Genome Biology

METHOD　　　　　　　　　　　　　　　　　　　　　　　　　　**Open Access**

CrossMark

DotAligner: identification and clustering of RNA structure motifs

Martin A. Smith[1,2*†], Stefan E. Seemann[3,4†], Xiu Cheng Quek[1,2] and John S. Mattick[1,2]

Abstract

The diversity of processed transcripts in eukaryotic genomes poses a challenge for the classification of their biological functions. Sparse sequence conservation in non-coding sequences and the unreliable nature of RNA structure predictions further exacerbate this conundrum. Here, we describe a computational method, DotAligner, for the unsupervised discovery and classification of homologous RNA structure motifs from a set of sequences of interest. Our approach outperforms comparable algorithms at clustering known RNA structure families, both in speed and accuracy. It identifies clusters of known and novel structure motifs from ENCODE immunoprecipitation data for 44 RNA-binding proteins.

Keywords: Functions of RNA structures, RNA structure clustering, Machine learning, RNA–protein interactions, Functional genome annotation, Regulation by non-coding RNAs

Background

As genomic technologies progress, an ever-increasing number of non-protein-coding RNAs (ncRNAs) are being discovered. Long non-coding RNAs (lncRNAs) are of particular interest for functional genome annotation given their abundance throughout the genome. So far, few lncRNAs have been functionally characterised, and those that have seem to be involved in regulation of gene expression and epigenetic states [1, 2]. Understanding the molecular mechanisms underlying the biological functions of lncRNAs – and how they are disrupted in disease – is required to improve the functional annotation of the human genome.

Many ncRNAs lack sequence conservation, in contrast to protein-coding genes. Most small ncRNAs have well characterised secondary and tertiary structures, as evidenced in Rfam, the largest collection of curated RNA families (2588 families as of version 12.2 [3]). In contrast, determining the structural features of lncRNAs is a complex problem given their size and, in general, faster evolutionary turnover. These challenges have raised

doubts concerning the prevalence of functional structural motifs in lncRNAs [4, 5], despite evolutionary and biochemical support for conserved base pairing interactions [6–8]. Nonetheless, the higher-order structure of RNA molecules is an essential feature of ncRNAs, which can be used for their classification and the inference of their biological function.

We, and others, hypothesise that lncRNAs act as scaffolds for the recruitment of proteins and assembly of ribonucleoproteins (RNPs), mediated by the presence of modular RNA structures, akin to the domain organisation of proteins [6, 9–14]. Protein-interacting regions of lncRNAs are likely to contain a combination of sequence and structure motifs that confer binding specificity, which may be present in multiple target transcripts. For example, there is evidence that sequence and structure components of transposable elements, which are frequent in lncRNAs [15, 16], have been co-opted into mammalian gene regulatory networks [17, 18]. Identifying and annotating the genomic occurrence of homologous RNA structure motifs from sets of biologically related sequences will improve our understanding of the structure–function relationship of lncRNAs and the molecular mechanisms underlying their regulatory features. Resolving this challenge can be beneficial for the analysis of high-throughput RNA sequencing experiments that measure how RNAs interact

*Correspondence: m.smith@garvan.org.au
†Equal contributors
[1]RNA Biology and Plasticity Group, Garvan Institute of Medical Research, 384 Victoria Street, NSW 2010 Sydney, Australia
[2]St Vincent's Clinical School, Faculty of Medicine, UNSW Australia, NSW 2010 Sydney, Australia
Full list of author information is available at the end of the article

Smith *et al. Genome Biology*   (2017) 18:244

Page 2 of 12

with other molecules, such as cross-linked RNA immuno-precipitation and RNAse footprinting methodologies.

The identification of RNAs with similar functions involves comparing both their primary sequence and higher-order structures simultaneously. However, sequence-based methods to identify common structural features perform poorly when sequence identity falls below 60 % [19]. Hence, methods are needed that find structural similarity independent from sequence conservation and freed from single RNA secondary structure predictions. The Sankoff algorithm resolves the optimal sequence-structure alignment of two RNAs [20], but its computational complexity limits its practicality. Its most comparable implementation, FoldAlign, employs a minimum free energy-based strategy with pruning of the associated dynamical programming matrix [21, 22]. Alternative strategies often employ pre-calculated secondary structure probability distributions (thermodynamically equilibrated canonical ensembles) for each sequence [23]. These can substantially speed up the calculation of structure-based alignments [24], of which there are many variants. The programs PMcomp [24], LocaRNA [25] and ProbAlign [26] use the pre-computed base pair probability matrices of two sequences and score the alignment based on the notion of a common secondary structure. The sequence-structure alignment problem is reduced to a two-dimensional problem in RNApaln [27] and StrAL [28], which derive probabilities for individual bases (such as the probability of being unpaired) from all base pairing probabilities. These methods all fail to consider explicitly suboptimal structures in the alignment. The pairwise alignment of entire base pairing probability matrices (RNA *dot plots*) was first introduced by CARNA [29, 30], which iteratively improves alignments using a constraint programming technique implementing a branch and bound scheme.

These pairwise RNA structure alignment algorithms can be used to identify clusters of homologous RNA structure motifs from a set of sequences of interest. Will et al. first showed that a (dis)similarity matrix can be constructed from all-vs-all pairwise RNA structure alignments with the pairwise alignment tool LocaRNA, identifying known and novel groups of homologous RNAs using hierarchical clustering [25]. However, this strategy involves applying a subjective threshold to the resulting dendrogram to extract structurally related sequences. Alternative approaches to all-vs-all pairwise comparisons for RNA structure clustering include NoFold, which clusters query sequences based on their relative similarity to a panel of reference structure motif profiles [31], and GraphClust, an alignment-free approach that decomposes RNA structures into graph-encoded features [32]. RNAscClust, an extension of GraphClust, utilises

