GigaScience

GraphClust2: annotation and discovery of structured RNAs with scalable and accessible integrative clustering

--Manuscript Draft--

Manuscript Number:	GIGA-D-19-00089R1
Full Title:	GraphClust2: annotation and discovery of structured RNAs with scalable and accessible integrative clustering
Article Type:	Research
Funding Information:	German Research Foundation (DFGgrant SFB992/1201) — Mr. Milad Miladi German Federal Ministry of Educationand Research (BMBF grant 031 A538A RBC) — Dr. Björn Andreas Grüning
Abstract:	RNA plays essential roles in all known forms of life. Clustering RNA sequences with common sequence and structure is an essential step towards studying RNA function. With the advent of high-throughput sequencing (HTS) techniques, experimental and genomic data are expanding to complement the predictive methods. However, the existing methods do not effectively utilize and cope with the immense amount of data becoming available. Hundreds of thousands of non-coding RNAs (ncRNAs) have been detected, however, the annotation of these ncRNAs is lacking behind. Here we present GraphClust2, a comprehensive approach for scalable clustering of RNAs based on sequence and structural similarities. GraphClust2 bridges the gap between HTS and structural RNA analysis, and provides an integrative solution by incorporating diverse experimental and genomic data in an accessible manner via the Galaxy framework. GraphClust2 can effciently cluster and annotate large datasets of RNAs and supports structure probing data. We demonstrate that the annotation performance of clustering functional RNAs can be considerably improved. Furthermore, an off-the-shelf procedure is introduced for identifying locally conserved structure candidates in long RNAs. We suggest the presence and the sparsity of phylogenetically conserved local structures for a collection of long non-coding RNAs. By clustering data from two CLIP experiments, we demonstrate the benefits of GraphClust2 for motif discovery under the presence of biological and methodological biases. Finally, we uncover prominent targets of double-stranded RNA binding protein Roquin-1, such as BCOR's 3'UTR that contains multiple binding stem-loops which are evolutionary conserved.
Corresponding Author:	Björn Grüning GERMANY
Corresponding Author Secondary Information:	
Corresponding Author's Institution:	
Corresponding Author's Secondary Institution:	
First Author:	Milad Miladi
First Author Secondary Information:	
Order of Authors:	Milad Miladi Eteri Sokhoyan Torsten Houwaart Steffen Heyne Fabrizio Costa Björn Andreas Grüning Rolf Backofen

Order of Authors Secondary Information:	
Response to Reviewers:	First, we would like to thank the reviewers and the editors for their encouraging and constructive comments on our manuscript. We have extended the manuscript, clarified the methodologies and performed additionally new experiments. Below we provide the point-to-point responses with referring to the corresponding modifications of the manuscript and the updated and extended results. To facilitate tracking the changes, paragraphs with a major amount of change are highlighted in blue within the manuscript. Reviewer #1: > In the presented manuscript, the authors describe a tool for the clustering of RNAs based on secondary structure similarities. Their approach can find application in the classification of RNAs and in finding structural motifs. The method has stand-alone implementations as well as it is integrated within the Galaxy framework, with the aim of facilitating and standardizing its usage. While the manuscript is globally well written, some aspects of the method could be better clarified. The authors might want to consider the following points: > 1. The clustering algorithm description is confusing. First, a graph kernel is used to identify RNAs forming initial clusters, which are then refined using UPGMA and CMfinder, and finally a covariance model is built for each cluster and used to scan the remaining RNAs. I don't understand how UPGMA and CMfinder are employed. #Authors: We thank the reviewer for their valuable comments. In this revision, the clustering description in "Methods overview" has been extensively updated and supplied with extended descriptions for better clarity. Inline below comes answers to the reviewer's question which are now mirrored in the manuscript as well. > Are they alternative to each other, or integrated in some undocumented way? #Authors: They are both used in our pipeline. First, a UPGMA is applied to each of the MinHash clusters to prune the set of sequences, then CMfinder is followed up on each of the pruned sets. > Which information provided by UPGMA and/or CMfinder is used to compute the covariance model? #Authors: The UPGMA step removes potential outliers within each cluster, which deemed similar according to the Minhash-Graph-kernel features but are later detected as dissimilar at the RNA-specific structural level according to the LocARNA scores. Afterward, the CMfinder level refines the structural alignment to improve identifying local structures. > Moreover, the iterative nature of the clustering algorithm is not evident from their description in the Materials and Methods section. I only realized that by looking at Fig. 2. I guess that the initial covariance models are progressively recomputed by adding new RNAs but, if it that's the case, a more detailed description of the procedure must be provided; #Authors: Thanks for this comment, we improved the corresponding section and made it more clear. The iteration step description is extended and moved to the beginning of the next paragraph, in the "Methods overview" (after cluster collection, before pre-clustering). The sequences, which are not assigned to a cluster until this round, are compared in the fast clustering step to identify new dense centers in the feature space. Eventually, new covariance models are generated and used to find further hits for the cluster.

> 2. Is the LocARNA score an ultrametric? Can the graph kernel similarity scores be converted into a distance to feed UPGMA?

#Authors:
For filtering the outliers within each cluster, we follow and use the strategy of the LocARNA tool, which is detailed in Will et al. work [Plos Comp. Bio. 2007. doi:10.1371/journal.pcbi.0030065]. The distance is approximated from the pairwise alignment score by LocARNA package. The UPGMA step is merely used to prune outlier sequences within each cluster and not for predicting the clusters. So the UPGMA and the distance approximation does not have a major effect on the clustering.
The kernel scores can be used to produce the distances with a lower runtime. However, here we use the LocARNA score since it is domain-specific and designed for structured RNAs. Because the quadratic pairwise comparison is only for the sequences within each cluster (usually ~10-100 sequences), the runtime is not a concern. We have extended the relevant description within the manuscript.

> 3. From the "Workflow output" section in M&M, it seems that fuzzy clusters (called soft clusters in the manuscript) can be obtained, but it is not explained how;

#Authors:
Thanks for the comment. An explanation is added to the "Methods overview" about the treatment of fuzzy/soft clusters.
A sequence can potentially match to multiple CMs. This would produce fuzzy/soft clusters, two clusters with overlap member ratio above the threshold are merged in the cluster collection step. A user-definable option to perform soft clustering is provided in the collection step.

4. In the Results and Discussion, section "Locally conserved candidates…", manual checking and filtering is reported. I wonder which is the impact of these expert manual screens, and which results a non-expert could expect to obtain;

#Authors:
In Figure 4. two tracks are reported. The first one, "candidate motifs", is the GraphClust2 automatically generated predictions and the second one, "manually curated subset", has been selected by screening candidates of the first track. So the non-expert gets the "candidate motifs" as a result which are supplemented with annotation scores from the three methods (RNAz, EvoFold2, R-scape). We have updated the text to clarify the difference.

> 5. It is not clear how RNAz, Evofold and R-scape are used, whether they provide filtering criteria or are just used to annotate and describe the results;

#Authors:
The three tools are invoked after clusters are predicted and collected. They are used to evaluate the structure conservation signals and identify highly conserved structural elements.
To clarify comments 4 & 5, we have revised the filtering section in the manuscript and added a specific reference to the M&M section.

> 6. Could running times for the described examples be provided?

#Authors:
We have added the runtime of the experiments as a supplementary Table S2. Furthermore, to demonstrate and verify GraphClust2's scalability, we have clustered a large metatranscriptome dataset. Please see the second results section about the runtime analysis. Besides, we have provided a new supplementary Figure S1.

Reviewer #2:
> The clustering of ncRNAs is an add-on to existing technologies of ncRNA annotation. This is done by allowing de-novo identification of ncRNA families and motifs, compared with the literature based family building process, starting from known ncRNA sequences which work as SEEDs. Also, tools that cluster ncRNA sequences are

scarce and, in most cases, publicly unavailable, therefore projects such as this are essential. Building a ncRNA family requires a number of repetitive curation steps aiming to improve the initial multiple sequence alignment and consensus secondary structure. For that reason, tools that omit the expert contribution require thorough assessment.

#Authors:
We sincerely thank the reviewer for their thoughtful comments and inspections of our submission. GraphClust2 tool is designed to complement available procedures and assist both non-experts as well as experts in speeding up the process of identifying and annotating ncRNAs.

> The authors nicely demonstrated that the inclusion of structure probing data such as SHAPE, can improve the clustering performance of GraphClust2 when compared with its preceding version, namely GraphClust. However, the results of the experiment based on eCLIP data left me questioning the quality of the clustering methodology and the test dataset.

#Authors:
We again thank the reviewer for their careful inspection. We have included new experiments and extended the eCLIP panels in Figure 5, which are detailed in our responses below. We believe that these enhancements would have resolved the false impression of performance, which was caused by having structures of two inherently different data sources (Rfam vs. eCLIP) side-by-side.

> The secondary structure generated from the largest cluster from the eCLIP data experiment, shows loss in base-pair covariation compared with the consensus secondary structure obtained from Rfam. It is very important to ensure the clustering works efficiently enough, as base-pair covariation is evidence that the secondary structure of a family of ncRNAs is correct.

#Authors:
The authors cannot agree more with the reviewer that the bp-covariation is strong evidence of structure-level conservation. For this reason, we have provided and highlighted the importance of base-pair covariations via R2R-Rscape plots, annotated structural alignments, conservation/covariation scores, which we have included in several plots. Following the reviewer's advice, we have performed additional experiments to validate the GraphClust2 performance and resolve the SLBP's covariation concern. Our response points are summarized here and the point-by-point answers come further below.

- The eCLIP data is not similar to the Rfam seed data. Therefore the structures are not expected to look the same. The eCLIP data is from a single human cell line and ortholog-only. This is contrary to the more diverse Rfam's seed data, which originates from 28 organisms and multiple experiments.

- The base-pair covariation of the applied human eCLIP data is inherently much less than Rfam.

- In line with the reviewer's suggestion, we have now validated the GraphClust2 performance for the eCLIP data by counterpart comparison with the Rfam's family covariance model. Rfam's CM is the ideal golden model since it has been built using the highly diverse set and covarying structures of Rfam's seed alignment. The Rfam's CM cmsearch hits are almost the same as (~96% overlap) the GraphClust2 prediction. Both GraphClust2 and Rfam showed the same (low) level of covariation (new Fig. 5-A).

- Following the reviewer's suggested experiment, we used GraphClust2 to cluster Rfam's seed sequences that were mixed up with 98.5% noise. GraphClust2 predicted and constructed the secondary structure with the same high level of covariation as the Rfam reference structure (new Fig. 5-B).

- The performance has been quantitatively measured and transparently demonstrated using the independently-designed Rfam-cliques and ProbeAlign dataset (Fig. 2 and

Table S1). Those datasets have multiple levels of sequence identity and designed to convey high covariation.

