar and W. Filipowicz. The host gene for intronic U17 small nucleolar RNAs

in mammals has no protein–coding potential and is a member of the 5’–terminal oligopyrimidine gene family. *Mol. Cell. Biol.*, 18:4509–4518, 1998.

[67] A. D. Omer, T. M. Lowe, A. G. Russell, H. Ebhardt, S. R. Eddy, and P. P. Dennis. Homologs of small nucleolar RNAs in Archaea. *Science*, 288:517–522, 2000.

[68] W. Filipowicz and V. Pogacić. Biogenesis of small nucleolar ribonucleoproteins. *Curr. Opin. Cell Biol.*, 14:319–327, 2002.

[69] J. C. Venter, M. D. Adams, E. W. Myers, P. W. Li, R. J. Mural, G. G. Sutton, H. O. Smith, M. Yandell, C. A. Evans, R. A. Holt, J. D. Gocayne, P. Amanatides, R. M. Ballew, D. H. Huson, J. R. Wortman, Q. Zhang, C. D. Kodira, X. H. Zheng, L. Chen, M. Skupski, G. Subramanian, P. D. Thomas, J. Zhang, G. L. Gabor Miklos, C. Nelson, S. Broder, A. G. Clark, J. Nadeau, V. A. McKusick, N. Zinder, A. J. Levine, R. J. Roberts, M. Simon, C. Slayman, M. Hunkapiller, R. Bolanos, A. Delcher, I. Dew, D. Fasulo, M. Flanigan, L. Florea, A. Halpern, S. Hannenhalli, S. Kravitz, S. Levy, C. Mobarry, K. Reinert, K. Remington, J. Abu-Threideh, E. Beasley, K. Biddick, V. Bonazzi, R. Brandon, M. Cargill, I. Chandramouliswaran, R. Charlab, K. Chaturvedi, Z. Deng, V. Di Francesco, P. Dunn, K. Eilbeck, C. Evangelista, A. E. Gabrielian, W. Gan, W. Ge, F. Gong, Z. Gu, P. Guan, T. J. Heiman, M. E. Higgins, R. R. Ji, Z. Ke, K. A. Ketchum, Z. Lai, Y. Lei, Z. Li, J. Li, Y. Liang, X. Lin, F. Lu, G. V. Merkulov, N. Milshina, H. M. Moore, A. K. Naik, V. A. Narayan, B. Neelam, D. Nusskern, D. B. Rusch, S. Salzberg, W. Shao, B. Shue, J. Sun, Z. Wang, A. Wang, X. Wang, J. Wang, M. Wei, R. Wides, C. Xiao, C. Yan, A. Yao, J. Ye, M. Zhan, W. Zhang, H. Zhang, Q. Zhao, L. Zheng, F. Zhong, W. Zhong, S. Zhu, S. Zhao, D. Gilbert, S. Baumhueter, G. Spier, C. Carter, A. Cravchik, T. Woodage,

F. Ali, H. An, A. Awe, D. Baldwin, H. Baden, M. Barnstead, I. Barrow, K. Beeson, D. Busam, A. Carver, A. Center, M. L. Cheng, L. Curry, S. Danaher, L. Davenport, R. Desilets, S. Dietz, K. Dodson, L. Doup, S. Ferriera, N. Garg, A. Gluecksmann, B. Hart, J. Haynes, C. Haynes, C. Heiner, S. Hladun, D. Hostin, J. Houck, T. Howland, C. Ibegwam, J. Johnson, F. Kalush, L. Kline, S. Koduru, A. Love, F. Mann, D. May, S. McCawley, T. McIntosh, I. McMullen, M. Moy, L. Moy, B. Murphy, K. Nelson, C. Pfannkoch, E. Pratts, V. Puri, H. Qureshi, M. Reardon, R. Rodriguez, Y. H. Rogers, D. Romblad, B. Ruhfel, R. Scott, C. Sitter, M. Smallwood, E. Stewart, R. Strong, E. Suh, R. Thomas, N. N. Tint, S. Tse, C. Vech, G. Wang, J. Wetter, S. Williams, M. Williams, S. Windsor, E. Winn-Deen, K. Wolfe, J. Zaveri, K. Zaveri, J. F. Abril, R. Guigo, M. J. Campbell, K. V. Sjolander, B. Karlak, A. Kejariwal, H. Mi, B. Lazareva, T. Hatton, A. Narechania, K. Diemer, A. Muruganujan, N. Guo, S. Sato, V. Bafna, S. Istrail, R. Lippert, R. Schwartz, B. Walenz, S. Yooseph, D. Allen, A. Basu, J. Baxendale, L. Blick, M. Caminha, J. Carnes-Stine, P. Caulk, Y. H. Chiang, M. Coyne, C. Dahlke, A. Mays, M. Dombroski, M. Donnelly, D. Ely, S. Esparham, C. Fosler, H. Gire, S. Glanowski, K. Glasser, A. Glodek, M. Gorokhov, K. Graham, B. Gropman, M. Harris, J. Heil, S. Henderson, J. Hoover, D. Jennings, C. Jordan, J. Jordan, J. Kasha, L. Kagan, C. Kraft, A. Levitsky, M. Lewis, X. Liu, J. Lopez, D. Ma, W. Majoros, J. McDaniel, S. Murphy, M. Newman, T. Nguyen, N. Nguyen, M. Nodell, S. Pan, J. Peck, M. Peterson, W. Rowe, R. Sanders, J. Scott, M. Simpson, T. Smith, A. Sprague, T. Stockwell, R. Turner, E. Venter, M. Wang, M. Wen, D. Wu, M. Wu, A. Xia, A. Zandieh, and X. Zhu. The sequence of the human genome. *Science*, 291:1304–1351, 2001.

[70] E. T. Prak and H. H. Kazazian Jr. Mobile elements and the human genome. *Nat. Rev. Genet.*, 1:134–144, 2000.

[71] H. Neil, C. Malabat, Y. d'Aubenton Carafa, Z. Xu andL. M. Steinmetz, and A. Jacquier. Widespread bidirectional promoters are the major source of cryptic transcripts in yeast. *Nature*, 457:1038–1042, 2009.

