Comparative genomics beyond sequence based alignments: RNA structures in the ENCODE regions

E. Torarinsson[1,2], Z. Yao[3], E. D. Wiklund[4], J. B. Bramsen[4] , C. Hansen[5], J. Kjems[4], N. Tommerup[5], W. L. Ruzzo[3,6] and J. Gorodkin[1*]

[1]Section for Genetics and Bioinformatics, IBVH, Faculty of Life Sciences, University of Copenhagen, 1870 Frederiksberg C, Denmark

[2]Department of Natural Sciences, Faculty of Life Sciences, University of Copenhagen, 1871 Frederiksberg C, Denmark

[3]Department of Computer Science and Engineering, University of Washington, Seattle WA 98195-2350, USA

[4]Department of Molecular Biology, University of Aarhus, Aarhus, Denmark

[5]Department of Cellular and Molecular Medicine, Wilhelm Johannsen Centre for Functional Genome Research, University of Copenhagen, 2200 Copenhagen N, Denmark

[6]Department of Genome Sciences, University of Washington Seattle WA 98195-5065, USA

Keywords: ENCODE, noncoding RNA, ncRNA

[*]Corresponding author

Section for Genetics and Bioinformatics
Department of Animal and Veterinary Sciences
And Center for Applied Bioinformatics
University of Copenhagen
Groennegaardsvej 3
1870 Frederiksberg C
Denmark
Phone: +45 3533 3578
Fax: +45 3528 3042
Email: gorodkin@genome.ku.dk

Abstract

Recent computational scans for noncoding RNAs (ncRNAs) in multiple organisms have relied on existing multiple sequence alignments. However, as sequence similarity drops, a key signal of RNA structure—frequent compensating base changes—is increasingly likely to cause sequence-based alignment methods to misalign, or even refuse to align, homologous ncRNAs, consequently obscuring that structural signal. We have used CMfinder, a structure-oriented local alignment tool, to search the ENCODE regions of vertebrate multiple alignments. In agreement with other studies, we find a large number of potential RNA structures in the ENCODE regions. We report 6,587 candidate regions with an estimated false positive rate of 50%. More intriguingly, many of these candidates may be better represented by alignments taking the RNA secondary structure into account than those based on primary sequence alone, often quite dramatically. For example, approximately one quarter of our predicted motifs show revisions in more than 50% of their aligned positions. Furthermore, our results are strongly complementary to those discovered by sequence-alignment-based approaches—84% of our candidates are not covered by Washietl et al., increasing the number of ncRNA candidates in the ENCODE region by 32%. In a group of eleven ncRNA candidates that were tested by RT-PCR, ten were confirmed to be present as RNA transcripts in human tissue, and most show evidence of significant differential expression across tissues. Our results broadly suggest caution in any analysis relying on multiple sequence alignments in less well-conserved regions, clearly support growing appreciation for the biological significance of ncRNAs, and strongly argue for considering RNA structure directly in any searches for these elements.

Introduction

The main objective of the ENCyclopedia Of DNA Elements (ENCODE) project is to identify all functional elements in the human genome sequence. For this purpose, 30MB, or roughly 1% of the total genome have been selected as ENCODE regions for this pilot project. The Pilot Project involves close interactions between computational and experimental scientists to evaluate a number of methods for annotating the human genome (The ENCODE Project Consortium, 2007). A major challenge in the project is to annotate the large number of noncoding RNAs (ncRNAs), which are difficult to find by computational or experimental means. The discovery of a steadily increasing number of untranslated RNAs since the late 1990s has dramatically changed views on the roles and importance of ncRNAs.

The task of computationally finding ncRNAs is difficult because one has to consider secondary structure as well as nucleotide sequence. With only one sequence available, one can fold the sequence using single sequence folding methods (Hofacker et al. 1994; Zuker, 2003; Ding and Lawrence, 2004), but structure can be detected more reliably from a set of related sequences, if available (Westhof and Michel, 1994; Westhof et al. 1996). Predicting the RNA secondary structure is a necessity when searching for structured ncRNAs, and this makes RNA search algorithms computationally expensive. The seminal approach of Sankoff (1985) performs simultaneous alignment and structure inference, but it remains too computationally expensive for broad use. Various approximations to it have been developed, including FOLDALIGN (Havgaard et al. 2007), Dynalign (Harmanci et al. 2007), Stemloc (Holmes, 2005) and Consan (Dowell and Eddy, 2006) all attempting to increase performance without sacrificing accuracy, but even these

procedures remain relatively computationally expensive. A natural alternative approach is to align the sequences first, and then do RNA structure inference based on the alignment. This strategy is particularly attractive now that high quality whole-genome multiple sequence alignments are available for 17 or more vertebrates (e.g., see Blanchette et al. 2004). Two recently developed programs, RNAz (Washietl et al. 2005a, 2005b) and EvoFold (Pedersen et al. 2006), exploited these alignments to search for ncRNAs. These timely scans resulted in thousands of putative novel structured ncRNAs. The initial RNAz and EvoFold scan restricted attention to those portions of the multiple alignments that were defined to be highly conserved (Siepel et al. 2005), thus minimizing the number of alignment errors. This year the RNAz and EvoFold developers joined forces to scan all multiple alignments in the ENCODE regions for putative ncRNAs, not only the most conserved ones (Washietl et al. 2007), resulting in many additional candidates in these regions (albeit with estimated false positive rates on the order of 50%).

Although these programs have significant strengths, their false negative rates and other limitations of these studies are essentially unexplored. A particular concern is exactly the reliance on existing multiple sequence alignments, which are based on DNA sequence similarity alone. Unfortunately, as sequence similarity drops, a key feature of RNA structure—frequent compensating base changes—is increasingly likely to cause sequence-based alignment methods that are ignorant of RNA structure to misalign, or even refuse to align, homologous ncRNAs, consequently obscuring that structural signal. As illustrated by an example below, even modest misalignments in moderately well-conserved sequences can have an adverse effect. Torarinsson et al. (2006) provide even deeper evidence, by using FOLDALIGN to show the apparent presence of thousands of RNA structures conserved between human and mouse in regions *not* aligned in the UCSC MULTIZ alignments. An additional concern is that RNAz and EvoFold generally assume that an RNA structure, if present, is present in all sequences in the alignment, ignoring the possibility of gain or loss on some branches of the phylogeny. Finally, both programs initially evaluate only *global* alignments within fixed width sliding windows, which further reduces sensitivity since a given placement of the window may include extraneous sequence flanking a given RNA structure, may include only part of the structure, or both.

In short, reliance on sequence-based alignments (and current tools) both biases away from regions that are conserved in structure but not sequence, while not fully protecting from alignment errors that also mask structure conservation. These observations lead us to apply CMfinder (Yao et al. 2006) to the ENCODE regions as a complement to the RNAz/EvoFold scans. CMfinder searches a set of (presumably) orthologous, unaligned sequences for local patterns indicative of conserved RNA sequence and structure. We do not rely on externally supplied alignments (except to indicate orthology), do not use a sliding window approach, and can ignore diverged sequences that do not appear to share the discovered RNA motif.

CMfinder has been very successfully used in discovering ncRNAs in bacteria. In a genome-wide study in the Firmicutes (Yao et al. 2007), CMfinder's top-ranking motifs included most known Firmicute RNA elements, and it achieved high accuracy in both membership prediction and secondary structure prediction in comparison to the hand-curated motif models from the Rfam database (Griffith-Jones, 2003). In addition, CMfinder predictions have led to discovery of many novel regulatory elements in this and other bacterial groups, including several new families of riboswitches (Weinberg et al. 2007).

In agreement with the previous studies, we find a large number of potential RNA structures in the ENCODE regions. We report 6,587 candidate regions with an estimated false positive rate of 50%. More intriguingly, many of our predicted motifs may be better represented by alignments taking the RNA secondary structure into account than those based on primary sequence alone, often quite dramatically. For example, approximately one quarter of our motifs show revisions in more than 50% of their positions, in comparison to the sequence-based MULTIZ alignments. Furthermore, our candidate regions are largely complementary to the results of the RNAz/EvoFold scans—while overlap with the candidates generated by those scans is much greater than would be expected by chance, 84% of our candidate regions do *not* overlap results of previous scans. These results broadly suggest caution in any analysis relying on multiple sequence alignments in less well-conserved regions, clearly support growing appreciation for the biological significance of ncRNAs, and strongly argue for taking RNA structure directly into account in any searches for these elements.

