Published online 28 June 2011

Nucleic Acids Research, 2011, Vol. 39, No. 17 ***7529–7547***
doi:10.1093/nar/gkr501

Exploiting *Oxytricha trifallax* nanochromosomes to screen for non-coding RNA genes

Seolkyoung Jung[1,2], Estienne C. Swart[3], Patrick J. Minx[4], Vincent Magrini[4], Elaine R. Mardis[4], Laura F. Landweber[3] and Sean R. Eddy[1,*]

[1]Janelia Farm Research Campus, Howard Hughes Medical Institute, Ashburn VA 20147, [2]Division of Biological and Biomedical Sciences, Washington University in St. Louis, St. Louis MO 63108, [3]Department of Ecology and Evolutionary Biology, Princeton University, Princeton NJ 08544 and [4]The Genome Center at Washington University, St. Louis MO 63108, USA

Received March 29, 2011; Revised May 31, 2011; Accepted June 1, 2011

ABSTRACT

We took advantage of the unusual genomic organization of the ciliate *Oxytricha trifallax* to screen for eukaryotic non-coding RNA (ncRNA) genes. Ciliates have two types of nuclei: a germ line micronucleus that is usually transcriptionally inactive, and a somatic macronucleus that contains a reduced, fragmented and rearranged genome that expresses all genes required for growth and asexual reproduction. In some ciliates including *Oxytricha*, the macronuclear genome is particularly extreme, consisting of thousands of tiny 'nanochromosomes', each of which usually contains only a single gene. Because the organism itself identifies and isolates most of its genes on single-gene nanochromosomes, nanochromosome structure could facilitate the discovery of unusual genes or gene classes, such as ncRNA genes. Using a draft *Oxytricha* genome assembly and a custom-written protein-coding genefinding program, we identified a subset of nanochromosomes that lack any detectable protein-coding gene, thereby strongly enriching for nanochromosomes that carry ncRNA genes. We found only a small proportion of non-coding nanochromosomes, suggesting that *Oxytricha* has few independent ncRNA genes besides homologs of already known RNAs. Other than new members of known ncRNA classes including C/D and H/ACA snoRNAs, our screen identified one new family of small RNA genes, named the Arisong RNAs, which share some of the features of small nuclear RNAs.

INTRODUCTION

RNAs have many important functions besides coding for proteins. One way of discovering new RNA functions is by identifying 'non-coding RNA' (ncRNA) transcripts that have no protein-coding role (1–3). There are many different kinds of ncRNAs, including independent functional RNA genes [such as genes for microRNAs (4) or bacterial small regulatory RNAs (5)], RNAs resulting from *cis*-regulatory control of mRNA transcription [such as riboswitches that prematurely attenuate transcription (6)], and RNAs that arise from processing of other RNA transcripts [such as the multitude of small Argonaute-bound RNAs acting in RNA interference and other RNAi-related pathways (7)].

Two main approaches have been used for systematic ncRNA identification, and both have resulted in controversial reports of surprisingly large numbers of ncRNAs in a wide variety of species (8–10). One approach is to experimentally enumerate RNA populations using high-throughput methods such as RNA-seq, cDNA sequencing and tiled microarrays (11–16). Transcriptomic results reporting large numbers of ncRNAs and pervasive non-coding transcription (17,18) have been challenged (19–21), because there are several other explanations for a putative ncRNA transcript to be seen. These alternative explanations include technical artifacts (20,22), fragments of the UTRs or introns of larger coding mRNAs (20), non-functional 'noise' in transcription and RNA processing (23) and coding mRNAs that escape detection by simple criteria of ORF length or homology (24). The other main approach has been computational prediction of regions of conserved RNA secondary structure (25–27), which identifies structural ncRNAs and *cis*-regulatory protein-binding structural motifs in mRNAs. One problem with the computational approach is that the

*To whom correspondence should be addressed. Tel: +01 571 209 4163; Fax: +01 571 209 4051; Email: eddys@janelia.hhmi.org

false positive prediction rates of the current programs are high enough to account for most of the observed predictions in large eukaryotic genomes (28).

A different systematic ncRNA identification approach, independent of transcriptomics and computational RNA structure prediction, might help address the controversies in this area. An example of a question that has not been addressed well by the current approaches is how many ncRNAs arise from independent RNA *genes* (i.e. independent loci with transcriptional initiation and termination signals), as opposed to ncRNAs obligately associated with nearby mRNA transcription or with non-genic RNA processing events (29). For example, there appear to be ncRNAs produced from enhancers of coding genes (30), but because enhancers can be distant from their gene, it is difficult to distinguish enhancer-associated ncRNA transcription from an independent ncRNA gene. Distinguishing independently functional ncRNA genes from other sources of putative ncRNAs would be one step toward focusing effort on specific classes of ncRNAs and RNA functions, rather than treating all 'ncRNAs' as a homogeneous class.

The unicellular eukaryotic ciliated protozoan *Oxytricha trifallax* and its relatives have an extraordinary genomic organization that offers an unusual opportunity for eukaryotic ncRNA gene discovery: single-gene 'nanochromosomes'. Ciliates (phylum Ciliophora) have two types of nuclei: a diploid mitotic germline micronucleus that is almost transcriptionally silent, and one or more somatic macronuclei that divide by amitotic fission (31,32). The macronuclear genome expresses all genes required for growth and asexual reproduction. Asexual variants of ciliates entirely lacking micronuclei occur in the wild (33). The macronuclear genome consists of many small, linear, acentric chromosomes which are produced from the micronuclear genome by a baroque ncRNA-dependent process of fragmentation, DNA elimination and (in some ciliates, including *Oxytricha*) rearrangement and unscrambling during sexual conjugation (33–38). These DNA processing events apparently depend both on non-genic transcription of long RNAs (39), and also (at least in *Tetrahymena*) involve large numbers of Argonaute-bound small RNAs (40–43). 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 (44,45). In *Oxytricha trifallax* [also known as *Sterkiella histriomuscorum* (46)], the micronuclear haploid DNA content of ~1 Gb is reduced by 95% to ~50–55 Mb of sequence complexity in the macronucleus. The macronucleus is thought to contain ~17 000–25 000 different nanochromosomes almost entirely in the range of 1–8 kb, with a mean of 2.2–2.5 kb (33,47,48). Each nanochromosome is amplified to an average copy number of ~1000. Remarkably, each nanochromosome usually contains only a single gene (49,50).

The unique nanochromosome structure of spirotrich ciliate genomes enables a systematic screen for new eukaryotic ncRNA genes that is essentially independent of either transcriptomics or computational RNA structure prediction. In effect, in these ciliates with gene-sized nanochromosomes, the organism itself has solved the hard eukaryotic genefinding problem for us. In *Oxytricha*, most genes and their *cis*-regulatory signals have been isolated on individual chromosomes, their locations demarcated by telomere addition, and most of their non-essential non-coding DNA has been removed (51). Given a macronuclear genome sequence, we can identify and discard nanochromosomes carrying protein-coding genes, 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 (44,49). The resulting subset of apparently non-coding 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 (44) to conduct such a screen.

MATERIALS AND METHODS

O. trifallax draft genome assembly

The WGS data set is a pre-publication whole genome shotgun draft assembly version 2.1.1 (June 2007), comprising 54 982 contig sequences (79.2 Mb) averaging 1.44 kb in length, using whole cell DNA prepared from vegetatively growing *O. trifallax* strain JRB310 (51) and size-selected by gel purification for <7 kb nanochromosomes to avoid the abundant rDNA nanochromosome. [This size fractionation nonetheless captures the great majority of the macronuclear genome, which is predominately in pieces of 1–8 kb (33,47,48)]. These data consist of all contigs of ≥2 reads (i.e. excluding singletons) from a PCAP (52) assembly of 728 035 ABI 3730 shotgun reads totalling 583.7 Mb of raw sequence. Overall assembly contiguity is less than expected from 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.

The 'pilot' data set is a collection of pilot sequencing data comprising 1976 complete nanochromosome sequences totalling 1.96 Mb consisting of 254 complete nanochromosome sequences from a Princeton/Utah pilot genome project (44,49,51), 1707 nanochromosomes generated by paired-end sequencing of full-length plasmid inserts cloned from a size-selected <1 kb nanochromosome fraction, the 7.6 kb ribosome DNA nanochromosome, and 14 additional full-length nanochromosome sequences.

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

Stage 1 data set: non-redundant 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 data sets, we extracted a non-redundant, merged set of apparently full-length *Oxytricha* nanochromosomes (the 'Stage 1' data set). All 1976 contigs in the pilot data set 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 ($[CCCCAAAA]_3$ at each contig's 5′-end and $[GGGGTTTT]_3$ 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. 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 (848 contain untrimmed flanking vector sequence ≥100 nt on one or both ends).

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

The Stage 1 data set consists of 9647 full-length nanochromosome sequences of average length 1.9 kb.

3× shotgun sequencing of *S. lemnae*

We surveyed 10 stichotrich ciliate isolates by PCR and sequencing of four conserved protein-coding genes: telomere-end binding proteins α and β, HSP70 and DNA polymerase α. The number of 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 (53). Two isolates (*O. fallax* and *Oxytricha* 'Bath') were too closely related, but eight isolates ('*Sterkiella histriomuscorum*', *O. nova*, *O.* Maryland, *Stylonychia lemnae*, *S. mytilus*, *Laurentellia* sp., *Paraurostyla* sp. and *Urostyla* sp.) were all suitable, ranging from 0.3 to 0.6 substitutions/4box-site. We chose *Stylonychia lemnae*, with an average of about 0.4 substitutions/4box-site.

Whole cell DNA obtained from *S. lemnae* strain 2x8/2 was kindly provided by Francziska Jönsson and Hans Lipps (University of Witten, Germany). A sample was sequenced in one 454FLX run. (Purification of macronuclear DNA away from micronuclear DNA is unnecessary, because macronuclear DNA is in vast excess.) This produced 568 094 reads totalling 146 Mb, about 3× average shotgun coverage of the presumed ∼50 Mb macronuclear genome in reads averaging 260 nt in length. The Newbler program (https://valicertext.roche.com/) was used to assemble these reads into 53 806 contigs ranging in size from 95 to 9947 nt.

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.

Estimation of genome assembly coverage

We attempted to gauge roughly how complete the *Stylonychia* and *Oxytricha* data sets are. Parra *et al.* (54) 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. We modified their method for use on a low-pass, low-contiguity shotgun assembly without full-length gene predictions. Using the same parameter settings described in (55), we searched each CEG with TBLASTX against each of our ciliate data sets, 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, 131 CEGs (∼53%) are present in the *Stylonychia* data set, 215 (∼87%) in the combined *Oxytricha* WGS+pilot data set, and 162 (∼65%) in the *Oxytricha* Stage 1 data set.