the evolutionary signatures of RNA structures (when available) as an additional classification feature [33].
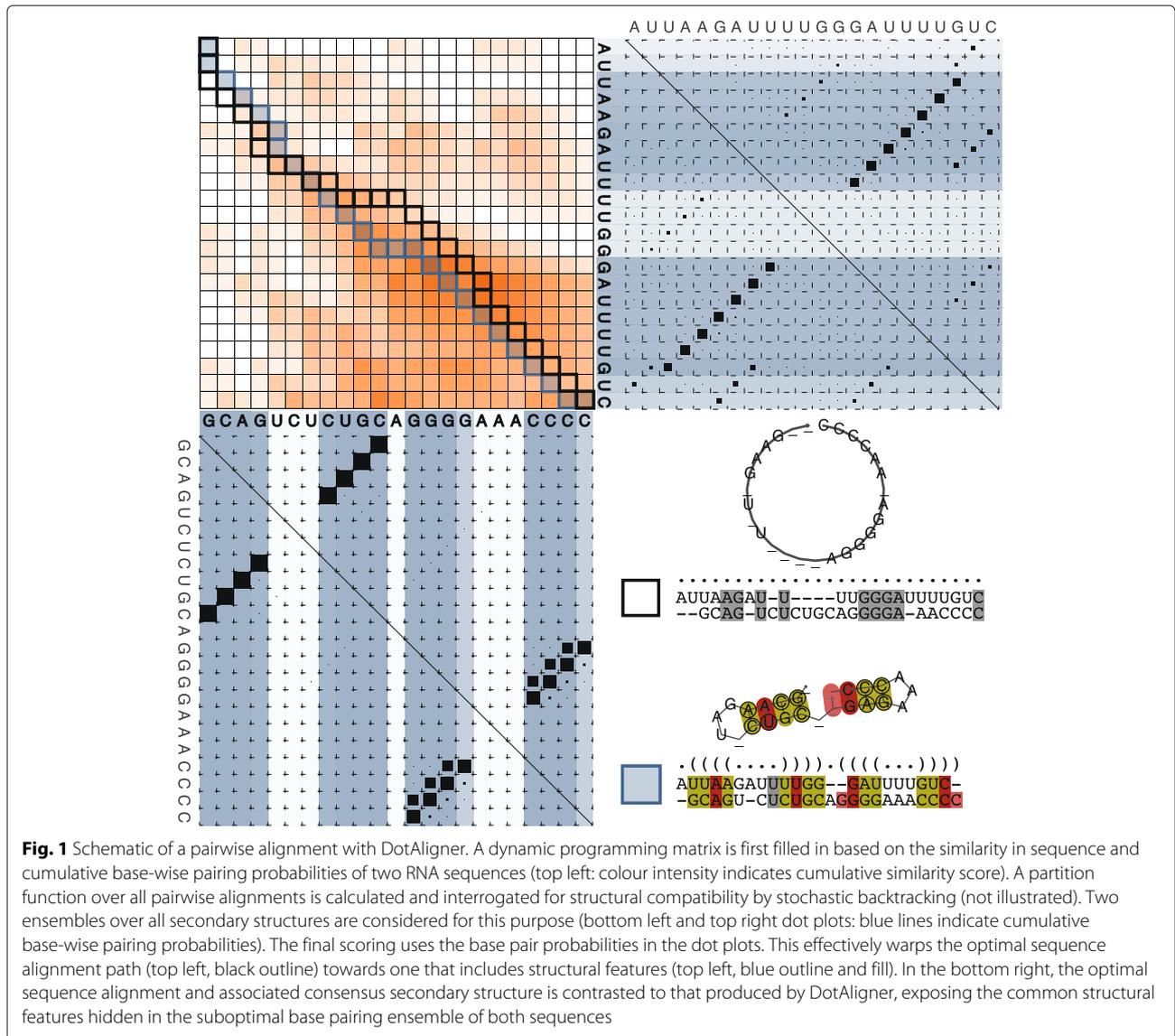
Here, we describe a computational pipeline for the identification and clustering of homologous RNA structures from a large set of query sequences. At its core lies DotAligner, a heuristic pairwise sequence alignment algorithm that considers the ensemble of base pair probabilities for each queried sequence. We benchmark the performance of DotAligner with other pairwise RNA structure alignment algorithms through several iterations of a stochastic sampling strategy across all Rfam seed alignments, highlighting the speed and accuracy of our method. We combine DotAligner with density-based clustering for the unsupervised identification of RNA structure motifs, which can identify both known Rfam families and novel RNA structure motifs from ENCODE enhanced cross-linked immunoprecipitation (eCLIP) data. Finally, we exemplify how clusters of homologous RNA structures identified by our method can be used to search for homologous structures across reference genomes and transcriptomes to generate a map of functionally related RNA structure motifs.

Results

Ensemble-guided pairwise RNA structure alignment

We developed an algorithm that leverages the diversity of suboptimal solutions from a partition function of RNA alignments to identify an optimal sequence-structure alignment of two RNAs. The algorithm, termed DotAligner, overcomes the limitations of comparing unique RNA secondary structures (such as minimum free energy predictions) to yield a pairwise alignment that considers mutual base pair probabilities. A schematic of how DotAligner functions is illustrated in Fig. 1.

DotAligner was developed with an emphasis on runtime performance to facilitate all-vs-all pairwise comparisons of RNA structures on large data sets. Consequently, it uses pre-calculated RNA dot plots to perform alignments. It also uses the observation that a significant subset of stochastic sequence alignments between two RNAs will overlap the correct structure-based alignment, even though the optimal sequence alignment deviates significantly from the structural alignment [34]. The algorithm combines an alignment-envelope heuristic with a fold-envelope heuristic, which impose constraints on suboptimal sequence alignments and pre-calculated base pair probabilities, respectively. The alignment procedure consists of two steps, each considering base pair probabilities: (1) generating a partition function of pairwise probabilistic string alignments and (2) stochastic sampling of string alignments and scoring of aligned dot plots. Existing building blocks are integrated to DotAligner in a novel way. A StrAL-like score is applied during the dynamic programming in step 1, then a CARNA-like score is used