[72] R. W. Holley, J. Apgar, G. A. Everett, J. T. Madison, M. Marquisee, S. H. Merrill, J. R. Penswick, and A. Zamir. Structure of a ribonucleic acid. *Science*, 14:1462–1465, 1965.

[73] C. Marker, A. Zemann, T. Terhorst, M. Kiefmann, J. P. Kastenmayer, P. Green, J. P. Bachellerie, J. Brosius, and A. Hüttenhofer. Experimental RNomics: Identification of 140 candidates for small non-messenger RNAs in the plant *Arabidopsis thaliana*. *Curr. Biol.*, 12:2002–2013, 2002.

[74] G. Yuan, C. Klambt, J. P. Bachellerie, J. Brosius, and A. Hüttenhofer. RNomics in *Drosophila melanogaster*: Identification of 66 candidates for novel non-messenger RNAs. *Nucl. Acids Res.*, 31:2495–2507, 2003.

[75] J. Vogel, V. Bartels, T. H. Tang, G. Churakov, J. G. Slagter-Jager, A. Hüttenhofer, and E. G. Wagner. RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucl. Acids Res.*, 31:6435–6443, 2003.

[76] M. Kawano, A. A. Reynolds, J. Miranda-Rios, and G. Storz. Detection of 5' and 3'-UTR-derived small RNAs and cis-encoded antisense RNAs in *Escherichia coli*. *Nucl. Acids Res.*, 33:1040–1050, 2005.

[77] T. H. Tang, N. Polacek, M. Zywicki, H. Huber, K. Brugger, R. Garrett, J. P. Bachellerie, and A. Hüttenhofer. Identification of novel non-coding RNAs as potential antisense regulators in the archaeon *Sulfolobus solfataricus*. *Mol. Microbiol.*, 55: 469–481, 2005.

[78] W. Deng, X. Zhu, G. Skogerbø, Y. Zhao, Z. Fu, Y. Wang, H. He, L. Cai, H. Sun, C. Liu, B. Li, B. Bai, J. Wang, D. Jia, S. Sun, H. He, Y. Cui, Y. Wang, D. Bu, and R. Chen. Organization of the *Caenorhabditis elegans* small non-coding transcriptome: Genomic features, biogenesis, and expression. *Genome Res.*, 16:20–29, 2006.

[79] M. Lagos-Quintana, R. Rauhut, A. Yalcin, J. Meyer, W. Lendeckel, and T. Tuschl. Identification of tissue-specific microRNAs from mouse. *Curr. Biol.*, 12:735–739, 2002.

[80] R. C. Lee and V. Ambros. An extensive class of small RNAs in *Caenohabditis elegans*. *Science*, 294:862–864, 2001.

[81] P. Y. Chen, H. Manninga, K. Slanchev, M. Chien, J. J. Russo, J. Ju, R. Sheridan, B. John, D. S. Marks, D. Gaidatzis, C. Sander, M. Zavolan, and T. Tuschl. The developmental miRNA profiles of zebrafish as determined by small RNA cloning. *Genes Dev.*, 19:1288–1293, 2005.

[82] C. Lu, S. S. Tej, S. Luo, C. D. Haudenschild, B. C. Meyers, and P. J. Green. Elucidation of the small RNA component of the transcriptome. *Science*, 309:1567–1569, 2005.

[83] P. Vitali, H. Royo, H. Seitz, J. P. Bachellerie, A. Hüttenhofer, and J. Cavaillé. Identification of 13 novel human modification guide RNAs. *Nucl. Acids Res.*, 31:6543–6551, 2003.

[84] X. Darzacq, B. E. Jdy, C. Verheggen, A. M. Kiss, E. Bertrand, and T. Kiss. Cajal body-specific small nuclear RNAs: a novel class of 2'-o-methylation and pseudouridylation guide RNAs. *EMBO J.*, 21:2746–2756, 2002.

[85] U. Nagalakshmi, Z. Wang, K. Waern, C. Shou, D. Raha, M. Gerstein, and M. Snyder. The transcriptional landscape of the yeast genome defined by RNA sequencing. *Science*, 320:1344–1349, 2008.

[86] R. D. Morin, M. D. O'Connor, M. Griffith, F. Kuchenbauer, A. Delaney, A. L. Prabhu, Y. Zhao, H. McDonald, T. Zeng, M. Hirst, C. J. Eaves, and M. A. Marra. Application of massively parallel sequencing to microRNA profiling and discovery in human embryonic stem cells. *Genome Res.*, 18:610–621, 2008.

[87] R. J. Taft, E. A. Glazov, N. Cloonan, C. Simons, S. Stephen, G. J. Faulkner, T. Lassmann, A. R. Forrest, S. M. Grimmond, K. Schroder, K. Irvine, T. Arakawa, M. Nakamura, A. Kubosaki, K. Hayashida, C. Kawazu, M. Murata, H. Nishiyori, S. Fukuda, J. Kawai, C. O. Daub, D. A. Hume, H. Suzuki, V. Orlando, P. Carninci, Y. Hayashizaki, and J. S. Mattick. Tiny RNAs associated with transcription start sites in animals. *Nat. Genet.*, 41:572–578, 2009.

[88] M. Guttman, M. Garber, J. Z. Levin, J. Donaghey, J. Robinson, X. Adiconis, L. Fan, M. J. Koziol, A. Gnirke, C. Nusbaum, J. L. Rinn, E. S. Lander, and A. Regev. Ab

initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, 28:503–510, 2010.

[89] M. Rederstorff, S. H. Bernhart, A. Tanzer, M. Zywicki, K. Perfler, M. Lukasser, I. L. Hofacker, and A. Hüttenhofer. RNPomics: defining the ncRNA transcriptome by cDNA library generation from ribonucleo-protein particles. *Nucl. Acids Res.*, 38: e113, 2010.

[90] M. Rederstorff and A. Hüttenhofer. cDNA library generation from ribonucleoprotein particles. *Nat. Protoc.*, 6:166–174, 2011.

