# Exploring genomic dark matter: homology search for non-coding RNA

Eva K. Freyhult[1], Jonathan P. Bollback[2] & Paul P. Gardner[2,3]

[1] *The Linnaeus Centre for Bioinformatics, Uppsala University, Box 598,75124 Uppsala, Sweden.*
[2] *Molecular Evolution Group, Institute of Molecular Biology and Physiology, University of Copenhagen, Universitetsparken 15, Copenhagen, Denmark.*
[3] *Correspondence should be addressed to PPGardner@bi.ku.dk*

**Homology search is one of the most common bioinformatic tasks, yet it is unknown how effective the currently available tools are for identifying noncoding RNAs (ncRNAs). In this work we use reliable ncRNA datasets to assess the effectiveness of methods such as BLAST, FASTA, HMMer and Infernal. Surprisingly, the most popular homology search methods are often among the least accurate. This means that many studies have used inappropriate tools for their analyses. Based upon our results we suggest homology search strategies using the currently available tools and some directions for future development.**

Compared to the relatively trivial task of protein homology search, ncRNA homology search is more challenging by the fact that intra- and inter- molecular basepairs are, in evolutionary terms, preserved to a higher degree than the sequence. The wobble $GU$ and other non-canonical basepairs allow RNA sequences to evolve seemingly unrelated sequences along nearly neutral paths through structure space (e.g. $A \cdot U \leftrightarrow G \cdot U \leftrightarrow G \cdot C$). Thus, specialized homology search techniques such as nucleotide specific scoring schemes[1], profile hidden Markov model methods (profile HMMs)[2,3] and covariance models (CMs)[4] are necessary for accurate ncRNA homology search.

The goal of this study is to identify programs which balance sensitivity (true predictions) and specificity (false predictions) for practical ncRNA homology search situations. We use large high quality ncRNA datasets and randomized controls to test the 12 homology search programs summarized in Table 1. These programs fall into one of three classes: sequence based methods, profile HMM methods and structure enhanced methods. In addition, we extend the use of ancestral sequence reconstructions (ASR) and predictive sequence reconstruction (PSR) for use in homology searches[5] to the RNA homology search problem (see Supplementary Fig. 1).

The most popular homology search methods are sequence based. The local matching of two sequences has been solved by Smith and Waterman in a mathematically optimal fashion using a dynamic programming procedure[6]. However, this method is too slow for the majority of practical homology search situations, where the database length is large. Hence heuristic methods such as BLAST and FASTA which speed the search procedure, at a cost to accuracy, are often used.

Profile hidden Markov model methods have been used for detecting patterns in multiple sequences[2,3]. An input alignment is used to build a probabilistic model which is used to scan a database for homologous sequences. The basic concept of profile HMMs can be understood by considering nucleotide frequencies in each column of an alignment. In the absence of gaps, the probability that a given sequence is generated by the same evolutionary processes as those in the alignment can be estimated by the product of position specific nucleotide frequencies. The architecture proposed by Krogh *et al.*[3] (see Supplementary Fig. 2) allows for insertions and deletions in the model and in addition deletions can be modeled in a position dependent manner. To account for overrepresented sequences in the input alignment, tree weighting schemes can be used[7] and there are schemes to avoiding overfitting and to account for unobserved data in the input[8].

Structure enhanced methods are frequently based on covariance models (CMs) which are an analogue of profile HMMs that include pairwise interactions due to RNA secondary structure. Where profile HMMs consist of a linear HMM architecture suitable for modeling linear protein sequences, tree-like CMs model tree-like RNA secondary structures that allow for basepairing interactions. States within the CM capture paired and unpaired regions while allowing insertions and deletions. To picture this, imagine the profile HMM model in Supplementary Figure 2 with basepairs between distant sites. Several new states need to be added to the model to accommodate this more complex structure. In the paired sites deletions now include either a single 5' or 3' base or the entire basepair and insertions can now be between either the 5' or 3' ends of a basepair. Bifurcation states are also included in the CM to allow for multiloops. The basic CM search procedure is analogous to the use of profile HMMs. An alignment replete with a structure annotation is provided by the user and used to train a CM that is specific to the input data which can be used to search a query database[4,7].

**Box 1: Caveats to algorithm assessments**

There are several limitations to any algorithm assessment. We outline the most important issues below.

*Test datasets:* We take for granted the accuracy of structural alignments taken from the literature, many of which have been constructed using the programs we are studying. However, given this limitation the analysis of a large and diverse dataset should outweigh any possible errors due to dataset inaccuracies. In addition, the datasets consist of a limited class of RNA families, the performance of some programs may improve on a different selection of datasets.

*Tool abuse:* Frequently researchers may apply a tool to a task that it is not designed for. For example, in this study we have applied sequence based tools to structured ncRNAs, assuming that sites are independent. This is a common but poor assumption.

*Tools improve:* Many of the tools tested here are recent developments and are still under active development. Hence, not all observations will remain reproducible. In fact, we hope this study helps improve future performance.

*Parameter settings:* The performance of some of the programs may benefit from optimizing program parameters. Here we have attempted to capture the essential features of each algorithm by using as many parameter combinations as was practical.

