# Computational analysis of conserved RNA secondary structure in transcriptomes and genomes

Sean R. Eddy

HHMI Janelia Farm Research Campus

19700 Helix Drive

Ashburn VA 20147 USA

eddys@janelia.hhmi.org

August 30, 2013

# Keywords

noncoding RNA, lncRNA, probing, SHAPE, statistical inference



# Abstract

Transcriptomics experiments and computational predictions both enable systematic discovery of new functional RNAs, but many putative noncoding transcripts arise instead from artifacts and biological noise, and current computational prediction methods have high false positive rates. I discuss prospects for improving computational methods for analyzing and identifying functional RNAs, with a focus on detecting signatures of conserved RNA secondary structure. An interesting new front is the application of chemical and enzymatic RNA structure probing experiments on a transcriptome-wide scale. I review several proposed approaches for incorporating structure probing data into computational RNA secondary structure prediction. Using probabilistic inference formalisms, I show how all these approaches can be unified in a well-principled framework. Using that framework, RNA probing data can easily be integrated into a wide range of different analyses that depend on RNA secondary structure inference, including homology search and genome-wide detection of new structural RNAs.

# Contents

INTRODUCTION ..... 4

HETEROGENEITY OF RNA FUNCTION AND BIOGENESIS ..... 7

TRANSCRIPTOMICS APPROACHES TO SYSTEMATIC ncRNA DISCOVERY ..... 8

COMPUTATIONAL CONSERVED RNA STRUCTURE DETECTION ..... 13

- Development of computational RNA structure detection methods ..... 13
- Current methods remain insufficiently reliable ..... 14
- Empirical FDRs are vulnerable to choice of negative controls ..... 16

PROBING-DIRECTED RNA STRUCTURE PREDICTION ..... 17

- Single sequence RNA structure prediction ..... 18
- SHAPE chemistry ..... 19
- SHAPE data analysis from a likelihood ratio perspective ..... 20
- Deigan's pseudoenergy approach ..... 20
- Sample and select approaches ..... 23
- Zarringhalam's pseudoenergy approach ..... 23
- Washietl's ensemble approach ..... 24

STATISTICAL INFERENCE FOR PROBING-DIRECTED STRUCTURE PREDICTION ..... 27

- Optimal structure prediction and a derivation of pseudoenergies ..... 28
- Ensemble prediction ..... 30

CONCLUSION ..... 31

# INTRODUCTION

Some of the most important and controversial questions in molecular biology and genomics today are about the biological functions of RNA (9, 36, 76, 107–109). Advances in sequencing technology have made it possible to survey RNA transcript populations comprehensively using cDNA sequencing (68), tiled microarrays (35), and now RNA-seq (62). As technology has become more sensitive, a large number of putatively noncoding RNA species have been detected, and the apparent complexity of RNA transcript populations has grown (16, 20, 30).

There are two fundamentally opposed views of this growing complexity. One view is that it indicates a vast unappreciated repertoire of functional noncoding RNAs (9, 15). Another view is that many supposed noncoding transcripts are the result of experimental artifacts, analysis errors, and transcriptional noise (4, 99, 108). On the one hand, the repertoire of functions for RNA certainly continues to expand, for noncoding RNA transcripts (2, 39, 40, 83, 118), for cis-regulatory RNA sequences in messenger RNAs (32, 42, 44, 80, 84, 92), and for catalytic RNAs (74, 90). On the other hand, high-throughput experiments and computational analysis pipelines have been found to suffer from serious systematic artifacts (65, 108, 132), and RNA biogenesis, like any biochemical process, must have some background level of infidelity (75, 99). The question is not whether all newly discovered RNA transcripts are functional or not. The question is, for any one of them, how to tell the difference.

Computational RNA sequence analysis methods are at the crux of addressing these questions, if only because the datasets are large. Historically, RNA computational analysis methods assume that the RNA to be analyzed is already known to be functional, and that an RNA secondary structure is involved. RNA secondary structure prediction (67, 134, 135), structure-guided sequence alignment (18, 54, 120, 124), and database similarity searching with RNA sequence/structure consensus models (8, 18, 49, 89) are examples of these classic computational problems. Now, it has also become important to be able to judge whether or not an RNA sequence is likely to have a

biological function, and whether or not it has an RNA secondary structure that plays a role in that function. For example, signatures of evolutionary sequence conservation help distinguish functional RNAs from transcriptional noise, and signatures of RNA secondary structure conservation help distinguish functional RNA sequence elements from functional RNA structural elements.

A class of computational RNA analysis methods has been developed that seeks to identify novel structural RNAs in genome sequences (6, 11, 13, 26, 72, 86, 102, 106, 113, 115, 123, 129). Structural RNA detection methods work by looking for evolutionarily conserved RNA secondary structure, by comparative analysis of patterns of covariation in homologous genome sequence alignments. As a result they detect structural RNAs, including both structural noncoding RNA genes and cis-regulatory RNA structures, but they do not detect functional RNAs that act as linear sequences.

Rather than helping to clarify the results coming from systematic transcriptomics, computational methods for structural RNA detection have sown a parallel line of confusion. These methods have been used to predict hundreds, thousands, even millions of novel structural ncRNAs, especially in large mammalian genomes (71, 72, 79, 91, 95, 114, 117, 123). With computational predictions and experimental transcriptomics both producing lots of candidate noncoding RNAs, this has sometimes been seen as independent confirmation of the existence of a vast hidden complexity of functional RNA, but the computational approaches are subject to their own list of potential artifacts.

The problem with computational RNA structure detection approaches is that they are unreliable (5). Their signal/noise ratio is poor and they are being used at a perilously ragged edge of statistical significance. Because of difficulties in establishing appropriate negative controls – adequately realistic homologous multiple genome alignments that are known *not* to be functional structural RNA – there are large uncertainties in calculating statistical significance. Small errors, well within these uncertainties, could erase the majority of the predictions. These methods need to be improved.

What might dramatically improve these methods is to identify new data sources that could be incorporated into genome-wide sequence analyses to increase the detectable signal for structural RNAs. One such data source has begun to look feasible. There is renewed interest in using chemical modification and enzymatic cleavage experiments to probe RNA secondary structure, both using well established reagents (such as RNases and dimethyl sulfate) (10, 38, 60, 77) and especially a powerful new class of reagents called SHAPE chemistry (59, 119). RNA structure probing experiments have been coupled to high throughput sequencing readouts, allowing these approaches to be applied at scale to probe many RNAs in parallel, including transcriptome-wide structure probing (37, 43, 48, 105, 133).

Structure probing data are noisy and statistical in nature, an informative but ambiguous signal of RNA structure. Several computational methods have been proposed already for incorporating probing data into single-sequence RNA secondary structure prediction methods (12, 29, 52, 53, 69, 77, 116, 131). As yet, it remains unclear which of these approaches is most powerful, most principled, or most generalizable to more complex problems than single sequence structure prediction.

In what follows, I expand on the above themes, and I end by showing how all the existing approaches for incorporating RNA structure probing data into RNA structure prediction can be viewed from a unified statistical inference perspective. This perspective suggests how RNA probing data can be naturally incorporated into all other classes of computational RNA analysis methods that depend on RNA secondary structure inference, including *de novo* genome-wide structure detection and homology search.

# HETEROGENEITY OF RNA FUNCTION AND BIOGENESIS

It is necessary to appreciate the extreme heterogeneity of RNA functions to understand the limitations of functional RNA discovery and analysis methods. RNAs can fold into complex three-dimensional structures. They can present sequence or structure motifs for binding regulatory macromolecules. They can use complementary base-pairing of linear sequence to recognize other nucleic acid sequences with exquisite specificity and efficiency. They can use complementarity to template nucleic acid synthesis. The act of transcription itself may have a function, rather than the RNA that is produced (55).

Different functional RNAs combine and deploy these modalities in a variety of different ways (28, 82). RNAs can serve as informational *messages*, as in protein-coding messenger RNAs. RNAs can act as structural and catalytic *machines*, much as protein enzymes and protein complexes do, as in ribosomal RNAs. RNAs can act as *scaffolds*, deploying a set of protein-binding motifs (either linear or structural) to facilitate assembly of a multiprotein complex, as in signal recognition particle RNA (73) or telomerase RNA (130). RNAs can act as *templates* for complementary RNA or DNA synthesis, as in the core of a telomerase RNA. RNAs can act as complementary *guides*, targeting a shared protein machine to a large number of different specific nucleic acid targets, like the small nucleolar guide RNAs that direct specific 2'-O-ribose methylations and pseudouridylations of other RNAs (81). Cis-regulatory RNA motifs act as post-transcriptional *signals* and *switches* (78), with roles in essentially every imaginable step of RNA biogenesis and trafficking, ranging from small linear sequence targets of an RNA-binding protein (80, 93) to complex RNA machines like riboswitches (92).

RNA biogenesis is also heterogeneous (51). Noncoding RNA genes may be transcribed by RNA polymerase I, II, or III, often as larger precursor transcripts that undergo trimming and pro-

cessing. Some functional RNAs are generated by processing of pre-mRNAs, including many intron-encoded miRNAs and snoRNAs (104, 121). Functional noncoding RNA transcripts are often not 5' capped and 3' polyA+ like mRNAs, but exhibit a variety of different types of 5' and 3' ends, including circular RNAs with no ends at all (58). A functional RNA can range anywhere in size from a 4-10nt protein-binding cis-regulatory site or a 20-25nt microRNA transcript to a large RNA catalyst or scaffold of several thousand nucleotides.

There is no such thing as an unbiased screen for functional RNAs. No one characteristic signal discriminates functional RNA from other sequences. Functional RNAs may have a conserved secondary and tertiary structure, but many do not, because a linear RNA sequence can function as a message, a guide, a scaffold, a template, or a signal. Functional RNAs may be independent transcripts arising from noncoding RNA genes, discoverable by transcriptomics, but at least as important (and still relatively understudied relative to transcriptional regulatory signals) are the roles of cis-regulatory RNA sequences in posttranscriptional gene regulation.

# TRANSCRIPTOMICS APPROACHES TO SYSTEMATIC ncRNA DISCOVERY

Powerful experimental transcriptomics approaches (35, 62, 68) have resulted in descriptions of large numbers of putative noncoding RNA transcripts, especially what are called long noncoding RNAs (lncRNAs). lncRNAs are loosely defined as apparently noncoding mRNA-like transcripts (5' capped, 3' polyA+ or polyA-, transcribed by RNA polymerase II) at least 200 nucleotides long.