Results

The candidates

We scanned 2 × 56,017 (forward/reverse) multiple alignment blocks from the UCSC MULTIZ multiple alignment (.maf) files, one block at a time (155nts long on average). Since previous studies were presumed to be effective in well-conserved regions, we restricted analysis to alignment blocks which overlap neither exons nor the most conserved elements (as defined by the PhastCons Conserved Elements (Siepel et al. 2005). These alignments covered 8.68 Mb of human sequence (out of the total of 30 Mb in the ENCODE regions), and included 3.87Mb of repetitive sequence as defined by the RepeatMasker (Smit et al. 1996) track of the UCSC alignments. We included alignments in repeat regions in human because many of the known ncRNAs are found there. This resulted in 10,106 predicted motifs which met our cutoff criteria: a composite score above 5 and Gibbs energy below −5 kcal/mol (see Methods). We estimated a false positive rate of 50% by repeating the analysis on shuffled alignments (see Methods). Composite score and energy distributions for randomized vs. original alignments are depicted in Fig. 1 showing a slight shift in the distribution towards lower energy and higher score for our native predictions. Some of these predicted motifs overlap or are sense/antisense to each other. Considering these as a single candidate region we have 6,587 candidate regions. Our candidate regions average 80nt in length, collectively covering a total of 0.53 MB, or 6.1% of our human input sequence. Candidate regions are approximately twice as dense (per nucleotide) in nonrepetitive regions (0.38 MB of 4.81 MB or 7.9%) than in repeat regions (0.15 MB of 3.87 MB or 3.9% of the repetitive input dataset).

Known ncRNAs

As noted by Washietl et al. (2007) the ENCODE regions are surprisingly poor in annotated ncRNAs. In fact, when studying Rfam (Griffiths-Jones et al. 2003), the Functional RNA project (www.ncRNA.org), and the snoRNA and miRNA tracks that have been mapped to the human genome by the UCSC Genome Browser (Kent et al. 2002), we could only find one ncRNA which fully overlapped our input alignments. This was the miRNA hsa-miR-483 on chromosome 11

identified by Fu et al. (2005) in fetal liver in human. In addition miR-483 has been annotated in mouse and rat "by similarity" in mirBase (Griffiths-Jones 2004; Griffiths-Jones et al. 2006). This miRNA was detected in our scan (composite score 8.6, energy -31.4) and was scored highly as an miRNA by RNAmicro (Hertel et al. 2006), which we ran on all our predictions. Our prediction, in addition to human, rat and mouse, also includes dog, cow and rabbit. Hsa-miR-483 was also detected by RNAz but was not in the input set for EvoFold (Washietl et al. 2007).

Transcription data and purifying selection

Using oligonucleotide tiling array techniques, transcription maps of TARs (transcriptionally active regions) (Bertone et al. 2004) and Transfrags (transcribed fragments) (Cheng et al. 2005) have been generated. We compared our predictions to TARs and Transfrags generated as a part of the ENCODE project, which used 11 human tissues (ENCODE Consortium, 2007). Note that these maps were derived from RNA fragments longer than 200 nucleotides. TARs and Transfrags were only generated for the repeat masked regions of the genome whereas we included the repeat regions, so candidates in repeat regions (25% of our total candidate regions) were ignored in calculating the following numbers. 16.9% of these candidate regions overlap TARs/Transfrags. At the nucleotide level, 11.8% of the bases in the predictions overlap a TAR or a Transfrag, compared to 7.0% of the input bases (that is, our whole repeat-masked input data). In a recent study by Kapranov et al. (2007) the genomic origins and the relations of human nuclear and cytosolic polyadenylated RNAs longer than 200 nucleotides (lRNA) in eight cell lines and whole-cell RNAs less than 200nt (sRNA) in two cell lines were investigated. Comparing our candidate regions to these new transfrags, on the nucleotide level, 3.0% and 27.4% of our candidates were overlapped by short and long RNAs respectively, compared to 1.5% and 16.0% of the input bases. The increased overlap with TARs/Transfrags, sRNA and lRNA is highly significant with p-values of 10^{-40}, 10^{-24} and 10^{-86}, respectively. Still, one has to be cautious since, as noted by Washietl et al (2007), the tiling-array studies may be more sensitive on G+C-rich regions and the TARs/Transfrags are very G+C-rich. With this in mind we divided our input data into five bins based on G+C content (0-35%, 35-40%, 40-45%, 45-50%, 50-100% G+C ranges, chosen to contain similar numbers of alignment blocks) and repeated our analysis on each bin separately. Surprisingly, none of the five G+C bins show statistically significant overlap with the tiling-array data. Basically, the explanation is that our predictions, the tiling-array predictions, and the observed overlap between them are all concentrated in the high G+C range, and controlling for this bias erases the apparent significance of the overall overlap. We did the same analysis for the RNAz and EvoFold candidates that are in our input data, and came to the same conclusion for their candidates. Although our analysis included only a portion of their candidates it does suggest that there is not a significant overlap with TARs/Transfrags when considering G+C content–the apparent overall significance of overlap with the tiling array data is seemingly explained by the G+C biases. However, Washietl et al. (2007) further point out that it is unclear whether the G+C bias for tiling-array data has a biological explanation or is a technical artifact. Additionally, they note that secondary structure may affect detection performance on tiling-arrays, considering the observation of several examples where highly stable ncRNAs result in negative signal “holes” in tiling-array data (Cheng et al. 2005). Together these observations leave open whether to expect tiling array technology to sensitively identify structured ncRNAs.

Lunter et al. (2006) have identified noncoding regions apparently under purifying selection on the basis of lack of indels. We compared our candidate regions to their set of Indel Purified Segments

(IPSs) on human assembly hg18. For our two most G+C-rich bins (where the majority of our candidate regions lie) there is a significant overlap to the IPSs ($P < 10^{-8}$ and $P < 10^{-31}$), indicating that many of our candidate regions are under purifying selection.

GENCODE

We also compared our candidate regions to the GENCODE annotations (Harrow et al. 2006), which aim to identify all human protein-coding genes in the ENCODE regions. We find that 40% of our candidates are intergenic whereas 60% overlap some non-exonic part of a protein coding gene (see Table 1). We also analyzed whether introns, 3' UTRs or 5' UTRs were enriched for our candidate regions, again stratified by G+C. Significant enrichment of predicted candidate regions is seen only in the highest G+C bin of 5' UTRs ($P < 10^{-6}$).

Table 1. GENCODE overlaps. Total number and percentage of candidates overlapping non-exonic GENCODE annotations.

Sense	Antisense	Both	Intron	5'UTR	3'UTR
1721 (43.7%)	1332 (33.8%)	884 (22.5%)	3274 (83.1%)	551 (14%)	89 (2.3%)

There are also 23 candidates that overlap with an exon, because we use the GENCODE annotation here, whereas our initial filtering was done with UCSC known genes annotation.

RNAz and EvoFold

As mentioned earlier, a similar scan to ours was performed with the global, alignment-dependent programs RNAz and EvoFold (Washietl et al. 2007). Note that they use the TBA (Threaded Blockset Aligner) repeat masked multiple sequence alignments with up to 28 species as prepared by the ENCODE alignment group (Margulies et al. 2007) whereas we used the MULTIZ alignments (with autoMZ driver) with up 17 species available at the UCSC genome browser. In both cases the alignments are prepared using the TBA/MULTIZ software (Blanchette et al. 2004). We used the latest assemblies (human hg18) whereas Washietl et al. (2007) use earlier assemblies (human hg17) because the TBA ENCODE alignments are only available for hg17. We used hg18 because it was the latest assembly with genome wide multiple alignments available. Furthermore, the input alignments for RNAz and EvoFold were pre-processed according to different preferences of these programs (Washietl et al. 2007).

To compare our predictions with those of RNAz and EvoFold, we used all their candidates (low and high confidence) that overlapped neither exons nor the PhastCons conserved elements (38% of their total predictions) (Siepel et al. 2005), and compared them to our 4933 (75% of our total candidate regions) candidates in nonrepetitive regions. Only 6.7% of these candidate regions overlap with EvoFold predictions, whereas 17.2% overlap with RNAz candidates (see Fig. 2). To estimate significance of this overlap we calculated p-values for our five G+C bins. For the two most G+C-rich G+C bins (45-50% and 50-100%, which contain the majority of our candidates) the overlap with EvoFold was significant ($P < 10^{-5}$ in both bins). The overlap with RNAz was significant in all five G+C bins ($P < 10^{-22}$, $P < 10^{-17}$, $P < 10^{-28}$, $P < 10^{-27}$ and $P < 10^{-39}$, ordered by increasing G+C%.) In the regions that do not overlap exons, PhastCons conserved elements, or repeat regions, we add 3861 new candidates to the 6071 RNAz or EvoFold candidates.