We contacted the authors of each of the programs included in this study in September 2005 (see Acknowledgments). We provided access to the datasets, scripts and a pre-print of the article for the authors from the BRaliBase website (www.binf.ku.dk/~pgardner/bralibase). We found their comments invaluable for minimizing the costs of the above caveats.

## Methods

In the following section we outline the datasets we used for this study and the approaches we used to compare sequence based, profile HMM and structure enhanced methods.

**Datasets** In order to test homology search tools we have obtained 602 5S rRNAs, 1114 tRNAs and 235 U5 spliceosomal RNAs. These hand-curated sequence alignments have mean lengths 117, 73 and 119 respectively[25–27]. From these, 583 subsets of five sequences and 360 subsets of twenty sequences have been sampled. The subsets were generated such that each sequence lies within one of the following five identity ranges from every other sequence in the subset: 40-60%, 50-70%, 60-80%, 70-85% and 80-95% (throughout this work identities are computed from the hand-curated alignments).

To measure algorithm sensitivity (defined in Box 2) we compute the number of matches to the database using only one subset of five or twenty sequences as input. Specificity (defined in Box 2) is measured using a shuffled database ten times larger than the curated databases, additionally the shuffling procedure preserves di-nucleotide frequencies[28].

**Thresholds** Score thresholds for each algorithm are optimized based on scans of the curated and shuffled tRNA, rRNA, or U5 databases with a smaller group of query sequences that uniformly covers the different RNA families and identity ranges. Example distributions and ROC plots are illustrated in Supplementary Figures 4&3. Raw scores rather than e-values are used here as there are a diverse number of methods implemented for computing e-values. In this study we are more concerned with the scores used by specific programs rather than the accuracy of the different e-value computations.

**65% scoring scheme** In order to compare the sequence based methods on an equal footing we have included a comparison of these using parameters optimized for datasets with 100-65% homology[1] (match=+5, mismatch=-4, gapopen=10, gapextension=10). Where a seed was required this was made as similar as possible, $W = 7$ for BLAST and `ktup` = 6 for FASTA.

**RNA centric scoring schemes** Several of the sequence based methods have associated scoring schemes which are designed for the unique problem of RNA homology search. Generally, these distinguish between transitions ($A \leftrightarrow G$ and $C \leftrightarrow U$), which are relatively frequent during RNA evolution, and transversions (the remaining mismatch types) which are relatively infrequent. WU-BLAST has a PUPY (purine-pyrimidine) score matrix (match=+4, transition=+2, transversion=-8). By default FASTA and SSEARCH score a $+5$ for matches and $-4$ for mismatches. Yet these tools have a "-U" scoring option that tolerates $G \cdot U$ wobble basepairs by scoring $G/A$ and $U/C$ mismatches as one less than a $G/G$ match in a strand specific manner. In addition, RSEARCH's RIBOSUM[22] and the more recent FoldAlign [29] score matrices use parameters estimated directly from the loop regions of large curated ncRNAs alignments. We have tested these as an alternative scoring scheme for FASTA homology searches.

**Genome scan** A ncRNA enriched region of the human genome was selected for further testing of representative homology search programs from each category. We identified a 40 MB region on chromosome 12 (coordinates 90,000,000-130,000,000; genome assembly NCBI35) that contains

5 5S rRNAs, 10 tRNAs (and 26 pseudo-tRNAs predicted by tRNAscan-SE) and 1 U5 spliceosomal RNA. We used input datasets containing ten sequences, each with a sequence identity to the associated target RNA in one of the following ranges 40-60%, 50-70%, 60-80% and 70-85%. The pairwise sequence identities within the datasets are between 60 and 90%.

**Timing** For the timing studies two databases of 166 megabases and 332 megabases, respectively, were used. Both databases contain 1114 tRNA sequences, the smaller database has one shuffled version of each of these, whereas the larger database has three. A single tRNA subset was used as a query for the timing study. The times for scanning the 2 databases are computed on (or calibrated to) a Sun Sparc v9 and 0.9 GHz CPU for each algorithm. From these values the algorithm speed (nts/s) and initialization times are computed.

**Phylogenetic sequence reconstruction** An empirical Bayesian approach[30] was used to stochastically sample ancestral sequences (ASR) from the internal nodes of a phylogeny describing the relationships among the query sequences (details of the sampling can be found in the Supplementary Methods). The phylogenetic tree and model parameters were estimated using MrBayes v3[31]. To accommodate the non-independence among sites arising from secondary structure an RNA doublet model was used to model substitutions in stem regions[31,32], while loop regions were modeled using the method of Hasegawa, Kishino, and Yano[33]. In addition to reconstructing ancestral sequences at the internal nodes of the phylogeny, a novel method was employed to sample ancestral sequences from un-observed lineages radiating from the internal nodes of the phylogeny using a Bayesian posterior predictive (PSR) approach (See Supplementary Methods)[34].

**Alignment and structure prediction** We use automatic structure prediction and alignment methods that previous studies have identified as being accurate[35,36]. The alignments are computed using ProAlign[37] and consensus structures are computed from these alignments using RNAalifold[38].

**Box 2: Performance Measures**

*Sensitivity* and *specificity* are common measures for determining the accuracy of homology search methods.

$$Sensitivity := \frac{TP}{TP + FN} \qquad Specificity := \frac{TN}{TN + FP}$$

where $TP$ is the number of "true positives", $TN$ is the number of "true negatives", $FN$ is the number of "false negatives" and $FP$ is the number of "false positives". $Sensitivity$ measures how much of the positive control dataset is recovered by the program in question, the $specificity$ measures what fraction of the rejected sequences that were correctly rejected.