[91] C. M. Sharma, S. Hoffmann, F. Darfeuille, J. Reignier, S. Findeiss, A. Sittka, S. Chabas, K. Reiche, J. Hackermüller, R. Reinhardt, P. F. Stadler, and J. Vogel. The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, 464:250–255, 2010.

[92] D. W. Selinger, K. J. Cheung, R. Mei, E. M. Johansson, C. S. Richmond, F. R. Blattner, D. J. Lockhart, and G. M. Church. RNA expression analysis using a 30 base pair resolution *Escherichia coli* genome array. *Nat. Biotechnol.*, 18:1262–1268, 2000.

[93] B. Tjaden, R. M. Saxena, S. Stolyar, D. R. Haynor, E. Kolker, and C. Rosenow. Transcriptome analysis of *Escherichia coli* using high-density oligonucleotide probe arrays. *Nucl. Acids Res.*, 30:3732–3738, 2002.

[94] V. Stolc, Z. Gauhar, C. Mason, G. Halasz, M. F. van Batenburg, S. A. Rifkin, S. Hua, T. Herreman, W. Tongprasit, P. E. Barbano, H. J. Bussemaker, and K. P. White.

A gene expression map for the euchromatic genome of *Drosophila melanogaster*. *Science*, 306:655–660, 2004.

[95] V. Stolc, M. P. Samanta, W. Tongprasit, H. Sethi, S. Liang, D. C. Nelson, A. Hegeman, C. Nelson, D. Rancour, S. Bednarek, E. L. Ulrich, Q. Zhao, R. L. Wrobel, C. S. Newman, B. G. Fox, , J. L. Markley, and M. R. Sussman. Identification of transcribed sequences in *Arabidopsis Thaliana* by using high-resolution genome tiling arrays. *Proc. Natl. Acad. Sci. USA*, 102:4453–4458, 2005.

[96] J. Cheng, P. Kapranov, J. Drenkow, S. Dike, S. Brubaker, S. Patel, J. Long, D. Stern, H. Tammana, G. Helt, V. Sementchenko, A. Piccolboni, S. Bekiranov, D. K. Bailey, M. Ganesh, S. Ghosh, I. Bell, D. S. Gerhard, and T. R. Gingeras. Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, 308:1149–1154, 2005.

[97] T. Babak, B. J. Blencowe, and T. R. Hughes. A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics*, 6:104, 2005.

[98] K. J. Nordström, M. A. Mirza, M. S. Almén, D. E. Gloriam, R. Fredriksson, and H. B. Schiöth. Critical evaluation of the FANTOM3 non-coding RNA transcripts. *Genomics*, 94:169–176, 2009.

[99] H. van Bakel and T. R. Hughes. Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genomic Proteomic.*, 8:424–436, 2009.

[100] 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,

and J. S. Mattick. Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.*, 16:11–19, 2006.

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

[102] K. Struhl. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, 14:103–105, 2007.

[103] H. van Bakel and T. R. Hughes. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.*, 8:e1000371, 2010.

[104] J. L. Rinn, G. Euskirchen, P. Bertone, R. Martone, N. M. Luscombe, S. Hartman, P. M. Harrison, F. K. Nelson, P. Miller, M. Gerstein, S. Weissman, and M. Snyder. The transcriptional activity of human chromosome 22. *Genes Dev.*, 17:529–540, 2003.

[105] 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. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. *Genome Res.*, 14:331–342, 2004.

[106] M. E. Dinger, K. C. Pang, T. R. Mercer, and J. S. Mattick. Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput. Biol.*, 4: e1000176, 2008.

[107] M. C. Frith, A. R. Forrest, E. Nourbakhsh, K. C. Pang, C. Kai, J. Kawai, P. Carnin-

ciand Y. Hayashizaki, T. L. Bailey, and S. M. Grimmond. The abundance of short proteins in the mammalian proteome. *PLoS Genet.*, 2:e52, 2006.

[108] Q. R. Li, A. R. Carvunis, H. Yu, J. D. Han, Q. Zhong, N. Simonis, S. Tam, T. Hao, N. J. Klitgord, D. Dupuy, D. Mou, I. Wapinski, A. Regev, D. E. Hill, M. E. Cusick, and W. Vidal. Revisiting the *Saccharomyces Cerevisiae* predicted ORFeome. *Genome Res.*, 18:1294–1303, 2008.

[109] M. I. Galindo, J. I. Pueyo, S. Fouix, S. A. Bishop, and J. P. Couso. Peptides encoded by short ORFs control development and define a new eukaryotic gene family. *PLoS Biol.*, 5:e106, 2007.

[110] C. A. Gleason, Q. L. Liu, and V. M. Williamson. Silencing a candidate nematode effector gene corresponding tothe tomato resistance gene Mi-1 leads to acquisition of virulence. *Mol. Plant Microbe Interact.*, 21:576–585, 2008.

[111] T. J. Strabala. CLE genes in plant development: Gain-of-function analyses, pleiotropy, hypermorphy and neomorphy. *Plant Signal. Behav.*, 3:457–459, 2008.

[112] P. Kapranov, J. Cheng, S. Dike, D. A. Nix, R. Duttagupta, A. T. Willingham, P. F. Stadler, J. Hertel, J. Hackermller, I. L. Hofacker, I. Bell, E. Cheung, J. Drenkow, E. Dumais, S. Patel, G. Helt, M. Ganesh, S. Ghosh, A. Piccolboni, V. Sementchenko, H. Tammana, and T. R. Gingeras. RNA maps reveal new RNA classes and a possible function for pervasive transcription. *Science*, 316:1484–1488, 2007.

[113] J. Berretta and A. Morillon. Pervasive transcription constitutes a new level of eukaryotic genome regulation. *EMBO Rep.*, 10:973–982, 2009.

[114] J. Ponjavic and C. P. Ponting. The long and the short of RNA maps. *Bioessays*, 29: 1077–1080, 2007.

[115] M. Lapidot and Y. Pilpel. Genome-wide natural antisense transcription: Coupling its regulation to its different regulatory mechanisms. *EMBO Rep.*, 7:1216–1222, 2006.