Smith *et al. Genome Biology*   (2017) 18:244

Page 3 of 12

Fig. 1 Schematic of a pairwise alignment with DotAligner. A dynamic programming matrix is first filled in based on the similarity in sequence and cumulative base-wise pairing probabilities of two RNA sequences (top left: colour intensity indicates cumulative similarity score). A partition function over all pairwise alignments is calculated and interrogated for structural compatibility by stochastic backtracking (not illustrated). Two ensembles over all secondary structures are considered for this purpose (bottom left and top right dot plots: blue lines indicate cumulative base-wise pairing probabilities). The final scoring uses the base pair probabilities in the dot plots. This effectively warps the optimal sequence alignment path (top left, black outline) towards one that includes structural features (top left, blue outline and fill). In the bottom right, the optimal sequence alignment and associated consensus secondary structure is contrasted to that produced by DotAligner, exposing the common structural features hidden in the suboptimal base pairing ensemble of both sequences
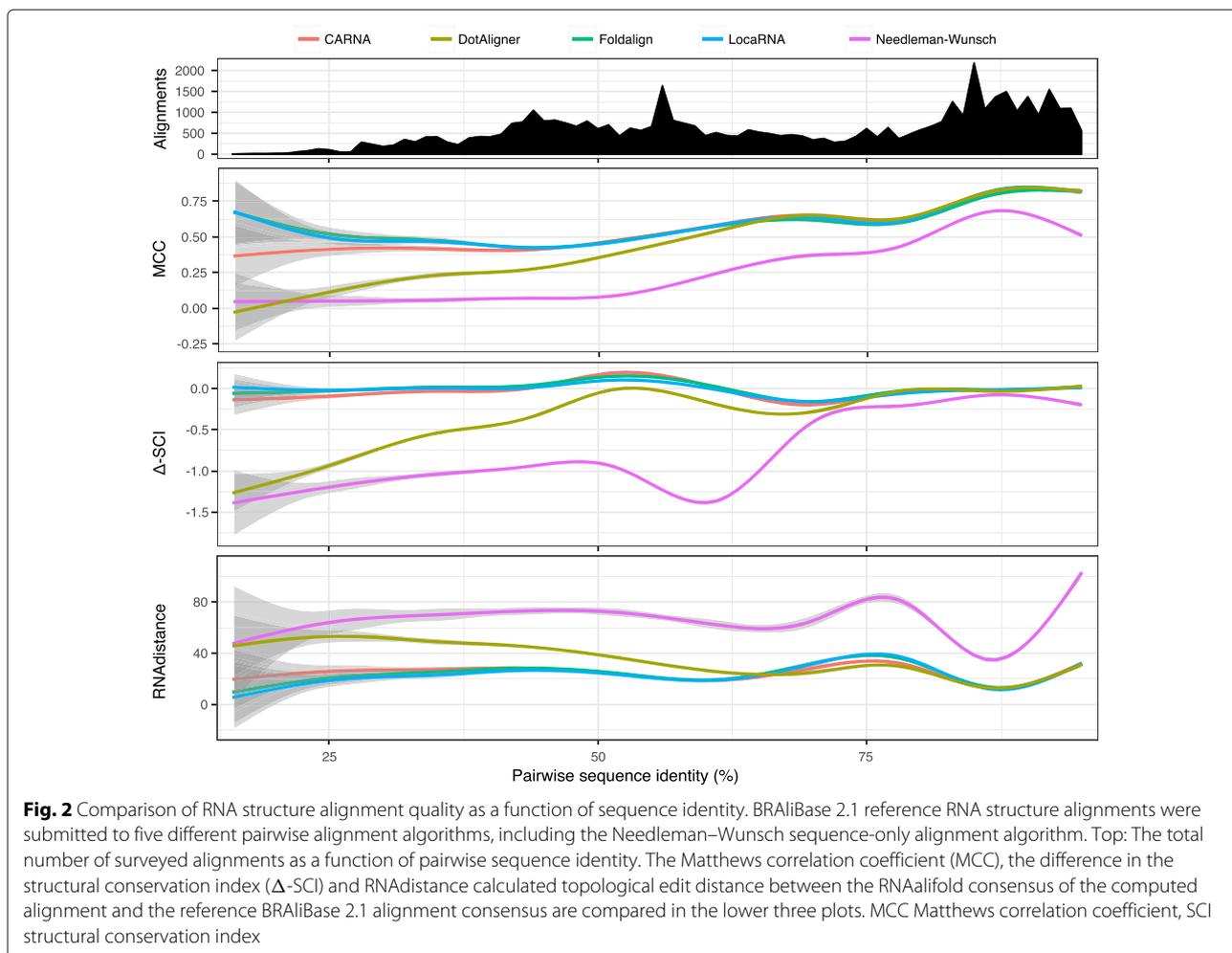
to score the aligned dot plots in step 2, and, lastly, the partition function in step 1 and sampling in step 2 are adapted from ProbA [34]. The detailed implementation and mathematical description of DotAligner can be found in Additional file 1.

Evaluation of pairwise alignment quality

We first tested DotAligner on BRAliBase 2.1 pairwise RNA structure alignments, a reference data set specifically designed for algorithm benchmarking [19, 35] (see 'Methods'). In this application, DotAligner seemingly performs worse than three other state-of-the-art algorithms, namely CARNA [30], FoldAlign [22, 36] and LocaRNA [25], as well as the Needleman–Wunsch pairwise sequence alignment algorithm, which ignores RNA structure content (Fig. 2). When comparing how well

the algorithms perform as a function of the pairwise sequence identity of BRAliBase 2.1 reference alignments, DotAligner produces alignments of lesser quality than comparable RNA structure alignment tools, particularly below 60 % sequence identity, albeit with better accuracy than sequence-only alignments. Upon closer inspection, DotAligner outperforms the other tools around the 65–80 % sequence identity range. As mentioned in the next section, this roughly corresponds to the average pairwise intra-family sequence identity of Rfam clans.

Interestingly, many of the pairwise structure alignments produced structural conservation index (SCI) scores above those from the BRAliBase 2.1 reference alignments (Fig. 2). The SCI represents the alignment consensus energy normalised by the average energy of the single

Smith *et al. Genome Biology* (2017) 18:244

Page 4 of 12

Fig. 2 Comparison of RNA structure alignment quality as a function of sequence identity. BRAliBase 2.1 reference RNA structure alignments were submitted to five different pairwise alignment algorithms, including the Needleman–Wunsch sequence-only alignment algorithm. Top: The total number of surveyed alignments as a function of pairwise sequence identity. The Matthews correlation coefficient (MCC), the difference in the structural conservation index (Δ-SCI) and RNAdistance calculated topological edit distance between the RNAalifold consensus of the computed alignment and the reference BRAliBase 2.1 alignment consensus are compared in the lower three plots. MCC Matthews correlation coefficient, SCI structural conservation index

sequences folded independently [37]. It has been shown to be one of the most reliable metrics for conserved RNA structure detection [38]. With the exception of DotAligner, the other RNA structure alignment tools display, on average, SCI values above 0 in the 45–60 % identity range, suggesting that the underlying optimisation algorithms tend to overestimate the number of paired bases in consensus RNA structure predictions.

DotAligner's capacity to produce competitive pairwise alignments is demonstrated via a 5S-adenosyl methionine riboswitch (Rfam clan RF00634, Additional file 2: Figure S1). In the Rfam alignment, the two representative sequences (AM420293_1 and CP000580_2_6) have a sequence identity of 55 %. Pure sequence alignment increases this to 69 %, but fails to align most structural features. DotAligner's pairwise probabilistic string alignment (step 1) creates an alignment with pairwise sequence identity (PID) = 56 %, which is increased to PID = 63 % through DotAligner's sampling. The number of correctly aligned suboptimal base pairs increases via DotAligner's sampling. In this example, the alignment scores do not

differ very much between DotAligner's optimal string alignment (step 1) and the best sample (step 2) (0.58 and 0.60, respectively), despite a $\sim 25\times$ increase of runtime through sampling ($s = 1000$ in this example). As justified below, the benefits of sampling are outweighed by other properties of the algorithm.

Fast and accurate classification of RNA structures
The intended application of DotAligner is the identification and clustering of RNA structural motifs from a large and diverse set of sequences of interest. Therefore, we evaluated the ability of DotAligner to distinguish between distinct structured RNA species from a heterogeneous sample of known RNA structure families. We performed all-vs-all pairwise structure alignments of stochastically sampled Rfam sequences, which were selected with constraints on their sequence composition (PID) to control for and ascertain any sequence-dependent biases (see 'Methods'). DotAligner alignment scores were then compared to a binary classification matrix representing the distinct Rfam families (Additional file 2: Figure S3).

Smith *et al. Genome Biology* (2017) 18:244

Page 5 of 12

Despite the seemingly poor quality of pairwise alignments generated by DotAligner, it reproduces the known classification of Rfam structures more accurately, in general, than the other surveyed pairwise RNA structure alignment tools (Fig. 3 and Additional file 2: Table S1). In fact, only when the average pairwise sequence identity drops below 55 % for a given set of homologous RNA structures do the other algorithms perform comparably to DotAligner (Fig. 3c). Interestingly, the sequence alignments produced by Needleman–Wunsch are able to cluster Rfam sequences into their respective clans well compared to more specialised RNA structure alignment tools, suggesting that most Rfam clans present sufficient stretches of local sequence identity

to cluster them appropriately. Indeed, realigning the sequences from Rfam seed alignments based on their sequence alone, while permitting free end gaps to evaluate local sequence similarity, shifts the median pairwise sequence identity from 59 % to 72 % (Additional file 2: Figure S4).