A measure combining both specificity and sensitivity is useful for ranking programs. In previous studies, the *error rate* ($\frac{FP+FN}{TP+TN+FP+FN}$)[22] has been used, we however, favor the more discriminative *Matthews correlation coefficient* ($MCC$) as defined below:

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

$MCC$ ranges from $-1$ for extremely inaccurate ($TP = TN = 0$) to 1 for very accurate predictions ($FP = FN = 0$).

In general, we measure TP as the number of known RNA sequences that overlap with predicted RNA sequences by at least one nucleotide. For the genome scan, TP is instead measured as the number of nucleotides that are in both a known and a predicted RNA sequence. From this it follows how TN, FP and FN are computed. To make a distinction between the regular performance measures defined in this box and the ones used for the genome scan, we call the latter *nSensitivity*, *nSpecificity* and *nMCC*.

To ease the comparison of the different measures we have ranked each program with representative parameter settings against its peers using MCC values for each subset of query sequences.

Table 1: Program descriptions, URLs and references for each of the 12 programs used in this study.

| Program | Description | URL | Reference |
|---|---|---|---|
| **Sequence based methods** | | | |
| **NCBI-BLAST**<br>**WU-BLAST** | The query database is tabulated as short sequences (seeds) which is scanned during the initial search phase, short matches are subsequently extended and scored. | www.ncbi.nlm.nih.gov/BLAST<br>blast.wustl.edu | [9,10] |
| **FASTA** | FASTA employs a lookup table to identify all matching words of length `ktup`. Diagonals of mutually supporting matches are located, linked and extended. High scoring matches are finally realigned using a local banded Smith-Waterman algorithm[11] | fasta.bioch.virginia.edu | [12] |
| **ParAlign** | A parallel computing technology, SIMD (Single Instruction Multiple Data), is used to compute exact un-gapped alignments. A novel heuristic is used to compute gapped alignments for high scoring un-gapped hits. The highest scoring database matches are realigned using a rigorous SIMD-based Smith-Waterman algorithm. | www.paralign.org | [13] |
| **SSEARCH** | Implements the Smith-Waterman local alignment algorithm[6]. | fasta.bioch.virginia.edu | [14] |
| **Profile HMM methods** | | | |
| **HMMer** | A profile HMM approach with a novel architecture (Plan 7) that distinguishes between global and local alignments probabilistically and excludes transitions from insert to delete states and vice versa. | hmmer.wustl.edu | [15] |
| **SAM** | This package uses the original profile HMM architecture[3] discussed in the text and displayed in Supplementary Figure 2. | www.cse.ucsc.edu/research/compbio/sam.html | [16,17] |

| Program | Description | URL | Reference |
|---|---|---|---|
| **Structure enhanced methods** | | | |
| **ERPIN** | An input alignment with structure annotation is converted into a combination of single sequence and helical lod-score based weight matrices. These profiles can then be used to rapidly screen a database for matching helical profiles and classical dynamic programming for the alignment of single stranded regions. | tagc.univ-mrs.fr/erpin | [18] |
| **Infernal** | Implements a covariance model (CM) as discussed in the text and illustrated in Supplementary Figure 2. | www.genetics.wustl.edu/eddy/infernal | [19] |
| RAVENNA | Converts a CM generated by Infernal into a profile HMM. This is used to rapidly filter the database for high scoring matches which can be aligned using the slower but more accurate Infernal package. | bio.cs.washington.edu/supplements/zasha-ravenna | [20] |
| **RSEARCH** | Implements a CM for a single input sequence and structure. BLOSUM-like score matrices[21] called RIBOSUM matrices are used to score database sequence matches to helical or single stranded regions of the query. | www.genetics.wustl.edu/eddy/software/#rsearch | [22] |
| **RSmatch** | Input and database sequences are folded using RNAfold[23] (or similar). The structures are decomposed into subcomponents which are organized into a tree model and the database is screened for significant hits using a tree alignment procedure. The alignment is scored using a combination of basepair and single strand score matrices. | exon.umdnj.edu/software/RSmatch | [24] |

## Results

The following discussion contains a detailed summary of the results presented in Figures 1-3 and the Supplementary Tables 1-4 and Supplementary Figures 5-7. We begin by outlining the results for each method with the parameter settings that had the optimal ranking. Unless stated otherwise we focus upon the results for the subsets with just 5 sequences. Secondly, we outline the results for our secondary tests which includes a comparison of the sequence based methods (NCBI-BLAST, WU-BLAST, FASTA, ParAlign and SSEARCH) with identical scoring parameters. These scoring parameters are optimized for sequence identities ranging from 100-65%[1] and is referred to the 65% scoring scheme throughout this section. The other secondary tests we present are of RNA centric scoring schemes for sequence based methods, the application of phylogenetic sequence reconstruction to homology searching and a scan of a section of the human genome.