[116] R. Louro, T. El-Jundi, H. I. Nakaya, E. M. Reis, and S. Verjovski-Almeida. Conserved tissue expression signatures of intronic noncoding RNAs transcribed from human and mouse loci. *Genomics*, 92:18–25, 2008.

[117] A. C. Seila, J. M. Calabrese, S. S. Levine, G. W. Yeo, P. B. Rahl, R. A. Flynn, R. A. Young, and P. A. Sharp. Divergent transcription from active promoters. *Science*, 322:1849–1851, 2008.

[118] L. J. Core, J. J. Waterfall, and J. T. Lis. Nascent RNA sequencing reveals widespread pausing and divergent initiation at human promoters. *Science*, 322:1845–1848, 2008.

[119] P. Preker, J. Nielsen, S. Kammler, S. Lykke-Andersen, M. S. Christensen, C. K. Mapendano, M. H. Schierup, and T. H. Jensen. RNA exosome depletion reveals transcription upstream of active human promoters. *Science*, 322:1851–1854, 2008.

[120] I. Shamovsky, M. Ivannikov, E. S. Kandel, and D. Gershon and. RNA-mediated response to heat shock in mammalian cells. *Nature*, 440:556–560, 2006.

[121] Carolyn J. Brown, Brian D. Hendrich, Jim L. Rupert, Ronald G. Lafreniere, Yigong Xing, Jeanne Lawrence, and Huntington F. Willard. The human Xist gene: Analysis

of a 17 kb inactive X-specific RNA that contains conserved repeats and is highly localized within the nucleus. *Cell*, 71:527–542, 1992.

[122] J. T. Lee, L. S. Davidow, and D. Warshawsky. Tsix, a gene antisense to Xist at the X-inactivation centre. *Nat. Genet.*, 21:400–404, 1999.

[123] J. L. Rinn, M. Kertesz, J. K. Wang, S. L. Squazzo, X. Xu, S. A. Brugmann, L. H. Goodnough, J. A. Helms, P. J. Farnham, E. Segal, and H. Y. Chang. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell*, 129:1311–1323, 2007.

[124] R. A. Gupta, N. Shah, K. C. Wang, J. Kim, H. M. Horlings, D. J. Wong, M. C. Tsai andT. Hung, P. Argani, J. L. Rinn, Y. Wang, P. Brzoska, B. Kong, R. Li, R. B. West, M. J. van de Vijver, S. Sukumar, and H. Y. Chang. Long non-coding RNA HOTAIR reprograms chromatin state to promote cancer metastasis. *Nature*, 464:1071–1076, 2010.

[125] T. R. Mercer, M. E. Dinger, S. M. Sunkin, M. F. Mehler, and J. S. Mattick. Specific expression of long noncoding RNAs in the mouse brain. *Proc. Natl. Acad. Sci. USA*, 105:716–721, 2008.

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

[127] S. Washietl, I. L. Hofacker, M. Lukasser, A. Hüttenhofer, and P. F. Stadler. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, 23:1383–1390, 2005.

[128] J. S. Pedersen, G. Bejerano, A. Siepel, K. Rosenbloom, K. Lindblad-Toh, E. S. Lander, J. Kent, W. Miller, and D. Haussler. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, 2: e33, 2006.

[129] E. P. Nawrocki, D. L. Kolbe, and S. R. Eddy. Infernal 1.0: Inference of RNA alignments. *Bioinformatics*, 25:1335–1337, 2009.

[130] R. J. Klein and S. R. Eddy. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics*, 4:44, 2003.

[131] T. M. Lowe and S. R. Eddy. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucl. Acids Res.*, 25:955–964, 1997.

[132] D. Laslett and B. Canback. ARAGORN, a program to detect tRNA genes and tm-RNA genes in nucleotide sequences. *Nucl. Acids Res.*, 32:11–16, 2004.

[133] M. Regalia, M. A. Rosenblad, and T. Samuelsson. Prediction of signal recognition particle RNA genes. *Nucl. Acids Res.*, 30:3368–3377, 2002.

[134] T. M. Lowe and S. R. Eddy. A computational screen for methylation guide snoRNAs in yeast. *Science*, 283:1168–1171, 1999.

[135] P. Schattner, S. Barberan-Soler, and T. M. Lowe. A computational screen for mammalian pseudouridylation guide H/ACA RNAs. *RNA*, 12:15–25, 2006.

[136] J. Hertel, I. L. Hofacker, and P. F. Stadler. SnoReport: computational identification of snoRNAs with unknown targets. *Bioinformatics*, 24:158–164, 2008.

[137] L. P. Lim, N. C. Lau, E. G. Weinstein, A. Abdelhakim, S. Yekta, M. W. Rhoades, C. B. Burge, and David P. Bartel. The microRNAs of *Caenorhabditis Elegans*. *Genes Dev.*, 17:991–1008, 2003.

[138] G. Terai, T. Komori, K. Asai, and T. Kin. miRRim: a novel system to find conserved miRNAs with high sensitivity and specificity. *RNA*, 13:2081–2090, 2007.

[139] J. Livny and M. K. Waldor. Identification of small RNAs in diverse bacterial species. *Curr. Opin. Microbiol.*, 10:96–101, 2007.

[140] P. S. Klosterman, A. V. Uzilov, Y. R. Bendaa, R. K. Bradley, S. Chao, C. Kosiol, N. Goldman, and I. Holmes. XRate: a fast prototyping, training and annotation tool for phylo-grammars. *BMC Bioinformatics*, 7:428, 2006.

[141] T. Babak, B. J. Blencowe, and T. R. Hughes. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, 8: 33, 2007.

[142] R. Upadhyay, P. Bawankar, D. Malhotra, and S. Patankar. A screen for conserved sequences with biased base composition identifies noncoding RNAs in the A-T rich genome of *Plasmodium Falciparum*. *Mol. Biochem. Parasitol.*, 144:149–158, 2005.

[143] P. Larsson, A. Hinas, D. H. Ardell, L. A. Kirsebom, A. Virtanen, and F. Söderbom. De novo search for non-coding RNA genes in the AT-rich genome of *Dictyostelium Discoideum*: Performance of Markov-dependent genome feature scoring. *Genome Res.*, 18:888–899, 2008.