> The following points could help investigate this further:
> 1. How taxonomically diverse is the dataset used? Although the dataset apart from human sequences also includes sequences from other species - which the authors do not mention in the manuscript - is likely not diverse enough. Histone3 family (RF00032) is built from 46 sequences coming from 28 distinct species

#Authors:
Following the reviewer's question, we have carefully analyzed the eCLIP and Rfam family datasets. The eCLIP data is human-only and cell-type-specific. From the eCLIP paper : "We generated 102 eCLIP experiments for 73 diverse RBPs in HepG2 and K562 cells" [(Van Nostrand et al. 2016)]. Therefore we have extracted regions from the human genome.
So it does not include other species and is therefore not comparable to the Rfam's RF00032 seed dataset. We have clarified this as well in the section "SLBP eCLIP" paragraph under Materials&Methods-Data.

> 2. What is the sequence identity threshold and how was it decided for the best clustering results? This is something the authors did not mention in the manuscript and testing different thresholds could potentially result in gain of base-pair covariation support

#Authors:
GraphClust2 has no explicit sequence identity threshold setting. The clustering procedure uses Graph Kernel for comparing secondary structure graphs, so there is no explicit definition of the sequence identity for the cluster identification. This procedure is akin to counting the matching-kmers of two sequences but in the graph 2D space. Also, in the cluster extension step, we use the CM bitscore/E-value for the hit thresholds.

> 3. Technical error: Eliminating possibilities 1 and 2 could point towards clustering issues the authors previously eluded

#Authors:
As we outlined above, (a) eCLIP data is human-only and much different than the Rfam's diverse set; (b) GraphClust2 doesn't have any explicit identity threshold and mainly relies on secondary structure level comparisons; (c) As shown in updated Fig.5-A, B, GraphClust2 has the same performance as the reference Rfam's CM. Rfam's CM is the ideal model since it has been built using the highly diverse set and covarying structures of Rfam's seed alignment.

> Would the authors be able to reconstruct the same secondary structure as in Rfam by using a simulated dataset composed of RF00032 sequences and noise?

#Authors:
We explicitly thank the reviewer for this suggestion. It helped to extend the evaluation and clarify a potential misinterpretation of the performance. Using RF00032 - 46 seed sequences combined with 2954 shuffled sequences of the same length and GC-content distribution. GraphClust2 was able to successfully predict the retrieve SLBP binding stem-loop with the same level of covariation as Rfam (new Fig. 5-B).

> Testing using real data:
> Another thing that I feel that needs to be answered is how well the tool is able to process a huge volume of real data. In a real case scenario, GraphClust2 would have to cluster millions of ncRNA sequences rather than just a few thousands mentioned in the paper. A possible dataset to benchmark the capabilities of the tool could be RNAcentral - the database of non-coding RNAs - currently containing almost 12 million sequences. This would raise the following questions:

#Authors:

We agree that demonstrating a very large scale dataset is a suitable add-on to the work. To answer this, we ran GraphClust2 on metatranscriptome dataset of ~3.6 million sequences and showed it's scalability and runtime linearity over the number of entries.
We would also like to highlight that structure-based clustering is a computationally-intense and cumbersome technique. We are not aware of any comparable tool (especially a publicly available & accessible one) that can de novo identification of even a thousand sequences. It should be also noted that identifying structural elements for the CLIP and lncRNAs are very demanding realistic scenarios and are quite large (e.g. XIST clustering was on 20,000 sequence fragments).

> 1. Would GraphClust2 be able to correctly classify the ncRNA sequences in their corresponding types?

#Authors:
While we agree that clustering and analyzing the entire RNAcental is an exciting study, we think that it is beyond the scope of this manuscript and deserves its own project. We believe that biologically novel and motivating questions should be first defined and prerequisites must be met (e.g. RNAcentral-Galaxy data interface), before investing such a large amount of effort on a project of that type. However, we hope we could convince the reviewer that GraphClust2 is scalable by analyzing a metatranscriptomics dataset with 3.6 Million of sequences.

> 2. Would the infrastructure be able to cope with such a dataset?

#Authors:
Yes. Galaxy and it's infrastructure is scalable to the degree that the computational resources behind Galaxy are scalable. The Europan Galaxy server has access to multiple clouds and HPC infrastructures, providing more than 5000 cores and 25TB of memory. We have performed the additional experiment on a very large dataset of transcriptomic sequences of marine community HTS data without facing any infrastructure challenge. Please refer to the new result section" Clustering runtime evaluation", where we have discussed the clustering of 3.6 million sequences from a metatranscriptomic sample, plus the new supplementary Fig. S1.

> 3. Graphclust2 runs on Galaxy platform via a GUI. What would the response time be in such cases?

#Authors:
We did not experience any lag or increase of the GUI response time while performing the million sequences metatranscriptome clustering. The interface response time is not expected to be affected by the computation load in a scalable Galaxy server. E.g. usegalaxy.eu is processing more than 150.000 jobs per month without any noticeable lag in response time.

> 4. Would the software be able to deal with noise that comes with real data?

#Authors:
The SLBP eCLIP and Roquin-1 PAR-CLIP data are both real data with an inherently large amount of noise. We have also shown that GraphClus2 can perform well with 98.5% noise of Rfam SLBP data. Again referring to the previous points, of course, biological interesting and narrowed-down questions and goals should be defined before answering this broad question.

> Technical issues with the dockerized version of GraphClust2:
> I was unable to pull the docker image and the command crashed with the error message "docker: unauthorized: authentication required." I tried a couple of probable solutions, but without any success. For this reason, I could not test the dockerized version of the tools and further testing would be required when the issue is resolved. It is important that the users do not experience this, especially when the software is targeting users with less technical expertise.

#Authors:
We are sorry that the reviewer could not use the docker instance. We have tested pulling the docker over several computers and different networks but have not faced the aforementioned problem.
Searching the web hints for a generic issue about the mentioned error. It is likely related to the clock misconfiguration of the client login session or due to pulling a firewall/VPN. We would suggest trying an alternative (direct) internet connection and maybe also trying to logout and in the docker client again (using docker logout and login commands). Discussions and potential solutions are provided here: https://github.com/jupyter/docker-stacks/issues/364 and https://github.com/jupyter/docker-stacks/issues/484
A straightforward alternative option would be to run GraphClust2 on European Galaxy server under https://graphclust.usegalaxy.eu. We highlight that this server is running with strict data privacy policies such that the user data and activities are protected and not discernible. The usegalaxy.eu server is GDPR compliant and you can find the terms of use at https://usegalaxy.eu/terms/.

> Manuscript text:

> From the first read-through it becomes apparent that a good amount of effort went into the writing of the manuscript. However, although the manuscript is well structured, there were sections where rephrasing is essential for the text to be more readable. I also identified various linguistic errors as well as typos, so I would suggest another read-through to correct those. For example, there are a couple of typos on Figures 1 and 2. Figure 1 - I think the authors meant clustering instead of "clusteting" and on Figure 2 - The title of the Y axis of both charts reads to "Adjusted Rand Inex" instead of Adjusted Rand Index, which is the term mentioned in the manuscript. I would also include the abbreviation ARI in parenthesis. I would also suggest avoiding strong words like "ultimate" and "superior" results (page 5).

#Authors:
We have read through the paper and revised the figures and resolved typos.

> Additionally, the background results should be analysed more thoroughly as these will give the users a better indication of how well GraphClust2 works and perhaps also provide answers to my previous questions.

#Authors:
We have also extended the discussions about the MALAT1 background data with discovery details. (Results, second paragraph of MALAT1)

> In addition to all the above, I also have the following suggestions and comments with respect to the tool usage and graphical user interface (GUI):

> 1. All the tools in GraphClust2 are accessible through a graphical user interface (GUI), which is dependent on the Galaxy Framework. On one hand this is nice, because it gives the users access to many other tools available through the Galaxy project, but limits the GraphClust usage to that. I believe that the provision of a command line version of the GraphClust2 toolset to enable batch processing of data would be very useful. This way the authors could also target more experienced users who prefer to use the CLIs of Linux based operating systems. Another benefit of a CLI compared to the current version of the GraphClust2 tools, is the parallelization of the data processing via utilization of HPC systems the users have access to. A command line version of the GraphClust2 toolset would also allow its integration with other systems as well

#Authors:
We agree that some users might feel more comfortable using a CLI and that is possible with Galaxy as well. Galaxy offers a RESTful API as well that can be targeted in various programming languages and that makes it possible to submit tools and workflows. However, we believe this is not the focus of the paper and is a more general Galaxy feature so we added a description of how GraphClust2 can be executed via a CLI in a new section of the readme on GitHub.
Parallelization is supported by Galaxy and the Docker Galaxy flavor which has

	GraphClust2 installed. In this sense, Galaxy is more or less just an abstraction layer to various HPC-schedulers and can be used to schedule jobs to your local HPC environment or Clouds. > 2. GraphClust Galaxy tools: The names aren't very explanatory, and it was slightly hard to navigate the first time. The names of the shared workflows aren't descriptive either and it isn't very clear what each workflow corresponds to. The users can find useful documentation by visiting the GitHub repository, but in my opinion the two shouldn't be dependent on one another. #Authors: We have extended the content of the front page. > The names of the individual tools are a bit confusing as well. A flowchart is provided in Figure 1, but I found it challenging trying to build a Galaxy workflow directly from that. One solution to this could be for the authors to include the names of the corresponding Galaxy tools in parenthesis, right within their matching component on the flowchart #Authors: The mapping between Figure 1 flowchart and matching Galaxy tools can be found within the GraphClust2 documentation and the front page of https://graphclust.usegalaxy.eu. Under the "GraphClust pipeline overview" section of the homepage, a number-annotated copy of Figure 1 is provided. The number-annotations are matched to the table at the bottom, where the components are described and linked. Furthermore, we provide implemented galaxy workflow instances (based on the flowchart), such that a user can reuse our pre-build workflow and build-upon on that. The workflows and all results of this paper are linked from the front page. > 3. GraphClust2 galaxy tour: I found the demo screen within the actual galaxy GraphClust screen confusing. When running in demo mode, from my experience with GraphClust2, Galaxy is basically mirroring the main webpage within the work panel of a tool with things overlapping. This makes the various demo steps hard to follow. I understand that sometimes issues like this one are due to framework limitations, but it would be nice if the authors could find an alternative solution to improve this #Authors: The html mirroring issue has been fixed, thanks for noticing.
Additional Information:	
Question	**Response**
Are you submitting this manuscript to a special series or article collection?	No
Experimental design and statistics Full details of the experimental design and statistical methods used should be given in the Methods section, as detailed in our Minimum Standards Reporting Checklist. Information essential to interpreting the data presented should be made available in the figure legends. Have you included all the information requested in your manuscript?	Yes