**Sequence based searches** The accuracies of WU-BLAST and NCBI-BLAST were similar when similar parameters were used, yet WU-BLAST was significantly faster (see Fig. 2). The default scoring schema used for NCBI-BLAST is tailored for sequences with 99% sequence homology, whereas WU-BLAST defaults are tailored for sequences with 65% sequence homology[1] which is more appropriate for our diverse ncRNA datasets (see Supplementary Tables 1 & 2). WU-BLAST has a more diverse array of options, including allowing a minimum seed length of 3 (W3) (compared to 7 (W7) for NCBI-BLAST). Hence the parameter settings producing the best accuracy for WU-BLAST are not implemented in NCBI-BLAST. However, shorter seed lengths did come at a significant cost to program speed.

A comparison of FASTA and WU-BLAST (W3) is complicated given that the median ranking of WU-BLAST (W3) was higher than that of FASTA for the 5 sequence input data, but the lower quartile of the WU-BLAST (W3) ranks was much lower than that of FASTA (see Fig. 2). Hence, most of the time WU-BLAST (W3) outperformed FASTA (the rankings reversed on the 20 sequence datasets). Yet, FASTA was significantly faster than WU-BLAST (W3) and compared well with NCBI-BLAST (W7,65%) in terms of speed.

ParAlign was the fastest of the homology search tools in this study. However, ParAlign had low sensitivity compared to both FASTA and BLAST, this was true also for the 65% scoring test (see Methods and Fig. 2 for results).

SSEARCH generally outperformed the other sequence based methods in terms of accuracy. However, WU-BLAST (W3) occasionally outperformed SSEARCH but WU-BLAST (W3) was significantly slower than SSEARCH. Both of these observations are surprising given that SSEARCH employs no heuristics to improve speed whereas WU-BLAST (W3) demands seeds matching at least three consecutive nucleotide positions, one would have expected the opposite to the results presented here.

**Profile HMMs** SAM usually outperformed HMMer in terms of accuracy, yet HMMer was significantly faster (see Fig. 1). Both HMMer and SAM outperformed the sequence based methods in terms of accuracy, furthermore, the speed of HMMer was comparable to some of the sequence based methods. The results for HMMer show that version 2.3.2 is slightly better than 1.8.4, this is in contrast to the HMMer documentation which suggests the opposite for nucleotide sequences (due to protein specific optimization). Earlier results based on protein datasets[39] showed that profile searches could be improved by using SAM models and HMMer searches. We observed no such improvement for nucleotide based results (see Supplementary Tables 1 & 2).

**Structure enhanced homology search** The CM based methods Infernal and RSEARCH both performed extremely well on these ncRNA datasets, providing predictions with very high sensitivity and specificity. These methods generally ranked either first or second in terms of the MCC for every search. However, there was a significant cost in terms of CPU, both take approximately 1 sec to search a kilobase using a 0.9 GHz processor. This is about 2 orders of magnitude slower than the profile HMM and sequence based methods.

The Infernal package was upgraded during the course of this study (version 0.55 was updated to 0.7). Sean Eddy and collaborators added Dirichlet mixtures[8] and effective sequence number scalings to the algorithm, resulting in a significant performance boost for both the 5 and 20 sequence datasets (see Supplementary Tables 1 & 2).

ERPIN predictions are very conservative for the small dataset or when sequence identity is high (frequently only the input dataset was recovered), resulting in high specificity yet low sensitivity predictions. However, the speed of ERPIN was comparable to the sequence based methods.

The results for RAVENNA were also rather good, the algorithm ranked third after Infernal and RSEARCH in terms of accuracy. The accuracy of RAVENNA when compared to the other profile HMM methods, HMMer and SAM, was excellent. The speed of RAVENNA was about the same magnitude as SAM, which is in good agreement with theory. However, RAVENNA requires a significant initialization time (approximately 25 minutes, see Supplementary Figure 5) from the overhead for calibrating the HMM to determining an appropriate threshold, therefore it is only economical to use RAVENNA on larger databases.

The speed of RSmatch was nearly an order of magnitude greater than that of the structure enhanced methods Infernal, RAVENNA and RSEARCH and the accuracy was much lower.

**65% scoring scheme** This study showed that the sequence based methods perform rather similarly when using comparable parameter settings (see Fig. 2, the results labeled “65%” in Supplementary Table 3, and Supplementary Figure 6). The non-heuristic method, SSEARCH, outranked

the other methods in all cases. This was followed by FASTA. The two incarnations of BLAST performed almost identically, however, WU-BLAST was significantly faster than NCBI-BLAST.

**RNA centric scoring schemes** Each of the RNA centric scoring schemes mentioned in the Methods section was trialed, the results are displayed in Supplementary Table 4, Figure 3 and supplementary figure 7. The scoring schemes we tested are the PUPY matrix that ships with WU-BLAST, the "-U" option for FASTA and the single sequence components of the score matrices used by RSEARCH [22] and FoldAlign [29]. These results were generally disappointing, none of the methods showed any improvement over less specific schemes when the RNA centric scores were used. In the case of the FoldAlign and RSEARCH score matrices this is justified as these matrices were built specifically for structural methods rather than the sequence based methods we have used here. We also trialed a transition/transversion scoring scheme optimized for 65% sequence identity[1] (data not shown), the results of this test were also disappointing. This indicates that a great deal more work is required before such scoring schemes can be used for practical RNA homology search.