Furthermore, we predict 1654 candidates in regions that are in repeat regions in human (excluded by Washietl et al. (2007)) and thereby add 5515 candidates to the 17,046 RNAz or EvoFold candidates in the ENCODE regions, corresponding to 32% of the total number of candidates.

EvoFold has a strong preference for TA-rich regions whereas RNAz prefers G+C-rich regions since the minimum free energy is important to RNAz. The CMfinder predictions are approximately normally distributed, centered on 53% GC content. Still, when considering that the background G+C content is 43% it is clear that CMfinder also prefers G+C rich regions which tend to be more structurally stable.

Candidate Database

All of our candidate regions are available in an online database (http://genome.ku.dk/resources/cmf_encode). The database includes a variety of additional annotations such as the overlaps described above, occurrences such as conserved tetraloop motifs and predicted microRNA using RNAmicro (Hertel et al. 2006). The database also supports easy access to subsets of the candidates with different features. For example, one can easily retrieve all candidates overlapping TARs/Transfrags or all miRNA predictions. Furthermore, each candidate region is linked directly to the UCSC genome browser. Despite the relatively high false positive rate, it is possible and simple to use the information in our database to select higher confidence predictions through the "Database Search" link. For example, one can choose predictions that overlap with EvoFold/RNAz predictions and/or overlap TARs/Transfrags.

Realigning parts of the genomes

A benchmark study by Gardner et al. (2005) compared the relative performances of structure-*versus* sequence-based methods when aligning pairs of known tRNAs. The study revealed a dramatic divergence in performance for sequences with identity below ~60%; i.e., sequence-based methods were dramatically worse below this threshold. Note that Gardner et al. define pairwise sequence identity as IDENTITIES/MIN(length A, length B) for sequences A and B (personal communication), whereas we, dealing with multiple alignments, define this as IDENTITIES/MAX(length A, length B). IDENTITIES is the number of identical positions in the alignment and the length is the gap-free length of the sequence. For example, the sequences ATGC and AG are 100% identical by the former definition, but only 50% identical by the latter. Applying our definition to Gardner et al.'s data lowers the pairwise sequence identities by 3% on average. Although Gardner et al.'s observation is based on pairwise alignments on tRNAs, it is reasonable to assume that there exists a sequence identity threshold, for sequence-based multiple alignment tools, below which the generated alignments will be sub-optimal when considering structured ncRNAs. This means that one should be careful when searching for structured ncRNAs in sequence-based alignments when the sequence similarity is below this threshold, because these alignments will contain many more errors which will propagate through alignment-dependent methods. CMfinder considers both sequence and structure information and is therefore expected to perform better on regions with low sequence similarity. Considering that our input alignments have 50% average pairwise sequence similarity, it is clear that when RNA secondary structure is of importance, these alignments will often benefit from being re-aligned, taking structure into account. We calculated how much of the sequence is being re-aligned by CMfinder, compared to the original sequence-

based alignment; as expected, the degree of re-alignment correlates with sequence similarity (Pearson correlation of −0.77) (see Fig. 3). Approximately one quarter of the alignments show re-alignment in more than 50% of positions (see Methods).

Most of the known ncRNA families probably exhibit artificially high sequence similarities due to ascertainment bias–members are often discovered based on sequence similarity. To demonstrate possible benefits of structure–aware alignment, we examined MULTIZ multiple alignment blocks identified by Wang et al. (2007) to contain matches to Rfam ncRNAs (Griffiths-Jones et al. 2003), with good matches to the Rfam model in all species in the same region of the alignment. In one example containing 10 mammals, with fairly high sequence identity (~72%), neither EvoFold nor RNAz report a candidate there. However, CMfinder identifies a candidate (composite score > 5 and energy < -5) in all 10 species in good general agreement with the H/ACA snoRNA known there (Rfam accession RF00402). CMfinder's alignment of the region differs from the MULTIZ alignment in only 13% of positions, yet this change is sufficient to flip the RNAz prediction from negative ("RNA probability" 0.11, based on using their script to select 6 organisms) to strongly positive (probability 0.98; see Supplement). EvoFold did not predict anything for either alignment. While this is just one example, it does highlight the fact that even reasonably solid sequence-based alignments may not suffice for RNA discovery. Considering the high number of ENCODE region alignments with relatively low sequence similarities, it is reasonable to expect CMfinder, in many cases, to perform better on these alignments than sequence-alignment-dependant tools.

Furthermore it should be noted that RNAz and EvoFold remove individual sequences with more than 25% and 20% gaps, respectively, as compared to human. This is not necessary when using CMfinder since it is alignment-independent. CMfinder found motifs in 1408 and 673 individual sequences that would have been removed because of too many gaps by EvoFold and RNAz, respectively. Also RNAz is limited to 4-6 sequences, so they sample 6 sequences (repeated 3 times if there are more than 10 sequences in the alignment), optimizing the selected sequences to have sequence similarity as close to 80% as possible. EvoFold considers every sequence in the alignment, resulting in a lower score if any sequence is missing the motif. In contrast, although number of species is a factor in its composite score, CMfinder can ignore a sequence if it does not contain the motif and still report a high scoring motif for the rest of the sequences.

Experimental verification

An increasing number of ncRNAs are reported to be implicated in tissue-specific developmental and disease processes (reviewed in (Costa 2005)), yet the precise biological function of most ncRNAs remains elusive. To further explore the biological relevance of our prediction method, we selected 11 high scoring ncRNA candidates for experimental verification. We selected high confidence predictions by setting stricter score cutoffs (composite score > 9 and energy < -15), by requiring a minimum length of 60 and required more than five compensating base changes, indicating a possible evolutionary pressure to maintain the structure. We tested the expression of these 11 candidates in human RNA pools using strand specific primers (see Methods). We found that 8 out of 11 ncRNA candidates indeed could be detected in human RNA samples by reverse transcription PCR (RT-PCR) (ncRNA candidate #1, #2, #4, #7, #8, #9, #10 and #11; Figure 4A). Such expression may simply reflect transcriptional noise, yet current literature suggests that mammalian ncRNAs exhibit highly tissue-specific expression profiles, which is likely to be indicative of specialized functions in the organism (Ravasi et al. 2006; Sasaki et al. 2007). Hence,

in order to expand our analysis and identify potential spatial and functional roles of our predicted set of ncRNAs, we performed an extensive expression analysis in 22 human tissues by RT-PCR totaling more than 250 separate duplicated reactions (see Methods). Our analysis demonstrated that 10 out of the 11 candidates are indeed expressed in one or more human tissues (Figure 4B). Interestingly, this analysis showed that 7 of 10 confirmed candidates exhibited a highly tissue-specific expression profile, whereas only two ncRNAs were more ubiquitously expressed (#10 and #11; Figure 4B). Hence, in agreement with the current consensus, we believe that the predicted ncRNAs may have highly defined biological roles (Ravasi et al. 2006; Sasaki et al. 2007). In addition, the highly differential expression patterns of the ncRNA candidates strongly suggest that the expression is real and not merely transcriptional noise, thus supporting the validity of our prediction method.

An interesting observation is that 9 out of 11 ncRNA candidates were detected in brain (Figure 4B). In fact, a similar enrichment of ncRNA expression in brain versus other tissues has previously been demonstrated in mouse (Ravasi et al. 2006) and several reports on the involvement and relative abundance of ncRNA in human CNS function and developing have recently emerged (Cavaille et al. 2001; French et al. 2001; Pollard et al. 2006; Sone et al. 2007). Further an RNAz screen of porcine EST sequences revealed that developmental brain tissue seems to contain more ncRNAs than other tissues (Seemann et al. 2007). In order to examine the expression profile of our CNS-expressed candidates in more detail, we performed RT-PCR analysis on human RNA purified from total brain, fetal brain, cerebellum, hippocampus and spinal cord (Figure 4C). Again, distinct expression profiles were identified. For example as observed in the other tissues candidate #11 was expressed in all the investigated nervous tissues (Figure 4C). Candidate #8, on the other hand, showed a more restricted expression profile, detected in fetal brain and, although less pronounced, hippocampus of adult brain (Figure 4C). Hence, even within a single organ, the predicted ncRNA candidates appear to have highly specialized expression profiles, which is suggestive of a distinct biological function.