Resources A description of all resources used, including antibodies, cell lines, animals and software tools, with enough information to allow them to be uniquely identified, should be included in the Methods section. Authors are strongly encouraged to cite Research Resource Identifiers (RRIDs) for antibodies, model organisms and tools, where possible. Have you included the information requested as detailed in our Minimum Standards Reporting Checklist?	Yes
Availability of data and materials All datasets and code on which the conclusions of the paper rely must be either included in your submission or deposited in publicly available repositories (where available and ethically appropriate), referencing such data using a unique identifier in the references and in the “Availability of Data and Materials” section of your manuscript. Have you have met the above requirement as detailed in our Minimum Standards Reporting Checklist?	Yes

GigaScience, 20xx, 1–16

doi: xx.xxxx/xxxx

Manuscript in Preparation

PAPER

GraphClust2: annotation and discovery of structured RNAs with scalable and accessible integrative clustering

Milad Miladi[1], Eteri Sokhoyan[1], Torsten Houwaart[2], Steffen Heyne[3], Fabrizio Costa[4], Björn Grüning[1,5,*] and Rolf Backofen[1,5,6,*]

[1]Bioinformatics Group, Department of Computer Science, University of Freiburg, Germany and [2]Institute of Medical Microbiology and Hospital Hygiene, University of Dusseldorf, Germany and [3]Max Planck Institute of Immunobiology and Epigenetics, Freiburg, Germany and [4]Department of Computer Science, University of Exeter, UK and [5]ZBSA Centre for Biological Systems Analysis, University of Freiburg, Germany and [6]Center for Biological Signaling Studies (BIOSS), University of Freiburg, Germany

[*]To whom correspondence should be addressed: backofen@informatik.uni-freiburg.de; gruening@informatik.uni-freiburg.de;

Abstract

RNA plays essential roles in all known forms of life. Clustering RNA sequences with common sequence and structure is an essential step towards studying RNA function. With the advent of high-throughput sequencing (HTS) techniques, experimental and genomic data are expanding to complement the predictive methods. However, the existing methods do not effectively utilize and cope with the immense amount of data becoming available.
Hundreds of thousands of non-coding RNAs (ncRNAs) have been detected, however, the annotation of these ncRNAs is lacking behind. Here we present GraphClust2, a comprehensive approach for scalable clustering of RNAs based on sequence and structural similarities. GraphClust2 bridges the gap between HTS and structural RNA analysis, and provides an integrative solution by incorporating diverse experimental and genomic data in an accessible manner via the Galaxy framework. GraphClust2 can efficiently cluster and annotate large datasets of RNAs and supports structure probing data. We demonstrate that the annotation performance of clustering functional RNAs can be considerably improved. Furthermore, an off-the-shelf procedure is introduced for identifying locally conserved structure candidates in long RNAs. We suggest the presence and the sparsity of phylogenetically conserved local structures for a collection of long non-coding RNAs. By clustering data from two CLIP experiments, we demonstrate the benefits of GraphClust2 for motif discovery under the presence of biological and methodological biases. Finally, we uncover prominent targets of double-stranded RNA binding protein Roquin-1, such as BCOR's 3'UTR that contains multiple binding stem-loops which are evolutionary conserved.

Key words: RNA secondary structure; structure-based clustering of RNAs; ncRNA annotation and discovery; Comparative RNA analysis;

Background

High throughput RNA sequencing and computational screens have discovered hundreds of thousands of non-coding RNAs (ncRNAs) with putative cellular functionality [1, 2, 3, 4, 5, 6, 7]. Functional analysis and validation of this vast amount of data demand a reliable and scalable annotation system for the ncRNAs, which is currently still lacking for several reasons. First, it is often challenging to find homologs even for many validated functional ncRNAs as sequence similarities can be very

Compiled on: August 23, 2019.
Draft manuscript prepared by the author.

low. Second, the concept of conserved domains, which is quite successfully applied for annotating proteins, is not well-established for ribonucleic acids.

For many ncRNAs and regulatory elements in messenger RNAs (mRNAs), however, it is well known that the secondary structure is better conserved than the sequence, indicating the paramount importance of structure for the functionality. This fact has promoted annotation approaches that try to detect structural homologs in the forms of RNA *families* and *classes* [8]. Members of an RNA family are similar and typically stem from a common ancestor, while RNA classes combine ncRNAs that overlap in function and structure. A prominent example of an RNA class whose members share a common function without a common origin is microRNA. One common approach to detect ncRNA of the same class is to align them first by sequence, then predict and detect functionally conserved structures by applying approaches like RNAalifold [9], RNAz [10], or Evofold [11]. A large portion of ncRNAs from the same RNA class, however, have a sequence identity of less than 70%. In this sequence identity range, sequence-based alignments are not sufficiently accurate [12, 13]. Alternatively, approaches for simultaneous alignment and folding of RNAs such as Foldalign, Dynalign, LocARNA [14, 15, 16] yield better accuracy.

Clusters of ncRNA with a conserved secondary structure are promising candidates for defining RNA families or classes. In order to detect RNA families and classes, Will et al. [17] and Havgaard et al. [14] independently proposed to use the sequence-structure alignment scores between all input sequence pairs to perform hierarchical clustering of putatively functional RNAs. However, their applicability is restricted by the input size, due to the high quartic computational complexity of the alignment calculations over a quadratic number of pairs. Albeit the complexity of similarity computation by pairwise sequence-structure alignment can been reduced to quadratic $O(n^2)$ of the sequence length [18], it is still infeasible for most of the practical purposes with several thousand sequence pairs. For the scenarios of this scale, alignment-free approaches such as GraphClust [19] and Nofold [20] propose solutions.

A stochastic context-free grammar (SCFG), also known as covariance model (CM), encodes the sequence and structure features of a family in a probabilistic profile. CM-base approaches have been extensively used, e.g. for discovering homologs of known families [21] or comparing two families [22]. Profile-based methods [20, 23] such as Nofold generally rely on a CM database of known families to annotate and cluster sequences by comparing against the profiles, therefore their applicability for *de novo* family or motif discovery is affected by the characteristics of the already known families and the provided models.

The GraphClust methodology uses a graph kernel approach to integrate both sequence and structure information into high-dimensional sparse feature vectors. These vectors are then rapidly clustered, with a linear-time complexity over the number of sequences, using a locality sensitive hashing technique. While this solved the theoretical problem, the use case guiding the development of the original GraphClust work, here as GraphClust1, was tailored for a user with in-depth experience in RNA bioinformatics that has already the set of processed sequences at hand, and now wants to detect RNA family and classes in this set. However, with the increasing amount of sequencing and genomic data, the tasks of detecting RNA family or classes and motif discovery have been broadened and are becoming a standard as well as appealing tasks for the analysis of high-throughput sequencing (HTS) data.

To answer these demands, here we propose GraphClust2 as a full-fledged solution within the *Galaxy* framework [24]. With the development of GraphClust2, we have materialized the following goals, GraphClust2 is: (i) allowing a smooth and seamless integration of high-throughput experimental data and genomic information; (ii) deployable by the end users less experienced with the field of RNA bioinformatics; (iii) easily expandable for up- and downstream analysis, and allow for enhanced interoperability; (iv) allowing for accessible, reproducible and scalable analysis; and (v) allowing for efficient parallelizations over different platforms; To assist the end users, we have developed auxiliary data processing workflows and integrated alternative prediction tools. The results are presented with intuitive visualizations and information about the clustering.

We show that the proposed solution has an improved clustering quality in the benchmarks. The applicability of GraphClust2 will be shown in some sought-after and prevailing domain scenarios. GraphClust2 supports structure probing data such as from SHAPE and DMS experiments. It will be demonstrated that the structure probing information assists in the clustering procedure and enhances the quality. By clustering ncRNAs from *Arabidopsis thaliana* with genome-wide in vivo DMS-seq data, we demonstrate that the genome-wide probing data can in practice be used for homologous discovery, beyond singleton structure predictions. Furthermore, an off-the-shelf procedure will be introduced to identify locally conserved structure candidates from deep genomic alignments, by starting from a custom genomic locus. By applying this methodology to a couple of well-studied long non-coding RNAs (lncRNAs), we suggest the presence and the sparsity of local structures with highly reliable structural alignments. GraphClust2 can be used as a structure motif-finder to identify the precise structural preferences of RNA binding proteins (RBPs) in cross-linking immunoprecipitation (CLIP) data. By comparing public CLIP data from two double-stranded RBPs SLBP and Roquin-1, we demonstrate the advantage of a scalable approach for discovering structured elements. Under subjective binding preferences of Roquin-1 and the protocol biases, a scaled clustering uncovers structured targets of Roquin-1 that are evolutionary conserved. Finally, we propose BCOR's mRNA as a prominent binding target of Roquin-1 that contains multiple stem-loop binding elements.

Materials and Methods

Methods overview

The clustering workflow. The GraphClust approach can efficiently cluster thousands of RNA sequences. This is achieved through a workflow with five major steps: (i) pre-processing the input sequences; (ii) secondary structure prediction and graph encoding; (iii) fast linear-time clustering; (iv) cluster alignment and refinement, with an accompanying search with alignment models for extra matches; and finally (v) cluster collection, visualization and annotation; An overview of the workflow is presented in Figure 1.

More precisely, the pre-processed sequences are individually folded according to the thermodynamic free energy models with the structure prediction tools RNAfold [25] or RNAshapes [26]. A decomposition graph kernel is then used to efficiently compute similarity according to the sequence and structure features of secondary structure graphs. The MinHash technique [27] and inverse indexing are used to identify the initial clusters, which correspond to dense neighborhoods of the graph feature space. Formal description and formulations of kernel and MinHash methods are provided in section S1 of the supplementary document. The MinHash clustering approach is very fast with a linear runtime complexity over the number of entries. This accordingly makes GraphClust2 much more efficient than the quadratic all-vs-all approach [19]. It

permits the clustering of up to hundreds of thousands of RNA sequences in a reasonable time frame.

After the MinHash clustering step, the *initial* clusters are refined using the RNA domain-specific tools; Firstly, from the sequences of each *initial cluster* a UPGMA tree is created to prune the clusters. The pairwise distances of the tree are approximated from LocARNA sequence-structure alignment scores, as is proposed and detailed in [17]. This pruning procedure keeps the subtree which has the highest average pairwise alignment on its leaves. Here we use the RNA domain-specific scores from LocARNA alignments, although it would have been possible to compute distances from the generic graph kernel scores. LocARNA scores are used, since the runtime complexity is not a concern here, as the pairwise alignments are only computed within each cluster. Each cluster has typically about 10-100 sequences, which is much smaller than the entire input data; In the second step after pruning, the multiple alignment of each pruned cluster is refined with CMfinder's expectation maximization algorithm [28]; Thirdly after the alignment is refined, a homology search using Infernal [21] tools is applied over the entire dataset. Such that for each cluster's refined alignment, a covariance model (CM) is built using cmbuild. The CM is then used to scan the entire sequence database using cmsearch. This CM homology search step extends the clusters with additional homologs that have been missed in the initial clustering; Finally, the sequences of each cluster are aligned with LocARNA, the consensus structures are predicted, visualized and annotated by conservation and covariation metrics.