**Application of phylogenetic sequence reconstruction to homology search** We found that the inclusion of ancestral sequence reconstructions (ASRs) improved the sensitivities of sequence based methods. This came at a cost to specificity, but in the cases of FASTA and particularly SSEARCH an overall increase in accuracy was observed (see Fig. 3). This proves that in principle this approach could be very useful in the future, particularly in those cases when high sensitivity searches are required.

We significantly improved the performance of ERPIN by including ASRs and particularly posterior predictive sequence reconstructions (PSRs). The optimal branch length for these searches was extremely long ($\approx 20$ substitutions per site). This implies that a good scheme for including priors could significantly improve ERPIN performance (the median sensitivity improved by about a factor of 17). Earlier versions (version$<= 0.55$) of Infernal also showed a significant improvement when PSRs were included (data not shown). However, once Dirichlet mixture priors[8] were added to the Infernal code (version$>= 0.6$) including the PSRs no longer boosted performance.

Generally, the inclusion of neither ASR nor PSR sequences improved the accuracy of the profile HMM or Infernal searches where tree weighting schemes and Dirichlet priors are used. The only sensible way to include ASR and PSR sequences with these methods is to replace the tree weighting and priors implemented with these. But this is outside the scope of this article.

**Genome scan** In order to test a representative set of algorithms in a realistic usage scenario we scan a 40 MB section of the human genome. This test had the additional advantage of providing a more accurate estimate of algorithm specificity. The results of the genome scan were in general agreement with the earlier results. This was a much harder test than than those presented earlier, this is compounded by the fact that the genome annotation we rely upon may not be completely accurate. However, HMMer performed surprisingly well on this test compared to both Infernal

and SAM. Infernal had the highest median MCC, yet had a slightly lower specificity than HMMer. Of the single sequence methods, both WU-BLAST and FASTA performed well. WU-BLAST had a higher sensitivity but a lower specificity (see Fig. 3 and Supplementary Fig. 8).

## Discussion

The most popular homology search methods did not necessarily perform the best in our study. These programs are optimized for rapid database searches with few false positives (high specificity), which is not always what the user requires. As a consequence many estimates of the amount of conserved DNA[40] and number of conserved ncRNAs[41,42] and non-conserved ncRNAs[43] are based on suboptimal homology search tools and hence likely to be inaccurate.

For reliable ncRNA searches using sequence based search methods RNA optimized PAM[44] and/or BLOSUM[21] style score matrices are desperately needed. There is sufficient data for computing these matrices freely available from sources such as the Rfam[26] and the rRNA databases[45,46] and the statistical methods for estimating these are well established. Additionally, given that base-pair stacking is important for RNA structure this signal may prove useful for RNA homology search and could be exploited by incorporating a di-nucleotide scoring scheme into the alignment procedure[47].

There are few heuristics at present for rapid profile HMM and CM based homology searches. One could, for example, apply the BLAST concept of a seed match to profile HMMs and CMs. A database could be rapidly scanned for short, ungapped matches to the model which could then be extended using the full profile HMM architecture, this should result in significant gains in speed at moderate costs to sensitivity. This, for example, could be achieved by using a local version of the rapid ERPIN program.

The results for the posterior predictive method we have outlined in this work are promising. The nucleotide distribution across the entire tree[48] could be used to replace the tree-weighting schemes currently employed by profile HMM methods[7]. Frequently the user knows the tree for the species in the database; To recover missing orthologs one could generate sequences at the missing tips conditioned on the observed data, performing homology searches with these sequences is likely to produce fruitful results.

The RSmatch algorithm relies on MFE structure predictions on a single sequence, which are known to be frequently inaccurate[35]. If the structure prediction phase for both the database and input sequences were based on comparative predictions, such as RNAalifold[38], the accuracy of this approach is likely to improve. In addition it could be used to cluster genome wide structure based ncRNA predictions[41,42].

Based upon our assessment of the currently available homology search tools we recommend a scheme using iterative rounds of the rapid sequence based methods such as WU-BLAST, FASTA or SSEARCH with sensible scoring schemes and a high threshold. These results can then be used to train CM models for Infernal searches to obtain more divergent sequences from lower scoring sequence based matches. Profile HMMs, particularly RAVENNA, could be used instead when CMs are not practical, for example, when the sequence length is greater than $\sim 200$ or the database is large. Throughout this work we have focused almost entirely on method accuracy, however, of frequent concern is the computation time for a search. For example, based on our timings with tRNA queries it would take Infernal approximately 96 days to search the human genome on a single 0.9 GHz processor, RAVENNA would take 40 hours, HMMer would take 9 hours, SSEARCH would take 4 hours and WU-BLAST (W7) just 4 minutes. Given the ready access many groups have to computing clusters it is reasonable to expect the more accurate methods to become popular in future.

Many of the currently available tools for ncRNA homology search tools are not yet performing as well as one would hope based on the results we have presented. Improvements in terms of accuracy and speed are needed. This is extremely important given the explosion of interest in ncRNAs generated by recently discovered ncRNAs such as miRNAs. Additionally, a current theory suggests that much of the apparent organism complexity not accounted for by corresponding expansions in the proteome can be attributed to regulation from the ncRNAome[49].