A small number of lncRNAs have known functions. One of the best studied is Xist, a very large (19 kb) ncRNA transcript that triggers the heterochromatization (Barr body formation) of one of the two X chromosomes in females. Among recent lncRNA discoveries, one of the best studied is HOTAIR, a $\sim$2kb RNA in the HOXC cluster that apparently regulates transcription of loci in the HOXD cluster in trans by a mechanism having to do with chromatin modifications and

association with both the PRC2 histone H3K27 methylation complex and the coREST complex (83, 103). Like Xist and HOTAIR, there is evidence for involvement of several other lncRNAs in chromatin modification. A few other lncRNAs have been proposed to have other functions.

For the most part, well studied lncRNAs like Xist, HOTAIR, and others such as MALAT1 (125) show substantial circumstantial evidence for their functionality, even leaving aside the detailed experimental studies that have been focused on them individually. They are highly expressed and nuclear localized. They contain evolutionarily conserved sequence regions. They are unique sequences, devoid or highly depleted of the transposable element (TE) remnants that are so abundant elsewhere in the human genome.

In contrast, in the large catalogs of lncRNAs that have yet to be experimentally characterized, most lack these expected characteristics of functional RNAs. For example, in one recent meta-analysis of 127 human RNA-seq libraries (30) – notable for the thoroughness of its data availability, which allowed me to reanalyze the work in considerable depth – Hangauer et al. identified 53,864 lncRNA loci expressed above a chosen expression level threshold. I have replotted two key observations from their paper in Figure 1 in order to make two points.

First, we must distinguish genome coverage from expression level when discussing pervasive transcription (9, 20, 108, 109) and lncRNAs. Figure 1A illustrates how most (here 67%) of the genome is detectable in cellular RNA (9), but only if we look at very low expression levels relative to mRNA transcripts. At expression levels more typical of known and annotated genes (coding and ncRNA both), only a small fraction of the genome is covered (108). For example, in Figure 1A, 85% of coding mRNA exons are covered at a read depth of at least 100, whereas only 5% of the genome and only 3% of the lncRNAs are covered at the same threshold. Functionally characterized lncRNAs are unlike the bulk of the lncRNA distribution because they tend to be expressed at levels comparable or even higher than coding mRNAs (look at H19, NEAT1, and MALAT1 for example).

Second, Figure 1B shows how most cataloged lncRNAs show sequence divergence comparable to repeat sequences which are assumed to be nonconserved and neutrally evolving. Only a small

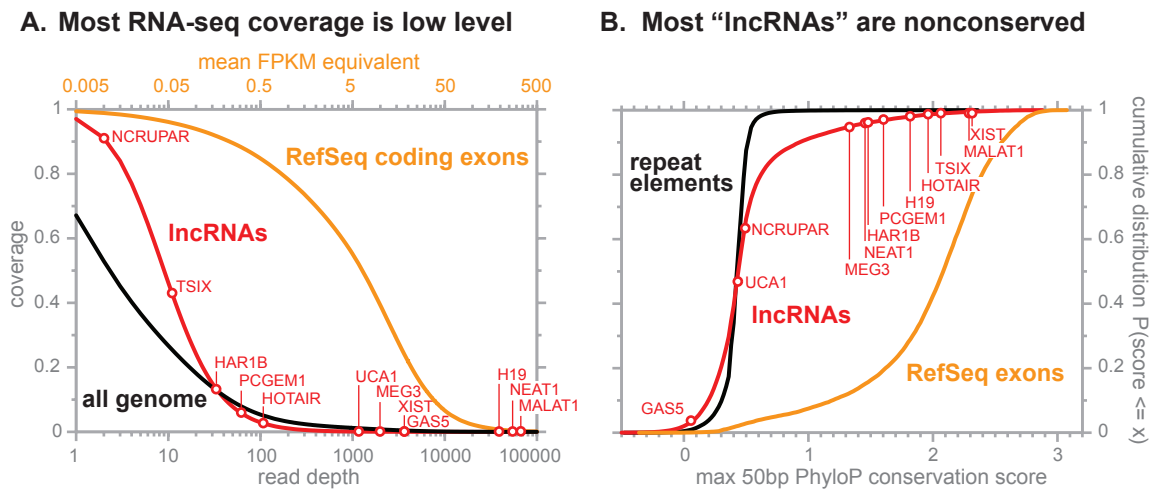


Figure 1: **Two key observations about pervasive transcription and lncRNA catalogs.** (30). Black line: fraction of the uniquely mappable genome (2570 Mb; hg18) covered at different thresholds of minimum read depth per genome position. Red line: coverage of the sequence of 53,864 lncRNAs (38 Mb). Orange line: coverage of the sequence of 364,265 coding exons of 34,978 RefSeq coding genes (34 Mb). Red circles placed along the lncRNA distribution mark the median read depth over 12 functionally characterized lncRNAs (y-axis position for these points has no meaning). Read depth units from (30) are roughly convertible to mean FPKM units (fragments per kilobase per million mapped reads; mean over 127 RNA-seq libraries), FPKM = 1000 * depth / (read length per frag) / (millions of fragments), where the aggregate dataset has 3.39 billion fragments with a mean read length per fragment of 60nt (x-axis labels at top). Data reanalyzed and replotted with permission from Hangauer and colleagues (Supplementary Figure S1 and Dataset S8 (30), plus a BED file of read depth coverage per genome position provided by M. Hangauer.) B. Cumulative distribution of sequence conservation (maximum in 50bp windows, as measured by PhyloP in a placental mammalian genome alignment) for human repeat elements (presumed to be neutrally evolving), exons of 31,204 RefSeq coding genes, and 53,864 lncRNAs defined by (30). Conservation values for 12 characterized functional lncRNAs are marked with circles, placed along the lncRNA cumulative distribution (y-axis position for these has no meaning). Figure redrawn with permission from the same data in Figure 3C in (30).

subset shows segments of sequence conservation, including most of the well-characterized functional lncRNAs. For example, only 7% of RefSeq exons fall below a threshold conservation score of 1 in Figure 1B, whereas 99.9% of repeats and 91% of lncRNAs are below this threshold. GAS5 is an exception proving the rule: GAS5 is an inside-out snoRNA carrier gene, whose function is to have conserved introns processed into snoRNAs (94). Moreover, according to my analysis with RepeatMasker, 56% of the sequence of these 53,864 lncRNAs consists of TE remnants, essentially indistinguishable from genomic background (53%).

An RNA does not necessarily have to be expressed at levels comparable to known mRNAs, nor evolutionarily conserved, nor devoid of TE remnants to be functional (15). However, there are other more likely explanations for low-level nonconserved transcripts with TE content similar to genomic background.

One source of "lncRNAs" is transcriptional noise (75, 99). Some authors have taken transcriptional noise to mean random transcription, a uniform haze across the genome (76), so that observing that a "lncRNA" is expressed in a tissue-specific manner has been taken as evidence of functionality (30). However, the neutral expectation is that cryptic RNA transcription and processing are driven by randomly occurring (specific and discrete, but cryptic) short binding sites for regulatory proteins. Expression patterns of these discrete cryptic transcripts will follow the specific spatiotemporal expression patterns of the regulatory proteins that activate them (17).

Other sources of "lncRNAs" are computational analysis errors, including failures to recognize predictable experimental artifacts. Half of the "noncoding RNAs" in a pioneering paper on lncRNAs from the FANTOM3 project (68) are cloning artifacts that arose by internal priming on poly-A tracts in pre-mRNA (65). False transcribed regions are created by cross-hybridization artifacts on genome tiling arrays (108) and by mismapping of RNA-seq reads (even uniquely mapped reads) (132).

Even defining a transcript as noncoding is surprisingly difficult (34). Many real proteins are shorter than the typical ORF length cutoff of $\geq$ 100aa for defining noncoding RNA (31). More

powerful methods for recognizing coding genes using comparative sequence analysis are often used (45, 112), but they are often trained and their accuracy evaluated on complete sequences of normal proteins, rather than on mRNAs expected to contaminate a lncRNA catalog, an extreme tail of the coding mRNA distribution enriched for short coding genes and partial transcript sequences. Even basic rules for defining noncoding RNA have proven inexplicably difficult to apply. The FANTOM3 bioinformatics pipeline failed to recognize that 27% of the "ncRNAs" contain an ORF of $\geq$ 100aa and 25% have a BLASTP similarity to the protein database of E $\leq 10^{-10}$ (65), even though these criteria were in the FANTOM3 analysis of coding potential.

Putative lncRNAs need to be treated as heterogeneous, not as a class. Only a subset are likely to be functional RNAs, and with a variety of functions. Careful computational analyses can help prioritize and sort them into different categories. Improved computational tools of all sorts will help these analyses – including read mappers with lower false positive mapping rates, spliced transcript assemblers that assemble longer and more complete RNA transcripts from RNA-seq data, more quantitative measurements of sequence conservation and evolutionary constraint, more powerful methods for detecting small coding regions, and better methods for detecting homology between RNA sequences.

Often these computational analyses are subtractive, looking for positive signals of something else (coding regions, experimental artifact), to winnow down a lncRNA candidate set and enrich for functional RNAs. One of the more interesting areas to me is the development of methods for detecting evolutionarily conserved RNA secondary structure, because this is one of the few affirmative signals we can look for in a functional RNA.

# COMPUTATIONAL CONSERVED RNA STRUCTURE DETECTION

An evolutionarily conserved RNA secondary structure might be the most general feature shared by many functional RNAs. Obviously, a drawback of using conserved structure as a detection strategy is that functional RNAs that act primarily by their linear sequence will be missed. In lncRNAs, even the best studied ones, it remains somewhat unclear whether there is much conserved RNA structure. RNA secondary structures have been proposed for parts of HOTAIR (37, 103), parts of Xist (50), and for other lncRNAs (66), and MALAT1 clearly has a fascinating tRNA-like structure at its 3' end (125, 126). But lncRNAs that act as scaffolds, for example for chromatin modification complexes, could well bind those complexes via single-stranded RNA sequence motifs. Nonetheless, computational detection of conserved secondary structure is a useful signal to positively identify at least a subset of functional RNAs against a background of other less interesting explanations, and an advantage (over transcriptomics, for example) is that cis-regulatory RNA structures in mRNAs can be detected too.

## Development of computational RNA structure detection methods

The first attempts to make a general genome-wide approach to detecting RNA structures looked for regions of a single sequence predicted to fold into more thermodynamically stable RNA structures than expected (41). However, random RNA sequences fold into secondary structures with predicted thermodynamic stabilities similar to functional RNAs, so this approach was deemed insufficiently powerful for genome-wide screens (85). At best about 30% of structural RNAs could be detected at an estimated false positive rate of about 10 per megabase of genome screened (85).

Attention moved to exploiting evolutionary conservation of RNA secondary structure in homologous sequence alignments as an additional source of signal to discriminate real functional

RNAs from background, with pairwise (86) or with multiple alignments (11, 13). The 2001 Rivas QRNA algorithm was estimated to detect about 80% of structural RNAs at an estimated false positive rate of about 20 per megabase, which made a screen of the small *E. coli* genome feasible (87).