The efficacy of the heuristics implemented in DotAligner is further accentuated by its runtime, which consistently lies between simple sequence alignment and more sophisticated RNA structure alignment algorithms (Fig. 3d and Additional file 2: Figure S5). The impact of sequence length does not correlate with area under the curve (AUC) scores, but it increases the runtime in a polynomial way (Additional file 2: Figure S6).
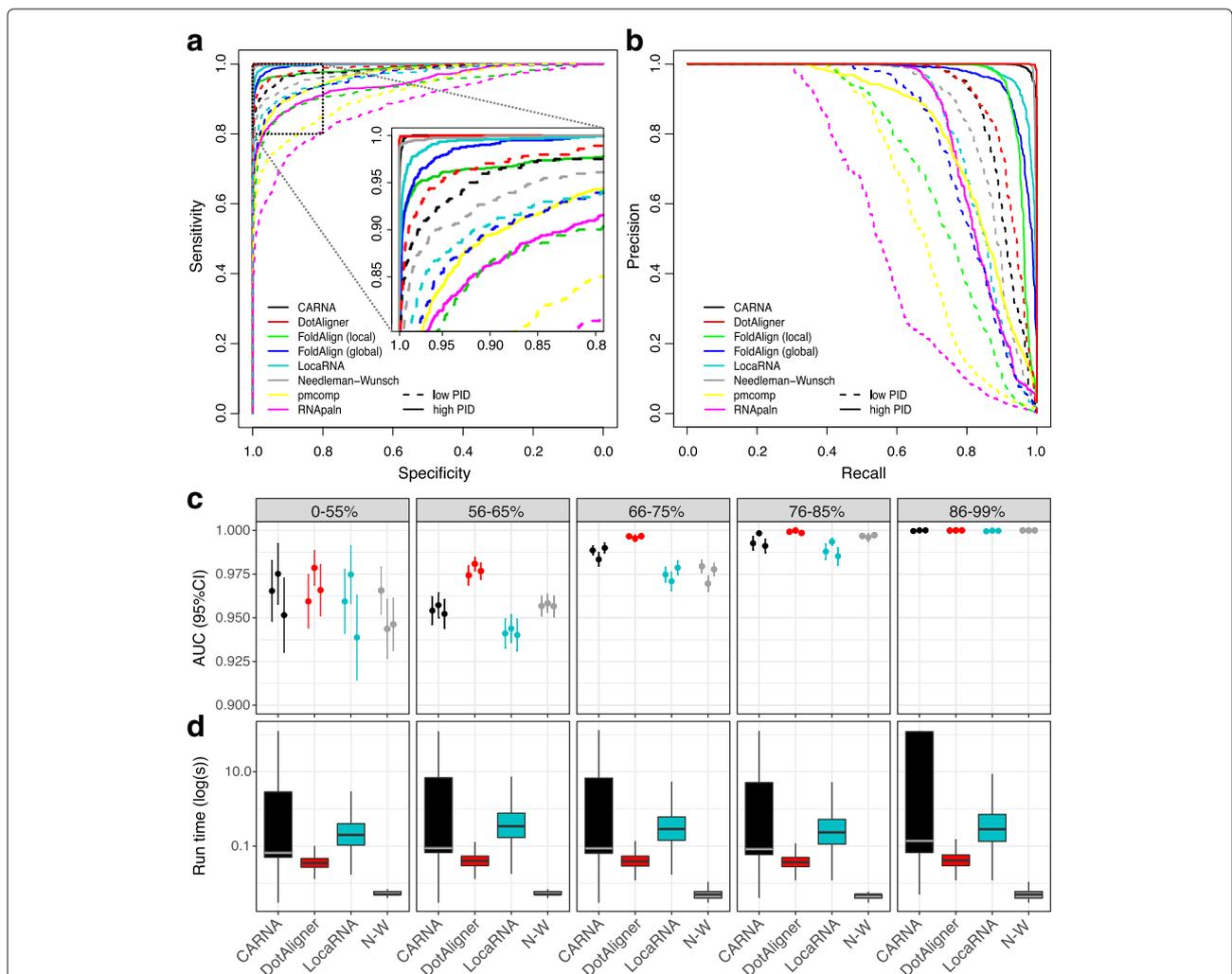

Fig. 3 Classification of known RNA structures. **a** Receiving operator characteristic (ROC) curves measuring the classification accuracy of the surveyed algorithms by contrasting their computed similarity matrices to a binary classification matrix of Rfam sequences (1 if the sequences are in the same family or 0 if different). High PID = 56–95 % pairwise sequence identity from the provided Rfam alignment; low PID = 1–55 %. **b** Precision vs recall curve. **c** Area under the curve (AUC) of ROC values with 95 % confidence intervals for the top four performing algorithms across five ranges of pairwise sequence identity, as calculated from a variant of the Needleman–Wunsch algorithm with free end gaps. The three replicates correspond to stochastically sampled sequences from Rfam 12.3 (see Additional file 2: Table S1). **d** Runtime distribution of single-thread computation on a 2.6 GHz AMD Opteron processor (note, a fixed upper limit of 120 s was imposed for CARNA). AUC area under the curve

Smith *et al. Genome Biology* (2017) 18:244

Page 6 of 12

Density-based clustering of homologous RNA structures

Given DotAligner's accurate clustering of known structured RNA using binary classification, we subjected its output to cluster analysis to identify and extract input sequences that display common sequence-structure motifs. Will et al. applied hierarchical clustering to the dissimilarity matrices produced by LocaRNA to organise sequences based on their structural homology [25]. However, this does not apply a cut-off that can be used to extract accurate novel clusters of structurally homologous sequences in an unsupervised manner. We attempted to achieve this by applying a statistical threshold derived from bootstrapping the underlying data using `pvclust` [39], but this generated clusters of variable size that often spanned across many disjoint families (data not shown).

We, therefore, opted for a density-based clustering strategy that, in theory, can decipher clusters of varying density (i.e. subsets of the data with significantly greater sequence-structure homology). The OPTICS (Ordering Points to Identify the Clustering Structure) algorithm [40] was chosen for this, as it has very few parameters to optimise. OPTICS is a derivative of the Density-Based Clustering for Application with Noise (DBSCAN) algorithm [41], which, as its name states, is suitable for noisy data, such as RNA immunoprecipitation followed by high-throughput sequencing. We benchmarked the two main OPTICS clustering parameters – the steepness threshold **xi** and the minimum number of points in a cluster (Additional file 2: Figure S7) – on a pooled set of 580 stochastically sampled Rfam sequences encompassing various ranges of sequence similarity, as well as a corresponding set of 580 dinucleotide shuffled controls (see 'Methods'). After performing all-vs-all pairwise alignments with DotAligner, we evaluated the effect of OPTICS parameters on clustering performance, revealing that a minimum of four points (or sequences) and a steepness threshold of 0.006 gave the best results (Additional file 2: Figure S7A).