[144] K. Numata, A. Kanai, R. Saito, S. Kondo, J. Adachi, L. G. Wilming, D. A. Hume,

Y. Hayashizaki, M. Tomita, RIKEN GER Group, and GSL Members. Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res.*, 13:1301–1306, 2003.

[145] 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-Impiombato, 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. Christoffels, 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. Schonbach, K. Sekiguchi, C. A. Semple, 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, and Y. Hayashizaki. The transcriptional landscape of the mammalian genome. *Science*, 309:1559–1563, 2005.

[146] T. K. Kim, M. Hemberg, J. M. Gray, A. M. Costa, D. M. Bear, J. Wu, D. A. Harmin, M. Laptewicz, K. Barbara-Haley, S. Kuersten, E. Markenscoff-Papadimitriou, D. Kuhl, H. Bito, P. F. Worley, G. Kreiman, and M. E. Greenberg. Widespread transcription at neuronal activity-regulated enhancers. *Nature*, 465:182–187, 2010.

[147] W. Foissner and H. Berger. Identification and ontogenesis of the *nomen nudum* hypotrichs (Protozoa: Ciliophora) *Oxytricha nova* (= *Sterkiella nova* sp. n.) and *O. trifallax* (= *S. histriomuscorum*). *Acta Protozoologica*, 38:215–248, 1999.

[148] K. Collins. Ciliate telomerase biochemistry. *Annu. Rev. Biochem.*, 68:187–218, 1999.

[149] J. A. Eisen, R. S. Coyne, M. Wu, D. Wu, M. Thiagarajan, J. R. Wortman, J. H. Bad-

ger, Q. Ren, P. Amedeo, K. M. Jones, L. J. Tallon, A. L. Delcher, S. L. Salzberg, J. C. Silva, B. J. Haas, W. H. Majoros, M. Farzad, J. M. Carlton, , J. Garg, R. E. Pearlman, K. M. Karrer, L. Sun, G. Manning, N. C. Elde, A. P. Turkewitz, D. J. Asai, D. E. Wilkes, Y. Wang, H. Cai, K. Collins, B. A. Stewart, S. R. Lee, K. Wilamowska, Z. Weinberg, W. L. Ruzzo, D. Wloga, J. Gaertig, J. Frankel, C. C. Tsao, M. A. Gorovsky, P. J. Keeling, R. F. Waller, N. J. Patron, J. M. Cherry, N. A. Stover, C. J. Krieger, C. del Toro, H. F. Ryder, S. C. Williamson, R. A. Barbeau, E. P. Hamilton, and E. Orias. Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol.*, 4:e286, 2006.

[150] J. M. Aury, O. Jaillon, L. Duret, B. Noel, C. Jubin, B. M. Porcel, B. Ségurens, V. Daubin, V. Anthouard, N. Aiach, O. Arnaiz, A. Billaut, J. Beisson, I. Blanc, K. Bouhouche, F. Câmara, S. Duharcourt, R. Guigo, D. Gogendeau, M. Katinka, A. M.Keller, R. Kissmehl, C. Klotz, F. Koll, A. Le Mouël, G. Lepère, S. Malinsky, M. Nowacki, J. K. Nowak, H. Plattner, J. Poulain andF. Ruiz, V. Serrano, M. Zagulski, P. Dessen, M. Bétermier, J. Weissenbach, C. Scarpelli, V. Schächter, L. Sperling, E. Meyer, J. Cohen, and P. Wincker. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, 444:171–178, 2006.

[151] D. Ammermann, G. Steinbrück, L. von Berger, and W. Hennig. The development of the macronucleus in the ciliated protozoan *Stylonychia mytilus*. *Chromosoma*, 45: 401–429, 1974.

[152] C. L. Jahn and L. A. Klobutcher. Genome remodeling in ciliated protozoa. *Annu. Rev. Microbiol.*, 56:489–520, 2002.

[153] D. M. Prescott. The DNA of ciliated protozoa. *Microbiol. Rev.*, 58:233–267, 1994.

[154] D. M. Prescott. Genome gymnastics: Unique modes of DNA evolution and processing in ciliates. *Nat. Rev. Genet.*, 1:191–198, 2000.

[155] L. A. Klobutcher, C. L. Jahn, and D. M. Prescott. Internal sequences are eliminated from genes during macronuclear development in the ciliated protozoan *Oxytricha Nova*. *Cell*, 36:1045–1055, 1984.

[156] J. D. Griffith, L. Comeau, S. Rosenfield, R. M. Stansel, A. Bianchi, H. Moss, and T. de Lange. Mammalian telomeres end in a large duplex loop. *Cell*, 97:503–514, 1999.

[157] K. G. Murti and D. M. Prescott. Telomeres of polytene chromosomes in a ciliated protozoan terminate in duplex DNA loops. *Proc. Natl. Acad. Sci. USA*, 96:14436–14439, 1999.

[158] F. Jönsson, J. Postberg, and H. J. Lipps. The unusual way to make a genetically active nucleus. *DNA Cell Biol.*, 28:71–78, 2009.

[159] S. A. Juranek and H. J. Lipps. New insights into the macronuclear development in ciliates. *Int. Rev. Cytol.*, 262:219–251, 2007.

[160] M. Nowacki, V. Vijayan, Y. Zhou, K. Schotanus, T. G. Doak, and L. F. Landweber. RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature*, 451:153–158, 2008.

[161] M. Nowacki, B. P. Higgins, G. M. Maquilan, E. C. Swart, T. G. Doak, and L. F.

Landweber. A functional role for transposases in a large eukaryotic genome. *Science*, 324:935–938, 2009.

[162] D. L. Chalker and M. C. Yao. Nongenic, bidirectional transcription precedes and may promote developmental DNA deletion in *Tetrahymena thermophila*. *Genes Dev.*, 15:1287–1298, 2001.

[163] K. Mochizuki, N. A. Fine, T. Fujisawa, and M. A. Gorovsky. Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell*, 110:689–699, 2002.