The general idea of detecting conserved RNA structure in multiple sequence (or multiple genome) alignments has now been extended and implemented in many ways, in RNAz (115), EvoFold (72), CMfinder (129), FOLDALIGN (102), and other approaches (6, 26, 106, 113, 123). These programs have been used to predict structural RNA regions in large eukaryotic genomes, especially in the human genome (71, 72, 79, 91, 95, 114, 117, 123). In one recent screen of the human genome, for example, Smith et al. predicted 4.1 million structural RNAs in the human genome, at an estimated sensitivity of about 30%, and an estimated false positive rate of about 170 per megabase. This was described as a "historically low" false positive rate.

In fact, the stringency demanded from these approaches has declined while the ambition to screen large mammalian genomes has increased. Moreover, there is substantial uncertainty in how false positive rates are estimated, either by shuffling or simulating negative multiple alignments. My laboratory abandoned attempts to extend QRNA screens to large genomes when we found that our rate of experimental confirmation of the expression of predicted intergenic RNA loci was far lower than the computationally predicted false positive rate (98). The Hughes laboratory reached the same conclusion in their experimental followup of a QRNA screen of the mouse genome (4). The current false positive rates from this class of methods remain too high to justify their use on large genomes (5).

## Current methods remain insufficiently reliable

Consider the recent computational screen by Smith et al. as a specific example (95). These authors applied two different approaches, RNAz 2.0 (27) and a new method called SISSIz (24), to the human genome, by comparative analysis of a multiple alignment of 35 mammalian genomes. RNAz

and SISSIz, like all methods in this class, work by scoring one small alignment window at a time (here 200 nucleotides) under a model that looks for RNA structure conservation, and classifying it as a structural RNA prediction if it passes a chosen score threshold. The whole genome alignment is scored in overlapping windows (in this case, 200nt windows overlapped by 100nt, both forward and reverse complement). They scored 50 million alignment windows.

The false positive rate – the fraction of windows that we incorrectly score as structural RNAs, even though they are not – is a critical number to estimate. To estimate a false positive rate, we have to devise a negative control – a way to obtain windows that are known not to contain a structural RNA, yet are matched controls for all other background properties of genomic alignment windows, such as sequence conservation, GC% composition, and indel pattern. This is a hard problem. Two main approaches are used. One approach is to shuffle alignments by columns, preserving properties like nucleotide composition in the window and primary sequence conservation in each column (1). Another approach is to simulate synthetic alignments according to a phylogenetic model (6, 23, 24).

It is easy to create poor negative controls. Naive shuffling of an alignment will do things like shatter indel patterns into a lot of single-base insertions and deletions, disrupt background dinucleotide composition (which tends to have a strong effect on RNA structure calculations (127)), and homogenize conservation and GC% composition across a window that might encompass a local region of high GC% or high conservation that tends to score highly (85). Real genome alignments may tend to score well only because of these confounding background effects, not because they contain RNA structures. On the other hand, the more a shuffling procedure tries to preserve more realistic background effects, the more it tends to preserve the original alignment. For example, a shuffling procedure used in (5) only altered the order of 53% of alignment columns on average, probably inadequate to destroy all the signal of a true RNA structure. Different methods produce very different predicted false positive rates. For example, Smith et al. show 10-fold differences in the false positive rates measured by simulations with SISSIz (24) versus shuffles with Multiperm

(1).

Smith et al. calibrated their score thresholds to allow a rate of 1% false positive predictions per 200nt alignment window, using both shuffled and simulated negatives. Therefore, when they score 50 million windows total, they expect about 500,000 false positives in total. The screen actually detects 4.1 million positive windows. Since we estimate that 500,000 of them are false, then all the excess (3.6 million) should be true. This gives us the so-called empirical false discovery rate (empirical FDR): $500,000/(4.1 \times 10^6) = 12\%$ FDR. (Varying how they generated negative control windows, Smith et al. reported their FDR to be 5-22%.)

## Empirical FDRs are vulnerable to choice of negative controls

Empirical FDRs are only as good as the estimate of the number of expected false positives under the assumed null hypothesis. If we underestimated the number of false positives by just 10-fold, a 12% false discovery rate might really be 100%. Essentially all our "statistically significant" candidates could be false.

Could the estimated FDR be off by 10-fold? Yes, easily. Consider the more familiar task of a BLASTN similarity search. We typically do not trust BLASTN E-values to better than a few orders of magnitude. BLASTN's estimated false positive rate, though quite good, is confounded by many nonrandom biases in real genome sequences – composition bias, repetitive sequence – that generate false positives at a higher rate than randomized expectation predicts. If we found 100 BLASTN hits in a database search at an E-value threshold of 10, we would not assume that 90 hits were true, but that is what an empirical FDR calculation does.

For conserved RNA structure detection, the null hypothesis is much more complex than BLASTN's, because of the need to match an even more complicated set of relevant properties of non-RNA genomic alignment windows. It would be prudent to have less confidence in the accuracy of these false positive estimates than in BLASTN's. It is all too easy to imagine background biological signals that could confound a structural RNA detector that are not taken into account in current

shuffled or simulated negative controls. One example is inverted DNA repeats, both short and long, which are abundant in genomes, partly because of the activity of DNA transposons. Inverted DNA repeats can look like RNA hairpins in a genome sequence analysis, even if they are never expressed as RNA.

Thus the application of these methods to large genomes seems premature and perilous, though the fundamental idea is sound. The discriminatory power of these methods needs to be increased substantially. One way to do this is by incorporating additional sources of information to increase signal/noise. Deeper multiple sequence alignments, analyzed with more powerful phylogenetic models of RNA structure-constrained sequence covariation patterns, is currently the main path forward in the field. Another interesting direction is opening up, because of dramatic improvements in the use of chemical and enzymatic RNA structure probing experiments.

# PROBING-DIRECTED RNA STRUCTURE PREDICTION

Chemical and enzymatic modification experiments have long been used to probe RNA structure. Various reagents differentially attack paired vs. unpaired nucleotides and generate cleavages or base modifications that can be assayed by sequencing (7, 19, 33, 61). Historically, interpreting chemical/enzymatic modification patterns has been something of a black art, and the experiments have been done on single RNAs at a time. Recently, better reagents have been developed including SHAPE chemistry (described below), and several genome-scale methods have coupled RNA structure probing to high throughput sequencing (37, 43, 48, 105, 111, 133). It has become feasible to probe simultaneously the structure of every RNA in a transcriptome. However, it remains unclear how structure probing data should best be incorporated into RNA structure analysis algorithms, even for the simplest problem of single-sequence RNA structure prediction.

Deigan et al. (2009) set a landmark in this area with a method for incorporating SHAPE probing data as soft constraints into single-sequence RNA structure prediction (12). This paper is a touch-

stone for understanding a growing body of work from several laboratories. Several papers have introduced alternative methods (69, 77, 116, 131); at least one paper has extended the method to another chemical probe, DMS (dimethyl sulfate) (10); and the method has been extended to pseudoknotted RNA structure prediction (29). In order to describe this work, it helps to give first some background on single sequence RNA secondary structure prediction, and on SHAPE chemistry.

## Single sequence RNA structure prediction

The most widely used methods for RNA secondary structure prediction work by free energy minimization. A nearest neighbor thermodynamic model (often called the "Turner rules") approximates the free energy ($\Delta G$) of an RNA secondary structure as a sum of individual free energy terms assigned to local features in an RNA structure, particularly to each base pair stack (i.e. neighboring base pairs, hence "nearest-neighbor model"), and also to hairpin, internal, and bulge loop lengths and various other elemental features (22, 47, 53, 128). Given the thermodynamic model, an efficient dynamic programming algorithm (the Nussinov/Zuker algorithm) guarantees finding the minimum energy RNA secondary structure (67, 134, 135). A related algorithm (the McCaskill algorithm) calculates the partition function, the sum over the ensemble of all possible secondary structures weighted by their predicted likelihood in solution, according to their estimated free energy (described in more detail below) (56). Using the McCaskill algorithm, alternative structures can be sampled from the ensemble according to their probability (14).

Although single sequence RNA secondary structure prediction has been useful, its accuracy remains unsatisfactory. Accuracy is limited by fundamental problems, including the fact that the residual error in the parameters (around 5% (128)) is greater than the typical free energy difference between quite different alternatives in the low free energy RNA landscape, and the fact that the model neglects the contribution of tertiary contacts and divalent cations to the overall free energy of an RNA's fold.

This tantalizing state of affairs – prediction accuracy that is useful but not reliable – has moti-

vated the search for additional information to constrain structure predictions, including data from chemical and enzymatic structure probing experiments (7, 19, 33, 61). Several past approaches for incorporating probing data in structure prediction have given uncompelling results (52, 53). Partly this is because probing experiments give noisy and ambiguous data (53), and partly it is because traditional probing reagents, such as DMS, have a complex dependence on sequence and local structure (19), making it difficult to parameterize an ad hoc approach. A breakthrough came from a probing reagent that acts in a much less context-dependent way, enabling simple ad hoc methods to be used.

## SHAPE chemistry

Introduced in 2005, SHAPE stands for "selective 2'-hydroxyl acylation analyzed by primer extension" (59). A SHAPE reagent acylates the 2'-hydroxyl position of a nucleotide's ribose sugar. The acylation impedes reverse transcription so its presence can be assayed by primer extension. The rate of the reaction depends on the local geometry of the nucleotide backbone (57). Nucleotides in Watson-Crick base pairs are constrained in an incompatible geometry, so they show low SHAPE reactivities. Unpaired nucleotides can show high SHAPE reactivities, presumably because a flexible nucleotide backbone can frequently visit a compatible geometry. Occasionally a nucleotide may happen to be constrained in the right geometry, making that nucleotide hyperreactive (57). Several different SHAPE reagents exist with different properties, such as reagents with fast reaction rates for probing kinetics (63), and reagents with properties better suited for in vivo SHAPE experiments (96).

The standard data processing protocol from a SHAPE experiment (3, 46, 70, 110) yields one normalized unitless number for each nucleotide in the probed RNA sequence. SHAPE values range from 0 to around 2, sometimes more (the upper bound is ill-defined because of the ad hoc nature of the "normalization").

## SHAPE data analysis from a likelihood ratio perspective

Like other structure probing reagents, SHAPE values do not unambiguously distinguish paired from unpaired bases. Rather, a SHAPE experiment confers probabilistic information about RNA secondary structure because the distribution of SHAPE reactivities for base-paired residues is different than for unpaired residues. For example, Figure 2A and 2B show empirical distributions of SHAPE values collected from *E. coli* SSU and LSU rRNA that were compiled by Sükösd et al. (101) from SHAPE experiments published by Deigan et al. (12).