In an iterative fashion, the steps downstream of the fast clustering can be repeated over the sequences which are not clustered in the previous iteration. GraphClust2 can also compute fuzzy soft overlapping clusters. The option to report overlapping clusters instead of a hard optimal partitioning can be set by the user at the cluster report step. Furthermore, a pre-clustering optional step can be invoked to remove near identical and redundant sequences using CD-HIT [29]. This pre-clustering would be beneficial for the datasets with high redundancy or very large number of sequences, e.g. metatranscriptomics data.

Workflow input. GraphClust2 accepts a set of RNA sequences as input. Sequences longer than a defined length are split and processed with a user-defined sliding window option. Two recommended settings are provided for ncRNA clustering and motif discovery as will be discussed in the workflow flavors section below. In addition to the standard FASTA formatted input, a collection of auxiliary workflows are implemented to allow the user to start from genomic coordinate intervals in BED format, or genomic alignments from orthologous regions in MAF format, or sequencing data from the structure probing experiments. Use case scenarios are detailed in the upcoming sections.

Workflow output. The core output of the workflow is the set of clustered sequences. Clusters can be chosen either as *hard partitions* having an empty intersection or as overlapping *soft partitions*, in the latter case elements can belong to more than one cluster. In-depth information and comprehensive visualizations about the partitions, cluster alignments and structure conservation metrics are produced (Figure 1). The consensus secondary structure of the cluster is annotated with base-pairing information such as statistically significant covariations that are computed with R-scape [30]. Evaluation metrics for structure conservation are reported. In the case of MAF input, color-coded UCSC tracks are automatically generated to locate and annotate conserved clusters in the genome browser. The in-browser integrated view of the clusterings makes it possible to quickly inspect the results. The Galaxy server keeps track of the input, intermediate and final outputs. The clustering results can be shared or downloaded to the client system.

Workflow flavors. Two preconfigured flavors of the workflow are offered for the local and the global scenarios, to facilitate the users without demanding an in-depth knowledge about configuring complex tools. The global flavor aims for clustering RNAs on the whole transcript, such as for annotating ncRNAs of short and medium lengths. The local flavor serves as the motif-finder. The motivation has been to orderly deal with putative genomic sequence contexts around the structured elements. Prediction methods usually require different settings in these two scenarios [31]. The main differences between the flavors are the pre-configured window lengths (~250 vs. ~100), the aligner parameters and the hit criteria of the covariance model search (E-value vs. *bit score*). The motif-finder flavor can be for example used to identify cis-regulatory elements, where it is expected to find structured motifs within longer sequences.

As a feature, the fast clustering can be tuned to weigh in sequence-based features. The graph for each entry consists of two disjoint parts. The primary part is the structure graph where the vertices are labeled with the nucleotides while the backbone and paired bases are connected by edges. Besides the primary part, a *path graph* can be included to represent the nucleotide string (option -seq-graph-t). By including the path graphs, sequence-only information would independently contribute to the feature vectors.

Integration of structure probing data

RNA structure probing is an emerging experimental technique for determining the RNA pairing states at nucleotide resolution. Chemical treatment with reagents like SHAPE (selective 2'-hydroxyl acylation analyzed by primer extension) and DMS (dimethyl sulfate) [32, 33] provide one-dimensional reactivity information about the accessibility of nucleotides in an RNA molecule. Structure probing (SP) can considerably improve the secondary structure prediction accuracy of RNAs [34, 35, 36]. SP-assisted computational prediction methods commonly incorporate the probing data by guiding the prediction algorithms via folding constraints and pseudo-energies [37, 38, 25]. Deigan et al. first introduced the position-specific pseudo-energy terms to incorporate the reactivity information alongside the free energy terms of thermodynamic models [39]. The pseudo-energy term for position i is defined as:

$$\Delta G_{\text{pseudo-energy}}(i) = m \ln(1 + \text{reactivity}(i)) + b$$

where parameters m and b determine a scaled conversion of the reactivities to the energy space. GraphClust2 supports structure probing data for enabling a guided structure prediction [25, 40]. The structure probing support is integrated into the pre-processing and the structure prediction steps to generate SP-directed structure graphs.

Implementation and installation

GraphClust2 is implemented within the Galaxy framework [24]. Galaxy offers several advantages to assist our goal of developing a scalable and user-friendly solution. The platform makes it convenient to deploy complex workflows with interoperable tools. Through the uniform user interface across different tools, it is easier for the users to work with new, unfamiliar tools and freely interchange them. Moreover, the standard-

ized data types will ensure that only inputs with valid types are passed to a tool. Interactive tutorial tours are produced to introduce the user interface and guide the user through sample clustering procedures.

GraphClust2's toolset has been made publicly available in Galaxy ToolShed [41] and can be easily installed into any Galaxy server instance. GraphClust2 is available also as a standalone container solution for a variety of computing platforms at: https://github.com/BackofenLab/GraphClust-2 and can be freely accessed on the European Galaxy server at: https://graphclust.usegalaxy.eu.

The workflow implementation. GraphClust2 workflow is composed of tools and scripts which are packaged in Bioconda and Biocontainers [42] and integrated into the Galaxy framework. This has enabled automatic installation of the tools in a version-traceable and reproducible way. All functional units and workflows are manually validated and are under extensive continuous integration tests. Strict versioning of tools and requirements ensures reproducible results over multiple different versions of a tool while delivering updates and enhancements.

Platform-independent virtualised container. GraphClust2 can be deployed on any Galaxy server instance, simply by installing the GraphClust tools from the Galaxy ToolShed. As a standalone solution, virtualised Galaxy instance based on Linux containers (Docker, rkt) [43] are provided that can be executed on Linux, OSX and Windows. This largely simplifies the deployment phase, guarantees a reproducible setup and makes it instantiable on numerous computation systems from personal computers to Cloud and high-performance computing (HPC) environments. The Docker image is based on the official Galaxy Docker image [44, 45] and is customized to integrate GraphClust2 tools, workflows and tutorial tours.

Data

Rfam-based simulated SHAPE. A set of Rfam [46] sequences and the associated SHAPE reactivities were extracted from the ProbeAlign benchmark dataset [47]. The simulated SHAPE reactivities have been generated according to the probability distributions that are fitted to experimental SHAPE data by Sükösd et al. methodology [48]. Rfam families containing at least ten sequences were used. A uniformly sized subset was also extracted, where exactly ten random sequences were selected per family to obtain a variation with a uniform unbiased contribution from each family.

Arabidopsis thaliana ncRNA DMS-seq. Arabidopsis DMS-seq reads were obtained from the structure probing experiment by Ding et al. [49] (NCBI SRA entries SRX321617 and SRX320218). The reads were mapped to TAIR-10 ncRNA transcripts (Ensembl release-38) [50]. Reactivities were computed for non-ribosomal RNAs based on the normalized reverse transcription stop counts using Structure-Fold tool in Galaxy [51]. We used Bowtie-2 [52] with the settings recommended by [53] (options –trim5=3, -N=1). Transcripts with poor read coverage tend to bias towards zero-valued reactivities [54]. To mediate this bias, low information content profiles with less than one percent non-zero reactivities were excluded. To focus on secondary structure predictions of the paralogs that can have high sequence similarity, the graphs were encoded with the primary part without path graphs. Information about the ncRNA families is available in the Supplementary Table S4.

Orthology sequence extraction from long RNA locus The genomic coordinates of the longest isoforms were extracted from RefSeq hg38 annotations [55] for FTL mRNA and lncRNAs NEAT1, MALAT1, HOTAIR and XIST. To obtain the orthologous genomic regions in other species, we extracted the genomic alignment blocks in Multiz alignment format (MAF) [56] for each gene using the UCSC table browser [57] (100way-vertebrate, extracted in Aug. 2018). Alignments were directly transferred to the Galaxy server via the UCSC-to-Galaxy data importer. MAF blocks were concatenated using MAF-Galaxy toolset [58] to obtain one sequence per species. An auxiliary workflow for this data extraction procedure is provided. This procedure is notably scalable and can be applied to any locus independent of the annotation availability. Alternatively, the user can provide e.g. full transcripts or synteny regions [59] for the downstream analysis. For the background shuffled input, Multiperm [60] was used to shuffle the Multiz alignment of MALAT1 locus.

SLBP eCLIP. Binding sites of SLBP were obtained from the ENCODE eCLIP project (experiment ENCSR483NOP) [61]. In the consortium's workflow, CLIPper [62] is used to extract peak regions of the read coverage data. The peaks are annotated with both p-values and log2-fold-change scores. These values are determined from the read-counts of the experiment compared with the read-counts of a size matched input. We extracted the peaks with a log2-fold-change of at least 4. To diminish the chance of missing the binding motif, the peak regions were extended by 60 nucleotides both up- and downstream. The sequences of the resulting 3171 binding target regions were used for clustering and motif analysis.

Roquin-1 PAR-CLIP. The 16000+ binding sites of Roquin-1 (RC3H1) were obtained from Murakawa et al. [63] (hg18 coordinates from the associated mdc-berlin web page). The 5000 target regions with the highest PAR-CLIP scores were used for the downstream analysis and structural clustering. The binding sites sequences were extracted using the *extract-genomic-dna* tool in Galaxy.

Structure conservation annotation with Evofold, RNAz and R-scape

For each of the studied long RNAs, the sequences were extracted from the orthologous genomic regions as detailed in the data preparation section. Clustering was performed using the motif-finder flavor. In the preprocessing step, the sliding window was set to 100b length and 70% shift. The LocARNA structural alignments of the predicted clusters were further processed using RNAz [64], Evofold [11] and R-scape [30], to *annotate* clusters with structure conservation potentials in the generated genomic browser tracks. RNAz uses a support vector machine (SVM) that is trained on structured RNAs and background to evaluate the thermodynamic stability of sequences folded freely versus constrained by the consensus structure. Evofold uses phylo-SCFGs to evaluate a conservation model for local structures against a competing nonstructural conservation model. R-scape quantifies the statistical significance of base-pair covariations as evidence of structure conservation, under the null hypothesis that alignment column pairs are evolved independently.

RNAz was invoked (option –locarnate) with the default 50% cutoff for SVM-class probability to annotate the clusters. Evofold was also run with the default parameters over the cluster alignments and supplied with the corresponding hg38-100way UCSC's phylogenetic tree [56]. Clusters that were predicted by Evofold to contain at least one conserved structure with more than three base-pairs were annotated as Evofold hits. R-scape was also applied with the default parameters (i.e. G-test statistics –GTp), clusters with at least two significant covariations

were annotated. Clusters are constrained to have a depth of at least 50 sequences. Alignments with spurious consensus structure or no conservation were excluded, using a structure conservation index (SCI) filter of one percent [64]. Clusters annotated by at least one of the three methods are designated as *locally conserved structure candidates*.