Table 1 Comparative clustering performance

Algorithm	Number of clusters	Sensitivity	Specificity	Accuracy
DotAligner+OPTICS	53	0.716	0.886	0.802
GraphClust	201	0.990	0.110	0.635
NoFold (all CMs)	62	0.866	0.965	0.916
NoFold (filtered)	45	0.674	0.976	0.826

In comparison to GraphClust, the combination of DotAligner and OPTICS performs comparably well (Fig. 4, Table 1, Additional file 2: Table S2). The default version of NoFold nonetheless outshines DotAligner at clustering known Rfam families. However, it intrinsically employs Rfam covariance models (CMs) that are also present in the test data; therefore, this specific application is likely to be subject to over-fitting. We, thus, removed 72 CMs associated with the Rfam sequences in our benchmarking data set from the NoFold algorithm, which yielded lower sensitivity and less accurate qualitative cluster metrics than the DotAligner and OPTICS combination, while its specificity increased slightly despite removing CMs from its classification set.

Identifying protein-binding RNA motifs from eCLIP data

The optimised parameters for OPTICS clustering of DotAligner output were incorporated into a high-performance computing pipeline that extracts clusters of homologous RNA structure motifs from a set of input sequences (see 'Methods'). This pipeline was applied to eCLIP sequencing data from 44 RNA binding proteins from the ENCODE consortium [42], with 100 positive control sequences from Rfam (Additional file 2: Table S3). From 2650 high-confidence eCLIP peaks (>eightfold-enrichment vs background, P value $< 10^{-4}$) that overlap evolutionarily conserved secondary structure predictions, 25 significant clusters of homologous RNA were detected, including all 11 positive controls (Fig. 5).
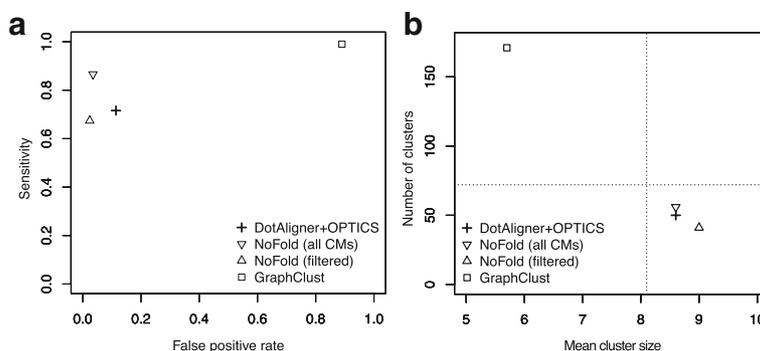

Fig. 4 Comparative clustering benchmark of Rfam sequences and their shuffled controls. Clustering performance metrics of three algorithms on 580 reference Rfam structures and their dinucleotide-shuffled controls. **a** Sensitivity vs false positive rate. **b** Qualitative cluster statistics (the horizontal dashed line indicates the real number of clusters from unique Rfam families). CM covariance model

Smith *et al. Genome Biology* (2017) 18:244

Page 7 of 12

Indeed, the *spike-in* Rfam sequences facilitate the identification of similar RNA structures, such as the homologues to SNORNA72 depicted in Fig. 5c, d. The four additional sequences that co-cluster with SNORNA72 controls are all associated with the DKC1 protein, which binds to H/ACA snoRNAs. Furthermore, three of the DKC1-bound peaks are annotated as snoRNAs in the Gencode 24 reference, while the fourth is not annotated as a snoRNA despite strong sequence and structure similarity, highlighting how this method can successfully identify and cluster new RNA structure motifs. Another example is the Y RNA cluster, which contains three sequences homologous to this Rfam family that are also associated with the TROVE2 protein, which binds to misfolded non-coding RNAs, pre-5S rRNA and Y RNAs.

Our method also identifies RNA structure families impartially, as exemplified by several clusters of DKC1-associated sequences, which present consensus secondary structures indicative of snoRNAs (Fig. 5e). Closer inspection of the corresponding eCLIP peaks revealed that these sequences are indeed annotated as snoRNAs in Gencode. There are also examples of de novo structural motifs that are associated with RNA-binding proteins with no previously known binding sites, such as an UPF1-dominated cluster (Fig. 5f) composed of a structure motif belonging to ALU repeats (Additional file 2: Figure S8). When searching the human genome for homology to the RNA structure motif derived from this cluster, most ALU elements are detected, as well as a few other repeat-containing sequences. Interestingly, 998 homologues to the motif did not localise to ALU elements (Additional file 2: Figure S8C, D), 58 % of which overlap miTranscriptome reference transcripts [43].

Fig. 5 De novo homologous RNA motif identification. **a,b** Reachability plots of OPTICS clustering display the OPTICS-derived ordering of points (*x*-axis) and their distance to the nearest neighbour (*y*-axis). Colours represent significant clusters. **a** Clustering of Rfam benchmarking data indicating the distance to the nearest neighbour. **b** Clustering of 2650 ENCODE eCLIP peaks + 100 Rfam controls that overlap evolutionarily conserved secondary structure predictions. The dominant RNA binding protein in each cluster is displayed next to significant clusters. Those with an asterisk are portrayed below. **c** Multiple structure alignment generated by mLocaRNA on the sequences from a cluster containing both Rfam SNORNA72 sequences and DKC1 (a snoRNA-binding protein) eCLIP peaks. An unannotated DKC1-bound sequence is marked with an asterisk. **d–f** RNAalifold-predicted consensus RNA secondary structures: **d** Structure of the alignment displayed in (**c**). **e** Structure of a cluster of impartially detected DKC1-bound snoRNAs. **f** Structure of a novel UPF1-bound motif. dist. distance

Smith *et al. Genome Biology*   (2017) 18:244

Page 8 of 12

Discussion

The increasing accessibility of next-generation sequencing and immunoprecipitation protocols provides large resources for in-depth transcriptome and interactome profiling. Elucidating the structural features of RNAs associated with RNA-binding proteins and RNP complexes, combined with the systematic classification of their genome- or transcriptome-wide occurrence, can identify recurrent functional motifs that may form components of regulatory networks. Pragmatically, the method we describe facilitates this process by enabling rapid and unsupervised clustering of RNA structure motifs with reasonable accuracy. We also show that clustering eCLIP sequences can identify new RNA structures and their homologues throughout the genome (Additional file 2: Figure S8A–C), which can be used to assign putative functions to non-coding loci and categorise them accordingly.

Given its relative speed and accuracy, DotAligner can be used to generate larger (dis)similarity matrices for cluster analysis than other pairwise structure alignment algorithms, or at least produce them with reasonable computational power. In addition to its speed, the strength of DotAligner lies in its capacity to accurately score structurally homologous RNA sequences and the suboptimal structural landscape of RNAs, reducing several dimensions of information into a single discriminative numeric value. Our results show that this can be sufficient to extract structurally and functionally related sequences from a large amount of noisy input. It is an ideal application for screening high-throughput sequencing data – such as RNA immunoprecipitation data – for common structural motifs.