To expand our analysis, Northern blot analysis was performed for the 10 ncRNA candidates, confirmed by RT-PCR, on human RNA from 15 different tissues (Figure 4D). In general, detection of ncRNAs by Northern blotting has proven very difficult as the majority of ncRNAs are low abundance transcripts (Sasaki et al. 2007). However, we were able to detect bands for ncRNA candidate #6 (Figure 4D), and the expression of candidate #6 was confirmed to be strictly brain-specific by the Northern blot analysis. The 2.8 kb long transcript is located within a 4 kb long intron of Synapsin 3. In figure 4D we have removed four tissues because of a high level of background noise, interfering with the results.

Next, we investigated the precise genomic locations of the ncRNAs; five of the ncRNA candidates (#1, #2, #6, #9 and #10) are located within intronic sequences of known genes, all but candidate #1 on the same strand. Overall, we find a good correlation between our ncRNA expression analysis and database searches for the predicted host mRNA; for instance candidate #6 is located within an intron of Synapsin 3 (SYN3) , which is neuron-specific and predominantly expressed in the brain (Kao et al. 1998). This expression profile is well confirmed by both our RT-PCR and Northern blot analysis showing a clear brain specific expression of ncRNA #6. Furthermore, candidate #9 is located within an intron of the GRM8 (glutamate receptor metabotropic 8) precursor encoding a G-protein-coupled metabotropic glutamate receptor expressed in the central nervous systems (Duvoisin et al. 1995). Again, our RT-PCR analysis confirms candidate #9 expression both in spinal cord and in most compartments of the brain and (Figure 4B and 4C). Finally, candidate #10

is located within the primary TIMP3 RNA transcript that encodes an inhibitor of matrix metalloproteinases (Genbank acc. NM_000362). TIMP3 mRNA is rather broadly expressed predominantly in brain, kidney and lung (Leco et al. 1994) which correlates well with the expression patterns of candidate #10 as evaluated by our RT-PCR analysis (Figure 4B). In conclusion, we find by both RT-PCR and Northern blot analysis that predicted ncRNA candidates are expressed in a highly tissue-specific manner which is likely indicative of specialized biological functions and thus supports the validity of our prediction method.

Discussion

Noncoding RNAs are receiving increasing attention in genome science. This paper describes the first large-scale search for structured ncRNAs in several vertebrate genomes using a local structural motif finding algorithm, which has identified several thousand novel candidate ncRNAs. Our work complements a previous pairwise scan for local structured RNA elements in corresponding unaligned regions of the human and mouse genomes (Torarinsson et al. 2006) by extending it to multiple genomes and including a wider range of sequence similarities. Furthermore, except to indicate orthology, the scan was not dependent on sequence-based pre-aligned genomic regions, as is the case with RNAz and EvoFold scans (Washietl et al. 2007), allowing us to increase the number of ncRNAs candidates in the ENCODE regions by 32%. With a growing number of sequenced genomes, and with improving genome alignment methods that are capable of capturing orthology among phylogenetically diverse species, analysis of syntenic yet diverse regions becomes more feasible (Margulies et al. 2006). Alignments of increasingly diverse regions often mean decreasing average pairwise sequence similarity. This is problematic for sequence-based alignment methods. When searching for structured ncRNAs one can therefore benefit from disregarding these alignments and re-align the regions considering sequence and structure, often resulting in better alignments. Indeed it has been shown, for pairwise alignments of tRNAs, that it is preferable also to consider structure when aligning these if sequence similarity is below ~60% (Gardner et al. 2005).

There are several remaining challenges in this field. Extending the analysis to (presumably) syntenic unaligned regions adjacent to aligned regions is one important direction. The main obstacles in doing this is data collection complexity and increased computation time. Candidate scoring is another challenge. Although useful, we don't feel that any of the methods used to date constitute the last word on this topic. Even seemingly simple issues like the dinucleotide composition of shuffled alignments used as null examples are problematic. Additionally, we expect many functionally important ncRNA motifs to be repeated in the genome, e.g., *cis*-regulatory elements controlling several genes in a common pathway, or multiple members of as-yet unknown RNA families. There has been limited work to date attempting to identify or cluster repeated motifs predicted by genome-scale RNA discovery approaches (Torarinsson et al. 2007; Will et al. 2007). The CMfinder-based approach we have described in this paper potentially provides an efficient alternative to these clustering approaches. Since each of our RNA motifs is described by a covariance model, in principle, we could use each to scan the genome for additional instances. Pragmatically, using each to scan the set of sequences representing each other motif should be effective and fast enough to be feasible (Weinberg et al. 2006), since we would expect reasonable cross-species conservation of each motif instance. However, completion of a full-genome CMfinder scan is a prerequisite. Finally, there is big need for high-throughput methods, computational and experimental, to identify a potential function for the tens of thousands of candidates that have resulted from scans like this.

Methods

Data

The multiple alignments from the ENCODE regions were obtained from the UCSC genome browser, more specifically, the multiple alignments of 16 vertebrate genomes with the human genome (assembly hg18, Mar. 2006). We post-processed these alignments to remove all alignments blocks that overlapped with exons of known genes (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/knownGene.txt.gz) or the highly conserved PhastCons elements (http://hgdownload.cse.ucsc.edu/goldenPath/hg18/database/phastConsElements17way.txt.gz) in human. Furthermore, we made an additional set with the reverse complementary sequences of each sequence in the alignment. GENCODE, TARs, Transfrags, EST and IPS data were obtained from UCSC's table browser (http://genome.ucsc.edu/cgi-bin/hgTables) and converted, when needed, from assembly hg17 to hg18 using their liftOver software (http://genome.ucsc.edu/cgi-bin/hgLiftOver). sRNA and lRNA data were obtained at http://transcriptome.affymetrix.com/publication/hs_whole_genome. EvoFold and RNAz candidates were obtained at http://www.tbi.univie.ac.at/papers/SUPPLEMENTS/ENCODE . Repetitive regions were defined by the UCSC RepeatMasker track for human (hg18).

False positive rate

In order to estimate the false positive rate we shuffled all of our input alignments and ran CMfinder on them. The alignments were shuffled as described by Washietl and Hofacker (2004) resulting in random alignments of the same base composition, sequence conservation and gap patterns. The shuffling method we used retains a coarse grained pattern of conservation (only columns with mean pairwise identity >0.5 and <0.5 were shuffled with each other, respectively) (Washietl et al. 2007). Note that this shuffling does not conserve the dinucleotide frequencies, which is an unsolved problem for shuffling multiple alignments. Dinucleotide frequencies have an effect on the Gibbs free energies due to stacking interactions. Since the Gibbs free energy plays a role in our scoring of the candidates, this has an unknown effect on our estimated false positive rate.

Running CMfinder

We ran CMfinder (version 0.2) separately on each alignment block in the MULTIZ alignment as well as the reverse complement of each such block. When running CMfinder we output up to 5 single stem predictions (size range 30-100bp) and 5 predictions containing two stems (size range 40-100bp). This corresponds to running CMfinder with the options "-n 5 -m 30 -M 100" and then with the options "-n 5 -s 2 -m 40 -M 100". Then we tried to combine the motifs using the greedy heuristics implemented in CMfinder's CombMotif.pl procedure, which estimates alignment scores for concatenation of all pairs of motifs, and combines them progressively by merging the two

motifs with the highest concatenation score. [0]See (Yao et al. 2006) for more details about these options.