Clustering performance metric

The clustering was benchmarked similarly to our previous work [65], such that the Rfam family where each input RNA belongs to is considered as the truth reference class. The performance is measured using the *adjusted Rand index* (ARI) [66] clustering quality metric, which is defined as:

$$\text{Adjusted Rand Index} = \frac{\text{Rand Index} - E[\text{Rand Index}]}{1 - E[\text{Rand Index}]}$$

The Rand Index [67] measures the fraction of the entry pairs that are related in the same way in both the predicted clustering and the reference assignment. E[Rand Index] is the expected *Rand Index* (for extended details please refer to [65]). The adjusted Rand index is the corrected-for-chance variation of the Rand Index with a maximum value of one. A better agreement between the predicted clustering and the reference assignment leads to a higher ARI value.

Results and Discussion

Clustering performance evaluation

Rfam-cliques benchmark. We evaluated GraphClust2 using known RNA families from the Rfam database [46]. The Rfam sequences were obtained from the *Rfam-cliques* benchmark introduced in our previous work [65]. The Rfam-cliques benchmark contains sets of RNA families at different sequence identity levels and allows for benchmarking a tool for the cases of low and high sequence identities (*Rfam-cliques-low* and *Rfam-cliques-high*). Each variation contains a collection of human members of the Rfam families together with homologs in the other species. As we wanted to evaluate the performance in a simulated scenario of genome-wide screening, we selected the human paralogs from the benchmarks and measured (using the adjusted Rand index metric) how well GraphClust1 and the new pipeline GraphClust2 correctly cluster members of the families together.

In comparison to GraphClust1, GraphClust2 provides alternative approaches for the identification of the secondary structures. Using similar configurations as in GraphClust1 [19], i.e. RNAshapes for structure prediction and bit score for CM search hits, the clustering performance of GraphClust2 is similar or better due to the integration of upgraded tools. However, the alternative configuration of RNAfold for structure prediction and E-value for CM search hits consequently improves the performance (ARI from 0.641 to 0.715 for Rfam-cliques-high, further details in Supplementary Table S1).

SHAPE-assisted clustering improves the performance. In the previous benchmark, the clustering relies on the free energy models for secondary structure prediction. A predicted structure sometimes deviates from the real functional structure due to the cellular context and folding dynamics. In this case, the structure probing SHAPE data associated with the real functional structure is expected to improve the quality of structure prediction, which in turn should improve the clustering. We wanted to investigate how an improvement in the structure prediction quality at the early clustering steps influences the final clustering results. To draw a conclusion, however, an extensive SHAPE data would be needed for a set of labeled homologous ncRNAs, ideally with different sequence identity and under similar experimental settings. Currently, such collection of data, especially over multiple organisms, is still unavailable. However, as the structure probing is turning into a standard and common procedure, data of such nature is expected to become available soon.

One solution to the mentioned data scarceness is provided in the literature [48], by simulating the experimental generation of a SHAPE profile from the real functional structure. Here, starting from a set of manually curated reference structures, the idea is to simulate SHAPE profiles that reflect the known reference structures. We used the benchmark from ProbeAlign [47] (see Material and Methods for details). Figure 2 shows the effect of incorporating simulated SHAPE data on clustering by guiding the structure prediction. As can be seen, the incorporation of SHAPE data has improved the clustering performance. Notably, an improvement can be achieved in fewer rounds of clustering iterations.

GraphClust2 is scalable.

To validate the GraphClust2's scalability and linearity claims, we used a millions-sequence biological dataset. GraphClust2 is implemented with a comprehensive support for parallel computation using the Galaxy framework. The MinHash-based clustering step is the only step where the entities are evaluated altogether to identify the dense neighborhoods as clusters. Thanks to the MinHash technique, this step has only a linear complexity (see methods and supplementary section S1). To empirically validate this, we clustered a large metatranscriptomic dataset of a marine sample from [68]. After merging the paired-end reads, the metatranscriptome contained 3,594,198 sequences with an average length 250 bases. To filter highly similar sequences, we performed sequence-based pre-clustering with CD-HIT set at a 90% similarity threshold. This produced about 913,000 sequences with a total of 195 million bases. GraphClust2 identified several large clusters of sizes larger than one hundred in one round. Translation-complex-related RNAs (tRNA, LSU and 5s rRNAs) were among the dominating ncRNA classes, matching the expectation due to the high expression levels of the families. Please refer to the supplementary Table S2 for further details. Clustering the entire 3.6 million sequences took less than a day on the European Galaxy server. To check the runtime growth over number of inputs, we measured the wall-clock runtime for sub-samples of various sizes on the European Galaxy server. GraphClust2 robustly scaled with a linear trend over the size of the input (Supplementary Figure S1).

Clustering *Arabidopsis* ncRNAs with DMS-seq in vivo structure probing data

As shown in the previous section, we expect structure probing information to improve the clustering. Information about the structure formations in vivo can be obtained from structure probing (SP) techniques. By determining the nucleotide-resolution base reactivities, where positions with high reactivity indicate unbound bases. Recently, high-throughput sequencing has enabled SP to be applied in a genome-wide manner, thus providing structure probing reactivities of an entire transcriptome [69]. In this way, large amount of SP data can be obtained. Despite the availability of genomic-wide SP data, its application for transcriptome-wide structure analysis is promising [70] but has remained largely underutilized.

Enhanced ncRNA annotation with in vivo SP data. We thus evaluated how the task of clustering and annotation of ncRNAs can

benefit from such type of genome-wide probing experiments. For this, we compared clusterings of *Arabidopsis thaliana* ncRNAs with and without considering the DMS-seq data by Ding et al. [49] (see Materials and Methods). Due to the relatively high sequence similarity of the annotated paralogous ncRNAs of *Arabidopsis thaliana*, the *Adjusted Rand Index* is high even when no SP data is considered (-DMS-seq mode ARI 0.88). Nonetheless, the quality metric is slightly improved by incorporating the SP data (+DMS-seq mode ARI 0.91). We further manually inspected the quality of the produced clusters. Figure 3 shows the enhanced results for identifying ncRNA classes by using GraphClust2 with in vivo probing data. In the +DMS-seq mode (Figure 3B), all detected clusters are pure RNA classes, while the -DMS-seq mode (Figure 3A) produces mixed-up clusters for *Group II Introns* family plus snoRNA, miRNA and U-snRNA classes. For example, as it can be seen in Figure 3C, the SP data improves the structure prediction by predicting a conserved stem for two of the *Group II Introns* only in the +DMS-seq mode, which leads to one pure cluster for the family.

Discovering locally conserved structured in long RNAs

RNA-seq experiments from biological conditions often result in differentially expressed transcripts, which are studied for functionality and regulatory features. A differential expression hints at putative regulatory effects. An orthogonal source of information for the functional importance of a transcript is phylogenetic conservation patterns. For long non-coding RNAs, however, sequence conservation is usually low, imposing limitations on the sequence level conservation analysis. This fact has been one motivating reason for a collection of recent studies to explore the conservation and functionality of lncRNAs at the secondary structure level [71, 72, 73]. A majority of the studies have been focusing on identifying widely spanned structures, postulating the existence of a to-be-discovered single global structure. However, some of the reported conservations have been challenged for lacking trustworthy base-pair covariations in the alignments [30].

Looking for locally conserved secondary structures in lncRNAs is alluring for several reasons. First, with an increase in the base pair span length the prediction quality decreases [31], which implies that global structure prediction for long RNAs tends to be inaccurate. Second, the structure of a transcribed RNA structure is influenced by RNA-binding proteins in vivo, and thus a predicted global structure likely deviates from the real functional structure. Third, in many cases and similar to the untranslated regions in mRNAs, only a locally conserved structural motif is expected to suffice to perform a function, independent of the precise global structure. We thus revert to a frequently used strategy in the RNA field, namely to look for locally conserved structural motifs. We wanted to evaluate whether we can use GraphClust2 for this purpose.

It should be noted that distinguishing conserved structures from background genomic sequence similarity using base-pair conservation signals is a challenging task. Genome-wide screening studies over genomic alignments require adjusted thresholds for statistical significance discovery and report up to 22% [4] false discovery rates that can be even higher [74]. Despite this and due to the persistent expansion of genomic data, the depth and quality of genomic alignments are continually increasing. Currently, there is a lack of off-the-shelf tools for comprehensively analyzing locally conserved structural elements of a specific locus. Here based on GraphClust2, we propose a data extraction and structure conservation detection methodology (as detailed in the Materials and Methods) that can readily be used for desired loci and genomic alignments to identify *candidates* with locally conserved structure potentials.

An advantage of this clustering approach over traditional screening methods is its ability as an unsupervised learning method, for not imposing explicit presumption on the depth or number of predicted motifs. This makes it possible to find the locally conserved structures also in the regions where a subset of species do not have a conserved structure. Furthermore, this approach does not require a precise co-location of the conserved elements within the transcript, in contrast to traditional alignment-based screening approaches. A further advantage is the availability of the solution in the Galaxy framework, as it provides a rich collection of assets for interactive data collection and analysis of genomic data. We used the 100way vertebrate alignments to extract the orthologous genomic regions for each of the studied RNAs in human and other vertebrates. Each of the orthologous sequences is split into windows, which are then clustered by GraphClust2. The alignment of each cluster has been further annotated with some of best practice complementary methods in assessing covariation patterns and structure conservation potentials, namely RNAz, Evofold and R-scape (see Material and Methods for details). In the following section, some example studies are presented.

Locally conserved candidates with reliable alignments are observable but uncommon

We investigated clustering of orthologous genomic regions of FTL mRNA and four well-studied lncRNAs, using the approach described before. The selected lncRNAs have been previously reported for having loss-of-function phenotypes [75, 76]. In Figures 4A-D and S3 the locations of locally conserved candidates are displayed. These locations are automatically generated by GraphClust2 from clusters with conserved structures (*candidate motifs* track). The track is automatically annotated and filtered using the computed metrics of Evofold, RNAz and R-scape tools (see methods and Figure 4 top-right legend). For these studied lncRNAs, an additional track (*manually curated subset*) is provided. The track is the selection subset of *candidate motifs* track which are manually further screened and selected by stringent expert criteria. The intention was to identify confident conserved elements that can be used e.g. for mutational experiments. The clusters were manually curated in a qualitative manner by inspecting the alignments, their consensus structures as well as the conservation metrics. Only the highly reliable structural alignments which posed a good level of covariation and were not deemed to be alignment artifacts were selected. The main filtering out criteria were: singleton compensatory mutations; avoidable column shifts producing artificial mutations; absence of any region with a basic level of sequence conservation; and similar frequencies of variations in both unpaired and compensatory mutated paired regions; Below we describe the observations from these lncRNAs' conservation analyses.