[164] O. Garnier, V. Serrano, S. Duharcourt, and E. Meyer. RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. *Mol. Cell. Biol.*, 24:7370–7379, 2004.

[165] D. L. Chalker, P. Fuller, and M. C. Yao. Communication between parental and developing genomes during *Tetrahymena* nuclear differentiation is likely mediated by homologous RNAs. *Genetics*, 169:149–160, 2005.

[166] H. M. Kurth and K. Mochizuki. Non-coding RNA: a bridge between small RNA and DNA. *RNA Biol.*, 6:138–140, 2009.

[167] J. L. Riley and L. A. Katz. Widespread distribution of extensive chromosomal fragmentation in ciliates. *Mol. Bio. Evol.*, 18:1372–1377, 2001.

[168] T. G. Doak, A. R. Cavalcanti, N. A. Stover, D. M. Dunn, R. Weiss, G. Herrick, and L. F. Landweber. Sequencing the *Oxytricha trifallax* macronuclear genome: a pilot project. *Trends Genet.*, 19:603–607, 2003.

[169] M. R. Lauth, B. B. Spear, J. Heumann, and D. M. Prescott. DNA of ciliated protozoa: DNA sequence diminution during macronuclear development of *Oxytricha*. *Cell*, 7: 67–74, 1976.

[170] M. T. Swanton, J. M. Heumann, and D. M. Prescott. Gene-sized DNA molecules of the macronuclei in three species of hypotrichs: Size distributions and absence of nicks. DNA of ciliated protozoa. VIII. *Chromosoma*, 77:217–227, 1980.

[171] D. C. Hoffman, R. C. Anderson, M. L. DuBois, and D. M. Prescott. Macronuclear gene-sized molecules of hypotrichs. *Nucl. Acids Res.*, 23:1279–1283, 1995.

[172] D. M. Prescott, J. D. Prescott, and R. M. Prescott. Coding properties of macronuclear DNA molecules in *Sterkiella nova* (*Oxytricha nova*). *Protist*, 153:71–77, 2002.

[173] A. R. Cavalcanti, D. M. Dunn, R. Weiss, G. Herrick, L. F. Landweber, and T. G. Doak. Sequence features of *Oxytricha trifallax* (class Spirotrichea) macronuclear telomeric and subtelomeric sequences. *Protist*, 155:311–322, 2004.

[174] A. R. Cavalcanti, N. A. Stover, L. Orecchia, T. G. Doak, and L. F. Landweber. Coding properties of *Oxytricha trifallax* (*Sterkiella histriomuscorum*) macronuclear chromosomes: Analysis of a pilot genome project. *Chromosoma*, 113:69–76, 2004.

[175] S. W. Cartinhour and G. A. Herrick. Three different macronuclear DNAs in *Oxytricha Fallax* share a common sequence block. *Mol. Cell Biol.*, 4:931–938, 1984.

[176] L. A. Klobutcher, M. E. Huff, and G. E. Gonye. Alternative use of chromosome fragmentation sites in the ciliated protozoan *Oxytricha Nova*. *Nucl. Acids Res.*, 16: 251–264, 1988.

[177] X. Huang, J. Wang, S. Aluru, S.-P. Yang, and L. Hillier. PCAP: A whole-genome assembly program. *Genome Res.*, 13:2164–2170, 2003.

[178] P. P. Gardner, J. Daub, J. G. Tate, E. P. Nawrocki, D. L. Kolbe, S. Lindgreen, A. C. Wilkinson, R. D. Finn, S. Griffiths-Jones, S. R. Eddy, and A. Bateman. Rfam: Updates to the RNA families database. *Nucl. Acids Res.*, 37:D136–D140, 2009.

[179] A. Seegmiller, K. R. Williams, R. L. Hammersmith, T. G. Doak, D. Witherspoon, T. Messick, L. L. Storjohann, and G. Herrick. Internal eliminated sequences interrupting the *Oxytricha* 81 locus: Allelic divergence, conservation, conversions, and possible transposon origins. *Mol. Bio. Evol.*, 13:1351–1362, 1996.

[180] D. M. Prescott and S. J. Dizick. A unique pattern of intrastrand anomalies in base composition of the DNA in hypotrichs. *Nucl. Acids Res.*, 28:4679–4688, 2000.

[181] A. Seegmiller, K. R. Williams, and G. Herrick. Two two-gene macronuclear chromosomes of the hypotrichous ciliates *Oxytricha fallax* and *O. trifallax* generated by alternative processing of the 81 locus. *Dev. Genet.*, 20:348–357, 1997.

[182] W. J. Chang, N. A. Stover, V. M. Addis, and L. F. Landweber. A micronuclear locus containing three protein-coding genes remains linked during macronuclear developmentin the spirotrichous ciliate *Holosticha*. *Protist*, 155:245–255, 2004.

[183] G. Parra, K. Bradnam, Z. Ning, T. Keane, and I. Korf. Assessing the gene space in draft genomes. *Nucl. Acids Res.*, 37:289–297, 2009.

[184] G. Parra, K. Bradnam, and I. Korf. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, 23:1061–1067, 2007.

[185] F. H. Stephenson. *Calculations for Molecular Biology and Biotechnology: A Guide to Mathematics in the Laboratory*. Academic Press, San Diego, CA, 2003.

[186] M. Guttman, I. Amit, M. Garber, C. French, M. F. Lin, D. Feldser, M. Huarte, O. Zuk, B. W. Carey, J. P. Cassady, M. N. Cabili, R. Jaenisch, T. S. Mikkelsen, T. Jacks, N. Hacohen, B. E. Bernstein, M. Kellis, A. Regev, J. L. Rinn, and E. S. Lander. Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, 458:223–227, 2009.

[187] G. Steinbrück. Overamplification of genes in macronuclei of hypotrichous ciliates. *Chromosoma*, 88:156–163, 1983.

[188] S. E. Baird and L. A. Klobutcher. Differential DNA amplification and copy number control in the hypotrichous ciliate *Euplotes crassus*. *J. Protozool.*, 38:136–140, 1991.