Estimation of actual nanochromosome length distribution

A histogram of nanochromosome lengths estimated from DNA contour lengths in electron microscope images was obtained from Figure 3 in Swanton *et al.* (48). We consider this to be the most direct measurement. Additionally, as a corroborating approach, pixel intensities were extracted 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 (ethidium is an intercalating dye) and converted to relative molar nanochromosome abundance by dividing by DNA length (56).

Computational 'nanoclassifier'

The computational nanoclassifier uses an HMM protein-coding gene model. 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 background (null hypothesis) model has the same HMM state-structure as the gene model, in order to match length distributions, but the emission statistics of all

states are changed to background: 5th order Markov (hexamer) background statistics in the exon states (estimated from the entire Stage 1 data set), and 0th order background nucleotide frequencies in all other states.

For positive (coding gene containing) training and test data, a set of 2520 nanochromosomes were identified in the Stage 1 data set 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 data sets 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 (57).

For additional training data, we obtained 24 annotated *Oxytricha* nanochromosome sequences from NCBI Genbank, and 33 nanochromosomes manually annotated using cDNA EST sequences.

For negative (non-coding) test data, we generated 2500 simulated nanochromosome sequences using an HMM consisting of three states (5′-telomere, sequence and 3′-telomere). 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 2nd order Markov statistics trained on the entire Stage 1 data set. Each was randomly assigned to one of ten jackknife training sets.

Analysis of nanochromosomes containing known ncRNA genes

The cmsearch program of Infernal 1.0.2 (58) was used to search 9647 Stage 1 nanochromosomes against 1372 ncRNA models in the Rfam 9.1 database (59) at an $E \leq 0.001$ threshold per query model. About 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), leaving 135 ncRNA homologs from 11 Rfam families on 134 different nanochromosomes.

ncRNA homology regions (as identified by the Infernal alignment, plus an extra 20 nt on each side) were masked by converting to N's, and any vector sequence was removed (using telomere endpoint coordinates described above). One nanochromosome carrying the ribosomal RNA genes (identified by the presence of 5.8S rRNA), was manually masked for SSU rRNA and LSU rRNA (Rfam does not include complete models of the large SSU and LSU rRNAs).

In the analyses in Table 1, WU-BLASTX comparisons to the NCBI NR database used options 'filter = seg filter = xnu C = 6' (C = 6 is the ciliate genetic code) with a $E < 10^{-5}$ threshold. WU-BLASTN comparisons to our *Stylonychia* shotgun data used options 'filter = seg filter = dust' 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 shown in Figure 2.

Comparative sequence analysis of coding conservation pattern

We produced pairwise *Oxytricha*/*Stylonychia* local alignments using WU-BLASTN with options 'filter = seg filter = dust maskextra = 10 M = 4 N = −5'. (These options improve BLASTN's ability to specifically detect relatively more distantly-related ncRNA homologies.) We selected alignments of ≥70% identity and $E \leq 10^{-5}$. These alignments were processed with 'blastn2qrnadepth.pl' and 'eqrna -a', from the the QRNA package (26).

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 shuffled the pairwise alignments by columns (thus preserving mean base composition and percent identity) before splitting.

O. trifallax culture and RNA extraction

O. trifallax strain JRB310 (60) was cultured to ≥5000 cells/ml density in 8 × 12 inch Pyrex dishes with 300 ml ciliate media. They were fed a mixture of algae (*C. elongatum*, University of Texas) grown in 500 ml flasks under light in Euglena media, and bacteria (*K. pneumoniae*). *Oxytricha* cells were collected on several layers of gauze to exclude clumps of algae, then filtered on a Nitex nylon membrane. The media is brought to 0.05 M EDTA to immobilize the motile cells and to reduce RNase activity, and cells are collected by centrifugation at 4°C. Total RNA was extracted by a standard Trizol (Invitrogen) protocol, and stored in 1 mM EDTA at −20°C.

Northern blots

A total of 2–10 μg of *Oxytricha* RNA were run on 4% acrylamide gels, electroblotted using a semi-dry electrophoretic transfer unit, and UV crosslinked to a ZetaProbe charged membrane (BioRad). DNA oligonucleotide probes (39–44 nt) were end-labeled with $\gamma^{32}P$-ATP using T4 polynucleotide kinase and hybridized to the northern blots in UltraHyb Oligo solution(Ambion) at 42°C for overnight (at least 15 h). Blots were washed twice in a solution of 2 × SSC and 0.1% SDS at 55°C for 5 and 15 min, then again twice in 0.1 × SSC and 0.1% SDS solution at 55°C for 5 and 15 min. Blots were either visualized by phosphorimager, or exposed at least 1 day

on X-ray film at −80°C. For some probes with lower calculated melting temperatures, an additional northern blot was done using less stringent hybridization and washing temperatures: 37°C for hybridization, 42°C for washing. A ^{32}P-labeled 50 bp dsDNA ladder (New England BioLabs) was used for molecular weight standards.

RACE-PCR

Poly-A tails were added to total RNA by terminal deoxynucleotidyl transferase. We used two different commercial RACE protocols: SMART-RACE (Clontech) and GeneRacer (Invitrogen). The SMART-RACE 5′-RACE protocol relies on addition of 3–5 untemplated C residues at the 3′-end of first strand synthesis by reverse transcriptase. This protocol is less efficient in our hands, but should be relatively insensitive to unusual 5′ RNA 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 be sure 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.

Programs and databases

Infernal 1.0.2 was used for RNA similarity searches (58). Infernal models of known ncRNA families were from the Rfam 9.1 database (59). 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. To evaluate the performance for nanochromosome classification, Genezilla (61), Unveil v1.0 (62), GeneID v1.2 (63) and Augustus 2.0 (64) were examined. Proximal Sequence Element (PSE) and 3′box consensus motif searches were done with HMMER 1.8.4 and HMMER 2.3.2. Multiple alignments were produced using MUSCLE (65) or CLUSTALW (66) and manually edited in Emacs using the RALEE alignment editing mode (67). Some screens for snoRNAs used snoGPS 0.2 (68) and snoscan 0.9 b (69). Analysis of cDNA/genome alignments used Exonerate 1.0.2 (57), and unpublished cDNA/EST data. For comparative analysis of coding gene sequence conservation patterns, we used QRNA 2.0.3 c (26). Sequence logos were generated with WebLogo 2.8.2 (70).

Data set 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 data sets described in the paper are available for download at http://selab.janelia.org/publications.html#JungEddy11.

RESULTS

A data set of 9647 full-length *O. trifallax* nanochromosomes

The *O. trifallax* macronuclear genome project is an ongoing collaboration between the Genome Center at Washington University and the Landweber laboratory at Princeton (http://genome.wustl.edu/genomes/view/oxytricha_trifallax/). We obtained two draft data sets: a 'WGS' data set consisting of the version 2.1.1 (June 2007) macronuclear genome shotgun assembly (54 982 contigs totalling 79.2 Mb), and a 'pilot' data set consisting of 1976 full-length nanochromosome sequences totalling 1.96 Mb.

Our screening strategy involves classification of full-length nanochromosomes as coding or non-coding. The pilot data consist of complete nanochromosomes, but the WGS data are a draft assembly with many incomplete contigs. We extracted presumptive full-length nanochromosomes in the WGS data set by identifying contigs with telomere repeats on both ends ('Materials and Methods' section). In cases where near-identical contigs were present in both the WGS and pilot data sets, we selected one at random. The combined data set (the 'Stage 1' data set) consists of 9647 presumptive full-length nanochromosomes.

Four typical examples of *Oxytricha* full-length nanochromosome organization are shown in Figure 1, including annotations by methods we describe below.

Characterizing completeness and bias of the Stage 1 data set

The Stage 1 data set 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 non-standard assembly effort beyond the scope of this work. However, because our main question is about the relative *proportion* of 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 sample, as follows.

We estimate that the data set includes 40–65% of the macronuclear genome, based on two different estimates. First, by dividing the complexity of the macronucleus as measured by reassociation kinetics (50–55 Mb) by the average nanochromosome size (2.2–2.5 kb) (33,47,48), *Oxytricha* is thought to contain about 20 000–25 000 different nanochromosomes; 9647 would represent around 40–50% coverage of the genome. Second, we measured coverage of a set of conserved core single-copy