We ranked all CMfinder motifs using a heuristic scoring function that favors motifs with instances in diverged species and stable consensus secondary structure. CMfinder sometimes identifies purely structural motifs (e.g., alignments of single hairpins) that could easily arise by chance. Such motifs are usually scored well by both EvoFold and RNAz. To discriminate against such, likely spurious, structural motifs with no sequence conservation, we consider local sequence conservation in the scoring function. This is based on the observation that most known ncRNA motifs, even the ones with low sequence conservation, contain mosaic patterns of local sequence conservation, which are plausibly interaction sites for other molecules under strong selection. On the other hand, we penalize global sequence conservation, as highly similar sequences are more likely to be conserved by selection pressure on primary sequence than on structure. The final score is defined as

$$r = sp * \sqrt{\frac{lc}{sid}} * \frac{bp}{len}$$

where *sp* is the number of species in which the motif occurs, *lc* the local sequence conservation score (see Supplement for details), *sid* the global average pairwise sequence identity, *bp* the number of base pairs in the consensus structure and *len* the alignment length. This score is referred to as the "composite score" (see Supplement for details). A variant of this somewhat ad hoc scoring scheme performed well on ncRNA discovery in Bacteria (Yao et al. 2007; Weinberg et al. 2007). The score used here is length normalized to favor motifs with compact RNA structure. We have tried a few alternatives, including RNAz and Evofold, both of which strongly favor short stable stemloop motifs with low sequence similarity that very likely to be aligned by chance. We have also tried to integrate our motif features for scoring by machine learning algorithms including support vector machine (SVM) and logistic regression, but these methods did not perform well, probably due to the heterogeneity of the features, and limitations of available training data.

After systematically studying various cutoffs we chose to focus on candidates with a composite score over 5 and Gibbs energy below -5, which resulted in a large number of candidates with a reasonable false positive rate (see Supplement for details). The energy is computed as the average energy of each sequence in the alignment as calculated by RNAfold (Hofacker et al. 1994) when constrained to the secondary structure annotated by CMfinder.

P-value calculation

To calculate the p-values we counted the number of candidate regions whose center nucleotide overlaps the data we are testing against, i.e. TARs. To get a p-value, we compare it to the null model that each candidate is a dart thrown randomly onto the genome. If the TARs cover a fraction P of the ENCODE nucleotides in MAF blocks (our input data), then it is a simple binomial model: each of the N darts has probability P of hitting a TAR. For N candidates, the expected number of hits is $\mu = N * P$, with a standard deviation $\sigma = \sqrt{N * P * (1 - P)}$. We then calculate the p-value using the normal approximation to the binomial distribution, pnorm function in R (pnorm(observed, μ, σ , lower.tail=F)). Out of a concern that various edge effects might distort the statistics, we also calculated the P-values using the leftmost and rightmost nucleotide, instead of the

center nucleotide. This gives very similar results, although, when comparing to RNAz and EvoFold the P-values were a bit worse, probably because they are global and use window lengths, whereas CMfinder is local, therefore an overlap with our candidates' central nucleotide to RNAz and EvoFold candidates seems more likely. See supplement for all the P-values.

Realignment calculation

To quantify how much has been realigned by CMfinder in a given motif compared to the original multiple alignment (see Fig. 3), we calculate the following quantities. Let *sp* be the number of sequences in the CMfinder alignment, and define *m* to be the number of matched positions in that alignment, i.e., the number of quadruples (*s, t, i, j*) with $1 \leq s < t \leq$ sp and such that position *i* of sequence *s* is aligned with position *j* of sequence *t*. Let *v* be the number of those matches that are realigned relative to the MULTIZ alignment, i.e., the number quadruples as above for which position *i* of *s* is matched to position *j* of *t* in the CMfinder alignment, but not in the MULTIZ alignment (either, *i* and *j* are aligned to nucleotides in different positions or to gaps). The overall re-alignment fraction we report is *v*/*m*. For example if we have two multiple alignments, A and B, of four sequences which are all 10bp long, we will compare all six possible sequence pairs (all pair-combinations of the four sequences). Say we have 6 columns that are aligned differently in alignment A and B between sequences 1 and 3 and that the rest is aligned alike. Then we would say that 10% (6/(6*10)) of alignment B is realigned compared to alignment A.

Experiments

The tissue specific expression profiles of 11 candidate ncRNAs were determined by RT-PCR using purified total RNA from 22 different human tissues (adrenal gland, bone marrow, brain (whole, fetal, cerebellum, and hippocampus), kidney, liver (fetal), lung, prostate, salivary gland, skeletal muscle, spleen, testis, thymus, thyroid gland, trachea, uterus, colon and small intestine). cDNA was generated by reverse transcription (RT) using M-MLV SuperScript® III Reverse Transcriptase (Invitrogen, Carlsbad CA). The RT was carried out according to the supplied standard protocol using either random hexamer primer (Figure 4B) or gene specific primers to test for strand specificity (Figure 4A) (see Supplement for primer list). A total of 5pmol primer and ca. 1µg RNA was used per 20µl RT reaction. Directly upon completion of the RT, the cDNA was amplified by PCR using HotStarTaq DNA polymerase (Qiagen, Valencia CA) according to the supplied protocol. The PCR was carried out on approx. 10% of the total cDNA (by mass per 20µl RT reaction) using the following program: 6min 95°C denaturing; [95°C denaturing – 0:30min; 54-56°C annealing – 0:30min; 72°C elongation 0:30min] (40 cycles); 72°C elongation 10:00min. A primer set for β-actin was used as a positive control. Blank and negative 'no RT' RNA controls (equal mass of RNA to cDNA) were also included to test for DNA contamination of the RNA samples. The PCR products were visualized by ethidium bromide staining on a 2% agarose gel. The complete procedure of RT-PCR and gel visualization was performed at least twice for each candidate in each individual tissue. The identity of the detected DNA fragments was confirmed by sequencing using the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems, Foster City CA) according to the supplied protocol.

For Northern blot analysis of ncRNA expression, Nylon membranes with pre-blotted human RNA samples (15µg/tissue) (Zyagen, San Diego CA) were hybridized at 37 °C in Ultrahyb hybridization

buffer (Ambion, Austin TX) with 80 nt. end-labeled probes antisense to the predicted ncRNAs. Upon overnight hybridization membranes were washed in 2xSSC, 0.1 % SDS and bands were visualized by phoshorimaging.

Acknowledgements

We thank Phil Green, Graham McVicker, Jakob H. Havgaard and Luc Jaeger for useful discussions We acknowledge funding from the Danish Research Council for production and technology and the Danish Center for Scientific Computation. Wilhelm Johannsen Centre for Functional Genome Research is established by the Danish National Research Foundation.

Figure legends

Figure 1. Score distribution of the full CMfinder input set (A) composite score and (B) consensus minimum free energies, for the native and random (shuffled) sequences. There is a slight shift towards lower energy and higher score for our native data.

Figure 2. Overlap of predictions made by CMfinder, RNAz and EvoFold. Only predictions that are not highly conserved (phastCons), outside exons and repeat regions are considered, since these regions are the common subset of the input regions to these three programs. The total number for each program is indicated in parenthesis below the label.

Figure 3. Average pairwise sequence similarity of the predicted motifs vs. the fraction that has been re-aligned compared to the original alignments.

Figure 4. Expression of predicted ncRNA candidates by RT-PCR and Northern blot analysis. (A) Strand-specific RT-PCR analysis of ncRNA candidates on human RNA pools (see Methods). β-actin was used as control yielding PCR products in the presence of reverse transcriptase (RT+), but not in its absence (RT-). (B) Tissue-specific expression of ncRNA candidates as evaluated by RT-PCR analysis of human RNA samples. The same β-actin controls as for A were used. (C) Expression of ncRNA candidates within the human CNS as evaluated by RT-PCR analysis. The same β-actin controls as for A and B were used. (D) Expression of ncRNA candidate #6 as evaluated by Northern blotting of human RNA samples from 11 tissues.

References

Bertone, P., Stoc, V., Royce, T.E., Rozowsky, J.S., Urban, A.E., Zhu, X., Rinn, J.L., Tongprasit, W., Samanta, M., Weissman, S., et al. 2004. Global identification of human transcribed sequences with genome tiling arrays. *Science* **306:** 2242-2246.

Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D., et al. 2004. Aligning mulitple genomic sequences with the threaded blockset aligner. *Genome Res.* **14(4):** 708-15.

Cavaille, J., P. Vitali, E. Basyuk, A. Huttenhofer, and J.P. Bachellerie. 2001. A novel brain-specific box C/D small nucleolar RNA processed from tandemly repeated introns of a noncoding RNA gene in rats. *J Biol Chem* **276**: 26374-26383.

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.

Costa, F.F. 2005. Non-coding RNAs: new players in eukaryotic biology. *Gene* **357**: 83-94.

Ding, Y., Chan, C.Y. and Lawrence, C.E. 2004. Sfold web server for statistical folding and rational design of nucleic acids. *Nucleic Acids Res.* **32:** W135-WI41.