Intuitively, we might imagine that probing data distributions would show distinct modes for unpaired versus paired bases – a high-reactivity peak for unpaired bases, a low-reactivity peak for paired. However, the mode of the distributions is low for both paired and unpaired bases. The information in SHAPE data comes from the increased variance at unpaired residues. An unpaired base is more likely to have a low SHAPE reactivity than a high one, but a base with high SHAPE accessibility is much more likely to be unpaired than paired. Figure 2C shows the paired/unpaired likelihood ratio as a function of the SHAPE value. A base with a low SHAPE value of 0 is about five-fold more likely to be paired than unpaired, and a base with a high SHAPE value of 2.0 is conversely about five-fold more likely to be unpaired. This is the information we want to incorporate into a SHAPE-directed structure prediction.

## Deigan's pseudoenergy approach

Deigan et al. proposed a particular pseudoenergy term for incorporating SHAPE data,

$$\Delta G'_i = m \log(\alpha_i + 1) + b,$$

where $\alpha_i$ is the SHAPE value for base $i$ in the RNA; $i = 1 \ldots n$, where $n$ is the sequence length; and $m$ and $b$ are free parameters with defaults set to $m = 2.6$ and $b = -0.8$ kcal mol$^{-1}$. This

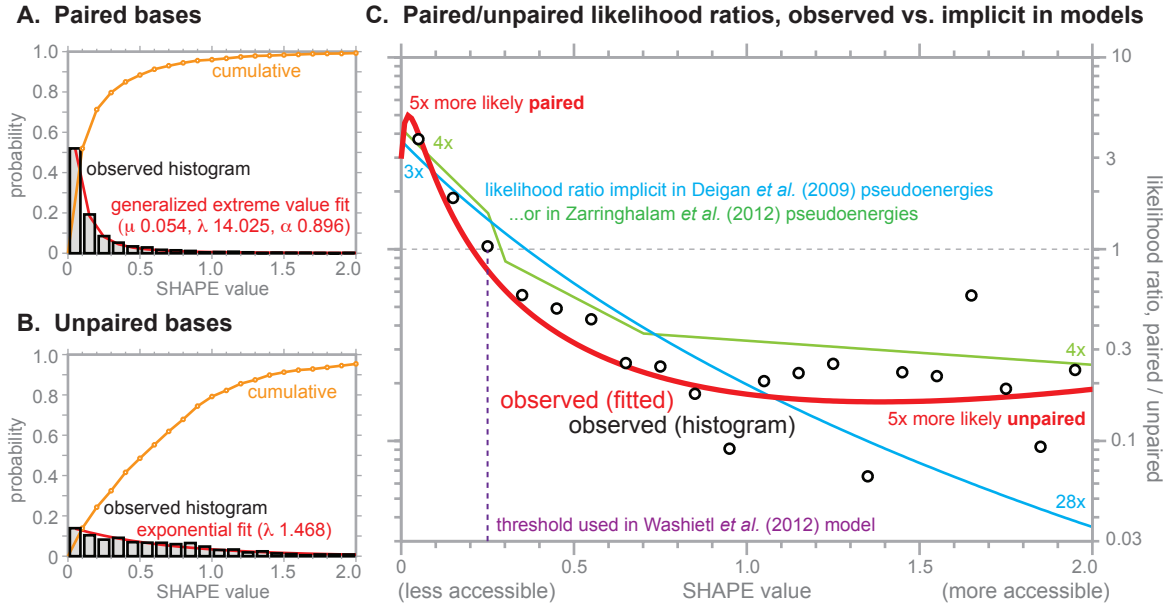


Figure 2: **Observed distributions of SHAPE values for paired vs. unpaired bases, compared to likelihood ratios implicit in different SHAPE-directed structure prediction methods.** (A, B): Distributions of SHAPE values observed for (A) 2531 paired nucleotides nd for (B) 1656 unpaired nucleotides in E. coli SSU and LSU rRNA in vitro SHAPE experiments of (12), collated by (101). Histograms, black bars; cumulative distributions, orange lines. Red lines show my maximum likelihood fits to the distributions chosen by (101), a generalized extreme value distribution (21) in (A) and an exponential in (B). Data replotted with permission from (101). Figure 1 in (101) distinguished helix end pairs from base pairs internal to a stacked stem, because helix ends are more flexible and accessible to SHAPE, but for simplicity I merged all base pairs here. (C) Paired/unpaired likelihood ratios: from the two fitted distributions in (A,B) (red line); from the corresponding histogram bins in (A,B) (black circles); implicit in the Deigan et al. pseudoenergy model (12) (blue line); and implicit in the Zarringhalam pseudoenergy model with their default $\beta = 0.89$ (131). See text for further explanation.

pseudoenergy term is applied to every residue $i$ involved in a base pair (and not to unpaired bases) in the calculations in the dynamic programming recursion.

If the SHAPE reactivity is minimal, $\alpha_i = 0$, then each base in a base pair is rewarded by an additional $-0.8$ kcal/mol. If reactivity is high, say $\alpha_i = 2.0$, then pairing base $i$ is disfavored by $+2.1$ kcal/mol. If SHAPE reactivity is $0.36$, there is no added pseudoenergy, and the SHAPE data are considered to equally favor pairing or unpairing base $i$.

Deigan et al. did not justify the choice of functional form, and they set $m$ and $b$ empirically by grid searching a wide range of parameter settings and finding values that optimized the accuracy of folding *E. coli* LSU rRNA. However, via statistical thermodynamics, we can observe that the pseudoenergy implies a likelihood model. The probability that base $i$ is paired, given the SHAPE data, is proportional to $e^{-\Delta G'_i/RT}$. Unpaired bases are implicitly assigned a pseudoenergy of 0, independent of SHAPE reactivity, so the probability that the base is unpaired is proportional to 1. The proportionality constant is the same (a simple partition function, the sum of the two terms), so we obtain:

$$\frac{P(\pi_i = \text{paired})}{P(\pi_i = \text{unpaired})} = e^{-\Delta G'_i/RT}.$$

The notation $\pi_i$ means the structural context of base $i$, which for the moment is just "paired" vs. "unpaired".

Thus the Deigan pseudoenergy term corresponds (at $37°$C) to saying that a maximally unreactive ($\alpha_i = 0; \Delta G'_i = -0.8$ kcal/mol) base is about 3-fold more likely to be paired than unpaired, and a highly reactive base $\alpha_i = 2.0; \Delta G'_i = +2.1$ kcal/mol) is about 28-fold more likely to be unpaired than paired. Figure 2C shows a plot of the paired/unpaired ratio implied by the Deigan pseudoenergy term, compared to the ratio implied by observed distributions (101).

## Sample and select approaches

It is not obvious that the thermodynamic RNA folding model can be combined in a mathematically defensible way with structure probing data. The use of arbitrary pseudoenergy parameters looks worryingly unprincipled. For this reason, an alternative approach called "sample and select" (69, 77) aims to keep the thermodynamic folding calculation separate from the probing data constraint. The idea is to sample suboptimal structures from the thermodynamic ensemble (14, 56), then rerank these sampled alternative structures by how well they agree with the structure probing data, according to some distance metric. The approach is not very powerful, because it relies on being able to sample the correct structure from the thermodynamic ensemble in the first place. If the correct structure has negligible posterior probability under the thermodynamic model, it is never sampled.

The approach requires choosing what distance should be calculated between the experimental probing data and a predicted structure. Sample and select approaches threshold the probing data to make discrete "paired" and "unpaired" calls for each base, then calculate the number of discrepancies from the predicted structure (a "Manhattan distance") (69, 77). From a statistical perspective a better-justified measure would be the log likelihood of the observed probing data given the structure, for example $\log P(\alpha \mid \pi) = \sum_i \log P(\alpha_i \mid \pi_i)$, using empirically collated $P(\alpha_i \mid \pi_i)$ distributions as in (101) (Figure 2A and 2B).

## Zarringhalam's pseudoenergy approach

Zarringhalam et al. use a distance-based argument to criticize the Deigan approach and develop a new one (131). They propose to optimize a distance between the SHAPE data and the structure, a Manhattan distance $\sum_i |\pi_i - a_i|$, where $\pi_i$ is the predicted structure of base $i$ which takes a value 1 if unpaired and 0 if paired, and $a_i$ is a modified SHAPE reactivity for position $i$, rescaled (by an ad hoc piecewise linear transformation of the original $\alpha_i$) to range $0 \ldots 1$. They add a pseudoenergy

of $\beta|\pi_i - a_i|$ to all bases $i$ (unpaired and paired). Impressively, they prove mathematically that this is guaranteed to improve (decrease) their calculated distance between the SHAPE data and the predicted structure, compared to the thermodynamic model alone, whereas the Deigan approach often yields higher distances for a SHAPE-constrained prediction compared to an unconstrained prediction.

Their argument hinges on whether we agree that we should want to minimize a Manhattan distance between the probing data and the predicted structure. In fact, this premise is probably not well justified, which is a shame, because of the strong theoretical proof that followed from their premise. As is apparent in Figure 2, a SHAPE reactivity $\alpha_i$ is not directly comparable to the probability that base $i$ is unpaired, because both paired and unpaired bases are more likely to have low $\alpha_i$ values.

In fact, in terms of the paired/unpaired likelihood ratios they imply, the Zarringhalam et al. and Deigan et al. approaches are rather similar (Figure 2). By using a symmetrical pseudoenergy function with the same $\beta$ for unpaired ($\pi_i = 1$) and paired ($\pi_i = 0$) bases, Zarringhalam et al. constrain the odds ratio to be symmetrical in the sense that at minimum SHAPE values, a base is about four-fold more likely to be paired than unpaired, and vice versa for maximum SHAPE values. The ad hoc piecewise linear mapping of the SHAPE value to the range 0..1 has the effect of rather closely approximating the empirical likelihood ratio distribution (compare the green and red lines).

## Washietl's ensemble approach

The above approaches assumed that base $i$ is either 100% paired or unpaired in the SHAPE experimental conditions. This corresponds to an assumption that a single RNA structure dominates in solution (even if the approach uses an ensemble calculation, as in (131)). What about an RNA that adopts two or more different structures in solution? Now the measured SHAPE data would be an ensemble-weighted average over these different structures: i.e. a function of the ensemble,

rather than of a single correct structure. Could SHAPE data be used not just to predict a single optimal structure, but to predict the ensemble? This is the point of an ensemble-based approach introduced by Washietl et al. (116). The basic idea is to perturb the energy parameters by the minimal amount needed to bring the ensemble base pairing probabilities into maximal agreement with the experimental SHAPE data.