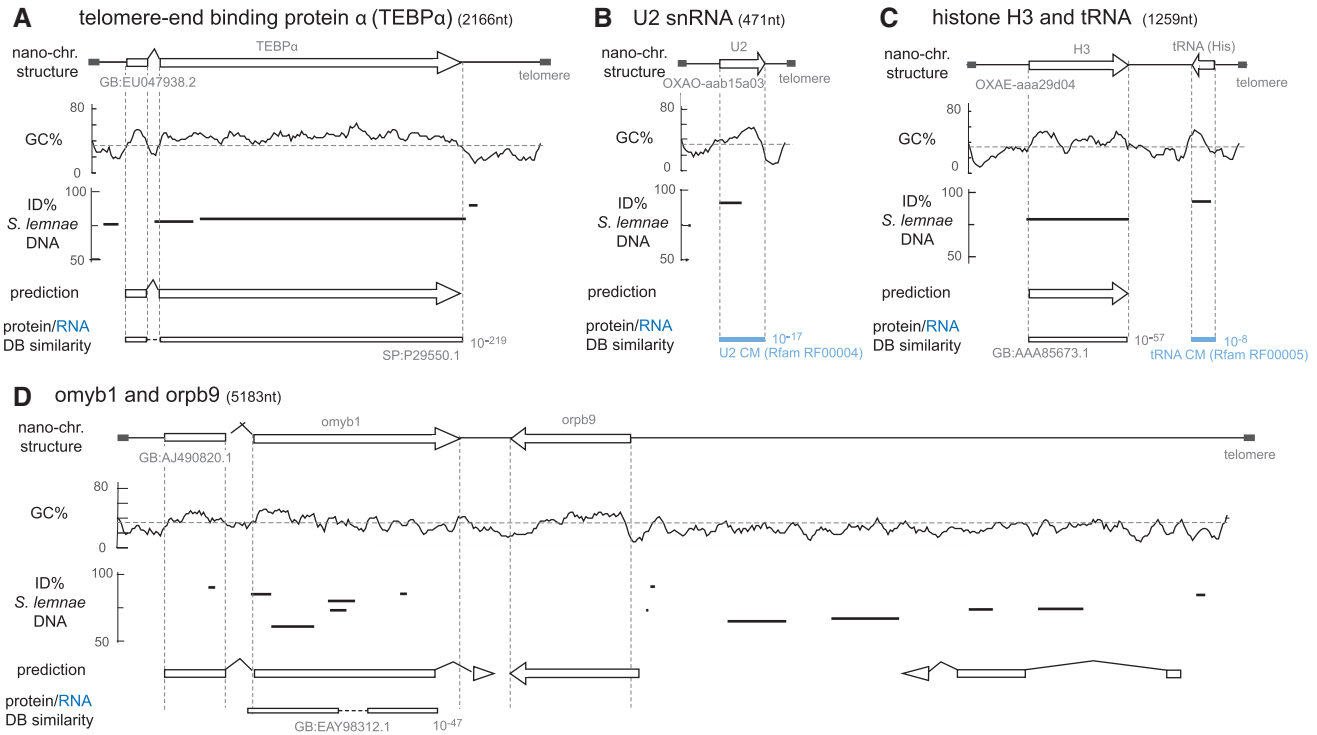

Figure 1. Examples of *O. trifallax* nanochromosomes. (**A**) A typical nanochromosome containing a single protein-coding gene (telomere-end binding protein α); (**B**) a typical nanochromosome containing a single ncRNA gene (U2 snRNA); (**C**) a nanochromosome containing both a protein-coding gene (histone 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 and D) or predicted by us by similarity (B and C). GC%: calculated GC% in sliding 50 nt windows with 10 nt 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 'Materials and Methods' section); prediction: coding gene prediction from our *Oxytricha* genefinding program; protein/RNA DB similarity: best significant WU-BLASTX matches to NCBI NR protein database excluding *O. trifallax* proteins (black) or Infernal cmsearch (58) matches to the Rfam RNA database (59) (blue).

eukaryotic protein genes (54), yielding an estimate of 65% coverage (see 'Materials and Methods' section).

We expect a bias toward 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 data set to two different estimates of the actual length distribution of the overall macronuclear genome. The actual distribution has been characterized previously by agarose gel electrophoresis, and also by measuring the contour lengths of approximately 1000 individual nanochromosomes in electron micrographs (48). We digitized an ethidium-stained agarose electropherogram and we extracted the EM contour length histogram from reference (48) (see 'Materials and Methods' section). Both methods gave 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 1.8 kb, ranging up to 13.8 kb. About 2% of nanochromosomes are larger than 7 kb, whereas only nine contigs in the Stage 1 data are longer than 7 kb (0.1%), indicating substantial (20×) undersampling of the 2% tail of longest nanochromosomes. Around 15% are 4–7 kb, where we have 202 contigs in the Stage 1 data (2%), indicating moderate (7–8×) undersampling in this length range. In the 80% of nanochromosomes that are <4 kb, there is only modest sampling bias (length histograms in this range are shown later in Figure 3).

We also expect a bias toward assembling more abundant nanochromosomes, which get higher sequence coverage. Each *Oxytricha* nanochromosome occurs with a mean of approximately 1000 copies per macronucleus (33,48), but some nanochromosomes are known to be maintained at different copy numbers. The most extreme case is the rDNA nanochromosome, found in about 100 000 copies. The rDNA appears as a prominent 7.6 kb band on agarose gels of macronuclear DNA (33), where a distinctive species-specific pattern of 100–200 overrepresented bands is also seen (71). Several examples of about 6-fold copy number differences have been observed when copy number of individual non-rDNA nanochromosomes has been measured (72,73), and a few cases of extreme overamplifications have been observed during prolonged vegetative growth (74). However, reassociation kinetics experiments have shown that bulk macronuclear DNA reanneals as if the great majority of sequences occur in roughly equal numbers (31,47,75). 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. As expected from previous published results, this coverage distribution is consistent with non-uniform copy number varying over perhaps an order of magnitude, and it appears we have likely sampled the bulk of that distribution.

A computational 'nanoclassifier' to detect coding nanochromosomes

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 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.

We need a coding genefinder to have high sensitivity to make our screen effective. We are less concerned with the comprehensiveness of our screen's ability to detect ncRNA genes—the Stage 1 data set is already only a sample—so we can tolerate somewhat low specificity (i.e. the rate of miscalling a non-coding nanochromosome as coding and throwing it away). We aimed to develop a coding gene classifier with ~98% 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 data set, with the rest (9547) containing one or more coding genes. At 98% sensitivity, about 200 (2%) of nanochromosomes carrying coding genes would be misclassified as non-coding. 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 280 'non-coding' candidate nanochromosomes, only 80 of which would really contain ncRNA genes (30%). This would be a tolerable signal/noise in a candidate set that we could sort out using further computational and experimental analysis. We would not want the signal/noise ratio of our ncRNA gene screen to drop much lower than this. 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.

To develop a high-sensitivity classifier, we first looked at using 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 (76). 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 data sets consisting of incomplete gene structures, because we have few cDNA-validated gene structures for *Oxytricha*. We chose Genezilla (61), Unveil (62), GeneID (63) and Augustus (64) for evaluation. Genezilla was the program used for genefinding by the *Tetrahymena thermophila* genome project (77), and GeneID was used by the *Paramecium tetraurelia* genome project (78). GeneID is trainable from partial gene structures.

We evaluated how accurately these programs could distinguish coding nanochromosomes from non-coding random sequences of the same size and composition. For positive training and test data, we identified a data set of 2520 nanochromosomes with significant BLASTX homology to known proteins, then defined partial exon/intron structures by protein/genomic DNA alignment with Exonerate (57). We also identified an additional training data set of all 24 annotated *Oxytricha* genes in Genbank, and 33 genes manually annotated using expressed sequence tag (EST) coverage. For negative data, we generated 2500 random nanochromosome-sized sequences according to *Oxytricha*'s overall 2nd-order Markov residue composition, flanked by simulated telomere repeats.

We jackknifed the positive and negative data sets to construct 10 different test sets of 252 positives and 250 negatives, leaving 90% of the positive data for training on Exonerate-annotated partial gene structures. We trained GeneID 10 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, and only 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 10 times on jackknifed sets of 252 positives and 250 negatives for its ability to discriminate coding nanochromosomes from synthetic non-coding nanochromosome-like sequences.

Figure 2A 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. This is probably due to the dearth of well-annotated *Oxytricha* gene structures for training data. It might 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.

We used hidden Markov model methodology (79) to specify a probabilistic model of *Oxytricha* nanochromosomes containing coding genes (see 'Materials and Methods' section). Figure 2B shows a schematic of our model. It includes standard statistical features for eukaryotic genefinding (80), such as 5th-order Markov (hexamer)

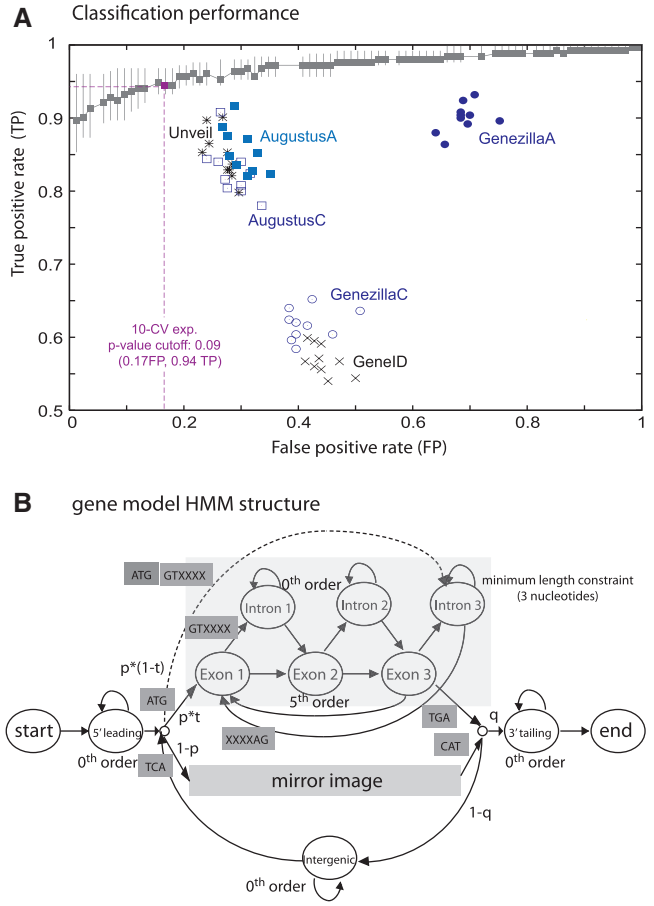

Figure 2. A hidden Markov model based coding nanoclassifier. (**A**) ROC curve of classification performance. 10CV exp: results of 10-fold cross-validation on jackknifed test sets of 252 positive sequences and 250 negative sequences, showing the average (gray box) and range of the 10 test results as *P*-value threshold is varied. GeneID: results of 10-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 data set, 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.

statistics for residues in coding exons and an intron 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 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 non-coding extragenic and intragenic DNA segments, so the overall model is that of a complete full-length nanochromosome containing one or more coding genes. 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.

A Viterbi (maximum likelihood) HMM parse of a nanochromosome sequence with this model would be a coding genefinding prediction of one or more coding gene structures. We will describe using the model for *Oxytricha* protein genefinding elsewhere (manuscript in preparation). Here we are interested in nanochromosome classification, not parsing. For classification, we need to set up a hypothesis test involving two specified models—the coding gene model just described, and a 'null hypothesis' that generates non-coding nanochromosomes. Our null model has an identical state and state transition structure as the coding model, with all residue emission probabilities (including start and stop codons and GT/AG splice sites) replaced by background probabilities. 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 might get classified spuriously by length rather than statistical coding signals. Given a full-length nanochromosome sequence, we calculate a log likelihood for both models (using the HMM Forward algorithm, summing over all possible parses), 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 non-coding nanochromosomes, but length and residue composition effects introduce biases into log-odds scores for individual nanochromosomes. 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 2nd 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 2A 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 data set (not just a jackknifed subset) for subsequent use.

Many *O. trifallax* nanochromosomes contain only a single gene

Previous studies indicate that *Oxytricha* nanochromosomes usually contain just a single gene (44,49,50,81) with a few exceptions (50,82–84), 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 identified homologs of known ncRNA genes in the Stage 1 data set and examined those ncRNA-containing nanochromosomes for protein-coding potential.

By searching the Stage 1 data set against the Rfam ncRNA database (59), we identified 135 putative non-coding RNA genes (on 134 different nanochromosomes) from 11 different families, including 106 transfer RNA (tRNA) genes (Table 1; 'Materials and Methods' section). In all but one case, we found a single ncRNA homolog per nanochromosome. Hundred and thirty-three nanochromosomes carried a single known ncRNA homolog, and one nanochromosome had homologs of two known ncRNA genes, RNase MRP and snoZ196.