The algorithm generates pairwise alignments that differ qualitatively to reference structural alignments at lower ranges of sequence identity, but it performs better than more complex algorithms within ranges of sequence similarity that substantially overlap those of functionally related RNAs, as presented in Rfam. This could be a consequence of refining the runtime parameters through testing on independently and stochastically sampled Rfam sequences. It is not impossible that other algorithms could undergo comparable parameter optimisation. However, the significantly higher computational complexity of other related tools compared to our method makes it fairly difficult (and resource intensive) to perform such brute-force parameter optimisation.

High-throughput CLIPseq data pose a challenge for the detection of consensus motifs since several molecules that are in close physical proximity to the target molecule can co-precipitate together. Consequently, other RNA sequences that do not directly bind to the target protein may be present. We have shown that our method is, nonetheless, suitable for such noisy biological data. For example, the UPF1 cluster we describe may be an example of an indirect binding event, as UPF1 directly interacts with STAU1, a double-stranded RNA-binding protein that has been reported to target ALU sequences [44]. Other clusters identified in our eCLIP analysis have sequences from more than one target protein clustered together, which raises the possibility that a common RNA structure motif may be bound by different proteins, either as part of a quaternary complex or as a common, competing binding target. We privilege this hypothesis over one of spurious false-positive clustering given our benchmark results and the observation that very few clusters were observed when analysing less stringently filtered eCLIP peaks (data not shown).

DotAligner has several variables that can influence the clustering results and speed depending on the type of input data. The most influential variables are the weight between sequence and structure similarity, and the exploration depth of suboptimal alignments in the stochastic backtracking. We have shown that stochastic sampling of suboptimal string alignments improves the alignment of RNA dot plots. However, the performance increase does not outweigh the increase in runtime associated with sampling suboptimal sequence alignments. Our Rfam clustering benchmark using a binary classification strategy has shown that the best trade-off between alignment accuracy and speed comes with the abandonment of sampling, as supported by the de novo structures identified from the ENCODE eCLIP data. Future optimisation of DotAligner parameters will likely increase its usability. For example, dynamic parameters could be implemented that adjust the degree of sampling diversity and number of samples based on the sequence identity obtained from step 1 of DotAligner. This could tune the algorithm's performance based on the nature of the input, potentially improving DotAligner's performance across all ranges of sequence identity. Another potential enhancement could be achieved in the stochastic sampling by considering only elements of the ensemble with probabilities larger than a threshold. By doing so, we could (1) reduce the number of useless samples, (2) guarantee that cells of high probability are passed (suboptimal structures) and (3) leave time/samples to explore the ensemble space (slightly modified alignments by limiting sample diversity) around these suboptimals.

Another great challenge lies in the accurate depiction of RNA structure motif boundaries. Whereas global structures may stabilise the RNA molecule, local structural domains are often sufficient for recognition by RNA binding proteins. A strategy to find optimal local alignments would be desirable for this. DotAligner can search for semi-local alignments by introducing penalty-free gaps at the sequence extremities (note, LocaRNA also supports this functionality). In this study, we did not investigate the optimisation of these local pairwise similarity scores,

Smith *et al. Genome Biology* (2017) 18:244

Page 9 of 12

because they may miss parts of the functional units (RNA structure) and, hence, hinder the search for optimal clusters. Instead, we circumvented this issue by overlapping eCLIP peaks to evolutionarily conserved RNA secondary structure predictions with well-characterised flanking helices supported by base pair covariation [6]. While preparing this manuscript, a complementary and comprehensive data set of evolutionarily conserved RNA secondary structures was published [45]. Its application could further increase the number of eCLIP peaks with accurate structural motif boundaries. Alternatively, RNA structure boundaries can be refined by, for example, using alternative strategies, such as computational boundary refinement with LocaRNA-P [46] or improving the biological data with enzymatic probing with the double-stranded RNase T1 endoribonuclease.

Conclusion

An efficient pairwise RNA sequence alignment heuristic, which intrinsically considers suboptimal base pairings, accurately discriminates between distinct structured RNA families. When combined with a noise-tolerant density-based clustering algorithm, this approach identifies known and novel RNA structure motifs from a set of input sequences without a priori knowledge. The resulting RNA structure motifs are subsequently used to identify homologues in the human genome, improving the annotation of lncRNAs and increasing the repertoire of functional genetic elements.

Methods

Benchmarking and parameter optimisation

The DotAligner algorithm implements several parameters that first need to be tuned before it can be applied to biological sequence analysis. All combinations of core parameters were tested on the 8976 pairwise RNA structure alignments curated in the BRAliBase 2.1 reference data set [35]. We first tested all combinations of the following parameters: *k* and *t* from 0 to 1 in increments of 0.1; *o* and *e* from 0.2 to 1 in increments of 0.2. For each set of parameter combinations, the number of alignments producing identical structural topologies to the reference alignment was determined using RNAdistance, SCI, a robust measure of RNA structural alignment integrity [38] based on the minimum free energy (MFE), and the Matthews correlation coefficient (MCC) of *predicted* and *reference* RNA secondary structure were also calculated for all resulting alignments:

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

$$\Delta \text{SCI} = \text{SCI}_{\text{predicted}}/\text{SCI}_{\text{reference}}.$$

Here, true positives (TP) is the number of representatives from the dominant Rfam family in a cluster. False positives (FP) is the number of non-dominant Rfam family representatives in a cluster, or the number of clusters where there is no dominant Rfam family (i.e. equally represented families), or the number of clusters where the dominant sequence is a negative control. False negatives (FN) is the number of Rfam sequences that fail to cluster. True negatives (TN) is the number of negative control sequences that fail to cluster. Moreover, SCI = $\text{MFE}_{\text{consensus}}/\overline{\text{MFE}}_{\text{single}}$.

The baseline parameters were then selected via a product rank of these two metrics and subjected to refinement using a binary classification approach, described in the next section.

Binary classification of RNA secondary structure families

Further refinement of the optimal parameters was performed using a binary classifier for two sets of 200 stochastically sampled Rfam entries with published structures: (i) a low pairwise sequence identity (PSI) set and (ii) a high PSI set, where any two sequences from the same family share between 0–55 % and 56–95 % PSI, respectively (Fig. 3a, b). The Java implementation of this algorithm, derived from [6], can be found in Additional file 1. Further investigation of the impact of local sequence similarity on algorithmic performance was done by sampling all seed alignments of Rfam version 12.3 via three replicates of our stochastic sampling procedure. The sequences were then stripped of gaps and pseudoknots (not present in the preliminary Rfam version 12.0 alignments), and realigned with a variant of Needleman–Wunsch enabling free end gaps. The samplings were limited to five ranges of sequence identity, as presented in Fig. 3c.