This study has serious implications for evolutionary studies that rely on programs tuned for high similarity sequences. Currently popular programs identify a biased set of homologs; only highly conserved homologs, that are under purifying or negative selection are likely to be discovered, while these fail to identify sequences under strong diversifying or positive selection. Clearly, this will result in underestimates of genome wide positive selection.

## Authors' contributions

PPG proposed the project. EKF and PPG contributed to dataset acquisition and algorithm testing. JPB developed the code for inferring ancestral and predictive sequences. PPG, EKF and JPB contributed to drafting the manuscript. All authors read and approved the final manuscript.

## Acknowledgments

The authors thank Sam Griffiths-Jones, David Ardell, Anders Krogh, Rasmus Nielsen and Zasha Weinberg for useful discussions. We also thank the homology-search algorithm developers Torbjørn Rognes, Bin Tian, William R. Pearson, Robert J. Klein, Zasha Weinberg, Stephen Altschul, Daniel Gautheret and Sean Eddy for taking the time to make useful comments on an early

draft of this manuscript, any remaining flaws are solely our responsibility. The high-performance computer clusters at UPPMAX and the University of Copenhagen Bioinformatics Centre were used to compute many of the results presented here. PPG is supported by a Carlsberg Foundation Grant (21-00-0680).

1. States, D. J., Gish, W. & Altschul, S. F. Improved sensitivity of nucleic acid database searches using application-specific scoring matrices. *METHODS: A companion to Methods in Enzymology* **3**, 66–70 (1991).

2. Haussler, D., Krogh, A., Mian, I. S. & Sjölander, K. Protein modeling using hidden Markov models: Analysis of globins. In *Proceedings of the Hawaii International Conference on System Sciences*, 792–802 (IEEE Computer Society Press, Los Alimitos, CA, 1993).

3. Krogh, A., Brown, M., Mian, I. S., Sjölander, K. & Haussler, D. Hidden Markov models in computational biology. Applications to protein modeling. *J. Mol. Biol.* **235**, 1501–1531 (1994).

4. Eddy, S. R. & Durbin, R. RNA sequence analysis using covariance models. *Nucleic Acids Res.* **22**, 2079–2088 (1994).

5. Collins, L. J., Poole, A. M. & Penny, D. Using ancestral sequences to uncover potential gene homologues. *Appl. Bioinformatics* **2**, 85–95 (2003).

6. Smith, T. & Waterman, M. Identification of common molecular subsequences. *J. Mol. Biol.* **147**, 195–197 (1981).

7. Durbin, R., Eddy, S. R., Krogh, A. & Mitchison, G. *Biological Sequence Analysis: Probabilistic models of protein and nucleic acids* (Cambridge University Press, 1998).

8. Sjölander, K. *et al.* Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. *Comput. Appl. Biosci.* **12**, 327–345 (1996).

9. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).

10. Gish, W. WU-BLAST 2.0 (1996-2005).

11. Chao, K. M., Pearson, W. R. & Miller, W. Aligning two sequences within a specified diagonal band. *Comput. Appl. Biosci.* **8**, 481–487 (1992).

12. Pearson, W. R. & Lipman, D. J. Improved tools for biological sequence comparison. *Proc. Natl. Acad. Sci. USA* **85**, 2444–2448 (1988).

13. Saebø, P. E., Andersen, S. M., Myrseth, J., Laerdahl, J. K. & Rognes, T. PARALIGN: rapid and sensitive sequence similarity searches powered by parallel computing technology. *Nucleic Acids Res.* **33**, 535–539 (2005).

14. Pearson, W. R. Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. *Genomics* **11**, 635–650 (1991).

15. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).

16. Hughey, R. & Krogh, A. Hidden Markov models for sequence analysis: extension and analysis of the basic method. *Comput. Appl. Biosci.* **12**, 95–107 (1996).

17. Karplus, K., Barrett, C. & Hughey, R. Hidden Markov models for detecting remote protein homologies. *Bioinformatics* **14**, 846–856 (1998).

18. Gautheret, D. & Lambert, A. Direct RNA definition and identification from multiple sequence alignments using secondary structure profiles. *J. Mol. Biol.* **313**, 1003–1011 (2001).

19. Eddy, S. R. A memory efficient dynamic programming algorithm for optimal structural alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3**, 18 (2002).

20. Weinberg, Z. & Ruzzo, W. L. Sequence-based heuristics for faster annotation of non-coding RNA families. *Bioinformatics* **22**, 445–452 (2006).

21. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **89**, 10915–10919 (1992).

22. Klein, R. J. & Eddy, S. R. RSEARCH: finding homologs of single structured RNA sequences. *BMC Bioinformatics* **4**, 44 (2003).

23. Hofacker, I. L., Fontana, W., Bonhoeffer, S. & Stadler, P. F. Fast folding and comparison of RNA secondary structures. *Monatshefte für Chemie* **125**, 167–188 (1994).

24. Liu, J., Wang, J. T., Hu, J. & Tian, B. A method for aligning RNA secondary structures and its application to RNA motif detection. *BMC Bioinformatics* **6**, 89 (2005).

25. Szymanski, M., Barciszewska, M. Z., Erdmann, V. A. & Barciszewski, J. 5S Ribosomal RNA Database. *Nucleic Acids Res.* **30**, 176–178 (2002).