There are usually several identical or near-identical copies of each locus in the assembly. We generally identify up to four apparent 'alleles' of any given sequence. The sequenced *Oxytricha* culture was an inadvertent mixture of two mating types, 310 and 510 (E.S. and L.F.L., unpublished data). There also appears to be a substantial fraction of alternative DNA processing (different macronuclear nanochromosome 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. Operationally, we manually grouped highly identical loci (roughly >85% identical in DNA sequence flanking each locus) into what we call 'quasiallele' groups. For each group, we assign a representative locus. In subsequent sections of the article 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* non-coding candidate'), and numbered each additional quasiallele 'Onc1.2', 'Onc1.3', etc. Names, coordinates, and other information for all loci described in the paper, including candidate loci described in the screen below, are collated in an electronically parsable table (Oxy_ncRNAs.list) included in additional data sets available for download (see 'Materials and Methods' section).

To estimate how many of these 134 nanochromosomes contain coding genes in addition to an ncRNA gene, we masked the homologous ncRNA regions (converted the sequence to N's) and used three different methods to look for possible coding genes: regions of significant similarity to known proteins (by BLASTX), significant BLASTN sequence conservation with *Stylonychia lemnae* (which will overestimate coding, by detecting all kinds of sequence conservation) (see 'Materials and Methods' section), and coding gene potential detected by our nanoclassifier. Results are summarized in Table 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%) as coding.

Each method for detecting coding genes has drawbacks, in terms of both sensitivity and specificity. In terms of sensitivity, some coding regions will not show BLASTX hits to the protein database because they are rapidly evolving or '*Oxytricha*-specific' genes. Some will not show BLASTN hits to *Stylonychia* because our *Stylonychia* shotgun data have partial coverage (see 'Materials and Methods' section). Our nanoclassifier has an estimated coding sensitivity of ~94%. 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

Table 1. Coding potential of ncRNA gene-containing nanochromosomes

ncRNA	Rfam accession	No. of 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 (59); the third column, 'No. of 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, and the number called coding by all three methods.

method. In terms of specificity for coding regions, some ncRNAs show BLASTX hits to the 'protein' databases because some ncRNA genes have been erroneously translated and deposited in the databases; BLASTN conservation to *Stylonychia* can mean many things besides a conserved coding region, including an ncRNA or a large regulatory DNA sequence; and our nanoclassifier has ~17% false positive rate.

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 non-coding 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 ~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.

A computational screen for non-coding nanochromosomes

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 non-coding nanochromosomes for further analysis. We applied our nanoclassifier to each of the 9647 presumptive full-length nanochromosomes in the Stage 1 data set (Figure 3). Unexpectedly, this identified a Stage 2 data set of only 507 non-coding 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. To further increase the stringency of the screen, we used BLASTX to identify nanochromosomes with significant similarity to known proteins. This removed 69 more contigs, leaving a Stage 3 data set of 438 non-coding 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 non-coding nanochromosomes at this stage of the screen, but we do not. (Indeed, the actual number of non-coding 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 coding '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.

3× genome sequence of *S. lemnae* for comparative analysis

We wanted to use 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, 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*. We chose the stichotrich *S. lemnae* after surveying 10 ciliate species for their evolutionary distance to *Oxytricha* by PCR-sequencing of four conserved coding genes (see 'Materials and Methods' section). *S. lemnae* 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 (53).

We sequenced whole cell DNA from *St. lemnae* to approximately 3× coverage in one Roche 454FLX run, yielding an assembly consisting of 53 806 contigs totaling 27.3 M residues. We estimate this assembly covers about 50% of the *Stylonychia* genome ('Materials and Methods' section). We therefore expect to be able to detect around 50% of single-copy evolutionary conserved regions in *Oxytricha* by comparison with the *Stylonychia* data set.

Sequence conservation in putatively non-coding nanochromosomes

We expect the steps to this point may have also enriched for artifactual 'non-coding' contigs that would arise either from sequence assembly errors or possible DNA processing errors in *Oxytricha*. Figure 3 shows length distributions for the contigs in each data set, showing the progressive enrichment of a peak of small (~100 nt) contigs in the Stage 3 data that we presume to be 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. This identified a data set of 127 conserved non-coding nanochromosomes (Stage 4; Figure 3). The peak of small contigs disappears.

The Stage 4 data set 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 non-coding nanochromosomes with no

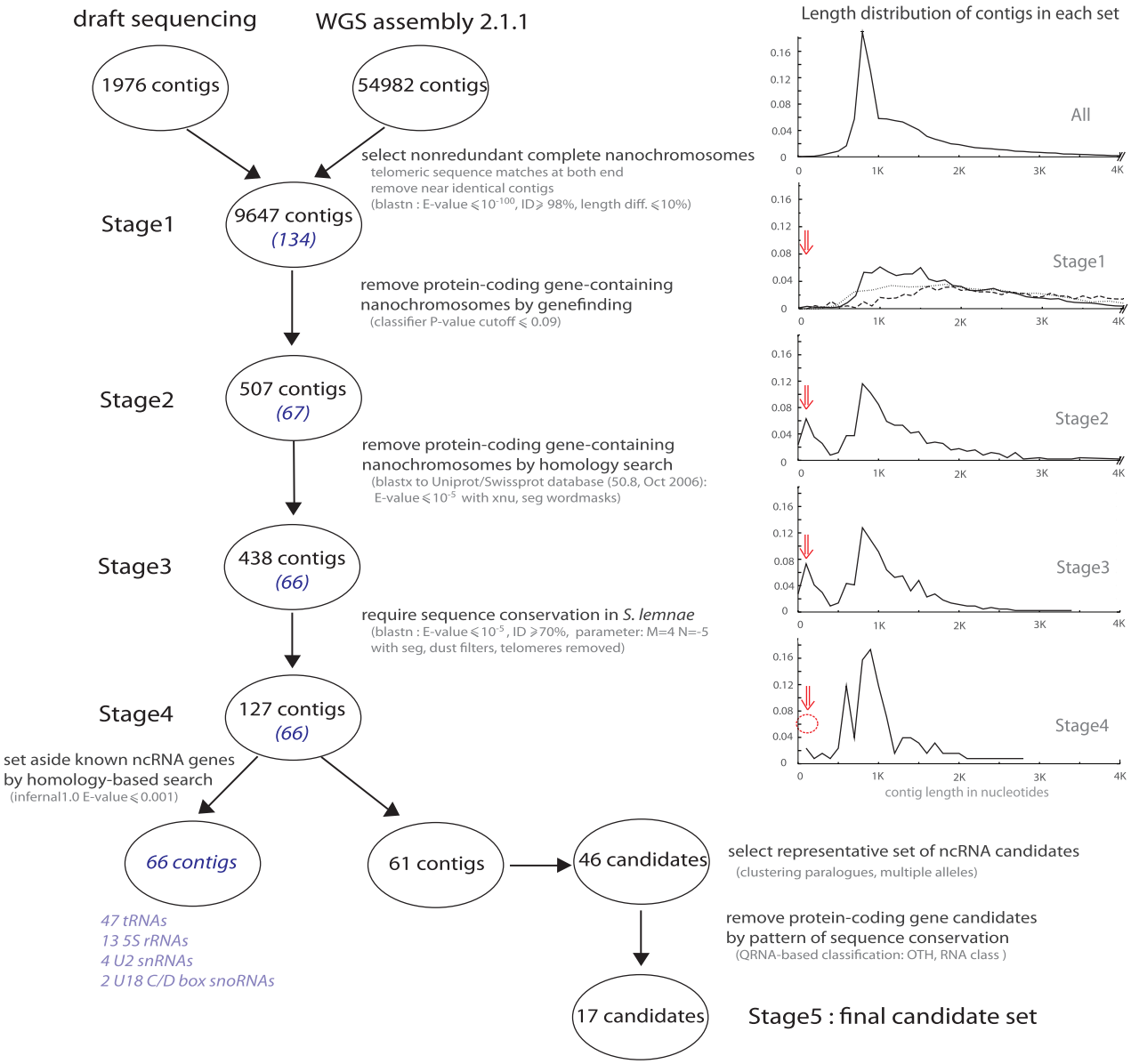

Figure 3. Flowchart of the screen for non-coding nanochromosomes. The graphs to the right show the length distribution of the data set at each stage of the screen. In the Stage 1 histogram, the dashed line shows the actual nanochromosome length distribution as estimated from an agarose gel electropherogram, and the dotted line shows the actual nanochromosome length distribution as estimated by Swanton *et al.* (48) from contour length in electron microscope images. Red arrows indicate a peak of small (presumably artifactual) non-coding contigs that is initially enriched, then removed when a requirement for DNA sequence conservation to *Stylonychia* is imposed.

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 (described below). 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).

Despite all the steps taken so far, we still expect that more than half of these 46 contigs 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 ~65% of *Oxytricha* proteins have significant similarity to known proteins, then we expect around 37 coding regions to pass into Stage 4. About 70% (26) of these would pass the Stage 4 conservation screen against the incomplete *Stylonychia* data set. 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 shows examples of this periodic pattern in a known coding region as opposed to known ncRNA genes. We scored this pattern in pairwise BLASTN alignments with the program QRNA (26). Although QRNA was originally designed to identify structural ncRNAs by comparative analysis (a task that remains difficult, with high false positive rate), it also includes an effective statistical model for discriminating conserved coding regions from other types of sequence conservation. On test data sets of coding and simulated non-coding *Oxytricha*/*Stylonychia* alignments, we estimated that QRNA has a true positive rate of 95% and a false positive rate of 3% for distinguishing conserved coding regions ('Materials and Methods' section). 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 non-coding nanochromosomes.