[189] G. Heyse, F. Jönsson, W.-J. Chang, and H. J. Lipps. RNA-dependent control of gene amplification. *Proc. Natl. Acad. Sci. USA*, 107:22134–22139, 2010.

[190] D. S. Harper, K. Song, and C. L. Jahn. Overamplification of macronuclear linear DNA molecules during prolonged vegetative growth of *Oxytricha nova*. *Gene*, 99: 55–61, 1991.

[191] G. Steinbrück, I. Haas, K.-H. Hellmer, and D. Ammermann. Characterization of macronuclear DNA in five species of ciliates. *Chromosoma*, 83:199–208, 1981.

[192] G. S. Slater and E. Birney. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, 6:31, 2005.

[193] C. A. Lozupone, R. D. Knight, and L. F. Landweber. The molecular basis of nuclear genetic code change in ciliates. *Curr. Biol.*, 11:65–74, 2001.

[194] J. E. Allen, W. H. Majoros, and S. L. Salzberg. JIGSAW, GeneZilla, and GlimmerHMM: Puzzling out the features of human genes in ENCODE regions. *Genome Biol.*, 7:S9.1–S9.13, 2006.

[195] W. H. Majoros, M. Pertea, C. Antonescu, and S. L. Salzberg. GlimmerM, Exonomy and Unveil: Three ab initio eukaryotic genefinders. *Nucl. Acids Res.*, 31:3601–3604, 2003.

[196] G. Parra, E. Blanco, and R. Guigó. GeneID in *Drosophila*. *Genome Research*, 10: 511–515, 2000.

[197] M. Stanke and S. Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19 Suppl 2:ii215–ii225, 2003.

[198] R. Durbin, S. R. Eddy, A. Krogh, and G. J. Mitchison. *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge UK, 1998. ISBN 0521629713.

[199] M. Q. Zhang. Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet.*, 3:698–709, 2002.

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

[201] M. Burset and R. Guigò. Evaluation of gene structure prediction programs. *Genomics*, 34:353–367, 1996.

[202] S. R. Eddy. A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.*, 3:e10, 2005.

[203] C. Maercker and H. J. Lipps. Analysis of the subtelomeric regions of macronuclear gene-sized DNA molecules of the hypotrichous ciliate *Stylonychia lemnae*: Implications for the DNA fragmentation process during macronuclear development? *Dev. Genet.*, 14:378–384, 1993.

[204] C. Eder, C. Maercker, J. Meyer, and H. J. Lipps. The processing of macronuclear DNA sequences during macronuclear development of the hypotrichous ciliate *Stylonychia lemnae*. *Int. J. Dev. Biol.*, 37:473–477, 1993.

[205] J. Wen, C. Maercker, and H. J. Lipps. Sequential excision of internal eliminated DNA sequences inthe differentiating macronucleus of the hypotrichous ciliate *Stylonychia lemnae*. *Nucl. Acids Res.*, 24:4415–4419, 1996.

[206] M. Kellis, N. Patterson, M. Endrizzi, B. Birren, and E. S. Lander. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423: 241–254, 2003.

[207] K. Williams, T. G. Doak, and G. Herrick. Developmental precise excision of *Oxytricha trifallax* telomere-bearing elements and formation of circles closed by a copy of the flanking target duplication. *EMBO J.*, 12:4593–4601, 1993.

[208] S. H. Bernhart, I. L. Hofacker, S. Will, A. R. Gruber, and P. F. Stadler. RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, 9:474, 2008.

[209] P. Schattner, W. A. Decatur, C. A. Davis, M. Ares, M. J. Fournier, and T. M. Lowe. Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucl. Acids Res.*, 32:4281–4296, 2004.

[210] P. Ganot, M. Caizergues-Ferrer, and T. Kiss. The family of box ACA small nucleolar RNAs is defined by an evolutionarily conserved secondary structure and ubiquitous sequence elements essential for RNA accumulation. *Genes Dev.*, 11:941–956, 1997.

[211] V. Atzorn, P. Fragapane, and T. Kiss. U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production. *Mol. Cell. Biol.*, 24: 1769–1778, 2004.

[212] G. L. Eliceiri. The vertebrate E1/U17 small nucleolar ribonucleoprotein particle. *J. Cell. Biochem.*, 98:486–495, 2006.

[213] E. Fayet-Lebaron, V. Atzorn, Y. Henry, and T. Kiss. 18S rRNA processing requires base pairings of snR30 H/ACA snoRNA to eukaryote-specific 18S sequences. *EMBO J.*, 28:1260–1270, 2009.

[214] 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, and D. G. Higgins. Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947–2948, 2007.

[215] O. Arnaiz, S. Cain, J. Cohen, and L. Sperling. ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. *Nucl. Acids Res.*, 35:D439–D444, 2007.

[216] G. Ricard, R. M. de Graaf, B. E. Dutilh, I. Duarte, T. A. van Alen, A. H. van Hoek, B. Boxma, G. W. van der Staay, S. Y. Moon van der Staay, W. J. Chang, L. F. Landweber, J. H. Hackstein, and M. A. Huynen. Macronuclear genome structure of the ciliate *Nyctotherus ovalis*: Single-gene chromosomes and tiny introns. *BMC Genomics*, 9:587, 2008.

[217] C. L. Chen, H. Zhou, J. Y. Liao, L. H. Qu, and L. Amar. Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of *Paramecium tetraurelia*. *RNA*, 15:503–514, 2009.

[218] G. E. Crooks, G. Hon, J. M. Chandonia, and S. E. Brenner. WebLogo: a sequence logo generator. *Genome Res.*, 14:1188–1190, 2004.

[219] J. Thomas, K. Lea, E. Zucker-Aprison, and T. Blumenthal. The spliceosomal snRNAs of *Caenorhabditis elegans*. *Nucl. Acids Res.*, 18:2633–2642, 1990.

[220] S. L. Stricklin, S. Griffiths-Jones, and S. R. Eddy. *C. elegans* noncoding RNA genes. In The *C. elegans* Research Community, editor, *WormBook*. doi/10.1895/wormbook.1.7.1, http://www.wormbook.org, 2005.