Explaining the approach requires more introduction to ensemble calculations. According to the Gibbs-Boltzmann equation of statistical thermodynamics, the probability that a system is in a given state $i$ with free energy $\Delta G_i$ is proportional to $e^{-\Delta G_i/RT}$, where $R$ is the gas constant (0.001986 kcal mol$^{-1}$ K$^{-1}$) and $T$ is the absolute temperature in Kelvin. If we can enumerate the free energies of all possible states of the system, then the probability that the system will be in state $i$ is

$$\frac{e^{-\Delta G_i/RT}}{\sum_j e^{-\Delta G_j/RT}}.$$

The summation over all states, $\sum_j e^{-\Delta G_j/RT}$, is the partition function, often abbreviated $Z$. This is the quantity that the McCaskill algorithm recursively calculates over all possible RNA secondary structures for a sequence (56).

Washietl et al. calculate a predicted ensemble base pairing probability $z_i(\theta, \epsilon)$ for each residue $i$ by using a partition function calculation, using thermodynamic model parameters $\theta$, perturbed by an error vector $\epsilon$ that describes the uncertainty inherent in the parameters. Perturbing the energy parameters amounts to treating the ensemble as a random variable, because the ensemble is completely determined by the energy parameters. For example, we might assume that every energy parameter $\theta_u$ for some element $u$ of RNA structure has a normally-distributed error $\epsilon_u \sim N(0, \tau_u^2)$ with variance $\tau_u^2$. This is an attractively explicit model of the uncertainty in the Turner rules. We could obtain $\tau_u^2$ variances corresponding to the different certainties of different parameters (base pair stacking parameters are better determined than loop parameters, for example). What Washietl

et al. actually implement is an alternative, where perturbations $\epsilon_i$ are assigned to each residue $i$, with one position-independent variance $\tau^2$. This choice somewhat weakens Washietl et al.'s argument that their model is more physically grounded than a pseudoenergy model, because these $\epsilon_i$ terms are now pseudoenergies, rather than an explicit error model for energy model parameters $\theta_u$.

Critically, Washietl et al. assume that the probing data $\alpha$ can be used to directly obtain an experimentally "observed" probability $p_i(\alpha)$ that base $i$ is paired. They assume that this "experimental measurement" is subject to experimental errors, with the discrepancy $z_i(\theta, \epsilon) - p_i(\alpha)$ normally distributed as $N(0, \sigma_i^2)$. They further assume that this error is position-independent so they use a single $\sigma^2$.

Under this formulation, both the energy parameters and the observed SHAPE data are assumed to be subject to unknown measurement errors parameterized by variances $\tau^2$ and $\sigma^2$, respectively, and they can write their problem as a least-squares optimization problem:

$$\min_{\epsilon} \sum_i \frac{\epsilon_i^2}{\tau^2} + \sum_i \frac{(z_i(\theta, \epsilon) - p_i(\alpha))^2}{\sigma^2}.$$

This is the maximum likelihood estimate for the perturbation vector $\epsilon$ under the assumption that both "errors" are normally distributed. The core of their paper then shows that this minimization can be done by gradient descent.

A big difference from other probing-directed structure prediction methods is that here the $\epsilon_i$ pseudoenergies are optimized for each particular RNA sequence and its SHAPE data. Each $\epsilon_i$ term essentially means, how hard does nucleotide $i$ have to be tweaked to get it to agree with the SHAPE data? The $\epsilon_i$ are a position-specific profile of where the RNA structure is discrepant from the prediction of thermodynamic model. Washietl et al. show an interesting example where $\epsilon_i$ terms tend to be high for nucleotides that are modified in vivo (the thermodynamic model does not take in vivo nucleotide modifications into account).

Because SHAPE data $\alpha$ do not directly report the probability that base $i$ is paired, the principal

weakness in the approach is in obtaining $p_i(\alpha)$. Washietl et al. tried many ways of mapping $\alpha_i$ to an "observed" pairing probability $p_i$, but in the end, simply thresholded at 0.25, setting $p_i = 1$ (paired) for $\alpha_i < 0.25$, $p_i = 0$ (unpaired) for $\alpha_i > 0.25$. By discretizing $p_i$ to 0 or 1, 100% paired or unpaired, the whole point of using an ensemble-averaged calculation gets thrown away. Moreover, when $\pi_i$ is discretized to 0 or 1, it is dubious whether the discrepancy $|z_i - p_i|$ should be treated as a normally-distributed error. Rather, many $p_i$ are just wrong.

# STATISTICAL INFERENCE FOR PROBING-DIRECTED STRUCTURE PREDICTION

A well-principled framework for combining the inherently statistical information from a probing experiment with the thermodynamic model of RNA folding is desirable. It might improve the accuracy of probing-directed structure, and it would also allow more subtle information to be extracted from structure probing data than just whether a base is paired or not. Reactivity depends on structural context, so reactivity carries statistical information about structural context. For example, bases in helix-ending pairs tend to be more reactive to SHAPE probing than bases in internally stacked stems (101). Chemical/enzymatic data from DMS modification and RNase cleavage mapping shows more complex sequence dependencies than SHAPE data, and the lack of principled approaches impedes analysis of more complicated data.

A general approach can be outlined using probabilistic inference. The observation that a pseudoenergy term implies the paired/unpaired odds ratio (Figure 2) essentially means that the reverse is true. We can use empirical likelihood distributions of SHAPE data values $\alpha_i$ to derive a principled approach in terms of probabilities.

A strong approach to any inference problem, especially one involving integration of different sources of evidence, starts with writing a generative probability model that specifies the joint probability distribution of all the data – including the observed data variables that are giving us

information, the hidden data variables that we seek to infer, and any additional hidden nuisance variables that our model needs to specify in order to calculate the joint probability. Here, what we need is a computable model of the joint probability $P(\alpha, \mathbf{x}, \pi, \theta, \psi)$, for the observed probing data $\alpha$, the RNA sequence $\mathbf{x}$, the RNA secondary structure $\pi$ we want to infer, parameters $\theta$ of an RNA folding model, and parameters $\psi$ for a likelihood model of generating SHAPE values from a particular structure.

## Optimal structure prediction and a derivation of pseudoenergies

Suppose we assume that a single correct RNA secondary structure dominates in solution. This critical assumption allows us to assume that the observed SHAPE data $\alpha$ arose directly from that single structure $\pi$. (Otherwise the observed data are an ensemble-weighted average $\langle\alpha\rangle$, over unknown $\alpha(\pi)$ for each structure in solution; more on this later.) We can factor the joint distribution into a product of independent terms, where the observed probing data is sampled as a function of that structure $\pi$, and the probability of $\pi$ is specified by the RNA folding model for sequence $\mathbf{x}$:

$$P(\alpha, \mathbf{x}, \pi, \psi, \theta) = P(\alpha \mid \mathbf{x}, \pi, \psi) P(\mathbf{x}, \pi \mid \theta) P(\psi) P(\theta)$$

To simplify things a bit further, we can assume that we will obtain fixed model parameters $\psi$ and $\theta$ from an outside source – for example, by fitting to previously known example SHAPE data to obtain $\psi$ (as in Figure 2), and by using the existing Turner energy model as $\theta$. That means we can drop both terms (because they equal 1).

Then, by Bayes' rule, the posterior probability of any particular structure $\pi$ is given by:

$$P(\pi \mid \alpha, \mathbf{x}, \psi, \theta) = \frac{P(\alpha \mid \mathbf{x}, \pi, \psi) P(\mathbf{x}, \pi \mid \theta)}{\sum_{\hat{\pi}} P(\alpha \mid \mathbf{x}, \hat{\pi}, \psi) P(\mathbf{x}, \hat{\pi} \mid \theta)}$$

Generative probability models for RNA structure prediction give us $P(\mathbf{x}, \pi \mid \theta)$ directly (88).

Indeed, Sükösd et al. already introduced an inference equation much like the above for incorporating SHAPE data in the probabilistic RNA structure prediction method PPFold (100). For the thermodynamic folding model we need a bit more algebra. Recall that the thermodynamic model gives us $P(\pi \mid \mathbf{x}, \theta)$ via Gibbs-Boltzmann:

$$P(\pi \mid \mathbf{x}, \theta) = \frac{e^{-\Delta G(\pi)/RT}}{\sum_{\hat{\pi}} e^{-\Delta G(\hat{\pi})/RT}}.$$

We need the joint probability (with $\mathbf{x}$) not the conditional (given $\mathbf{x}$), but we can expand $P(\mathbf{x}, \pi \mid \theta) = P(\pi \mid \mathbf{x}, \theta) P(\mathbf{x} \mid \theta)$, and since we are only dealing with a single given sequence $\mathbf{x}$, we can cancel the $P(\mathbf{x} \mid \theta)$ term out of the posterior probability equation. Likewise the thermodynamic partition function $\sum_{\hat{\pi}} e^{\Delta G(\hat{\pi})/RT}$ is the same for the numerator and denominator of the posterior probability equation, so it cancels too. This leaves our posterior probability as:

$$P(\pi \mid \cdot) = \frac{P(\alpha \mid \mathbf{x}, \pi, \psi) \quad e^{-\Delta G(\pi)/RT}}{\sum_{\hat{\pi}} P(\alpha \mid \mathbf{x}, \hat{\pi}, \psi) \quad e^{-\Delta G(\hat{\pi})/RT}}$$

Since the denominator, summed over all possible structures $\hat{\pi}$, behaves as a normalization constant with respect to any individual structure $\pi$, let's just call it $Z'$ by analogy to a partition function – but remember it isn't the same as the thermodynamic partition function, because it includes the probability of the observed SHAPE data. Then take the logarithm of both sides, since probability model calculations are generally done as sums of logarithms rather than products of probabilities, to avoid numerical underflows:

$$\log P(\pi \mid \cdot) = \log P(\alpha \mid \mathbf{x}, \pi, \psi) - \frac{\Delta G(\pi)}{RT} - \log Z'.$$

Finally, if we're only interested in inferring the optimal (maximum probability) structure, we can drop the constant $Z'$ and just maximize:

$$\text{argmax}_\pi \log P(\pi \mid \cdot) = \text{argmax}_\pi \left[ \log P(\alpha \mid \mathbf{x}, \pi, \psi) - \frac{\Delta G(\pi)}{RT}. \right]$$

This means that all we need to add to the thermodynamic folding calculation, to make it a probing-directed calculation, is the log probability of observing the probing data $\alpha$ given the RNA structure.

So long as the $\psi$ parameterization of the probing data likelihoods is factored such that it maps well onto the folding model's energy parameterization $\theta$, in terms of having similarly factored dependencies on elements of RNA structure and sequence context, then this equation is readily implemented in the dynamic programming recursions of existing RNA structure prediction programs. For example, we might simply assume that the observed SHAPE data $\alpha_i$ for base $i$ are independent of the sequence and only depend on what structural context $i$ is in $\pi_i$, which might be simply "paired" vs. "unpaired". At every step of the dynamic programming recursion that adds base $i$ to a growing substructure, depending on whether $i$ is unpaired or paired in that substructure term in the recursion, the appropriate $\log P(\alpha_i \mid \pi_i)$ term is just added to the appropriate free energy parameter (scaled by $1/RT$).