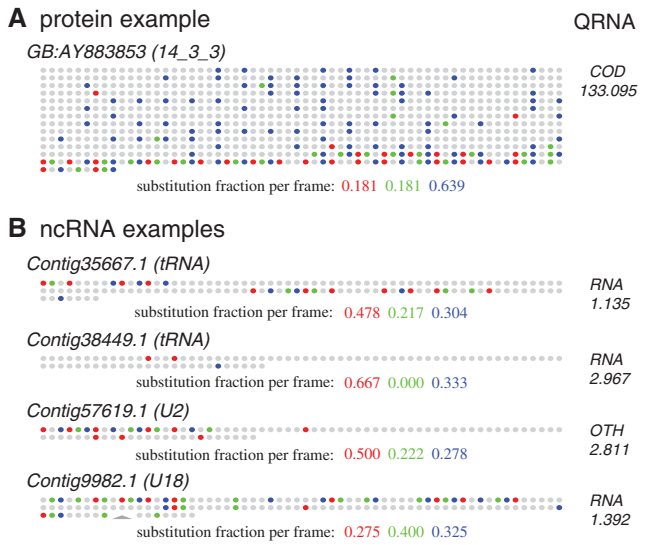

Figure 4. Comparative sequence analysis of sequence regions conserved with *Stylonychia*, showing examples of one known protein and four ncRNA genes. A BLASTN alignment of an *Oxytricha* reference sequence to its *Stylonychia* homolog is represented graphically, with each colored dot representing one *Oxytricha* residue, with identities to *Stylonychia* colored gray and substitutions in three different frames colored red, green and blue. Each alignment is wrapped across multiple lines [i.e. (**A**) shows one *Oxytricha* coding sequence wrapped into 14 lines, not an alignment of 14 sequences]. QRNA (26) 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 in eye 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).

Spliceosomal small nuclear RNAs flanked by conserved motifs

Seventeen seemed like a surprisingly small number of new ncRNA genes to find in a eukaryotic genome, given some of the current literature on eukaryotic ncRNAs (8–10). We sought to study in more detail how some of the largest known ncRNA gene families behaved in the screen, in order to be sure that we were sampling them at the expected frequency, and to look for any unexpected reasons why we could miss ncRNA genes. For example, we were concerned that only the U2 and U6atac spliceosomal small nuclear RNAs (snRNAs) were identified in the Stage 1 data set. 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 (85). We analyzed the incomplete contigs of the WGS+pilot data set 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 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 found in the Stage 1 data set. We expected about half of them, given our coverage estimate of 40–65%. This might 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.

Manual analysis of the pattern of sequence conservation flanking snRNA loci revealed two conserved motifs. A 17 nt motif TgACCCATnAAnnnTTA occurs about 50–60 nt upstream of the putative 5′-end of all snRNAs and some other ncRNAs (RNase P, telomerase and SRP). This motif is likely to be the homolog of the 'proximal sequence element' (PSE) found upstream of snRNA genes in many organisms (86,87) including other ciliates (88). A 19–20 nt motif AAAnGAAAnnGTTTGA TTAg occurs 8–12 nt downstream of the putative 3′-end of most snRNAs (except for U6 and U6atac, which show the hallmark T_n terminator of polIII-transcribed small RNAs). This motif is likely the functional analog (if not the homolog) of the 3′box motif responsible for 3′-end processing in snRNAs and other small RNAs in many organisms (89). These putative transcriptional signals gave us additional means to analyze the sequences of the novel ncRNA loci that the screen identifies, as described later.

Small nucleolar RNAs often intron-encoded, and underrepresented

We expect that like other eukaryotes, *Oxytricha* has tens to hundreds of small nucleolar RNAs (snoRNAs) (90). 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. *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 (91) in *Oxytricha* by TBLASTN. However, the similarity search analysis in Table 1 only identified four 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. We used a variety of automated and manual approaches to identify a set of probable *Oxytricha* snoRNAs in the WGS+pilot data set, to see how snoRNA loci would behave in our screen. These methods included the snoscan and snoGPS search programs (68,89); *Stylonychia* sequence conservation; low-stringency Rfam/Infernal homology searches; searches for conserved regions flanked by the PSE and 3′box motifs identified above; and manual sequence analysis.

Overall, in analyses of the entire assembly (not only Stage 1 data), and including the results of the screen described below, we predicted 35 distinct snoRNA loci, including 29 distinct methylation guide C/D snoRNAs, five distinct H/ACA snoRNAs, and one distinct U3 snoRNA locus, on 20 different nanochromosomes. Only 4 of the 20 contigs (20%) are incomplete and fail to reach the Stage 1 data set, somewhat fewer than expected from 40% to 65% coverage. Nine of the 16 (56%) Stage 1 nanochromosomes are classified as coding, about what is expected from 50% ncRNAs being on non-coding nanochromosomes. 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 described below.

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, sometimes because an entire cluster is carried on one long precursor ncRNA that is processed to release multiple snoRNAs. In both identified arrays, the C/D snoRNAs appear to be intronic in a carrier transcript, judging from the presence of strongly conserved 5′-splice sites and lack of other conservation in the contigs (3′-splice sites are less conserved and more difficult to identify in AT-rich *Oxytricha* sequence.). Another 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 non-coding transcripts. Large arrays may be on large contigs that are not fully assembled in the current data set (and thus less likely to appear in our Stage 1 data set), and intron-encoded ncRNAs in coding genes will be screened out by the coding gene classifier step. Although the screen successfully detects both C/D and H/ACA snoRNAs (described below), they are likely underrepresented for these reasons. This illustrates a weak point in the screen.

RNA expression assayed by northern and RACE-PCR

To test whether our 17 candidate nanochromosomes express RNA transcripts from the identified regions of sequence conservation, we performed northern blots using ~40 nt single-stranded oligonucleotide probes directed against the most conserved region of each candidate. We probed each strand separately, using total RNA extracted from a single growth condition, vegetative growth in standard culture. For 7 of the 17 candidates we detected a small RNA transcript (Figure 5). 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 are part of the additional data sets available for download; see 'Materials and Methods' section.).

For six of the seven candidates detected by northern, we performed 5′- and 3′-RACE-PCRs and sequenced multiple clones from each in order to identify complete transcript sequences. 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 below, transcript sequence(s) implied by RACE-PCR sequencing were consistent with the band(s) observed by northern (Figure 5).

Manual analysis

We analyzed each of the 17 candidate loci in detail, taking particular advantage of the pattern of *Stylonychia* conservation (including multiple alignments where possible). For ncRNA loci that appear to conserve an intramolecular RNA secondary structure, we used manual comparative analysis to infer the structure.

Of the 10 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

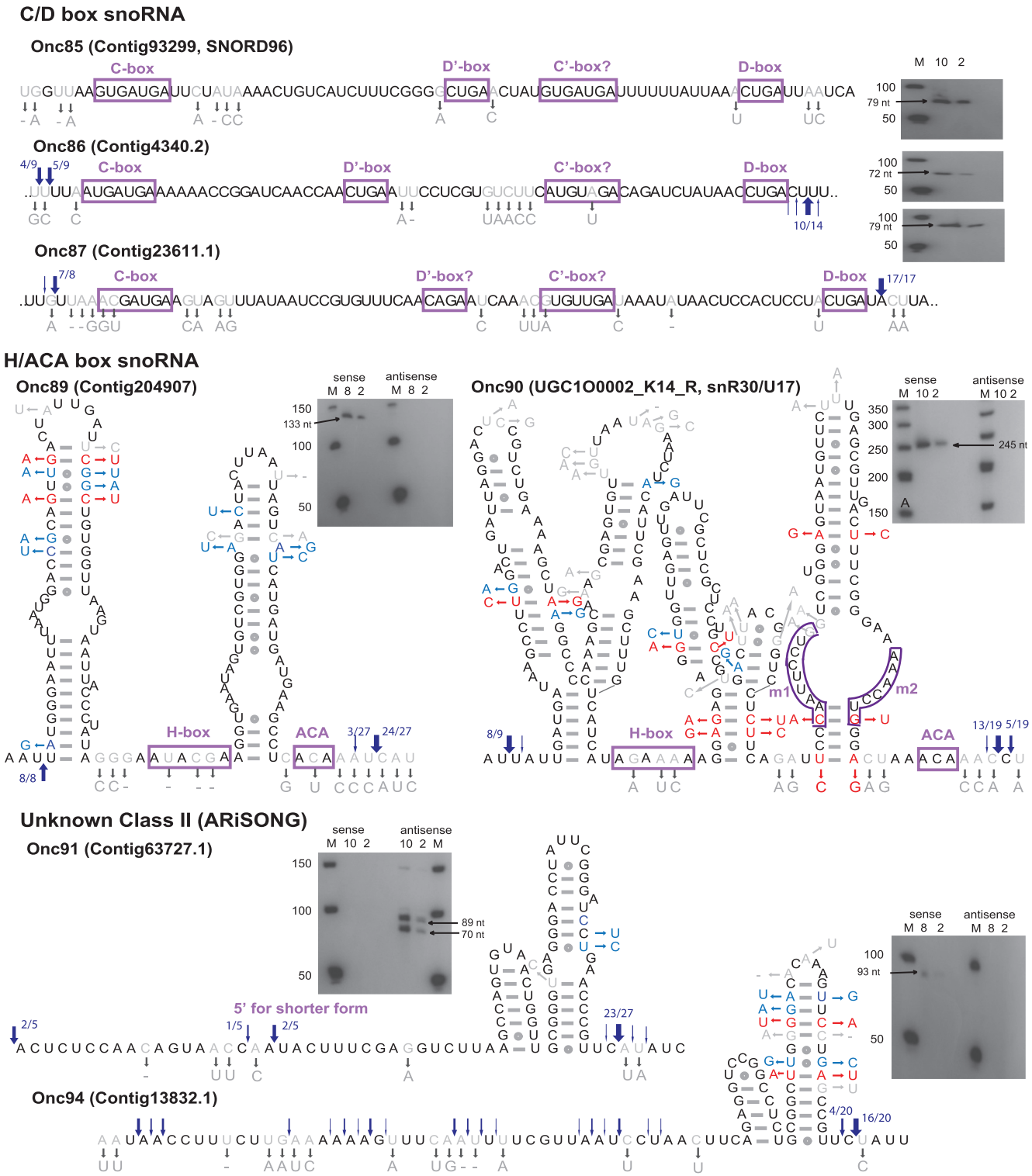

Figure 5. Experimental confirmation of small RNA transcripts. Sequences, predicted secondary structures, and northern and RACE data for seven candidates with detected transcripts. Genomic sequences are shown for each locus, with 5′- and 3′-ends of transcripts determined by RACE indicated by dark blue arrows. Arrows pointing between nucleotides indicate an unambiguously determined end; arrows pointing at an 3′-end A indicate an ambiguity, where we cannot distinguish an 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 50 bp 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 (104) then manually modified based on comparative sequence analysis and other features (such as the predicted target sites for the two H/ACA 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.

(Onc89, Onc90) (Figure 5). The C/D snoRNAs and the Onc89 H/ACA snoRNA have typical structures for these classes of eukaryotic snoRNAs. 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; (92)]. 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 (93,94). The proposed interaction for yeast snR30 and human U17 with their cognate SSU rRNAs is conserved for Onc90 with *Oxytricha* SSU rRNA (data not shown) (95).