Dowell, R.D. and Eddy, S.R. 2006. Efficient pairwise RNA structure prediction and alignment using sequence alignment constraints. *BMC Bioinformatics* **7:** 400.

Duvoisin, R.M., C. Zhang, and K. Ramonell. 1995. A novel metabotropic glutamate receptor expressed in the retina and olfactory bulb. *J Neurosci* **15**: 3075-3083.

Eddy, S.R. and Durbin, R. 1994. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22:** 2079-2088.

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.

French, P.J., T.V. Bliss, and V. O'Connor. 2001. Ntab, a novel non-coding RNA abundantly expressed in rat brain. *Neuroscience* **108**: 207-215.

Fu, H., Tie, Y., Xu, C., Zhang, Z., Zhu, J., Shi, Y., Jiang, H., Sun, Z. and Zheng, X. 2005. Identification of human fetal liver miRNAs by a novel method. *FEBS Lett.* **579:** 3849-3854.

Gardner, P.P., Wilm, A., and Washietl, S. 2005. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nuc.Acid.Res.* **33(8):** 2433-2439.

Griffiths-Jones, S., Bateman, A., Marshall, M., Khanna, A., Eddy, S.R. 2003. Rfam: an RNA family database. *Nucleic Acids Res.* **31:** 439-441.

Griffiths-Jones, S. 2004. The microRNA Registry. *Nucleic Acids Res.* **32:** Database Issue, D109-D111

Griffiths-Jones, S., Grocock, R.J., van Dongen, S., Bateman, A. and Enright, A.J. 2006. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res.* **34:** Database Issue, D140-D144

Harmanci, A.O., Sharma, G. and Mathews, D.H. 2007. Efficient pairwise RNA structure prediction using probabilistic alignment constraints in Dynalign. *Bioinformatics.* 8:130.

Harrow, J., F. Denoeud, A. Frankish, A. Reymond, C.K. Chen, J. Chrast, J. Lagarde, J.G. Gilbert, R. Storey, D. Swarbreck, C. Rossier, C. Ucla, T. Hubbard, S.E. Antonarakis, and R. Guigo. 2006. GENCODE: producing a reference annotation for ENCODE. *Genome Biol.* (Suppl 1) **7**: S4 1-9.

Havgaard, J.H., Torarinsson, E. and Gorodkin, J. 2007. Fast Pairwise Structural RNA Alignments by Pruning of the Dynamical Programming Matrix. *PLoS Comput. Biol.*, 3(10):e193.

Hertel, J. and Stadler, P.F. 2006. Hairpins in a Haystack: Recognizing miRNA Precursors in Comparative Genomics Data. *Bioinformatics* **22(14):** e197-202.

Holmes, I. 2005. Accelerated probabilistic inference of RNA structure evolution. *BMC Bioinformatics* **6:** 73.

Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M. and Schuster, P. 1994. Fast folding and comparison of RNA secondary structures. *Monatsh. Chem.* **125**:167-188.

Kao, H.T., B. Porton, A.J. Czernik, J. Feng, G. Yiu, M. Haring, F. Benfenati, and P. Greengard. 1998. A third member of the synapsin gene family. *Proc Natl Acad Sci U S A* **95**: 4667-4672.

Kent, W.J., Sugnet, C. W., Furey, T. S., Roskin, K.M., Pringle, T. H., Zahler, A. M., and Haussler, D. 2002. The Human Genome Browser at UCSC. *Genome Res.* **12(6)**, 996-1006.

Kapranov, P., Cheng, J., Dike, S., Nix, D.A., Duttagupta, R., Willingham, A. T., Stadler, P. F., Hertel, J., Hackermüller, J., Hofacker, I.L., Bell, I., Cheung, E., Drenkow, J., Dumais, E., Patel, S., Helt, G., Ganesh, M., Ghosh, S., Piccolboni, A., Sementchenko, V., Tammana, H., and Gingeras, T.R. 2007. RNA Maps Reveal New RNA Classes and a Possible Function for Pervasive Transcription. *Science* **316**(5830):1484-1488.

Leco, K.J., R. Khokha, N. Pavloff, S.P. Hawkes, and D.R. Edwards. 1994. Tissue inhibitor of metalloproteinases-3 (TIMP-3) is an extracellular matrix-associated protein with a distinctive pattern of expression in mouse cells and tissues. *J Biol Chem* **269:** 9352-9360.

Lunter, G., Ponting, C.P., and Hein, J. 2006. Genome-wide identification of human functional DNA using a neutral indel model. *PLoS Comput. Biol.* 2(1): e5.

Margulies, E.H., Chen, C.W. and Green, E.D. 2006. Differences between pair-wise and multi-sequence alignment methods affect vertebrate genome comparisons. *Trends Genet.* **22:** 187-193.

Margulies, E.H., Cooper, G.M., Asimenos, G., Thomas, D.J., Dewey, C.N., Siepel, A., Birney, E., Keefe, D., Schwartz, A.S., Hou, M., et al. 2007. Analyses of deep mammalian sequence alignments and constraint predictions for 1% of the human genome. *Genome Res*. **17:** 746-759.

Pedersen, J.S., Bejerano, G., Siepel, A., Rosenbloom, K., Lindblad-Toh, K., Lander, E., Rogers, J., 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.

Pollard, K.S., S.R. Salama, N. Lambert, M.A. Lambot, S. Coppens, J.S. Pedersen, S. Katzman, B. King, C. Onodera, A. Siepel et al. 2006. An RNA gene expressed during cortical development evolved rapidly in humans. *Nature* **443**: 167-172.

Ravasi, T., H. Suzuki, K.C. Pang, S. Katayama, M. Furuno, R. Okunishi, S. Fukuda, K. Ru, M.C. Frith, M.M. Gongora 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.

Sankoff, D. 1985. Simultaneous solution of the RNA folding, alignment and protosequence problems. *SIAM. J. Appl. Math.* **45:** 810-825.

Sasaki, Y.T., M. Sano, T. Ideue, T. Kin, K. Asai, and T. Hirose. 2007. Identification and characterization of human non-coding RNAs with tissue-specific expression. *Biochem Biophys Res Commun.* **357**: 991-996.

Seemann, S.E., Gilchrist, M.J., Hofacker, I.L, Stadler, P.F. and Gorodkin, J. 2007. Detection of RNA structures in porcine EST data and related mammals. *BMC Genomics* **8**:316.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L. W., Richards, S., et al. 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. **15:** 1034-1050.

Smit, A.F.A., Hubley, R. and Green, P. 1996-2004. RepeatMasker Open-3.0. <http://www.repeatmasker.org>.

Sone, M., T. Hayashi, H. Tarui, K. Agata, M. Takeichi, and S. Nakagawa. 2007. The mRNA-like noncoding RNA Gomafu constitutes a novel nuclear domain in a subset of neurons. *J Cell Sci* **120**: 2498-2506.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice, *Nucleic Acids Res*. **22:** 4673-4680.

Torarinsson E., Havgaard, J.H., Gorodkin, J. 2007. Multiple structural alignment and clustering of RNA sequences. *Bioinformatics*, **23**(8):926-32.

Torarinsson, E. Sawera, M., Havgaard, J.H., Fredholm, M. and Gorodkin, J. 2006. Thousand of corresponding human and mouse genomic regions unalignable in primary sequence contain common RNA strucuture. *Genome Res.* **16:** 885-889.

Wang, A.X., Ruzzo, W.L. and Tompa, M. 2007. How accurately is ncRNA aligned within whole-genome multiple alignments? *BMC Bioinformatics*, To Appear.

Washietl, S. and Hofacker, I.L. 2004. Consensus Folding of Aligned Sequences as a New Measure for the Detection of Functional RNAs by Comparative Genomics. *JMB*. **342:** 19-30.

Washietl, S., Hofacker, I.L. and Stadler, P.F. 2005a. Fast and reliable prediction of noncoding RNAs. *Proc. Natl. Acad. Sci. USA* **102:** 2454-2459.

Washietl, S., Hofacker, I.L., Lukasser, M., Hüttenhofer, A. and Stadler, P.F. 2005b. Mapping of conserved RNA secondary structures predicts thousands of functional non-coding RNAs in the human genome. *Nature Biotech.* **23:** 1383-1390.

Washietl, S., Pedersen, J.S., Korbel, J.O., Gruber, A.R., Hackermuller, J., Hertel, J., Lindemeyer, M., Reiche, K., Stocsits, C., Tanzer, A., et al. 2007. Structured RNAs in the ENCODE Selected Regions of the Human Genome. *Genome Research* **17:** 852-864.