In summary, this derivation suggests that an appropriate SHAPE "pseudoenergy" term for base $i$ is $\Delta G'_i = RT \log P(\alpha_i \mid \pi_i)$. This term should not be viewed as an energy at all, but as a log probability in a statistical inference approach. More sequence and structure context could easily be incorporated into this model as desired, by relaxing any of the simplifying assumptions.

## Ensemble prediction

Deriving the ensemble-based approach of of Washietl et al. (116) in terms of statistical inference is also possible, but more difficult, and I will only sketch the main issues. Now we want to infer the posterior distribution over ensembles, as opposed to just a single structure. This is equivalent

to inferring the posterior distribution $P(\theta \mid \cdot)$, as opposed to $P(\pi \mid \cdot)$, because the ensemble probabilities are completely determined as a function of the the folding model parameters $\theta$, for a given sequence $\mathbf{x}$. As in (116), we might specify a prior distribution $P(\theta)$ by assuming $\theta \sim N(\hat{\theta}, \tau^2)$, i.e. normally-distributed perturbations around the standard Turner parameters $\hat{\theta}$. The difficulty comes from the fact that the observed SHAPE data need to be treated as an ensemble average $\langle \alpha \mid \mathbf{x}, \theta \rangle = \sum_\pi \alpha(\pi) P(\pi \mid \mathbf{x}, \theta)$. The likelihood term $P(\langle \alpha \rangle \mid \cdot)$ unfortunately becomes a nasty multiple integral over all possible unknown $\alpha(\pi)$ vectors for all the individual structures in the ensemble, subject to the constraint that their ensemble-weighted average equals $\langle \alpha \rangle$. Under simplifying independence assumptions that the SHAPE data $\alpha_i$ for each position only depend on a small number $K$ of structural contexts for $(x_i, \pi_i)$, we could obtain a tractable $K$-dimensional integration over those states (for example $K = 2$ for $\pi_i =$ unpaired vs. paired). All this should be do-able, resulting in a strongly grounded version of the Washietl et al. approach where we could avoid directly comparing SHAPE values $\alpha_i$ to base pairing probabilities $z_i$, and instead utilize an empirical likelihood model for the observed SHAPE data.

# CONCLUSION

A statistical inference approach for incorporating structure probing data is easily generalized beyond single sequence structure prediction. More complicated RNA structure analysis problems, including algorithms for *de novo* conserved structure detection and for sequence/structure homology search (64, 122) also depend on scoring schemes that require inference of an unknown secondary structure. It would be straightforward to extend the approach described here to any of these methods. Essentially all that needs to be done is to include an empirical log probability term for the observed probing data at each base $i$, given the unknown structural context that an algorithm is trying to put $i$ in. If transcriptome-wide structure probing data become readily available in a variety of organisms (37, 105), we can imagine using these experimental data systematically across a

variety of tasks in computational RNA structure analysis (122).

Computational RNA sequence and structure analysis is a broad topic. There are many areas that I have not done justice to in this review. In particular, I have tended to focus on functional RNA analysis and discovery in multicellular eukaryotes, especially human, because this is where so much current controversy exists about pervasive transcription and lncRNAs. However, arguably, the richest hunting grounds for new functional RNAs are not in multicellular eukaryotes but in bacteria, where small RNAs are used extensively for post-transcriptional regulation. There is an excellent body of literature on bacterial regulatory RNAs that I lacked space to delve into (25, 97).

# Summary points

1. Functional RNAs are heterogeneous, and no one characteristic suffices to detect them all in an unbiased fashion.

2. Putative long noncoding RNAs (lncRNAs) are likely to be a heterogeneous population that includes analysis artifacts and transcriptional noise, but a subset of lncRNAs are well expressed and evolutionarily conserved.

3. One useful positive signal that helps distinguish many functional noncoding RNAs from other explanations is the presence of an evolutionarily conserved RNA secondary structure.

4. Genome-wide computational screens for regions of conserved secondary structure are a promising means for detecting functional structural RNAs (both RNA genes and cis-regulatory RNA motifs), but current methods have false positive rates that are too high.

5. RNA structure probing experiments help constrain secondary structure prediction, and have recently been adapted to systematic transcriptome-wide measurements, offering a way to increase the signal/noise of any computational analysis that depends on inferring RNA structure.

6. Several proposed methods for probing-directed RNA secondary structure prediction can be unified and rationalized using principles of probabilistic inference.

7. Using similar probabilistic inference principles, structure probing data could be used to quantitatively constrain and improve other computational RNA analyses, including homology search, alignment, and conserved structure detection.

# DISCLOSURE

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

# ACKNOWLEDGEMENTS

My thanks to Matt Hangauer, Christine Heitsch, Michael McManus, Martin Smith, and Ian Vaughn for providing data; to Michael Clark, Tanja Gesell, Jan Gorodkin, Ivo Hofacker, Zsuzsanna Sükösd, Eric Westhof, and Michael Zuker for feedback and discussions; to members of my laboratory including Fred Davis, Tom Jones, Seolkyoung Jung, Eric Nawrocki, Elena Rivas, and Travis Wheeler for critical comments on the manuscript; and to the hospitality of the beautiful Centro de Ciencias de Benasque Pedro Pascual in Benasque, Spain, where most of this article was drafted.

# Literature Cited

1. Anandam P, Torarinsson E, Ruzzo WL. 2009. Multiperm: Shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. *Bioinformatics* 25:668–69

2. Argaman L, Hershberg R, Vogel J, Bejerano G, Wagner EG, et al. 2001. Novel small RNA-encoding genes in the intergenic regions of *Escherichia coli*. *Curr. Biol.* 11:941–50

3. Aviran S, Trapnell C, Lucks JB, Mortimer SA, Luo S, et al. 2011. Modeling and automation of sequencing-based characterization of RNA structure. *Proc. Natl. Acad. Sci. USA* 108:11069–74

4. Babak T, Blencowe BJ, Hughes TR. 2005. A systematic search for new mammalian noncoding RNAs indicates little conserved intergenic transcription. *BMC Genomics* 6:104

5. Babak T, Blencowe BJ, Hughes TR. 2007. Considerations in the identification of functional RNA structural elements in genomic alignments. *BMC Bioinformatics* 8:33

6. Bradley RK, Uzilov AV, Skinner ME, Bendaña YR, Barquist L, Holmes I. 2009. Evolutionary modeling and prediction of non-coding RNAs in *Drosophila*. *PLoS ONE* 4:10

7. Brunel C, Romby P, Westhof E, Ehresmann C, Ehresmann B. 1991. Three-dimensional model of *Escherichia coli* ribosomal 5S RNA as deduced from structure probing in solution and computer modeling. *J. Mol. Biol.* 221:293–308

8. Bussotti G, Raineri E, Erb I, Zytnicki M, Wilm A, et al. 2011. BlastR–fast and accurate database searches for non-coding RNAs. *Nucl. Acids Res.* 39:6886–95

9. Clark MB, Amaral PP, Schlesinger FJ, Dinger ME, Taft RJ, et al. 2011. The reality of pervasive transcription. *PLoS Biol.* 9:e1000625

10. Cordero P, Kladwang W, VanLang CC, Das R. 2012. Quantitative dimethyl sulfate mapping for automated RNA secondary structure inference. *Biochemistry* 51:7037–39

11. Coventry A, Kleitman DJ, Berger B. 2004. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proc. Natl. Acad. Sci. USA* 101:12102–7

12. Deigan KE, Li TW, Mathews DH, Weeks KM. 2009. Accurate SHAPE-directed RNA structure determination. *Proc. Natl. Acad. Sci. USA* 106:97–102

13. di Bernardo D, Down T, Hubbard T. 2003. ddbRNA: detection of conserved secondary structures in multiple alignments. *Bioinformatics* 19:1606–11

14. Ding Y, Lawrence CE. 2003. A statistical sampling algorithm for RNA secondary structure prediction. *Nucl. Acids Res.* 31:7280–7301

15. Dinger ME, Amaral PP, Mercer TR, Mattick JS. 2009. Pervasive transcription of the eukaryotic genome: Functional indices and conceptual implications. *Brief. Funct. Genomic. Proteomic.* 8:407–23

16. Djebali S, Davis CA, Merkel A, Dobin A, Lassmann T, et al. 2012. Landscape of transcription in human cells. *Nature* 489:101–8

17. Eddy SR. 2013. The ENCODE project: Mistakes overshadowing a success. *Curr. Biol.* 23:R259–261

18. Eddy SR, Durbin R. 1994. RNA sequence analysis using covariance models. *Nucl. Acids Res.* 22:2079–88

19. Ehresmann C, Baudin F, Mougel M, Romby P, Ebel JP, Ehresmann B. 1987. Probing the structure of RNAs in solution. *Nucl. Acids Res.* 15:9109–28

20. ENCODE Project Consortium. 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489:57–74

21. Evans M, Hastings N, Peacock B. 2000. *Statistical Distributions*. New York, NY: John Wiley & Sons, third edition

22. Freier SM, Kierzek R, Jaeger JA, Sugimoto N, Caruthers MH, et al. 1986. Improved free-energy parameters for predictions of RNA duplex stability. *Proc. Natl. Acad. Sci. USA* 83:9373–77

23. Gesell T, von Haeseler A. 2006. *In silico* sequence evolution with site-specific interactions along phylogenetic trees. *Bioinformatics* 22:716–22

24. Gesell T, Washietl S. 2008. Dinucleotide controlled null models for comparative RNA gene prediction. *BMC Bioinformatics* 9:10

25. Gottesman S, Storz G. 2011. Bacterial small RNA regulators: Versatile roles and rapidly evolving variations. *Cold Spring Harb. Perspect. Biol.* 3:10

26. Gruber AR, Bernhart SH, Hofacker IL, Washietl S. 2008. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics* 9:10

27. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. 2010. RNAz 2.0: Improved noncoding RNA detection. *Pac. Symp. Biocomput.* 15:69–79

28. Guttman M, Rinn JL. 2012. Modular regulatory principles of large non-coding RNAs. *Nature* 482:339–46

29. Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. 2013. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. *Proc. Natl. Acad. Sci. USA* 110:5498–5503

30. Hangauer MJ, Vaughn IW, McManus MT. 2013. Pervasive transcription of the human genome produces thousands of previously unidentified long intergenic noncoding RNAs. *PLoS Genet.* 9:10

31. Hobbs EC, Fontaine F, Yin X, Storz G. 2011. An expanding universe of small proteins. *Curr. Opin. Microbiol.* 14:167–73

32. Hogan DJ, Riordan DP, Gerber AP, Herschlag D, Brown PO. 2008. Diverse RNA-binding proteins interact with functionally related sets of RNAs, suggesting an extensive regulatory system. *PLoS Biol.* 6:10