The remaining two candidates that show small RNA expression (Onc91 and Onc94) share a well-conserved predicted RNA structure (Figure 5), but are not detectably homologous to any well-known eukaryotic small RNA families. We refined a multiple alignment of these loci and predicted their conserved secondary structure. Although we do not know their function, we provisionally named these the 'Arisong' family of ncRNAs. (In Korean, 'arisong hada' is to be something of unsure or confusing status.) We used Infernal and BLAST to iteratively search for additional Arisong RNA homologs in the *Oxytricha* genome, our *Stylonychia* genome and other available ciliate genome sequences: *T. thermophila* (Nov06 version) (96), *P. tetraurelia* [Dec06(v1) version] (78,97) and *Nyctotherus ovalis* (98). We refined our consensus secondary structure prediction as new homologs were identified. We found a total of 15 Arisong loci in the entire *Oxytricha* data set (the 'all' data set, including partially assembled nanochromosomes). These cluster into 7 distinct loci (Onc91, Onc92, Onc94, Onc95, Onc96, Onc155, Onc156) which appear to be paralogous (as opposed to allelic). We find 8 loci in *Stylonychia*, 8 in *Paramecium* and 1 in *Nyctotherus*. Six of the eight *P. tetraurelia* loci have been previously predicted to be small RNA genes called PM01_1-6 by identifying a PSE motif upstream of conserved sequence (99).

All these loci are predicted to share a consensus secondary structure consisting of a coaxially-stacked dumbbell, with highly conserved sequences at the stacked junction, a highly conserved 3′ GUUC tail, and a highly variable 5′-end (Figure 6). This structure is well-supported by a number of compensatory base pairs observed in the multiple alignment (A Stockholm format multiple alignment file of the Arisong RNAs, Arisong.sto, is included in the additional data sets available for download; see 'Materials and Methods' section.).

The variable sequence at the 5′-end is curious. Although the 3′-ends of Onc91 and Onc94 were readily mapped by RACE-PCR and both RNAs exhibit well-defined bands on northern blots, we had difficulty obtaining 5′-RACE-PCR products for Arisong loci, 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; see 'Materials and Methods' section). We also observed that Onc91 shows two northern bands of approximately equal intensity.

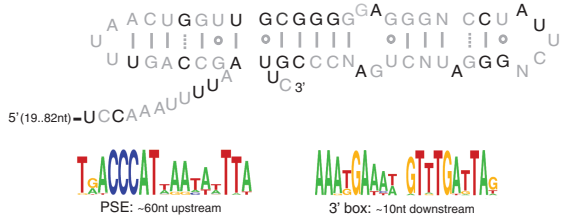

Figure 6. Consensus secondary structure of the Arisong RNAs and their flanking regulatory elements. The structure shown is drawn based on the individual structures of Arisong RNAs shown in Figure 5. The sequence is the majority-rule sequence consensus of a multiple alignment of 32 Arisong RNAs. Highly conserved residues (identical in ≥80% of aligned sequences) are shown in black; variable residues (identical in <50%) are shown as 'N'; weakly conserved residues are in gray. 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 (PSE.sto and 3box.sto, included in additional data sets available for download; see 'Materials and Methods' section) using the WebLogo program (70), after removing columns containing >50% gaps.

All seven distinct *Oxytricha* Arisong loci are flanked on their nanochromosome by a consensus PSE element about 50–60 nt upstream, and a consensus 3′box element about 5–10 nt downstream. This suggests that the Arisong loci are transcribed and processed similarly to spliceosomal snRNAs and U3 snoRNA, which are also flanked by PSE and 3′box elements.

DISCUSSION

Our screen identified a family of at least 32 small Arisong RNA genes in four ciliate species, encompassing a previous prediction of six small RNA loci in *P. tetraurelia* called PM01_1-6 (99). Chen *et al.* (99) noted the primary sequence conservation at these *Paramecium* loci and the fact that they were flanked by typical transcription and processing signals of polII-transcribed small RNAs, an upstream PSE and a downstream 3′box (the PSE in *Oxytricha* shares essentially the same consensus, while the 3′box consensus in the two species is different). Our work confirms Chen *et al.*'s (99) prediction, and extends it by: (a) expanding the Arisong RNA family to include homologs in four ciliate species; (b) confirming the expression of 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.

We can only speculate about the function of the Arisong RNAs. 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. However, these signals are not entirely unique to snRNAs. We identified U4atac and U6atac snRNAs in *Oxytricha*, suggesting the presence of a minor spliceosome, but U11 or U12 homologs remain unidentified. However, the consensus structure of the Arisong RNAs does not appear to resemble U11 or U12. The 5′ sequence variability of Arisong loci, and the two different sizes of the Onc91 Arisong RNA are somewhat evocative of spliceosomal snRNAs involved in trans-splicing that donate a 5′ splice leader while conserving an snRNA-like structure in their 3′ part. However, the structure of the Arisong RNAs does not resemble other known splice leader RNAs, we do not see a convincing conserved Sm binding site (although we do identify conserved putative Sm binding sites in *Oxytricha* snRNAs), and we do not see the 5′ sequence of Arisong RNAs on *Oxytricha* ESTs or cDNAs.

In a broader context, one possible conclusion of this study is a negative result. If novel independently-transcribed ncRNA *genes* were numerous in all eukaryotes, we expected to see many non-coding 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. The only novel ncRNAs detected by the screen overall were the Arisong family. The remaining non-coding nanochromosomes consist mostly of nanochromosomes carrying known ncRNA genes, and a number of small (~100 nt) nanochromosomes that show no sequence conservation with *Stylonychia* and that we currently presume to be artifacts either of nanochromosome formation or of sequence assembly. Our screen identified no long, mRNA-like ncRNA genes, other than probable non-coding host transcripts for arrays of intronic snoRNAs. However, our screen does identify representatives of most well-known structural ncRNA gene families, such as transfer RNAs and spliceosomal RNAs. This result suggests that if *Oxytricha* has large numbers of undiscovered ncRNAs encoded in its macronuclear genome, their genomic location must be systematically biased in ways that homologs of most well-studied ncRNA genes are not—for example, that they do not arise from independently transcribed ncRNA genes, but will instead come from non-genic transcription, or from processes obligately 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 result is also consistent with recent arguments that most of the 'ncRNA' that has been observed in mammalian systems is a mix of technical artifact and RNAs arising from *cis*-acting processes associated with transcription of nearby coding genes (22). However, the extent to which our largely negative results in *Oxytricha* sheds light on the current controversy about ncRNA abundance and function in eukaryotes in general must be couched with a number of limitations and caveats of our approach, which we enumerate as follows.

This conclusion depends on a sampling argument, because the data set of full length nanochromosomes is estimated to be only 40–65% complete. In principle, even just a small sample (a hundred or so) would suffice to conclude that the proportion of non-coding nanochromosomes is quite small, so long as that sample were random and unbiased. However, there are two important sources of bias to consider the Stage 1 data set, a bias toward shorter nanochromosomes, and a bias toward more abundant (higher copy number) nanochromosomes. We believe that neither bias is sufficient to account for the negative result, as follows.

The principal concern with a bias toward shorter nanochromosomes is that we could miss ncRNA genes like the recently described mammalian long intergenic ncRNAs (lincRNAs) (13,100). However, lincRNAs are only 'long' relative to other previously well-studied ncRNAs, which are often 100–400 nt. Mammalian lincRNAs have about the same length distribution as coding mRNAs (mean lengths of 2.5 kb versus 2.4 kb, respectively, according to GENCODE v4 transcript annotation; http://www.sanger.ac.uk/PostGenomics/encode/). The length distribution of the Stage 1 data set covers the great majority of coding nanochromosomes, so it is also expected to cover lincRNA-like ncRNA genes. Second, and more generally, the results of our screen show that non-coding nanochromosomes are systematically the smaller nanochromosomes. All 211 nanochromosomes longer than 4 kb were classified as coding. If a class of large ncRNA-containing nanochromosomes were present even at a few percent, we would have expected to sample some non-coding nanochromosomes in those 211 large contigs.

Abundance bias also seems unlikely to affect the conclusion. Our analysis of read coverage statistics, combined with published results of reassociation kinetics experiments (31,47,75), suggest that copy number variation generally appears to be modest, probably generally within an order of magnitude of the mean copy number of approximately 1000 per macronucleus. Our assembly coverage ranges over two orders of magnitude, up to 87× per contig. This should be sufficient to sample the bulk of the range of copy number variation. Combined with the estimate of 40–65% completeness of the Stage 1 data set, it seems unlikely that a population of low-copy ncRNA-carrying nanochromosomes exists that has been entirely missed, as opposed to somewhat undersampled.

Because we have only sampled the genome, we expect there are a few more undiscovered ncRNA genes in the *Oxytricha* macronuclear genome besides the Arisong family. We estimate that the overall probability of sampling any given ncRNA gene in the complete screen is roughly 10%. This comes from multiplying ~50% completeness of the genome, ~50% of ncRNA genes found on non-coding nanochromosomes, ~80% specificity of the computational nanoclassifier, and ~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 homologs of known RNAs such as the spliceosomal RNAs made it to Stage 4 in our screen.

Although *Oxytricha*'s unusual macronuclear genome enables this screen, this unusualness itself also limits extrapolation of our results to other genomes. For example, our approach does not look for the possibility of ncRNA genes in the much larger *micronuclear* genome. 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) (37). 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. Although ciliates clearly utilize functional (nongenic) ncRNA transcripts extensively in DNA elimination and rearrangements (36,39–43), 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 (8,101). It should, however, bear on the question of whether there are large numbers of undiscovered ncRNA genes in eukaryotes in general.

Our study might also serve to illustrate some of the difficulties in distinguishing ncRNA 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 non-coding conserved nanochromosomes, 5/17 still appear to us upon manual analysis to contain at least fragments of conserved coding sequence. This multistep analysis might be contrasted to studies that have identified large numbers of putative 'non-coding' RNAs using simplistic definitions such as lack of ORF >100 amino acids (14,102), or finding cDNA transcripts that do not overlap with Ensembl gene predictions (103). We believe one reason that contributes to us finding so few ncRNA genes, whereas some other studies find so many, results from different standards in computational analysis of coding genes. Especially in light of our results here, we believe that 'non-coding' RNA loci in other organisms merit careful reexamination, as others have argued (19,21,28).

ACCESSION NUMBER

Stylonychia lemnae WGS sequence: DDBJ/EMBL/GenBank accession ADNZ 01000000.