FTL: The cis-acting *Iron Response Element* (IRE) is a conserved structured element, that is located on the 5'UTR of *FTL* (*ferritin light chain*) and several other genes. Mutations that disrupt the hairpin structure of IRE cause disease phenotypes by changing the binding affinity of a regulatory Iron Response Protein [77, 78]. As a proof of concept, we applied GraphClust2 to discover structural motifs in the homologous regions of the *FTL* mRNA. The IRE element was identified as one of the three clusters detected by Evofold (Figure 4A).

NEAT1: The NEAT1 analysis suggests very limited but also very reliable structure conservation at the 3' end of the transcript that is consensually detected by the three evaluated tools.

MALAT1: MALAT1 has relatively a higher level of sequence conservation among the four studied lncRNAs. A higher number of clusters were predicted with a couple of reliable candidates that lean towards the 3' side of the transcript.

To examine how many of the detected motifs are expected to be false positive predictions, we ran the pipeline on ten shufflings of the MALAT1 100way alignment. For the shuffled background, we used Multiperm to preserve the gap structure, local conservation structure patterns and the relative dinucleotide frequencies of the MALAT1 alignment [60]. On average 16.7 candidates were reported for the shuffled genomic alignments, in comparison to the 23 candidates reported for the genomic alignment (Figure 4F). In the predicted candidates set from background, none was commonly annotated by the three methods. For the applied alignment depths and thresholds, Evofold had a considerably higher discovery rate than R-scape and RNAz. In total out of 10 shuffles 167, 9 and 0 clusters were predicted to have a conserved structure by Evofold, R-scape and RNAz respectively.

HOTAIR: The predicted candidates for HOTAIR are all located on the intronic regions of the precursor lncRNA. Clustering from the second exon, through skipping the first exon and intron, did not change this observation. A dense number of candidates can be noticed on the first intron that is overlapping with the promoter region of HOXC11 on the opposite strand. Most notably is the candidate cluster HOTAIR-C29, which is highly enriched in G-U wobble base pairs (Figure 4E). In contrast to Watson-Crick GC and AU base pairings, the GU reverse complement AC is not a canonical base pair [79]. Therefore, this structure can only be formed on the antisense RNA and not on the HOXC11's sense strand.

XIST: The XIST candidates are mainly located on the repeat regions and are paralog-like (Figure S3). The manual evaluation of the cluster structural alignments was inconclusive. In the mixture of paralog-like and homolog-like sequences of the cluster alignments, it was not possible to conclude whether the structural variations are merely artifacts of sequence repetition or compensatory mutations of hypothetical structure conservation.

Clustering RNA binding protein target sites

SLBP eCLIP. A well-characterized example of an RBP with specific structural preferences is SLBP (*Histone Stem-Loop-Binding Protein*). We clustered target sites of human SLBP using the publicly available eCLIP data [61]. The largest cluster with a defined consensus structure bears statistically significant base-pair covariations. The structure matches the SLBP's Rfam family *"histone 3'UTR stem-loop"* (RF00032) Using the family CM to identify SLBPs on the eCLIP data, we were able to predict exactly the same stem loop structure with the same level of base-pair-covariation (Figure 5A). GraphClust2 and Rfam's CM hits have more than 95% overlap. These correspondences demonstrate that GraphClust2 can identify the consensus structure element from CLIP data with a high sensitivity.

The stem-loop structure of the eCLIP data has a lower covariation-level than Rfam's seed alignment (Figure 5 A vs. B). This is because the Rfam data is phylogenetically diverse (RF00032 seed: 28 species) while eCLIP data is only for Human (eCLIP: K562 cell-line). We checked how GraphClust2 would perform if the eCLIP data from diverse organisms were available. To simulate an eCLIP data with high covariation level, we mixed up the 46 seed sequences of RF00032 family with 2954 shuffled sequences to obtain 3000 sequences such that SLBP is convoluted with 98.5% background. The sequences from the *full* RF00032 set were shuffled to obtain the background of same length and nucleotide content distribution. As can be seen in Figure 5B, GraphClust2 successfully managed to cluster the family entries as one cluster. Here, the cluster has the same stem-loop in the consensus secondary structure with the same covariation level as Rfam's reference structure.

Scalable clustering identifies novel CDE-like elements in Roquin-1 PAR-CLIP data. Roquin-1 is a protein with conserved double-stranded RNA binding domains that binds to a constitutive decay element (CDE) in TNF-alpha 3'UTR and several other mRNAs [80, 81]. Roquin-1 promotes mRNA decay and plays an essential role in the post-transcriptional regulation of the immune system [82]. We clustered the binding sites of the publicly available Roquin-1 PAR-CLIP data [63] with GraphClust2. Clustering identified structured elements in three dominant clusters with defined consensus structures. Figure 5C shows the alignments and consensus structures of the three clusters. The consensus structures are similar to the previously reported CDE and CDE-like elements [83].

It should be noted that the union of Roquin-1's CDE-like motifs have a lower enrichment score based on the PAR-CLIP ranks, in comparison to the SLBP motif based on the eCLIP ranks (Figures 5D,E and S4). For example, only 6 of CDE-like motifs are within the top 100 PAR-CLIP binding sites. Therefore, only the clustering of a broader set of binding targets, with a permissive score threshold, allows identifying the CDE-like elements reliably. We hypothesize that two reasons contribute to the observed distinction. Firstly, eCLIP is an improved protocol with a size-matched input to capture background RNAs of the CLIP protocol [61]. On the other hand, PAR-CLIP is known to have relatively higher false positive rates [84]. Secondly, the ROQ domain of Roquin-1 has two RNA binding sites, one that specifically recognizes CDE-like stem-loops and one that binds to double-stranded RNAs [83, 85]. This would likely broaden the Roquin-1 binding specificity beyond CDE-like stem-loops.

BCOR 3'UTR is a prominent conserved target of Roquin-1. We performed a follow-up conservation study over the identified CDE-like motifs from the clustering of Roquin-1 binding sites (Figure 6A). By investigating RNAalifold consensus structure predictions for Multiz alignments of the top 10 binding sites of the conserved candidates, the BCOR's CDE-like motif was observed to have a highly reliable consensus structure with supporting levels of compensatory mutations. Interestingly the reported CLIP binding site region contains two conserved stem-loops (Figure 6B,C). The shorter stem-loop has a double-sided base-pair covariation and the longer stem-loop contains bulges and compensatory one-sided mutations (Figure 6D,E). Downstream of this site, further binding sites with lower affinities can be seen, where one contains another CDE-like motif. So in total BCOR's 3'UTR contains three CDE-like motifs (Figure 6F). BCOR has been shown to be a corepressor of BCL6 which is a major sequence-specific transcription repressor. BCL6 expression is tightly regulated and induced by cytokines signaling like Interleukins IL4/7/21 [86, 87]. Overall these results propose BCOR to be a functionally important target of Roquin-1 and assert the role of Roquin-1 in regulating follicular helper T cells differentiation and immune homeostasis pathways [81].

Conclusion

We have presented a method for structural clustering of RNA sequences with a web-based interface within the Galaxy framework. The linear-time alignment-free methodology of GraphClust2, accompanied by cluster refinement and extension using

RNA comparative methods and structure probing data, were shown to improve the detection of ncRNA families and structurally conserved elements. We have demonstrated on real-life and complex application scenarios that GraphClust2 provides an accessible and scalable way to perform RNA structure analysis and discovery.

GraphClust2 provides an integrative solution, which can start from raw HTS and genomic data and ends with predicted motifs with extensive visualizations and evaluation metrics. The users can benefit from the vast variety of the bioinformatics tools integrated by the Galaxy community and extend these applications in various ways. Thus, it will be for the first time possible to start from putative ncRNAs in transcriptomic RNA-seq studies and immediately cluster the identified transcripts for annotation purposes in a coherent manner.

Availability of source code and requirements

- Project name: GraphClust2
- Project repository: https://github.com/BackofenLab/GraphClust-2
- Project home page: https://graphclust.usegalaxy.eu
- Galaxy tools repository: https://github.com/bgruening/galaxytools/tools/GraphClust/
- Operating system(s): Unix (Platform independent with Docker)
- GraphClust2 Docker image: https://hub.docker.com/r/backofenlab/docker-galaxy-graphclust
- License: GNU GPL-v3
- RRID: SCR_017286

Availability of supporting data and materials

The data presented here that illustrates our work is available from Zenodo [88] and all steps taken for data analysis are accessible via a collection of Galaxy histories from the project homepage at the European Galaxy server (https://graphclust.usegalaxy.eu).

Declarations

List of abbreviations

ARI: adjusted Rand index; CDE: constitutive decay element; CLIP: cross-linking immunoprecipitation; CM: covariance model; DMS: dimethyl sulfate; HPC: high-performance computing; HTS: high-throughput sequencing; lncRNA: long non-coding RNA; MAF: Multiz alignment format; mRNA: messenger RNA; ncRNA: non-coding RNA; RBP: RNA binding protein; SCFG: stochastic context-free grammar; SHAPE: selective 2'-hydroxyl acylation analyzed by primer extension; SP: Structure probing;

Funding.
This work was supported by German Research Foundation Collaborative Research Centre 992 Medical Epigenetics (DFG grant SFB 992/1 2012) and German Federal Ministry of Education and Research (BMBF grant 031 A538A RBC (de.NBI)).

Competing Interests

The author(s) declare that they have no competing interests

Acknowledgements

We thank Freiburg Galaxy team for their support. We thank Sean Eddy for the helpful comments and discussions. We also thank Sita J. Saunders and Mehmet Tekman for providing feedback about this manuscript.