33. Inoue T, Cech TR. 1985. Secondary structure of the circular form of the *Tetrahymena* rRNA intervening sequence: a technique for RNA structure analysis using chemical probes and reverse transcriptase. *Proc. Natl. Acad. Sci. USA* 82:648–52

34. Kageyama Y, Kondo T, Hashimoto Y. 2011. Coding vs non-coding: Translatability of short ORFs found in putative non-coding transcripts. *Biochimie* 93:1981–86

35. Kapranov P, Cawley SE, Drenkow J, Bekiranov S, Strausberg RL, et al. 2002. Large-scale transcriptional activity in chromosomes 21 and 22. *Science* 296:916–19

36. Kapranov P, Laurent GS. 2012. Dark matter RNA: existence, function, and controversy. *Front. Genet.* 3:10

37. Kertesz M, Wan Y, Mazor E, Rinn JL, Nutter RC, et al. 2010. Genome-wide measurement of RNA secondary structure in yeast. *Nature* 467:103–7

38. Kladwang W, VanLang CC, Cordero P, Das R. 2011. A two-dimensional mutate-and-map strategy for non-coding RNA structure. *Nat. Chem.* 3:954–62

39. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T. 2001. Identification of novel genes coding for small expressed RNAs. *Science* 294:853–58

40. Lau NC, Lim LP, Weinstein EG, Bartel DP. 2001. An abundant class of tiny RNAs with probable regulatory roles in *Caenorhabditis elegans*. *Science* 294:858–62

41. Le SY, Chen JH, Currey KM, Maizel JV. 1988. A program for predicting significant RNA secondary structures. *Comput. Applic. Biosci.* 4:153–59

42. Lécuyer E, Yoshida H, Parthasarathy N, Alm C, Babak T, et al. 2007. Global analysis of mRNA localization reveals a prominent role in organizing cellular architecture and function. *Cell* 131:174–87

43. Li F, Zheng Q, Ryvkin P, Dragomir I, Desai Y, et al. 2012. Global analysis of RNA secondary structure in two metazoans. *Cell Rep.* 1:69–82

44. Li S, Breaker RR. 2013. Eukaryotic TPP riboswitch regulation of alternative splicing involving long-distance base pairing. *Nucl. Acids Res.* 41:3022–31

45. Lin MF, Jungreis I, Kellis M. 2011. PhyloCSF: a comparative genomics method to distinguish protein coding and non-coding regions. *Bioinformatics* 27:i275–82

46. Low JT, Weeks KM. 2010. SHAPE-directed RNA secondary structure prediction. *Methods* 52:150–58

47. Lu ZJ, Turner DH, Mathews DH. 2006. A set of nearest neighbor parameters for predicting the enthalpy change of RNA secondary structure formation. *Nucl. Acids Res.* 34:4912–24

48. Lucks JB, Mortimer SA, Trapnell C, Luo S, Aviran S, et al. 2011. Multiplexed RNA structure characterization with selective 2'-hydroxyl acylation analyzed by primer extension sequencing (SHAPE-seq). *Proc. Natl. Acad. Sci. USA* 108:11063–68

49. Macke TJ, Ecker DJ, Gutell RR, Gautheret D, Case DA, Sampath R. 2001. RNAMotif, an RNA secondary structure definition and search algorithm. *NAR* 29:4724–35

50. Maenner S, Blaud M, Fouillen L, Savoye A, Marchand V, et al. 2010. 2-D structure of the A region of Xist RNA and its implication for PRC2 association. *PLoS Biol.* 8:10

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

52. Mathews DH, Disney DH, Childs MD, Schroeder JL, Zuker M, Turner DH. 2004. Incorporating chemical modification constraints into a dynamic programming algorithm for prediction of RNA secondary structure. *Proc. Natl. Acad. Sci. USA* 101:7287–92

53. Mathews DH, Sabina J, Zuker M, Turner DH. 1999. Expanded sequence dependence of thermodynamic parameters improves prediction of RNA secondary structure. *J. Mol. Biol.* 288:911–40

54. Mathews DH, Turner DH. 2002. Dynalign: an algorithm for finding the secondary structure common to two RNA sequences. *J. Mol. Biol.* 317:191–203

55. Mazo A, Hodgson JW, Petruk S, Sedkov Y, Brock HW. 2007. Transcriptional interference: an unexpected layer of complexity in gene regulation. *J. Cell Sci.* 120:2755–61

56. McCaskill JS. 1990. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. *Biopolymers* 29:1105–19

57. McGinnis JL, Dunkle JA, Cate JH, Weeks KM. 2012. The mechanisms of RNA SHAPE chemistry. *J. Am. Chem. Soc.* 134:6617–24

58. Memczak S, Jens M, Elefsinioti A, Torti F, Krueger J, et al. 2013. Circular RNAs are a large class of animal RNAs with regulatory potency. *Nature* 495:333–38

59. Merino EJ, Wilkinson KA, Coughlan JL, Weeks KM. 2005. RNA structure analysis at single nucleotide resolution by selective 2'-hydroxyl acylation and primer extension (SHAPE). *J. Am. Chem. Soc.* 127:4223–31

60. Mitra S, Shcherbakova IV, Altman RB, Brenowitz M, Laederach A. 2008. High-throughput single-nucleotide structural mapping by capillary automated footprinting analysis. *Nucl. Acids Res.* 36:10

61. Moazed D, Stern S, Noller HF. 1986. Rapid chemical probing of conformation in 16 S ribosomal RNA and 30 S ribosomal subunits using primer extension. *J. Mol. Biol.* 187:399–416

62. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. 2008. Mapping and quantifying mammalian transcriptomes by RNA-seq. *Nat. Methods.* 5:621–28

63. Mortimer SA, Weeks KM. 2007. A fast-acting reagent for accurate analysis of RNA secondary and tertiary structure by SHAPE chemistry. *J. Am. Chem. Soc.* 129:4144–45

64. Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics, in press

65. Nordström KJ, Mirza MA, Almén MS, Gloriam DE, Fredriksson R, Schiöth HB. 2009. Critical evaluation of the FANTOM3 non-coding RNA transcripts. *Genomics* 94:169–76

66. Novikova IV, Hennelly SP, Sanbonmatsu KY. 2012. Structural architecture of the human long non-coding RNA, steroid receptor RNA activator. *Nucl. Acids Res.* 40:5034–51

67. Nussinov R, Pieczenik G, Griggs JR, Kleitman DJ. 1978. Algorithms for loop matchings. *SIAM J. Appl. Math.* 35:68–82

68. Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, et al. 2002. Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cDNAs. *Nature* 420:563–73

69. Ouyang Z, Snyder MP, Chang HY. 2013. SeqFold: genome-scale reconstruction of RNA secondary structure integrating high-throughput sequencing data. *Genome Res.* 23:377–87

70. Pang PS, Elazar M, Pham EA, Glenn JS. 2011. Simplified RNA secondary structure mapping by automation of SHAPE data analysis. *Nucl. Acids Res.* 39:10

71. Parker BJ, Moltke I, Roth A, Washietl S, Wen J, et al. 2011. New families of human regulatory RNA structures identified by comparative analysis of vertebrate genomes. *Genome Res.* 21:1929–43

72. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, et al. 2006. Identification and classification of conserved RNA secondary structures in the human genome. *PLoS Comput. Biol.* 2:e33

73. Peluso P, Herschlag D, Nock S, Freymann DM, Johnson AE, Walter P. 2000. Role of 4.5S RNA in assembly of the bacterial signal recognition particle with its receptor. *Science* 288:1640–43

74. Perreault J, Weinberg Z, Roth A, Popescu O, Chartrand P, et al. 2011. Identification of hammerhead ribozymes in all domains of life reveals novel structural variations. *PLoS Comput. Biol.* 7:10

75. Pickrell JK, Pai AA, Gilad Y, Pritchard JK. 2010. Noisy splicing drives mRNA isoform diversity in human cells. *PLoS Genet.* 6:10

76. Ponting CP, Belgard TG. 2010. Transcribed dark matter: Meaning or myth? *Hum. Mol. Genet.* 19:R162–68

77. Quarrier S, Martin JS, Davis-Neulander L, Beauregard A, Laederach A. 2010. Evaluation of the information content of RNA structure mapping data for secondary structure prediction. *RNA* 16:1108–17

78. Rabani M, Kertesz M, Segal E. 2008. Computational prediction of RNA structural motifs involved in posttranscriptional regulatory processes. *Proc. Natl. Acad. Sci. USA* 105:14885–90

79. Rabani M, Kertesz M, Segal E. 2011. Computational prediction of RNA structural motifs involved in post-transcriptional regulatory processes. *Methods Mol. Biol.* 714:467–79

80. Ray D, Kazan H, Cook KB, Weirauch MT, Najafabadi HS, et al. 2013. A compendium of RNA-binding motifs for decoding gene regulation. *Nature* 499:172–77

81. Reichow SL, Hamma T, Ferré-D'Amaré AR, Varani G. 2007. The structure and function of small nucleolar ribonucleoproteins. *Nucl. Acids Res.* 35:1452–64

82. Rinn JL, Chang HY. 2012. Genome regulation by long noncoding RNAs. *Annu. Rev. Biochem.* 81:145–66

83. Rinn JL, Kertesz M, Wang JK, Squazzo SL, Xu X, et al. 2007. Functional demarcation of active and silent chromatin domains in human HOX loci by noncoding RNAs. *Cell* 129:1311–23

84. Riordan DP, Herschlag D, Brown PO. 2011. Identification of RNA recognition elements in the *Saccharomyces cerevisiae* transcriptome. *Nucl. Acids Res.* 39:1501–9

85. Rivas E, Eddy SR. 2000. Secondary structure alone is generally not statistically significant for the detection of noncoding RNAs. *Bioinformatics* 6:583–605

86. Rivas E, Eddy SR. 2001. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* 2:8

87. Rivas E, Klein RJ, Jones TA, Eddy SR. 2001. Computational identification of noncoding RNAs in *E. coli* by comparative genomics. *Curr. Biol.* 11:1369–73

88. Rivas E, Lang R, Eddy SR. 2012. A range of complex probabilistic models for RNA secondary structure prediction that include the nearest neighbor model and more. *RNA* 18:193–212

89. Sakakibara Y, Brown M, Hughey R, Mian IS, Sjölander K, et al. 1994. Stochastic context-free grammars for tRNA modeling. *Nucl. Acids Res.* 22:5112–20

90. Salehi-Ashtiani K, Lupták A, Litovchick A, Szostak JW. 2006. A genomewide search for ribozymes reveals an HDV-like sequence in the human CPEB3 gene. *Science* 313:1788–92

91. Seemann SE, Sunkin SM, Hawrylycz MJ, Ruzzo WL, Gorodkin J. 2012. Transcripts with in silico predicted RNA structure are enriched everywhere in the mouse brain. *BMC Genomics* 13:10