ACKNOWLEDGEMENTS

Thanks to Mariuscz Nowacki for assistance with experiments at Princeton; to Franziska Jönsson and Hans Lipps for *Stylonychia lemnae* strain 2x8/2 and DNA; to the production and informatics staff at the Genome Center at Washington University for contract *Stylonychia* sequencing and assembly; and to the molecular biology and scientific computing cores at Janelia Farm, especially Goran Ceric, for superb technical assistance.

FUNDING

Howard Hughes Medical Institute; National Institutes of Health (grant number R01-HG01363); by an endowment from Alvin Goldfarb; and by a graduate student fellowship from the Samsung Scholarship Foundation. Funding for open access charge: HHMI Janelia Farm.

Conflict of interest statement. None declared.

REFERENCES

1. Hogg,J.R. and Collins,K. (2008) Structured non-coding RNAs and the RNP renaissance. *Curr. Opin. Chem. Biol.*, **12**, 684–689.
2. Nilsen,T.W. (2008) RNA 1997–2007: a remarkable decade of discovery. *Mol. Cell*, **28**, 715–720.
3. Wilusz,J.E., Sunwoo,H. and Spector,D.L. (2009) Long noncoding RNAs: Functional surprises from the RNA world. *Genes Dev.*, **23**, 1494–1504.
4. Bartel,D.P. (2004) MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell*, **116**, 281–297.
5. Waters,L.S. and Storz,G. (2009) Regulatory RNAs in bacteria. *Cell*, **136**, 615–628.
6. Tucker,B.J. and Breaker,R.R. (2005) Riboswitches as versatile gene control elements. *Curr. Opin. Struct. Biol.*, **15**, 342–348.
7. Farazi,T.A., Juranek,S.A. and Tuschl,T. (2008) The growing catalog of small RNAs and their association with distinct Argonaute/Piwi family members. *Development*, **135**, 1201–1214.
8. Mattick,J.S. (2001) Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Rep.*, **2**, 986–991.
9. Mattick,J.S. (2004) The hidden genetic program of complex organisms. *Sci. Am.*, **291**, 60–67.
10. Mattick,J.S. (2009) The genetic signatures of noncoding RNAs. *PLoS Genetics*, **5**, e1000459.
11. Cawley,S., Bekiranov,S., Ng,H.H., Kapranov,P., Sekinger,E.A., Kampa,D., Piccolboni,A., Sementchenko,V., Cheng,J., Williams,A.J. *et al.* (2004) Unbiased mapping of transcription factor binding sites along human chromosomes 21 and 22 points to widespread regulation of noncoding RNAs. *Cell*, **116**, 499–509.
12. Cheng,J., Kapranov,P., Drenkow,J., Dike,S., Brubaker,S., Patel,S., Long,J., Stern,D., Tammana,H., Helt,G. *et al.* (2005) Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution. *Science*, **308**, 1149–1154.
13. Guttman,M., Garber,M., Levin,J.Z., Donaghey,J., Robinson,J., Adiconis,X., Fan,L., Koziol,M.J., Gnirke,A., Nusbaum,C. *et al.* (2010) Ab initio reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nat. Biotechnol.*, **28**, 503–510.
14. Okazaki,Y., Furuno,M., Kasukawa,T., Adachi,J., Bono,H., Kondo,S., Nikaido,I., Osato,N., Saito,R., Suzuki,H. *et al.* (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature*, **420**, 563–573.
15. Rederstorff,M. and Hüttenhofer,A. (2011) cDNA library generation from ribonucleoprotein particles. *Nat. Protoc.*, **6**, 166–174.
16. Sharma,C.M., Hoffmann,S., Darfeuille,F., Reignier,J., Findeiss,S., Sittka,A., Chabas,S., Reiche,K., Hackermüller,J., Reinhardt,R.

et al. (2010) The primary transcriptome of the major human pathogen *Helicobacter pylori*. *Nature*, **464**, 250–255.
17. Ravasi,T., Suzuki,H., Pang,K.C., Katayama,S., Furuno,M., Okunishi,R., Fukuda,S., Ru,K., Frith,M.C., Gongora,M.M. *et al.* (2006) Experimental validation of the regulated expression of large numbers of non-coding RNAs from the mouse genome. *Genome Res.*, **16**, 11–19.
18. The ENCODE Project Consortium (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
19. Babak,T., Blencowe,B.J. and Hughes,T.R. (2005) A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics*, **6**, 104.
20. Nordström,K.J., Mirza,M.A., Almén,M.S., Gloriam,D.E., Fredriksson,R. and Schiöth,H.B. (2009) Critical evaluation of the FANTOM3 non-coding RNA transcripts. *Genomics*, **94**, 169–176.
21. van Bakel,H. and Hughes,T.R. (2009) Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genomic Proteomic.*, **8**, 424–436.
22. van Bakel,H. and Hughes,T.R. (2010) Most "dark matter" transcripts are associated with known genes. *PLoS Biol.*, **8**, e1000371.
23. Struhl,K. (2007) Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.*, **14**, 103–105.
24. Dinger,M.E., Pang,K.C., Mercer,T.R. and Mattick,J.S. (2008) Differentiating protein-coding and noncoding RNA: challenges and ambiguities. *PLoS Comput Biol.*, **4**, e1000176.
25. Pedersen,J.S., Bejerano,G., Siepel,A., Rosenbloom,K., Lindblad-Toh,K., Lander,E.S., Kent,J., Miller,W. and Haussler,D. (2006) Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.*, **2**, e33.
26. Rivas,E. and Eddy,S.R. (2001) Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics*, **2**, 8.
27. Washietl,S., Hofacker,I.L., Lukasser,M., Hüttenhofer,A. and Stadler,P.F. (2005) Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.*, **23**, 1383–1390.
28. Babak,T., Blencowe,B.J. and Hughes,T.R. (2007) Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics*, **8**, 33.
29. Carninci,P. (2010) RNA dust: where are the genes? *DNA Res.*, **17**, 51–59.
30. Kim,T.K., Hemberg,M., Gray,J.M., Costa,A.M., Bear,D.M., Wu,J., Harmin,D.A., Laptewicz,M., Barbara-Haley,K., Kuersten,S. *et al.* (2010) Widespread transcription at neuronal activity-regulated enhancers. *Nature*, **465**, 182–187.
31. Ammermann,D., Steinbrück,G., von Berger,L. and Hennig,W. (1974) The development of the macronucleus in the ciliated protozoan *Stylonychia mytilus*. *Chromosoma*, **45**, 401–429.
32. Jahn,C.L. and Klobutcher,L.A. (2002) Genome remodeling in ciliated protozoa. *Annu. Rev. Microbiol.*, **56**, 489–520.
33. Prescott,D.M. (1994) The DNA of ciliated protozoa. *Microbiol. Rev.*, **58**, 233–267.
34. Jönsson,F., Postberg,J. and Lipps,H.J. (2009) The unusual way to make a genetically active nucleus. *DNA Cell Biol.*, **28**, 71–78.
35. Juranek,S.A. and Lipps,H.J. (2007) New insights into the macronuclear development in ciliates. *Int. Rev. Cytol.*, **262**, 219–251.
36. Nowacki,M., Vijayan,V., Zhou,Y., Schotanus,K., Doak,T.G. and Landweber,L.F. (2008) RNA-mediated epigenetic programming of a genome-rearrangement pathway. *Nature*, **451**, 153–158.
37. Nowacki,M., Higgins,B.P., Maquilan,G.M., Swart,E.C., Doak,T.G. and Landweber,L.F. (2009) A functional role for transposases in a large eukaryotic genome. *Science*, **324**, 935–938.
38. Prescott,D.M. (2000) Genome gymnastics: unique modes of DNA evolution and processing in ciliates. *Nat. Rev. Genet.*, **1**, 191–198.
39. Chalker,D.L. and Yao,M.C. (2001) Nongenic, bidirectional transcription precedes and may promote developmental DNA deletion in *Tetrahymena thermophila*. *Genes Dev.*, **15**, 1287–1298.
40. Chalker,D.L., Fuller,P. and Yao,M.C. (2005) Communication between parental and developing genomes during *Tetrahymena* nuclear differentiation is likely mediated by homologous RNAs. *Genetics*, **169**, 149–160.
41. Garnier,O., Serrano,V., Duharcourt,S. and Meyer,E. (2004) RNA-mediated programming of developmental genome rearrangements in *Paramecium tetraurelia*. *Mol. Cell. Biol.*, **24**, 7370–7379.
42. Kurth,H.M. and Mochizuki,K. (2009) Non-coding RNA: a bridge between small RNA and DNA. *RNA Biol.*, **6**, 138–140.
43. Mochizuki,K., Fine,N.A., Fujisawa,T. and Gorovsky,M.A. (2002) Analysis of a piwi-related gene implicates small RNAs in genome rearrangement in *Tetrahymena*. *Cell*, **110**, 689–699.
44. Doak,T.G., Cavalcanti,A.R., Stover,N.A., Dunn,D.M., Weiss,R., Herrick,G. and Landweber,L.F. (2003) Sequencing the *Oxytricha trifallax* macronuclear genome: a pilot project. *Trends Genet.*, **19**, 603–607.
45. Riley,J.L. and Katz,L.A. (2001) Widespread distribution of extensive chromosomal fragmentation in ciliates. *Mol. Biol. Evol.*, **18**, 1372–1377.
46. Foissner,W. and Berger,H. (1999) 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.
47. Lauth,M.R., Spear,B.B., Heumann,J. and Prescott,D.M. (1976) DNA of ciliated protozoa: DNA sequence diminution during macronuclear development of *Oxytricha*. *Cell*, **7**, 67–74.
48. Swanton,M.T., Heumann,J.M. and Prescott,D.M. (1980) 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.
49. Cavalcanti,A.R., Stover,N.A., Orecchia,L., Doak,T.G. and Landweber,L.F. (2004) Coding properties of *Oxytricha trifallax* (*Sterkiella histriomuscorum*) macronuclear chromosomes: analysis of a pilot genome project. *Chromosoma*, **113**, 69–76.
50. Prescott,D.M., Prescott,J.D. and Prescott,R.M. (2002) Coding properties of macronuclear DNA molecules in *Sterkiella nova* (*Oxytricha nova*). *Protist*, **153**, 71–77.
51. Cavalcanti,A.R., Dunn,D.M., Weiss,R., Herrick,G., Landweber,L.F. and Doak,T.G. (2004) Sequence features of *Oxytricha trifallax* (class Spirotrichea) macronuclear telomeric and subtelomeric sequences. *Protist*, **155**, 311–322.
52. Huang,X., Wang,J., Aluru,S., Yang,S.-P. and Hillier,L. (2003) PCAP: a whole-genome assembly program. *Genome Res.*, **13**, 2164–2170.
53. Eddy,S.R. (2005) A model of the statistical power of comparative genome sequence analysis. *PLoS Biol.*, **3**, e10.
54. Parra,G., Bradnam,K., Ning,Z., Keane,T. and Korf,I. (2009) Assessing the gene space in draft genomes. *Nucleic Acids Res.*, **37**, 289–297.
55. Parra,G., Bradnam,K. and Korf,I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.
56. Stephenson,F.H. (2003) *Calculations for Molecular Biology and Biotechnology: A Guide to Mathematics in the Laboratory*. Academic Press, San Diego, CA.
57. Slater,G.S. and Birney,E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, **6**, 31.
58. Nawrocki,E.P., Kolbe,D.L. and Eddy,S.R. (2009) Infernal 1.0: inference of RNA alignments. *Bioinformatics*, **25**, 1335–1337.
59. Gardner,P.P., Daub,J., Tate,J.G., Nawrocki,E.P., Kolbe,D.L., Lindgreen,S., Wilkinson,A.C., Finn,R.D., Griffiths-Jones,S., Eddy,S.R. *et al.* (2009) Rfam: Updates to the RNA families database. *Nucl. Acids Res.*, **37**, D136–D140.
60. Williams,K., Doak,T.G. and Herrick,G. (1993) 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.
61. Allen,J.E., Majoros,W.H. and Salzberg,S.L. (2006) JIGSAW, GeneZilla, and GlimmerHMM: puzzling out the features of human genes in ENCODE regions. *Genome Biol.*, **7**, S9.1–S9.13.
62. Majoros,W.H., Pertea,M., Antonescu,C. and Salzberg,S.L. (2003) GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders. *Nucleic Acids Res.*, **31**, 3601–3604.