[221] B. W. Hargrove, A. Bhattacharyya, A. M. Domitrovich, G. M. Kapler, K. Kirk, D. E. Shippen, and G. R. Kunkel. Identification of an essential proximal sequence element in the promoter of the telomerase RNA gene of *Tetrahymena thermophila*. *Nucl. Acids Res.*, 27:4269–4275, 1999.

[222] S. Egloff, D. O'Reilly, and S. Murphy. Expression of human snRNA genes from beginning to end. *Biochem. Soc. Trans.*, 36:590–594, 2008.

[223] R. D. Hinrichsen, D. Fraga, and M. W. Reed. 3’-modified antisense oligodeoxyribonucleotides complementary to calmodulin mRNA alter behavioral responses in *Paramecium*. *Proc. Natl. Acad. Sci. USA*, 89:8601–8605, 1992.

[224] A. Galvani and L. Sperling. RNA interference by feeding in *Paramecium*. *Trends Genet.*, 18:11–12, 2002.

[225] R. Sweeney, Q. Fan, and M. C. Yao. Antisense in abundance: the ribosome as a vehicle for antisense RNA. *Genet. Eng. N.Y.*, 20:143–151, 1998.

[226] A. G. Paschka, F. Jönsson, V. Maier, M. Möllenbeck, K. Paeschke, J. Postberg, S. Rupprecht, and H. J. Lipps. The use of RNAi to analyze gene function in spirotrichous ciliates. *Europ. J. Protistol.*, 39:449–454, 2003.

[227] K. Paeschke, T. Simonsson, J. Postberg, D. Rhodes, and H. J. Lipps. Telomere end-binding proteins control the formation of G-quadruplex DNA structures in vivo. *Nat. Struct. Mol. Biol.*, 12:847–854, 2005.

[228] W. J. Murphy, K. P. Watkins, and N. Agabian. Identification of a novel Y branch structure as an intermediate in *Trypanosome* mRNA processing: Evidence for trans splicing. *Cell*, 47:517–525, 1986.

[229] R. E. Sutton and J. C. Boothroyd. Evidence for trans splicing in *Trypanosomes*. *Cell*, 47:527–535, 1986.

[230] J. P. Bruzik, K. Van Doren, D. Hirsh, and J. A. Steitz. Trans splicing involves a novel form of small nuclear ribonucleoprotein particles. *Nature*, 335:559–562, 1988.

[231] A. Rajkovic, R. E. Davis, J. N. Simonsen, and F. M. Rottman. A spliced leader is present on a subset of mRNAs from the human parasite *Schistosoma mansoni*. *Proc. Natl. Acad. Sci. USA*, 87:8879–8883, 1990.

[232] C. L. Will and R. Lührmann. Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol. Chem.*, 386:713–724, 2005.

[233] S. L. Reichow, T. Hamma, A. R. Ferré-D'Amaré, and G. Varani. The structure and function of small nucleolar ribonucleoproteins. *Nucl. Acids Res.*, 35:1452–1464, 2007.

[234] I. Myslyuk, T. Doniger, Y. Horesh, A. Hury, R. Hoffer, Y. Ziporen, S. Michaeli, and R. Unger. Psiscan: a computational approach to identify H/ACA-like and AGA-like non-coding RNA in *Trypanosomatid* genomes. *BMC Bioinformatics*, 9:471, 2008.

[235] D. J. Leader, G. P. Clark, J. Watters, A. F. Beven, P. J. Shaw, and J. W. Brown. Clusters of multiple different small nucleolar RNA genes in plants are expressed as and processed from polycistronic pre-snoRNAs. *EMBO J.*, 16:5742–5751, 1997.

[236] G. Chanfreau, G. Rotondo, P. Legrain, and A. Jacquier. Processing of a dicistronic small nucleolar RNA precursor by the RNA endonuclease rnt1. *EMBO J.*, 17:3726–3737, 1998.

[237] C. M. Smith and J. A. Steitz. Classification of *gas5* as a multi-small-nucleolar-RNA (snoRNA) host gene and a member of the 5'-terminal oligopyrimidine gene family reveals common features of snoRNA host genes. *Mol. Cell. Biol.*, 18:6897–6909, 1998.

[238] J. S. Mattick. Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, 25:930–939, 2003.

[239] W. C. Warren, D. F. Clayton, H. Ellegren, A. P. Arnold, L. W. Hillier, A. Künstner, S. Searle, S. White, A. J. Vilella, S. Fairley, A. Heger, L. Kong, C. P. Ponting, E. D. Jarvis, C. V. Mello, P. Minx, P. Lovell, T. A. Velho, M. Ferris, C. N. Balakrishnan, S. Sinha, C. Blatti, S. E. London, Y. Li, Y. C. Lin, J. George, J. Sweedler, B. Southey, P. Gunaratne, M. Watson, K. Nam, N. Backström, L. Smeds, B. Nabholz, Y. Itoh, O. Whitney, A. R. Pfenning, J. Howard, M. Völker, B. M. Skinner, D. K. Griffin, L. Ye, W. M. McLaren, P. Flicek, V. Quesada, G. Velasco, C. Lopez-Otin, X. S. Puente, T. Olender, D. Lancet, A. F. Smit, R. Hubley, M. K. Konkel, J. A. Walker, M. A. Batzer, W. Gu, D. D. Pollock, L. Chen, Z. Cheng, E. E. Eichler, J. Stapley, J. Slate, R. Ekblom, T. Birkhead, T. Burke, D. Burt, C. Scharff, I. Adam, H. Richard, M. Sultan, A. Soldatov, H. Lehrach, S. V. Edwards, S. P. Yang, X. Li, T. Graves, L. Fulton, J. Nelson, A. Chinwalla, S. Hou, E. R. Mardis, and R. K. Wilson. The genome of a songbird. *Nature*, 464:757–762, 2010.

[240] R. C. Edgar. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, 5:113, 2004.

[241] S. Griffiths-Jones. RALEE–RNA ALignment editor in Emacs. *Bioinformatics*, 21: 257–259, 2005.