92. Serganov A, Nudler E. 2013. A decade of riboswitches. *Cell* 152:17–24

93. Silverman IM, Li F, Gregory BD. 2013. Genomic era analyses of RNA secondary structure and RNA-binding proteins reveal their significance to post-transcriptional regulation in plants. *Plant Sci.* 205:55–62

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

95. Smith MA, Gesell T, Stadler PF, Mattick JS. 2013. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res., in press. (doi: 10.1093/nar/gkt596)

96. Spitale RC, Crisalli P, Flynn RA, Torre EA, Kool ET, Chang HY. 2013. RNA SHAPE analysis in living cells. *Nat. Chem. Biol.* 9:18–20

97. Storz G, Vogel J, Wassarman KM. 2011. Regulation by small RNAs in bacteria: Expanding frontiers. *Mol. Cell.* 43:880–91

98. Stricklin SL. 2006. *Noncoding RNA Genes in* Caenorhabditis elegans. Ph.D. thesis, Washington University School of Medicine

99. Struhl K. 2007. Transcriptional noise and the fidelity of initiation by RNA polymerase II. *Nat. Struct. Mol. Biol.* 14:103–5

100. Sükösd Z, Knudsen B, Kjems J, Pedersen CN. 2012. PPfold 3.0: Fast RNA secondary structure prediction using phylogeny and auxiliary data. *Bioinformatics* 28:2691–92

101. Sükösd Z, Swenson MS, Kjems J, Heitsch CE. 2013. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. *Nucl. Acids Res.* 41:2807–16

102. Torarinsson E, Sawera M, Havgaard JH, Fredholm M, Gorodkin J. 2006. Thousands of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA structure. *Genome Res.* 16:885–89

103. Tsai MC, Manor O, Wan Y, Mosammaparast N, Wang JK, et al. 2010. Long noncoding RNA as modular scaffold of histone modification complexes. *Science* 329:689–93

104. Tycowski KT, Shu MD, Steitz JA. 1996. A mammalian gene with introns instead of exons generating stable RNA products. *Nature* 379:464–66

105. Underwood JG, Uzilov AV, Katzman S, Onodera CS, Mainzer JE, et al. 2010. FragSeq: transcriptome-wide RNA structure probing using high-throughput sequencing. *Nat. Methods.* 7:995–1001

106. Uzilov AV, Keegan JM, Mathews DH. 2006. Detection of non-coding RNAs on the basis of predicted secondary structure formation free energy change. *BMC Bioinformatics* 7:173

107. van Bakel H, Hughes TR. 2009. Establishing legitimacy and function in the new transcriptome. *Brief. Funct. Genomic. Proteomic.* 8:424–36

108. van Bakel H, Hughes TR. 2010. Most "dark matter" transcripts are associated with known genes. *PLoS Biol.* 8:e1000371

109. van Bakel H, Nislow C, Blencowe BJ, Hughes TR. 2011. Response to "the reality of pervasive transcription". *PLoS Biol.* 9:e1001102

110. Vasa SM, Guex N, Wilkinson KA, Weeks KM, Giddings MC. 2008. ShapeFinder: a software system for high-throughput quantitative analysis of nucleic acid reactivity information resolved by capillary electrophoresis. *RNA* 14:1979–90

111. Wan Y, Qu K, Ouyang Z, Chang HY. 2013. Genome-wide mapping of RNA structure using nuclease digestion and high-throughput sequencing. *Nat. Protoc.* 8:849–69

112. Washietl S, Findeiss S, Müller SA, Kalkhof S, von Bergen M, et al. 2011. RNAcode: robust discrimination of coding and noncoding regions in comparative sequence data. *RNA* 17:578–94

113. Washietl S, Hofacker IL. 2004. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. *J. Mol. Biol.* 342:19–30

114. Washietl S, Hofacker IL, Lukasser M, Hüttenhofer A, Stadler PF. 2005. Mapping of conserved RNA secondary structures predicts thousands of functional noncoding RNAs in the human genome. *Nat. Biotechnol.* 23:1383–90

115. Washietl S, Hofacker IL, Stadler PF. 2005. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* 102:2454–59

116. Washietl S, Hofacker IL, Stadler PF, Kellis M. 2012. RNA folding with soft constraints: Reconciliation of probing data and thermodynamic secondary structure prediction. *Nucl. Acids Res.* 40:4261–72

117. Washietl S, Pedersen JS, Korbel JO, Stocsits C, Gruber AR, et al. 2007. Structured RNAs in the ENCODE selected regions of the human genome. *Genome Res.* 17:852–64

118. Wassarman KM, Repoila F, Rosenow C, Storz G, Gottesman S. 2001. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* 15:1637–51

119. Watts JM, Dang KK, Gorelick RJ, Leonard CW, Bess JW Jr, et al. 2009. Architecture and secondary structure of an entire HIV-1 RNA genome. *Nature* 460:711–16

120. Wei D, Alpert LV, Lawrence CE. 2011. RNAG: a new Gibbs sampler for predicting RNA secondary structure for unaligned sequences. *Bioinformatics* 27:2486–93

121. Westholm JO, Lai EC. 2011. Mirtrons: microRNA biogenesis via splicing. *Biochimie* 93:1897–1904

122. Will S, Siebauer MF, Heyne S, Engelhardt J, Stadler PF, et al. 2013. LocARNAscan: incorporating thermodynamic stability in sequence and structure-based RNA homology search. *Algorithms Mol. Biol.* 8:14

123. Will S, Yu M, Berger B. 2013. Structure-based whole-genome realignment reveals many novel noncoding RNAs. *Genome Res.* 23:1018–27

124. Wilm A, Higgins DG, Notredame C. 2008. R-Coffee: a method for multiple alignment of non-coding RNA. *Nucl. Acids Res.* 36:10

125. Wilusz JE, Freier SM, Spector DL. 2008. 3' end processing of a long nuclear-retained noncoding RNA yields a tRNA-like cytoplasmic RNA. *Cell* 135:919–32

126. Wilusz JE, Spector DL. 2010. An unexpected ending: Noncanonical 3' end processing mechanisms. *RNA* 16:259–66

127. Workman C, Krogh A. 1999. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucl. Acids Res.* 27:4816–22

128. Xia T, , Burkard ME, Kierzek R, Schroeder SJ, et al. 1998. Thermodynamic parameters for an expanded nearest-neighbor model for formation of RNA duplexes with Watson-Crick base pairs. *Biochemistry* 37:14719–35

129. Yao Z, Weinberg Z, Ruzzo WL. 2006. CMfinder – a covariance model based RNA motif finding algorithm. *Bioinformatics* 22:445–52

130. Zappulla DC, Cech TR. 2004. Yeast telomerase RNA: a flexible scaffold for protein subunits. *Proc. Natl. Acad. Sci. USA* 101:10024–29

131. Zarringhalam K, Meyer MM, Dotu I, Chuang JH, Clote P. 2012. Integrating chemical footprinting data into RNA secondary structure prediction. *PLoS ONE* 7:10

132. Zhang Z, Huang S, Wang J, Zhang X, de Villena FPM, et al. 2013. GeneScissors: a comprehensive approach to detecting and correcting spurious transcriptome inference owing to RNA-seq reads misalignment. *Bioinformatics* 29:10

133. Zheng Q, Ryvkin P, Li F, Dragomir I, Valladares O, et al. 2010. Genome-wide double-stranded RNA sequencing reveals the functional significance of base-paired RNAs in *Arabidopsis*. *PLoS Genet.* 6:10

134. Zuker M. 1989. On finding all suboptimal foldings of an RNA molecule. *Science* 244:48–52

135. Zuker M, Stiegler P. 1981. Optimal computer folding of large RNA sequences using thermodynamics and auxiliary information. *Nucl. Acids Res.* 9:133–48

# Annotations for selected citations

**(12)** Pioneering paper on SHAPE-directed RNA secondary structure prediction.

**(30)** Comprehensive survey and meta-analysis of human transcriptomic data from 127 different RNA-seq libraries.

**(37)** Describes PARS (parallel analysis of RNA structure), a transcriptome-wide application of RNA structure probing.

**(65)** A thorough re-analysis of FANTOM3 "noncoding RNAs", showing that they were contaminated with several artifacts.

**(95)** The most recent human genome-wide computational screen for conserved RNA structure detection.

**(101)** One of few places where empirical distributions for SHAPE data have been shown.

**(105)** Describes FragSeq, a transcriptome-wide application of RNA structure probing.

**(116)** A conceptually different approach for SHAPE-directed structure prediction, focused on ensemble calculations.

**(131)** An approach for SHAPE-directed secondary structure prediction that contrasts to (12).

# Acronyms and definitions

**Transcriptomics** Systematic discovery, quantitation, and cataloging of individual RNA transcripts.

**RNA secondary structure** An essentially two-dimensional representation of an RNA in terms of its intramolecular nested base pairing interactions that form stems and loops.

**Noncoding RNA (ncRNA)** RNA that does not code for protein. Depending on context, may include mRNA untranslated regions (UTRs) and/or nonfunctional RNA.

**Long noncoding RNA (lncRNA)** Noncoding RNA transcripts longer than (i.e. other than) the abundant classes of small RNAs such as microRNAs, snoRNAs, snRNAs, and tRNAs.

**Transcriptional noise** RNA species produced by background error rates of other normal RNA biogenesis processes.

**Pervasive transcription** Especially in mammalian genomes, the observation that most of a genome is transcribed at a detectable level.

**False discovery rate (FDR)** The fraction of a set of predictions that are statistically expected to be false positives.

**Expectation value (E-value)** The number of false positives expected at or above some score threshold.

**RNA structure probing** Using enzymatic or chemical modification and/or cleavage experiments to differentiate base-paired from single-stranded nucleotides in an RNA.

**SHAPE** Acronym for "selective 2'-hydroxyl acylation analyzed by primer extension"; a structure probing chemistry with favorable sequence-independent properties.

**Ensemble** The set of all possible secondary structures for an RNA sequence; often in the context of assigning each structure a probability.

**Partition function** A sum over an ensemble; normalization factor for converting free energies of individual structures to probabilities in the ensemble.

**Manhattan distance** Distance between two vectors computed as the sum of absolute differences of each element, like walking distance on a city grid.

**Gibbs-Boltzmann equation** Statistical thermodynamics equation relating free energy of individual states to the probability of each state in an ensemble.

**Bayes' rule** A basic equation in probability calculus for calculating a posterior probability: $P(B \mid A) = P(A \mid B)P(B)/P(A)$.

**Posterior probability** The conditional probability of a random variable of interest, given observed data, often obtained by Bayes' rule.

**Joint probability** The probability of two or more random variables together, as in $P(A, B)$.

**Conditional probability** The probability of one random variable given the value of another, as in $P(A \mid B)$.