References

1. Uzilov AV, Keegan JM, Mathews DH. Detection of noncoding RNAs on the basis of predicted secondary structure formation free energy change. BMC Bioinformatics 2006;7(1):173.
2. Weinberg Z, Wang JX, Bogue J, Yang J, Corbino K, Moy RH, et al. Comparative genomics reveals 104 candidate structured RNAs from bacteria, archaea, and their metagenomes. Genome biology 2010;11(3):R31.
3. Will S, Yu M, Berger B. Structure-based whole-genome realignment reveals many novel noncoding RNAs. Genome research 2013;.
4. Smith MA, Gesell T, Stadler PF, Mattick JS. Widespread purifying selection on RNA structure in mammals. Nucleic Acids Research 2013;41(17):8220–8236.
5. 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 research 2017;p. gr-208652.
6. Weinberg Z, Lünse CE, Corbino KA, Ames TD, Nelson JW, Roth A, et al. Detection of 224 candidate structured RNAs by comparative analysis of specific subsets of intergenic regions. Nucleic Acids Research 2017;45(18):10811–10823.
7. Torarinsson E, Yao Z, Wiklund ED, Bramsen JB, Hansen C, Kjems J, et al. Comparative genomics beyond sequence-based alignments: RNA structures in the ENCODE regions. Genome research 2008;18(2):242–251.
8. Stadler PF. Class-specific prediction of ncRNAs. In: RNA Sequence, Structure, and Function: Computational and Bioinformatic Methods Springer; 2014.p. 199–213.
9. Bernhart SH, Hofacker IL, Will S, Gruber AR, Stadler PF. RNAalifold: improved consensus structure prediction for RNA alignments. BMC bioinformatics 2008;9(1):474.
10. Washietl S, Hofacker IL, Stadler PF. Fast and reliable prediction of noncoding RNAs. Proceedings of the National Academy of Sciences 2005;102(7):2454–2459.
11. Pedersen JS, Bejerano G, Siepel A, Rosenbloom K, Lindblad-Toh K, Lander ES, et al. Identification and classification of conserved RNA secondary structures in the human genome. PLoS computational biology 2006;2(4):e33.
12. Gardner PP, Giegerich R. A comprehensive comparison of comparative RNA structure prediction approaches. BMC bioinformatics 2004;5(1):140.
13. Washietl S, Hofacker IL. Consensus folding of aligned sequences as a new measure for the detection of functional RNAs by comparative genomics. Journal of molecular biology 2004;342(1):19–30.
14. Havgaard JH, Torarinsson E, Gorodkin J. Fast pairwise structural RNA alignments by pruning of the dynamical programming matrix. PLOS computational biology 2007;3(10):e193.
15. Fu Y, Sharma G, Mathews DH. Dynalign II: common secondary structure prediction for RNA homologs with domain insertions. Nucleic Acids Research 2014;42(22):13939–13948.
16. Will S, Siebauer MF, Heyne S, Engelhardt J, Stadler PF, Reiche K, et al. LocARNAscan: Incorporating thermodynamic stability in sequence and structure-based RNA homology

search. Algorithms for Molecular Biology 2013;8(1):14.
17. Will S, Reiche K, Hofacker IL, Stadler PF, Backofen R. Inferring Non-Coding RNA Families and Classes by Means of Genome-Scale Structure-Based Clustering. PLoS Comput Biol 2007;3(4):e65.
18. Will S, Otto C, Miladi M, Möhl M, Backofen R. SPARSE: quadratic time simultaneous alignment and folding of RNAs without sequence-based heuristics. Bioinformatics 2015;31(15):2489–2496.
19. Heyne S, Costa F, Rose D, Backofen R. GraphClust: alignment-free structural clustering of local RNA secondary structures. Bioinformatics 2012;28(12):i224–i232.
20. Middleton SA, Kim J. NoFold: RNA structure clustering without folding or alignment. RNA 2014;20(11):1671–1683.
21. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. Bioinformatics 2013;29(22):2933–2935.
22. Eggenhofer F, Hofacker IL, Backofen R. CMV-Visualization for RNA and Protein family models and their comparisons. Bioinformatics 2018;1:3.
23. Pignatelli M, Vilella AJ, Muffato M, Gordon L, White S, Flicek P, et al. ncRNA orthologies in the vertebrate lineage. Database 2016;2016.
24. Afgan E, Baker D, Batut B, Van Den Beek M, Bouvier D, Čech M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic Acids Research 2018;46(W1):W537–W544.
25. Lorenz R, Luntzer D, Hofacker IL, Stadler PF, Wolfinger MT. SHAPE directed RNA folding. Bioinformatics 2016;32(1):145.
26. Steffen P, Voß B, Rehmsmeier M, Reeder J, Giegerich R. RNAshapes: an integrated RNA analysis package based on abstract shapes. Bioinformatics 2005;22(4):500–503.
27. Broder AZ. On the resemblance and containment of documents. In: Compression and complexity of sequences 1997. proceedings IEEE; 1997. p. 21–29.
28. Yao Z, Weinberg Z, Ruzzo WL. CMfinder—a covariance model based RNA motif finding algorithm. Bioinformatics 2005;22(4):445–452.
29. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 2006;22(13):1658–9.
30. Rivas E, Clements J, Eddy SR. A statistical test for conserved RNA structure shows lack of evidence for structure in lncRNAs. Nature Methods 2017;14(1):45–48.
31. Lange SJ, Maticzka D, Möhl M, Gagnon JN, Brown CM, Backofen R. Global or local? Predicting secondary structure and accessibility in mRNAs. Nucleic Acids Research 2012;40(12):5215–5226.
32. Wilkinson KA, Merino EJ, Weeks KM. Selective 2'-hydroxyl acylation analyzed by primer extension (SHAPE): quantitative RNA structure analysis at single nucleotide resolution. Nature protocols 2006;1(3):1610.
33. Tijerina P, Mohr S, Russell R. DMS footprinting of structured RNAs and RNA–protein complexes. Nature protocols 2007;2(10):2608.
34. Kutchko KM, Laederach A. Transcending the prediction paradigm: novel applications of SHAPE to RNA function and evolution. Wiley Interdisciplinary Reviews: RNA 2017;8(1).
35. Hajdin CE, Bellaousov S, Huggins W, Leonard CW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA secondary structure modeling, including pseudoknots. Proceedings of the National Academy of Sciences 2013;110(14):5498–5503.
36. Miladi M, Montaseri S, Backofen R, Raden M. Integration of accessibility data from structure probing into RNA-RNA interaction prediction. Bioinformatics 2018;.
37. Eddy SR. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annual review of biophysics 2014;43:433–456.
38. Spasic A, Assmann SM, Bevilacqua PC, Mathews DH. Modeling RNA secondary structure folding ensembles using SHAPE mapping data. Nucleic Acids Research 2018;46(1):314–323.
39. Deigan KE, Li TW, Mathews DH, Weeks KM. Accurate SHAPE-directed RNA structure determination. Proceedings of the National Academy of Sciences 2009;106(1):97–102.
40. Grüning BA, Fallmann J, Yusuf D, Will S, Erxleben A, Eggenhofer F, et al. The RNA workbench: best practices for RNA and high-throughput sequencing bioinformatics in Galaxy. Nucleic Acids Research 2017;45(W1):W560–W566.
41. Blankenberg D, Kuster GV, Bouvier E, Baker D, Afgan E, Stoler N, et al. Dissemination of scientific software with Galaxy ToolShed. Genome Biology 2014;15(2):403.
42. Grüning B, Dale R, Sjödin A, Chapman B, Rowe J, Tomkins-Tinch C, et al. Bioconda: sustainable and comprehensive software distribution for the life sciences. Nature Methods 2018;15(7):475–476.
43. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux Journal 2014;2014(239):2.
44. Grüning B, Chilton J, van den Beek M, Batut B, Chambers M, ISHII M, et al., bgruening/docker-galaxy-stable: Galaxy Docker Image 18.09; 2018. https://doi.org/10.5281/zenodo.1251998.
45. Grüning B, Chilton J, Köster J, Dale R, Soranzo N, van den Beek M, et al. Practical computational reproducibility in the life sciences. Cell systems 2018;6(6):631–635.
46. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al. Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Research 2018;46(D1):D335–D342.
47. Ge P, Zhong C, Zhang S; BioMed Central. ProbeAlign: incorporating high-throughput sequencing-based structure probing information into ncRNA homology search. BMC bioinformatics 2014;15(9):S15.
48. Sükösd Z, Swenson MS, Kjems J, Heitsch CE. Evaluating the accuracy of SHAPE-directed RNA secondary structure predictions. Nucleic Acids Research 2013;41(5):2807–2816.
49. Ding Y, Tang Y, Kwok CK, Zhang Y, Bevilacqua PC, Assmann SM. In vivo genome-wide profiling of RNA secondary structure reveals novel regulatory features. Nature 2014;505(7485):696.
50. Zerbino DR, Achuthan P, Akanni W, Amode MR, Barrell D, Bhai J, et al. Ensembl 2018. Nucleic Acids Research 2017;46(D1):D754–D761.
51. Tang Y, Bouvier E, Kwok CK, Ding Y, Nekrutenko A, Bevilacqua PC, et al. StructureFold: genome-wide RNA secondary structure mapping and reconstruction in vivo. Bioinformatics 2015;31(16):2668–2675.
52. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nature methods 2012;9(4):357–359.
53. Ding Y, Kwok CK, Tang Y, Bevilacqua PC, Assmann SM. Genome-wide profiling of in vivo RNA structure at single-nucleotide resolution using structure-seq. Nature protocols 2015;10(7):1050.
54. Choudhary K, Deng F, Aviran S. Comparative and integrative analysis of RNA structural profiling data: current practices and emerging questions. Quantitative Biology 2017;5(1):3–24.
55. Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, et al. GENCODE: the reference human genome annotation for The ENCODE Project. Genome research 2012;22(9):1760–1774.