A binary classification matrix was then constructed, where sequences x and y have a score of 1 if they belong to the same Rfam family, or a score of 0 if they do not. The similarity matrix resulting from all-vs-all pairwise comparisons with DotAligner was tested for accuracy using the AUC of the receiving operator characteristic, as calculated by the R package pROC [47]. A more restricted range of parameter values was then tested on both data sets. A ranked sum for both data sets of the AUC was performed to determine the default runtime parameters for DotAligner, namely $\theta = 0.5$, $\kappa = 0.3$, $g_{\text{o}} = 1$ and $g_{\text{ext}} = 0.05$ (Additional file 2: Table S4). Parameter θ (or -t in the command line) is the weight of sequence similarity compared to the similarity of unpaired probabilities, κ (or -k) is the weight between sequence-based similarity and dot plot similarity, g_{o} (or -go) is the gap opening penalty and g_{ext} (-go) is the gap extension penalty. Sampling-specific parameters s (number of samples) and T (measure of sampling diversity) had minimal impact in the refined parameter optimisation from sampled Rfam

Smith *et al. Genome Biology* (2017) 18:244

Page 10 of 12

clans, although the parameters can increase alignment scores in low and medium pairwise sequence identity ranges (Additional file 2: Figures S1 and S2A). We also show that, on average, the alignment score saturates after 1000 samples of the stochastic backtracking for $T = 0.25$ (Additional file 2: Figure S2B). CARNA version 1.2.5 was run with parameters `--write-structure --noLP --time-limit=120000`. LocaRNA version 1.7.13 was run with parameter `--noLP`. FoldAlign version 2.1.1 was run with and without parameter `-global`. Default parameters were used for `pmcomp`, downloaded from https://www.tbi.univie.ac.at/RNA/PMcomp/, and RNA-paln version 2.3.5. The custom implementation of Needleman–Wunsch can be found in the GitHub repository associated with this work, as can the benchmark implementation scripts.

Clustering RNA structures with randomised controls

OPTICS benchmarking was performed by stochastically sampling the collection of Rfam 12.0 seed alignments using the Java program `GenerateRfamsubsets.java` (see Additional file 1) with three ranges of pairwise sequence identity: 1–55 %, 56–75 % and 75–95 %, a minimum of five representative sequences per family, and sizes ranging between 70 and 170 nt. The resulting 580 unique sequences were then shuffled while controlling their dinucleotide content with the `easel` program included in the Infernal (v1.1.2) software package [48] with option `-k 2`. The 1160 sequences were submitted to all-vs-all pairwise comparisons with DotAligner and the scores were inverted and normalised (`min=1, max=0`) into a dissimilarity matrix, which was then imported into the R statistical programming language, converted into a 'dist' object without transformation, and subjected to OPTICS clustering as implemented in the `dbscan` CRAN repository with a range of parameters (see Fig. 4a, b).

Other tested RNA clustering approaches were Graph-Clust and NoFold. We ran GraphClust version 0.7.6 inside the Docker image provided with RNAscClust with default parameters. NoFold version 1.0.1 uses 1973 Rfam CMs by default as empirical feature space. For the NoFold (all CMs) variant, we ran the program with default parameters, whereas for the NoFold (filtered) variant, we reduced the feature space to 1902 CMs by removing Rfam families from our benchmark set.

The following clustering performance metrics were used:

$$\text{Sensitivity (recall)} = \text{TP}/(\text{TP} + \text{FN}),$$
$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP}),$$
$$\text{False positive rate} = 1 - \text{Specificity},$$
$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}),$$
$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN}).$$

Clustering of protein-bound evolutionarily conserved RNAseq reads

The genomic coordinates of ENCODE eCLIP peaks were downloaded in bed format from the April 2016 release via the ENCODE portal (https://www.encodeproject.org/search). The resulting 5,040,096 peaks were filtered to keep only those with ≥eightfold enrichment over the total input background and an associated P value $\leq 10^{-4}$. Furthermore, peaks were merged if they overlapped by more than 50 nt to avoid over-representing the same sequence (Additional file 1). The remaining peaks were subsequently filtered by retaining only those that have a same-strand overlap with any evolutionarily conserved structure predictions from [6]. Finally, the associated genomic sequences were extracted into a fasta file, which was supplemented with 100 reference RNA structures from 11 Rfam families (see Additional file 2: Table S3). Merging, overlap and sequence extraction operations were performed with Bedtools version v2.26.0.

The normalised similarity matrix resulting from all-vs-all pairwise comparisons with DotAligner was then subjected to clustering with the `dbscan` 1.1-1 R package from Michael Hahsler (https://github.com/mhahsler/dbscan) using the command `opticsXi( optics(D, eps=1, minPts=4, search="dist"), xi = 0.006, minimum=T)`. The sequences for each cluster were then extracted and submitted to multiple structure alignment with mLocaRNA version 1.9.1 using parameters `--probabilistic --iterations=10 --consistency-transformation --noLP`.

Additional files

Additional file 1: Supplementary methods describing the DotAligner implementation in detail, RNA structure clustering and eCLIP data processing methodologies. (PDF 602 kb)

Additional file 2: Supplementary tables and figures with descriptions. (PDF 3890 kb)

Smith *et al. Genome Biology*   (2017) 18:244

Page 11 of 12

and developed the analytic pipelines. SES created and implemented the DotAligner source code. XQ assisted in DotAligner parameter optimisation and benchmarking. All authors read and approved the final manuscript.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]RNA Biology and Plasticity Group, Garvan Institute of Medical Research, 384 Victoria Street, NSW 2010 Sydney, Australia. [2]St Vincent's Clinical School, Faculty of Medicine, UNSW Australia, NSW 2010 Sydney, Australia. [3]Center for non-coding RNA in Technology and Health (RTH), University of Copenhagen, Groennegaardsvej 3, 1870 Frederiksberg, Denmark. [4]Department of Veterinary and Animal Sciences, Faculty of Health and Medical Sciences, University of Copenhagen, DK-1870, Frederiksberg, Denmark.