Weinberg, Z., Barrick, J.E., Yao, Z., Roth, A., Kim, J. N. , Gore, J. , Wang, J.X., Lee, E.R., Block, K.F., Sudarsan, N., Neph, S., Tompa, M., Ruzzo, W.L., and Breaker, R.R. 2007. Identification of 22 candidate structured RNAs in bacteria using the CMfinder comparative genomics pipeline. *Nucleic Acids Research* **35(14)**: 4809-19.

Weinberg, Z. and Ruzzo, W.L. 2006. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics,* **22**(1):35-39.

Westhof, E. and Michel, F. 1994. In Nagai, K. and Mattaj, I.W. (eds.) *RNA-Protein Interactions*. Oxford University Press, pp. 26-51.

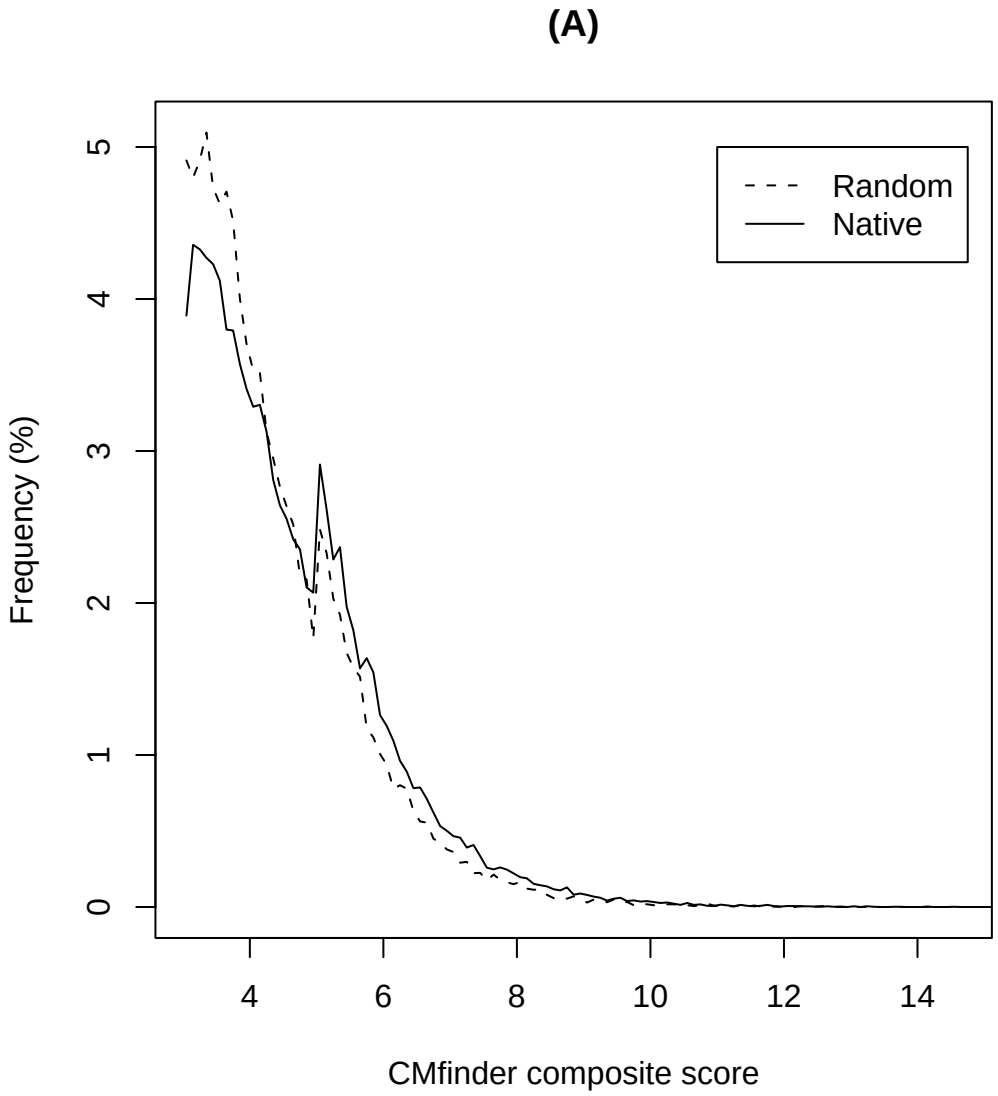
Westhof, E. et al. 1996 In Bihop, M.J. and Rawlings, C. J. (eds.) *DNA-Protein Sequence Analysis.* Oxford University Press, pp. 255-278.

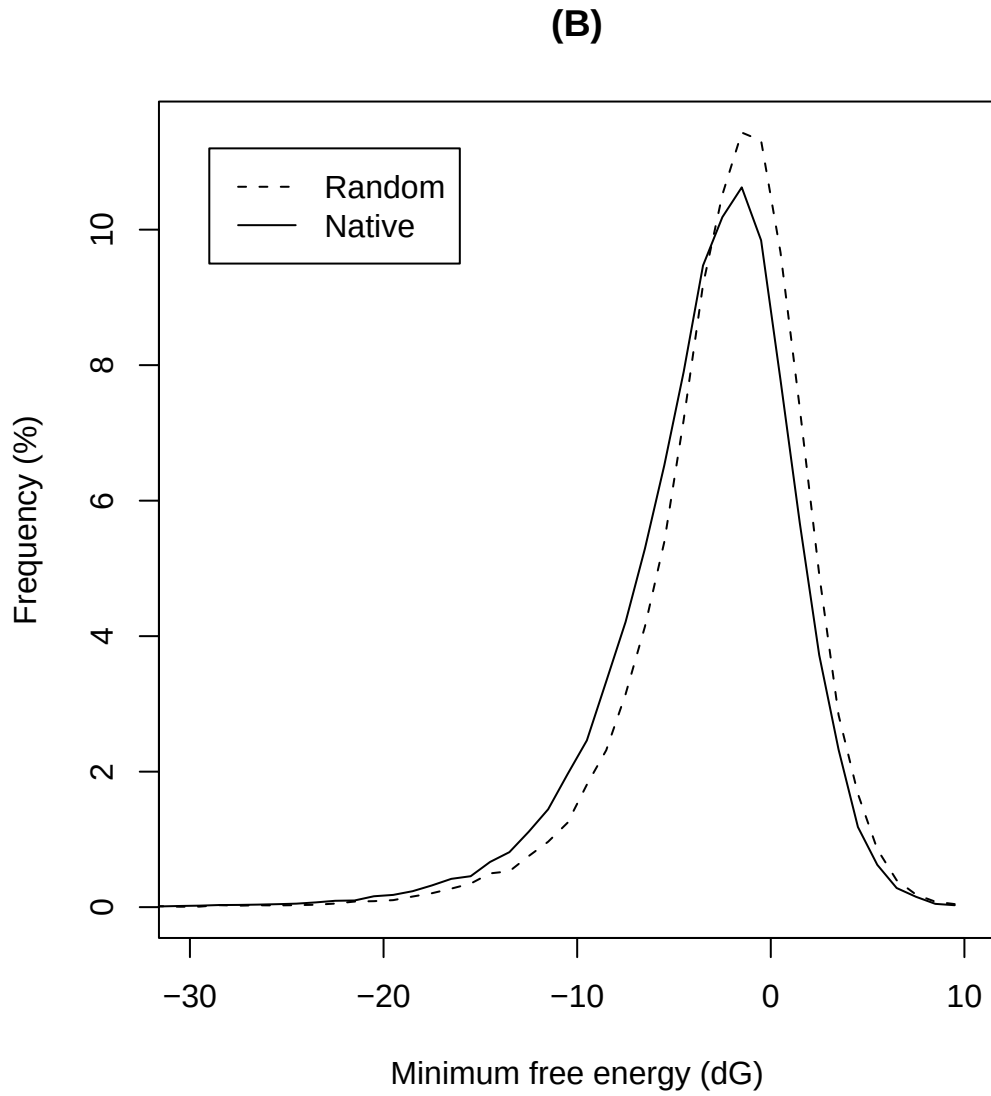
Will, S., Reiche, K., Hofacker, I.L., Stadler, P.F. and Backofen, R., 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS Comput Biol.* **3**(4):e65.