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

27. Zwieb, C. The uRNA database. *Nucleic Acids Res.* **25**, 102–103 (1997).

28. Workman, C. & Krogh, A. No evidence that mRNAs have lower folding free energies than random sequences with the same dinucleotide distribution. *Nucleic Acids Res.* **27**, 4816–4822 (1999).

29. Hull Havgaard, J. H., Lyngsø, R., Stormo, G. D. & Gorodkin, J. Pairwise local structural alignment of RNA sequences with sequence similarity less than 40%. *Bioinformatics* **21**, 1815–1824 (2005).

30. Huelsenbeck, J. P. & Bollback, J. P. Empirical and hierarchical Bayesian estimation of ancestral states. *Syst. Biol.* **50**, 351–366 (2001).

31. Huelsenbeck, J. P. & Ronquist, F. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics* **17**, 754–755 (2001).

32. Schöniger, M. & von Haeseler, A. A stochastic model for the evolution of autocorrelated DNA sequences. *Mol. Phylogenet. Evol.* **3**, 240–247 (1994).

33. Hasegawa, M., Kishino, H. & Yano, T. Dating of the human-ape splitting by molecular clock of mitochondrial DNA. *J. Mol. Evol.* **21**, 160–174 (1985).

34. Bollback, J. P. Posterior mapping and predictive distributions. In Nielsen, R. (ed.) *Statistical Methods in Molecular Evolution*, 189–203 (Springer Verlag New York, Inc. New York, USA, 2005).

35. Gardner, P. P. & Giegerich, R. A comprehensive comparison of comparative RNA structure prediction approaches. *BMC Bioinformatics* **5**, 140 (2004).

36. Gardner, P. P., Wilm, A. & Washietl, S. A benchmark of multiple sequence alignment programs upon structural RNAs. *Nucleic Acids Res.* **33**, 2433–2439 (2005).

37. Löytynoja, A. & Milinkovitch, M. C. A hidden Markov model for progressive multiple alignment. *Bioinformatics* **19**, 1505–1513 (2003).

38. Hofacker, I., Fekete, M. & Stadler, P. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* **319**, 1059–1066 (2002).

39. Madera, M. & Gough, J. A comparison of profile hidden Markov model procedures for remote homology detection. *Nucleic Acids Res.* **30**, 4321–4328 (2002).

40. Hillier, L. W. *et al.* Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. *Nature* **432**, 695–716 (2004).

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

42. Pedersen, J. S. *et al.* Identification and Classification of Conserved RNA Secondary Structures in the Human Genome. *PLoS Comput. Biol.* **2**, 251–262 (2006).

43. Pang, K. C., Frith, M. C. & Mattick, J. S. Rapid evolution of noncoding RNAs: lack of conservation does not mean lack of function. *Trends Genet.* **22**, 1–5 (2006).

44. Dayhoff, M., Schwartz, R. & Orcutt, B. *Atlas of Protein Sequence and Structure*, vol. 5, chap. A model of evolutionary change in proteins, 345–352 (1978).

45. Wuyts, J., Van de Peer, Y., Winkelmans, T. & De Wachter, R. The European database on small subunit ribosomal RNA. *Nucleic Acids Res.* **30**, 183–185 (2002).

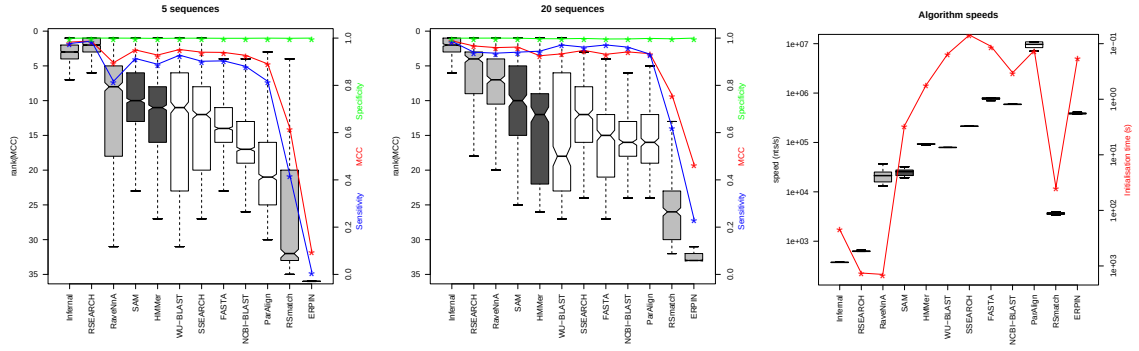
46. Cannone, J. *et al.* The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC Bioinformatics* **3**, 2 (2002).

47. Lunter, G. & Hein, J. A nucleotide substitution model with nearest-neighbour interactions. *Bioinformatics* **20**, 216–216 (2004).