63. Parra,G., Blanco,E. and Guigó,R. (2000) GeneID in *Drosophila*. *Genome Research*, **10**, 511–515.
64. Stanke,M. and Waack,S. (2003) Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, **19(Suppl. 2)**, ii215–ii225.
65. Edgar,R.C. (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics*, **5**, 113.
66. Larkin,M.A., Blackshields,G., Brown,N.P., Chenna,R., McGettigan,P.A., McWilliam,H., Valentin,F., Wallace,I.M., Wilm,A., Lopez,R. *et al*. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, **23**, 2947–2948.
67. Griffiths-Jones,S. (2005) RALEE–RNA ALignment editor in Emacs. *Bioinformatics*, **21**, 257–259.
68. Schattner,P., Decatur,W.A., Davis,C.A., Ares,M., Fournier,M.J. and Lowe,T.M. (2004) Genome-wide searching for pseudouridylation guide snoRNAs: analysis of the *Saccharomyces cerevisiae* genome. *Nucleic Acids Res*., **32**, 4281–4296.
69. Lowe,T.M. and Eddy,S.R. (1999) A computational screen for methylation guide snoRNAs in yeast. *Science*, **283**, 1168–1171.
70. Crooks,G.E., Hon,G., Chandonia,J.M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res*., **14**, 1188–1190.
71. Steinbrück,G. (1983) Overamplification of genes in macronuclei of hypotrichous ciliates. *Chromosoma*, **88**, 156–163.
72. Baird,S.E. and Klobutcher,L.A. (1991) Differential DNA amplification and copy number control in the hypotrichous ciliate *Euplotes crassus*. *J. Protozool*., **38**, 136–140.
73. Heyse,G., Jönsson,F., Chang,W.-J. and Lipps,H.J. (2010) RNA-dependent control of gene amplification. *Proc. Natl Acad. Sci. USA*, **107**, 22134–22139.
74. Harper,D.S., Song,K. and Jahn,C.L. (1991) Overamplification of macronuclear linear DNA molecules during prolonged vegetative growth of *Oxytricha nova*. *Gene*, **99**, 55–61.
75. Steinbrück,G., Haas,I., Hellmer,K.-H. and Ammermann,D. (1981) Characterization of macronuclear DNA in five species of ciliates. *Chromosoma*, **83**, 199–208.
76. Lozupone,C.A., Knight,R.D. and Landweber,L.F. (2001) The molecular basis of nuclear genetic code change in ciliates. *Curr. Biol*., **11**, 65–74.
77. Eisen,J.A., Coyne,R.S., Wu,M., Wu,D., Thiagarajan,M., Wortman,J.R., Badger,J.H., Ren,Q., Amedeo,P. and Jones,K.M. (2006) Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol*., **4**, e286.
78. Aury,J.M., Jaillon,O., Duret,L., Noel,B., Jubin,C., Porcel,B.M., Ségurens,B., Daubin,V., Anthouard,V., Aiach,N. *et al*. (2006) Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature*, **444**, 171–178.
79. Durbin,R., Eddy,S.R., Krogh,A. and Mitchison,G.J. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge UK.
80. Zhang,M.Q. (2002) Computational prediction of eukaryotic protein-coding genes. *Nat. Rev. Genet*., **3**, 698–709.
81. Prescott,D.M. and Dizick,S.J. (2000) A unique pattern of intrastrand anomalies in base composition of the DNA in hypotrichs. *Nucleic Acids Res*., **28**, 4679–4688.
82. Chang,W.J., Stover,N.A., Addis,V.M. and Landweber,L.F. (2004) A micronuclear locus containing three protein-coding genes remains linked during macronuclear developmentin the spirotrichous ciliate *Holosticha*. *Protist*, **155**, 245–255.
83. Seegmiller,A., Williams,K.R., Hammersmith,R.L., Doak,T.G., Witherspoon,D., Messick,T., Storjohann,L.L. and Herrick,G. (1996) Internal eliminated sequences interrupting the *Oxytricha* 81 locus: allelic divergence, conservation, conversions, and possible transposon origins. *Mol. Biol. Evol*., **13**, 1351–1362.
84. Seegmiller,A., Williams,K.R. and Herrick,G. (1997) 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.
85. Will,C.L. and Lührmann,R. (2005) Splicing of a rare class of introns by the U12-dependent spliceosome. *Biol. Chem*., **386**, 713–724.
86. Stricklin,S.L., Griffiths-Jones,S. and Eddy,S.R. (2005) *C. elegans* noncoding RNA genes, In WormBook (ed.) *The* C. elegans *Research Community*. http://www.wormbook.org pdf (15 June 2011, date last accessed).
87. Thomas,J., Lea,K., Zucker-Aprison,E. and Blumenthal,T. (1990) The spliceosomal snRNAs of *Caenorhabditis elegans*. *Nucl. Acids Res*., **18**, 2633–2642.
88. Hargrove,B.W., Bhattacharyya,A., Domitrovich,A.M., Kapler,G.M., Kirk,K., Shippen,D.E. and Kunkel,G.R. (1999) Identification of an essential proximal sequence element in the promoter of the telomerase RNA gene of *Tetrahymena thermophila*. *Nucl. Acids Res*., **27**, 4269–4275.
89. Egloff,S., O'Reilly,D. and Murphy,S. (2008) Expression of human snRNA genes from beginning to end. *Biochem. Soc. Trans*., **36**, 590–594.
90. Kiss,T. (2002) Small nucleolar RNAs: an abundant group of noncoding RNAs with diverse cellular functions. *Cell*, **109**, 145–148.
91. Reichow,S.L., Hamma,T., Ferré-D'Amaré,A.R. and Varani,G. (2007) The structure and function of small nucleolar ribonucleoproteins. *Nucleic Acids Res*., **35**, 1452–1464.
92. Ganot,P., Caizergues-Ferrer,M. and Kiss,T. (1997) 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.
93. Atzorn,V., Fragapane,P. and Kiss,T. (2004) U17/snR30 is a ubiquitous snoRNA with two conserved sequence motifs essential for 18S rRNA production. *Mol. Cell. Biol*., **24**, 1769–1778.
94. Eliceiri,G.L. (2006) The vertebrate E1/U17 small nucleolar ribonucleoprotein particle. *J. Cell. Biochem*., **98**, 486–495.
95. Fayet-Lebaron,E., Atzorn,V., Henry,Y. and Kiss,T. (2009) 18S rRNA processing requires base pairings of snR30 H/ACA snoRNA to eukaryote-specific 18S sequences. *EMBO J*., **28**, 1260–1270.
96. Eisen,J.A., Coyne,R.S., Wu,M., Wu,D., Thiagarajan,M., Wortman,J.R., Badger,J.H., Ren,Q., Amedeo,P. and Jones,K.M. (2006) Macronuclear genome sequence of the ciliate *Tetrahymena thermophila*, a model eukaryote. *PLoS Biol*., **4**, e286.
97. Arnaiz,O., Cain,S., Cohen,J. and Sperling,L. (2007) ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. *Nucleic Acids Res*., **35**, D439–D444.
98. Ricard,G., de Graaf,R.M., Dutilh,B.E., Duarte,I., van Alen,T.A., van Hoek,A.H., Boxma,B., van der Staay,G.W., van der Staay,S.Y.M., Chang,W.J. *et al*. (2008) Macronuclear genome structure of the ciliate *Nyctotherus ovalis*: Single-gene chromosomes and tiny introns. *BMC Genomics*, **9**, 587.
99. Chen,C.-L., Zhou,H., Liao,J.-Y., Qu,L.-H. and Amar,L. (2009) Genome-wide evolutionary analysis of the noncoding RNA genes and noncoding DNA of *Paramecium tetraurelia*. *RNA*, **15**, 503–514.
100. Guttman,M., Amit,I., Garber,M., French,C., Lin,M.F., Feldser,D., Huarte,M., Zuk,O., Carey,B.W., Cassady,J.P. *et al*. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. *Nature*, **458**, 223–227.
101. Mattick,J.S. (2003) Challenging the dogma: the hidden layer of non-protein-coding RNAs in complex organisms. *Bioessays*, **25**, 930–939.
102. Numata,K., Kanai,A., Saito,R., Kondo,S., Adachi,J., Wilming,L.G., Hume,D.A., Hayashizaki,Y., Tomita,M.RIKEN GER Group *et al*. (2003) Identification of putative noncoding RNAs among the RIKEN mouse full-length cDNA collection. *Genome Res*., **13**, 1301–1306.
103. Warren,W.C., Clayton,D.F., Ellegren,H., Arnold,A.P., Hillier,L.W., Künstner,A., Searle,S., White,S., Vilella,A.J., Fairley,S. *et al*. (2010) The genome of a songbird. *Nature*, **464**, 757–762.
104. Bernhart,S.H., Hofacker,I.L., Will,S., Gruber,A.R. and Stadler,P.F. (2008) RNAalifold: improved consensus structure prediction for RNA alignments. *BMC Bioinformatics*, **9**, 474.