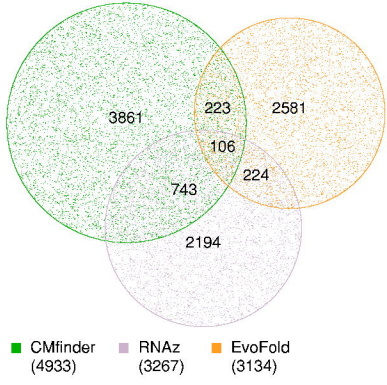
Yao, Z., Barrick, J.E., Weinberg, Z., Neph, S., Breaker, R.R, Tompa, M. and Ruzzo, W.L. 2007. A Computational Pipeline for High Throughput Discovery of *cis*-Regulatory Noncoding RNA in Prokaryotes. *PLoS Computational Biology* 3(7): e126.

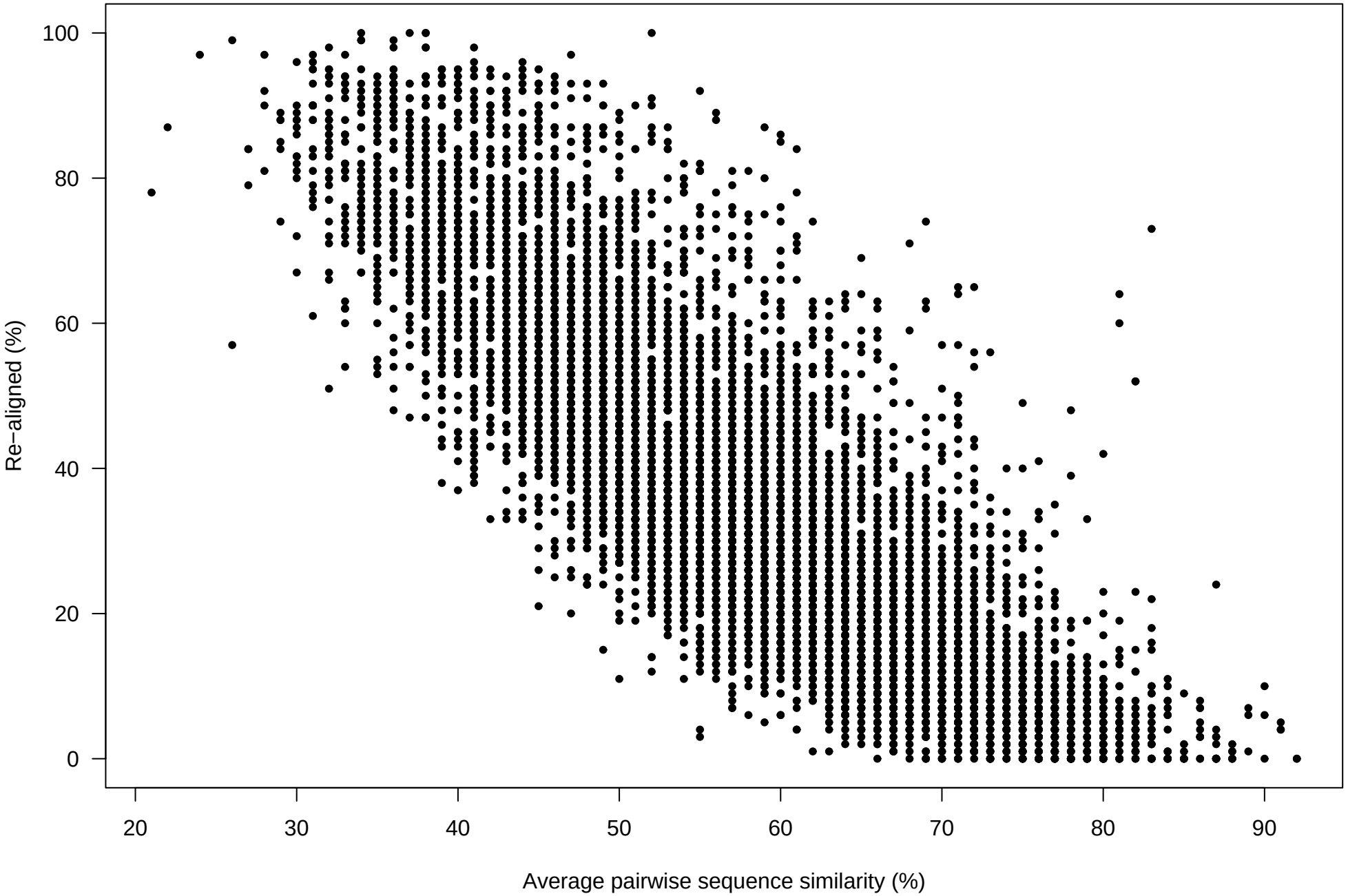
Yao, Z., Weinberg, Z. and Ruzzo, W.L. 2006. CMfinder - A Covariance Model Based RNA Motif Finding Algorithm. *Bioinformatics* **22:** 445-452.

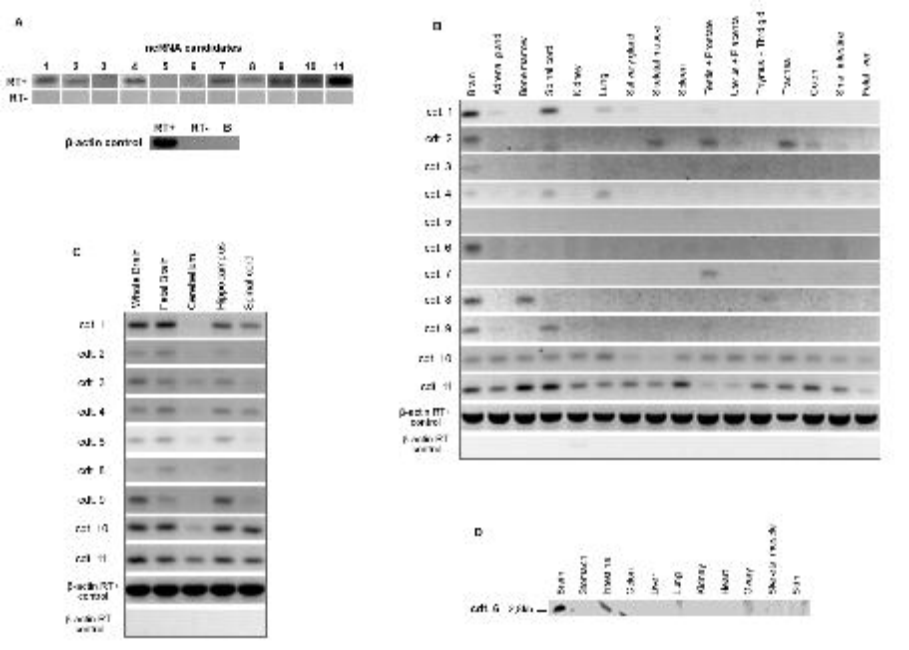
Zuker, M. 2003. Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Res.* **31:** 3406-3415.


(A)
Random
Native
Frequency (%)
0
1
2
3
4
5
4
6
8
10
12
14
CMfinder composite score


(B)
Random
Native
Frequency (%)
0
2
4
6
8
10
−30
−20
−10
0
10
Minimum free energy (dG)


3861
223
2581
106
224
743
2194
CMfinder
(4933)
RNAz
(3267)
EvoFold
(3134)


100
80
60
40
20
0
Re−aligned (%)
20
30
40
50
60
70
80
90
Average pairwise sequence similarity (%)


A
mRNA candidates
1
2
3
4
5
6
7
8
9
10
11
RT+
RT-
β-actin control
RT+
RT-
B
Brain
Adrenal gland
Bone marrow
Cerebellum
Kidney
Lung
Mammary gland
Skeletal muscle
Spleen
Testis
Thymus
Trachea
Uterus
Whole embryo
Pooled
cat 1
cat 2
cat 3
cat 4
cat 5
cat 6
cat 7
cat 8
cat 9
cat 10
cat 11
β-actin RT+ control
β-actin RT- control
C
Whole Brain
Fetal Brain
Cerebellum
Hippocampus
Skeletal muscle
cat 1
cat 2
cat 3
cat 4
cat 5
cat 6
cat 7
cat 8
cat 9
cat 10
cat 11
β-actin RT+ control
β-actin RT- control
D
Brain
Stomach
Intestine
Colon
Liver
Lung
Kidney
Heart
Ovary
Skeletal muscle
Skin
cat 6