References
1. Morris KV, Mattick JS. The rise of regulatory RNA. Nat Rev Genet. 2014;15(6):423–37.
2. Engreitz JM, Ollikainen N, Guttman M. Long non-coding RNAs: spatial amplifiers that control nuclear structure and gene expression. Nat Rev Mol Cell Biol. 2016;17(12):756–70.
3. Nawrocki EP, Burge SW, Bateman A, Daub J, Eberhardt RY, Eddy SR. Rfam 12.0: updates to the RNA families database. Nucleic Acids Res. 2015;43(D1):130–7.
4. Eddy SR. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annu Rev Biophys. 2014;43:433–56.
5. Rivas E, Clements J, Eddy SR. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nat Methods. 2016;14(1):45–4.
6. Smith MA, Gesell T, Stadler PF, Mattick JS. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Res. 2013;41:8220–36.
7. Spitale RC, Flynn RA, Zhang QC, Crisalli P, Lee B, Jung JW, et al. Structural imprints in vivo decode RNA regulatory mechanisms. Nature. 2015;519(7544):486–90.
8. Lu Z, Zhang QC, Lee B, Flynn RA, Smith MA, Robinson JT, et al. RNA duplex map in living cells reveals higher-order transcriptome structure. Cell. 2016;165(5):1267–79.
9. Zappulla D, Cech T. RNA as a flexible scaffold for proteins: yeast telomerase and beyond. Cold Spring Harb Symp Quant Biol. 2006;71:217–24.
10. Hogg JR, Collins K. Structured non-coding RNAs and the RNP Renaissance. Curr Opin Chem Biol. 2008;12(6):684–9.
11. Rinn JL, Chang HY. Genome regulation by long noncoding RNAs. Ann Rev Biochem. 2012;81:145–66.
12. Mercer TR, Mattick JS. Structure and function of long noncoding RNAs in epigenetic regulation. Nat Struct Mol Biol. 2013;20(3):300–7.
13. Chujo T, Yamazaki T, Hirose T. Architectural RNAs (arcRNAs): a class of long noncoding RNAs that function as the scaffold of nuclear bodies. Biochim Biophys Acta Gene Regul Mech. 2016;1859(1):139–46.
14. Blythe AJ, Fox AH, Bond CS. The ins and outs of lncRNA structure: how, why and what comes next? Biochim Biophys Acta Gene Regul Mech. 2016;1859(1):46–58.
15. Kapusta A, Kronenberg Z, Lynch VJ, Zhuo X, Ramsay L, Bourque G, et al. Transposable elements are major contributors to the origin, diversification, and regulation of vertebrate long noncoding RNAs. PLoS Genet. 2013;9(4):1003470.
16. Hezroni H, Koppstein D, Schwartz MG, Avrutin A, Bartel DP, Ulitsky I. Principles of long noncoding RNA evolution derived from direct comparison of transcriptomes in 17 species. Cell Rep. 2015;11(7):1110–22.
17. Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, et al. Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet. 2010;42(7):631–4.
18. Kelley D, Rinn J. Transposable elements reveal a stem cell-specific class of long noncoding RNAs. Genome Biol. 2012;13(11):107.
19. Gardner PP, Wilm A, Washietl S. A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res. 2005;33(8):2433–9. doi:10.1093/nar/gki541.
20. Sankoff D. Simultaneous solution of the RNA folding, alignment and protosequence problems. SIAM J Appl Math. 1985;45:810–25.
21. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol. 2007;3(10):1896–908. doi:10.1371/journal.pcbi.0030193.
22. Sundfeld D, Havgaard JH, de Melo AC, Gorodkin J. Foldalign 2.5: multithreaded implementation for pairwise structural RNA alignment. Bioinformatics. 2016;32(8):1238–40. doi:10.1093/bioinformatics/btv748.
23. McCaskill JS. The equilibrium partition function and base pair binding probabilities for RNA secondary structure. Biopolymers. 1990;29(6–7):1105–19. doi:10.1002/bip.360290621.
24. Hofacker IL, Bernhart SH, Stadler PF. Alignment of RNA base pairing probability matrices. Bioinformatics. 2004;20(14):2222–7. doi:10.1093/bioinformatics/bth229.
25. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. PLoS Comput Biol. 2007;3(4):65. doi:10.1371/journal.pcbi.0030065.
26. Roshan U, Livesay DR. Probalign: multiple sequence alignment using partition function posterior probabilities. Bioinformatics. 2006;22(22):2715–21. doi:10.1093/bioinformatics/btl472.
27. Lorenz R, Bernhart SH, Honer Zu Siederdissen C, Tafer H, Flamm C, Stadler PF, et al. ViennaRNA package 2.0. Algorithms Mol Biol. 2011;6:26. doi:10.1186/1748-7188-6-26.
28. Dalli D, Wilm A, Mainz I, Steger G. STRAL: progressive alignment of non-coding RNA using base pairing probability vectors in quadratic time. Bioinformatics. 2006;22(13):1593–9. doi:10.1093/bioinformatics/btl142.
29. Palù A, Möhl M, Will S. A propagator for maximum weight string alignment with arbitrary pairwise dependencies. In: Cohen D, editor. Principles and practice of constraint programming – CP 2010. 2010. p. 167–75. doi:10.1007/978-3-642-15396-9 16.
30. Sorescu DA, Möhl M, Mann M, Backofen R, Will S. CARNA – alignment of RNA structure ensembles. Nucleic Acids Res. 2012;40(Web Server issue):49–53. doi:10.1093/nar/gks491.
31. Middleton SA, Kim J. Nofold: RNA structure clustering without folding or alignment. RNA. 2014;20(11):1671–83. doi:10.1261/rna.041913.113.
32. Heyne S, Costa F, Rose D, Backofen R. GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinformatics. 2012;28(12):224–32. doi:10.1093/bioinformatics/bts224.
33. Miladi M, Junge A, Costa F, Seemann SE, Hull Havgaard J, Gorodkin J, et al. RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. Bioinformatics. 2017. doi:10.1093/bioinformatics/btx114.
34. Muckstein U, Hofacker IL, Stadler PF. Stochastic pairwise alignments. Bioinformatics. 2002;18(Suppl 2):153–60.
35. Wilm A, Mainz I, Steger G. An enhanced RNA alignment benchmark for sequence alignment programs. Algorithms Mol Biol. 2006;1(1):1.
36. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLoS Comput Biol. 2007;3(10):193.
37. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proc Natl Acad Sci USA. 2005;102(7):2454–9.
38. Gruber AR, Bernhart SH, Hofacker IL, Washietl S. Strategies for measuring evolutionary conservation of RNA secondary structures. BMC Bioinform. 2008;9(1):122.
39. Suzuki R, Shimodaira H. Pvclust: an R package for assessing the uncertainty in hierarchical clustering. Bioinformatics. 2006;22(12):1540–2.
40. Ankerst M, Breunig M, Kriegel H, et al. Ordering points to identify the clustering structure. In ACM Sigmod record ACM. 1999;28(2):49–60.
41. Ester M, Kriegel HP, Sander J, Xu X, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. Kdd. 1996;96:226–31.
42. Van Nostrand EL, Pratt GA, Shishkin AA, Gelboin-Burkhart C, Fang MY, Sundararaman B, et al. Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced clip (eclip). Nat Methods. 2016;13(6):508–14.

Smith *et al. Genome Biology* (2017) 18:244

Page 12 of 12

43. Iyer MK, Niknafs YS, Malik R, Singhal U, Sahu A, Hosono Y, et al. The landscape of long noncoding RNAs in the human transcriptome. Nat Genet. 2015;47(3):199–208.

44. Gong C, Maquat LE. LncRNAs transactivate STAU1-mediated mRNA decay by duplexing with 3′ UTRs via Alu elements. Nature. 2011;470(7333):284.

45. Seemann SE, Mirza AH, Hansen C, Bang-Berthelsen CH, Garde C, Christensen-Dalsgaard M, et al. The identification and functional annotation of RNA structures conserved in vertebrates. Genome Res. 2017;27:1371–83.

46. Will S, Joshi T, Hofacker IL, Stadler PF, Backofen R. LocARNA-P: accurate boundary prediction and improved detection of structural RNAs. RNA. 2012;18(5):900–14.

47. Robin X, Turck N, Hainard A, Tiberti N, Lisacek F, Sanchez JC, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinform. 2011;12(1):1.

48. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics. 2013;29(22):2933–5.

49. Smith MS, Seemann SE. GitHub repository for DotAligner, including source code, pipelines, and data (bigredbutton). doi:10.5281/zenodo.1066258.