48. Nielsen, R. Mapping mutations on phylogenies. *Syst. Biol.* **51**, 729–732 (2002).

49. Mattick, J. Non-coding RNAs: the architects of eukaryotic complexity. *EMBO Reports* **2**, 986–991 (2001).

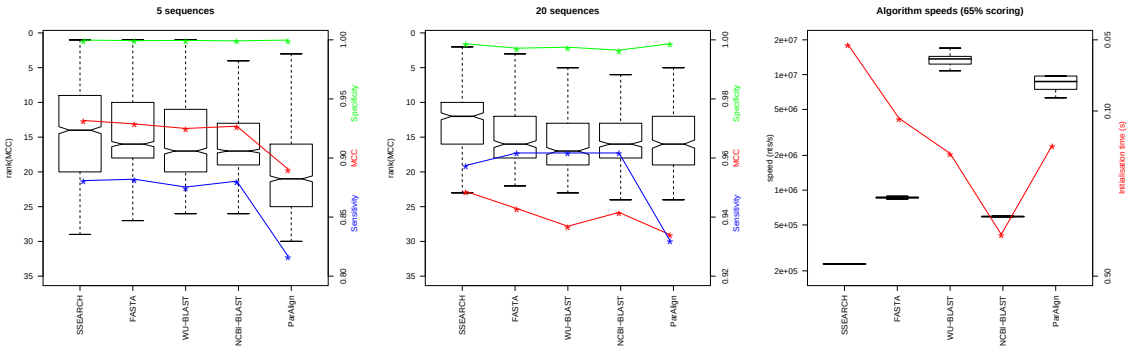
## Figures




(a) (b) (c)

Figure 1: A comparison of the accuracy and efficiencies of homology search methods showing only the highest ranking parameter settings for each algorithm from Supplementary Table 1. These were NCBI-BLAST (W7,65%), WU-BLAST (W3), FASTA, ParAlign (65%), SSEARCH, HMMer (2.3.2,local), SAM (3.5,local), ERPIN, Infernal (0.7,local), RAVENNA, RSEARCH, and RSmatch. (a) and (b) Boxplots of algorithm ranks for the 5 and 20 sequence subsets respectively. The blue curves show the median sensitivity, the green curve the median specificity and the red curve the median MCC for each of the 12 programs. (c) Boxplots of algorithm speeds in nucleotides per second. The red curve shows median initialization times for the different programs.

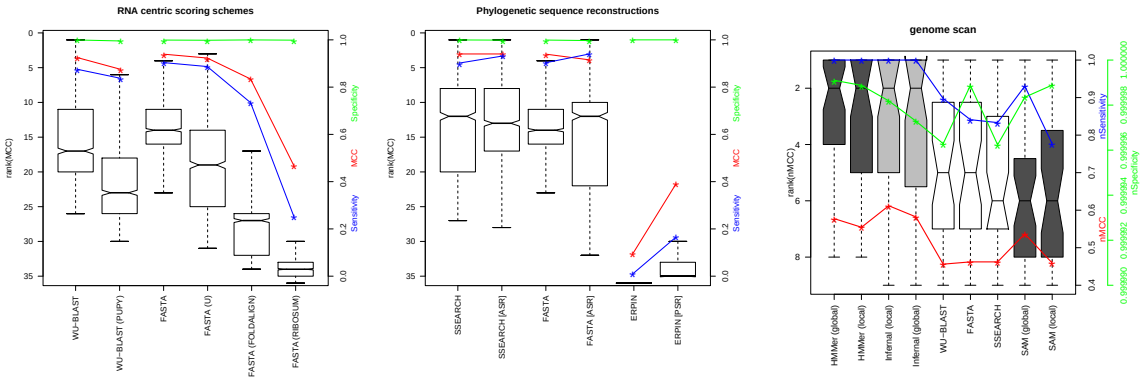
**Figure 1 - A comparison of the accuracy and speed of homology search methods**




(a) (b) (c)

Figure 2: A comparison of the accuracy of sequence based methods with identical scoring parameters. These boxplots show the distributions of the ranks on MCC and timing data for each of the homology search methods when using a scoring scheme optimized for nucleotide sequences with 65% identity (match=+5, mismatch=-4, gapopen=10, gapextension=10). (a) and (b) Boxplots of algorithm ranks for the 5 and 20 sequence subsets respectively. The blue curves show the median sensitivity, the green curve the median specificity and the red curve the median MCC for each of the 12 programs. (c) Boxplots of algorithm speeds in nucleotides per second. The red curve shows median initialization times for the different programs.

**Figure 2 - A comparison of the accuracy of sequence based methods with the 65% scoring scheme**




(a) (b) (c)

Figure 3: (a) A comparison of the accuracy of sequence based methods with score matrices optimized for ncRNA. These boxplots show the distributions of the ranks on MCC for each of the homology search methods when using one of WU-BLAST (W7), WU-BLAST (W7,PUPY), FASTA, FASTA (U), FASTA (RIBOSUM) or FASTA (FOLDALIGN). These matrices are discussed in more detail in the text. (b) A comparison of FASTA, SSEARCH and ERPIN with and without phylogenetic sequence reconstructions included in the input. Ancestral sequence reconstruction (ASR) were used in the case of FASTA and SSEARCH and posterior predictive sequences in the case of ERPIN. Both (a) and (b) show results using 5 query sequences. (c) A set of representative programs from each category were run on human chromosome 12 (coordinates 90,000,000-130,000,000; ver NCBI35). The boxplot displays algorithm ranks, additionally median nMCC, median nSensitivity and median nSpecificity for each algorithm are displayed using the y-axis on the right.

**Figure 3 - A comparison of the accuracy of methods using RNA centric scoring matrices, phylogenetic sequence reconstructions and the genome scan results.**