56. Blanchette M, Kent WJ, Riemer C, Elnitski L, Smit AF, Roskin KM, et al. Aligning multiple genomic sequences with the threaded blockset aligner. Genome research 2004;14(4):708–715.
57. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Research 2004;32(suppl_1):D493–D496.
58. Blankenberg D, Taylor J, Nekrutenko A, Team G. Making whole genome multiple alignments usable for biologists. Bioinformatics 2011;27(17):2426–2428.
59. 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 reports 2015;11(7):1110–1122.
60. Anandam P, Torarinsson E, Ruzzo WL. Multiperm: shuffling multiple sequence alignments while approximately preserving dinucleotide frequencies. Bioinformatics 2009;25(5):668–669.
61. 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). Nature methods 2016;13(6):508.
62. Lovci MT, Ghanem D, Marr H, Arnold J, Gee S, Parra M, et al. Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges. Nature structural & molecular biology 2013;20(12):1434.
63. Murakawa Y, Hinz M, Mothes J, Schuetz A, Uhl M, Wyler E, et al. RC3H1 post-transcriptionally regulates A20 mRNA and modulates the activity of the IKK/NF-κB pathway. Nature communications 2015;6:7367. http://bimsbstatic.mdc-berlin.de/landthaler/RC3H1/.
64. Gruber AR, Findeiß S, Washietl S, Hofacker IL, Stadler PF. RNAz 2.0: improved noncoding RNA detection. In: Biocomputing 2010 World Scientific; 2010.p. 69–79.
65. Miladi M, Alexander J, Fabrizio C, Stefan E S, Jakob Hull H, Gorodkin J, et al. RNAscClust: clustering RNA sequences using structure conservation and graph based motifs. Bioinformatics 2017;33(14):2089–2096.
66. Hubert L, Arabie P. Comparing partitions. Journal of Classification 1985;2(1):193–218.
67. Rand WM. Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 1971;66(336):846–850.
68. Pfreundt U, Wambeke FV, Caffin M, Bonnet S, Hess WR. Succession within the prokaryotic communities during the VAHINE mesocosms experiment in the New Caledonia lagoon. Biogeosciences 2016;13(8):2319–2337.
69. Strobel EJ, Angela MY, Lucks JB. High-throughput determination of RNA structures. Nature Reviews Genetics 2018;p. 1.
70. Ledda M, Aviran S. PATTERNA: transcriptome-wide search for functional RNA elements via structural data signatures. Genome biology 2018;19(1):28.
71. Smola MJ, Christy TW, Inoue K, Nicholson CO, Friedersdorf M, Keene JD, et al. SHAPE reveals transcript-wide interactions, complex structural domains, and protein interactions across the Xist lncRNA in living cells. Proceedings of the National Academy of Sciences 2016;113(37):10322–10327.
72. Kaushik K, Sivadas A, Vellarikkal SK, Verma A, Jayarajan R, Pandey S, et al. RNA secondary structure profiling in zebrafish reveals unique regulatory features. BMC genomics 2018;19(1):147.
73. Zhang B, Mao YS, Diermeier SD, Novikova IV, Nawrocki EP, Jones TA, et al. Identification and characterization of a class of MALAT1-like genomic loci. Cell reports 2017;19(8):1723–1738.
74. Eddy SR. Computational analysis of conserved RNA secondary structure in transcriptomes and genomes. Annual Review of Biophysics 2014;43(1):433–456.
75. Ulitsky I. Evolution to the rescue: using comparative genomics to understand long non-coding RNAs. Nature Reviews Genetics 2016;17(10):601.
76. Gutschner T, Hämmerle M, Eißmann M, Hsu J, Kim Y, Hung G, et al. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer research 2013;73(3):1180–1189.
77. Allerson CR, Cazzola M, Rouault TA. Clinical severity and thermodynamic effects of iron-responsive element mutations in hereditary hyperferritinemia-cataract syndrome. Journal of Biological Chemistry 1999;274(37):26439–26447.
78. Solem AC, Halvorsen M, Ramos SB, Laederach A. The potential of the riboSNitch in personalized medicine. Wiley Interdisciplinary Reviews: RNA 2015;6(5):517–532.
79. Reiche K, Stadler PF. RNAstrand: reading direction of structured RNAs in multiple sequence alignments. Algorithms for Molecular Biology 2007;2(1):6.
80. Leppek K, Schott J, Reitter S, Poetz F, Hammond MC, Stoecklin G. Roquin promotes constitutive mRNA decay via a conserved class of stem-loop recognition motifs. Cell 2013;153(4):869–881.
81. Fu M, Blackshear PJ. RNA-binding proteins in immune regulation: a focus on CCCH zinc finger proteins. Nature Reviews Immunology 2017;17(2):130.
82. Maeda K, Akira S. Regulation of mRNA stability by CCCH-type zinc-finger proteins in immune cells. International immunology 2017;29(4):149–155.
83. Tan D, Zhou M, Kiledjian M, Tong L. The ROQ domain of Roquin recognizes mRNA constitutive-decay element and double-stranded RNA. Nature structural & molecular biology 2014;21(8):679.
84. Wheeler EC, Van Nostrand EL, Yeo GW. Advances and challenges in the detection of transcriptome-wide protein–RNA interactions. Wiley Interdisciplinary Reviews: RNA 2018;9(1):e1436.
85. Schlundt A, Niessing D, Heissmeyer V, Sattler M. RNA recognition by Roquin in posttranscriptional gene regulation. Wiley Interdisciplinary Reviews: RNA 2016;7(4):455–469.
86. Chevrier S, Kratina T, Emslie D, Tarlinton DM, Corcoran LM. IL4 and IL21 cooperate to induce the high Bcl6 protein level required for germinal center formation. Immunology and cell biology 2017;95(10):925–932.
87. Nurieva RI, Chung Y, Martinez GJ, Yang XO, Tanaka S, Matskevitch TD, et al. Bcl6 mediates the development of T follicular helper cells. Science 2009;325(5943):1001–1005.
88. Miladi M, Grüning B, Sokhoyan E, BackofenLab/docker-galaxy-graphclust: December 2017; 2018. https://doi.org/10.5281/zenodo.1135094.
89. Weinberg Z, Breaker RR. R2R-software to speed the depiction of aesthetic consensus RNA secondary structures. BMC bioinformatics 2011;12(1):3.
90. Kerpedjiev P, Hammer S, Hofacker IL. Forna (force-directed RNA): Simple and effective online RNA secondary structure diagrams. Bioinformatics 2015;31(20):3377–3379.
91. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Research 2015;43(7):e47–e47.
92. Lai D, Proctor JR, Zhu JYA, Meyer IM. R-CHIE: a web server and R package for visualizing RNA secondary structures. Nucleic Acids Research 2012;40(12):e95–e95.

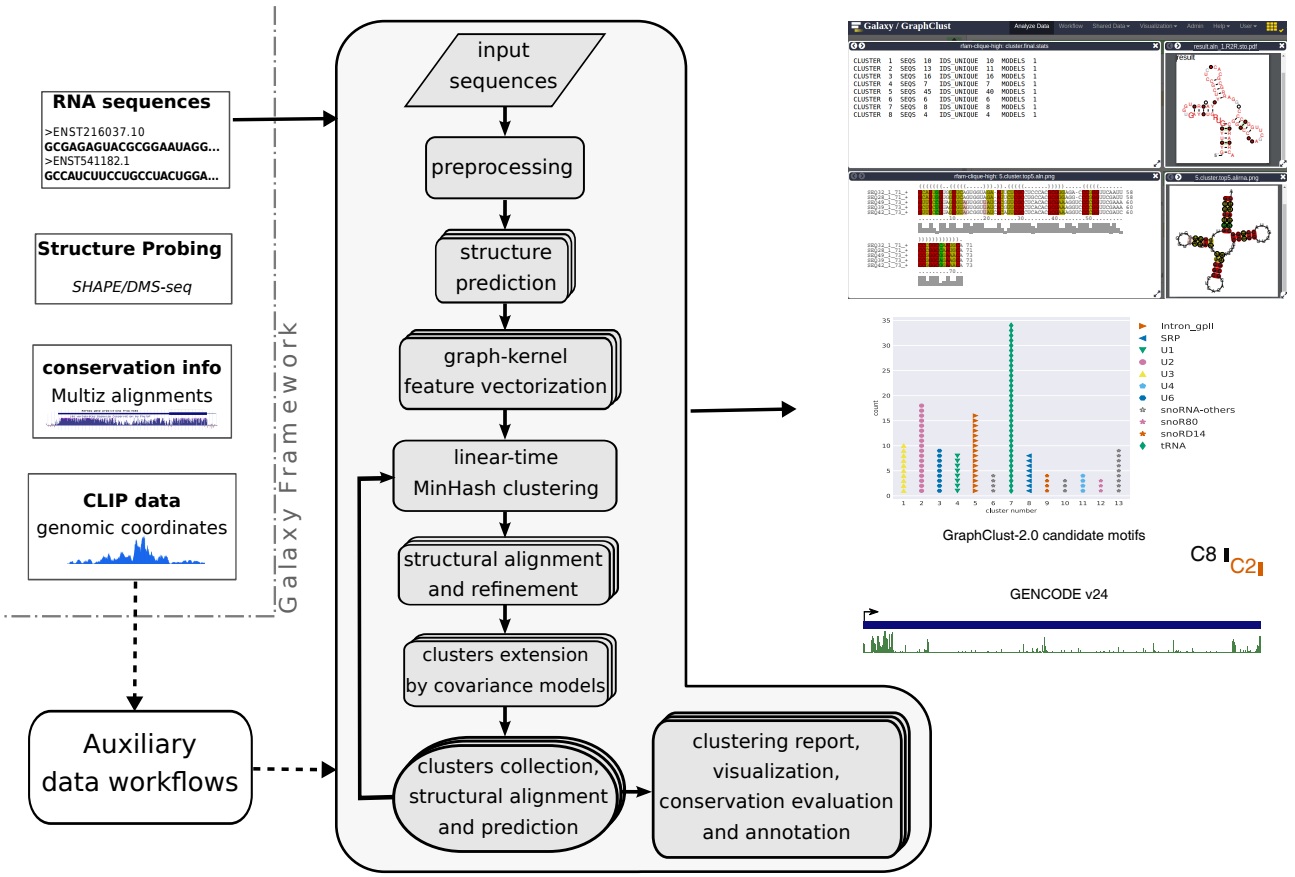

Figure 1. Overview of the GraphClust2 methodology. The flowchart represents the major clustering steps and is supplemented by graphical representations of the associated output data entries. The dashed arrows indicate optional data paths. Auxiliary workflows facilitate integrative clustering of experimental and genomic data including structure probing raw reads or processed reactivities, genomic alignments and conservation information, and genomic intervals e.g. from the CLIP experiments. On the right, a sample selection of the clustering outputs including the overview of the clusters, cluster alignment with LocARNA, RNAalifold consensus structure, and R2R [89] visualization and annotation of the cluster structure by R-scape. Clusters can also be visualised and annotated for the orthology structure conservation predictions.

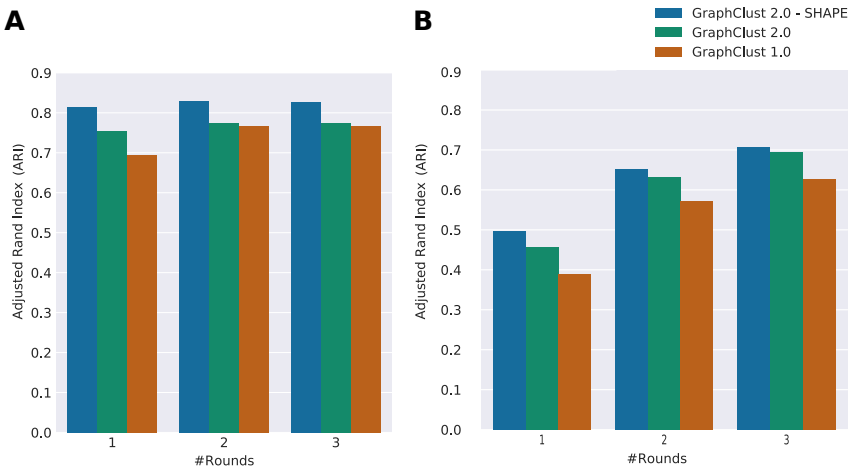

Figure 2. Clustering quality performance over Rfam-based ProbeAlign benchmark dataset and the associated simulated SHAPE data. For comparison, GraphClust2 and GraphClust1 performances are also shown. Incorporating the simulated SHAPE data assists in the clustering performance. (**A**) ARI clustering quality metric for 1-3 rounds of iterative clustering. ARI of the clusterings did not have noticeable improvements after three rounds. (**B**) Similar to (A) but for uniformly sized families, such that precisely ten sequences are randomly extracted per family.

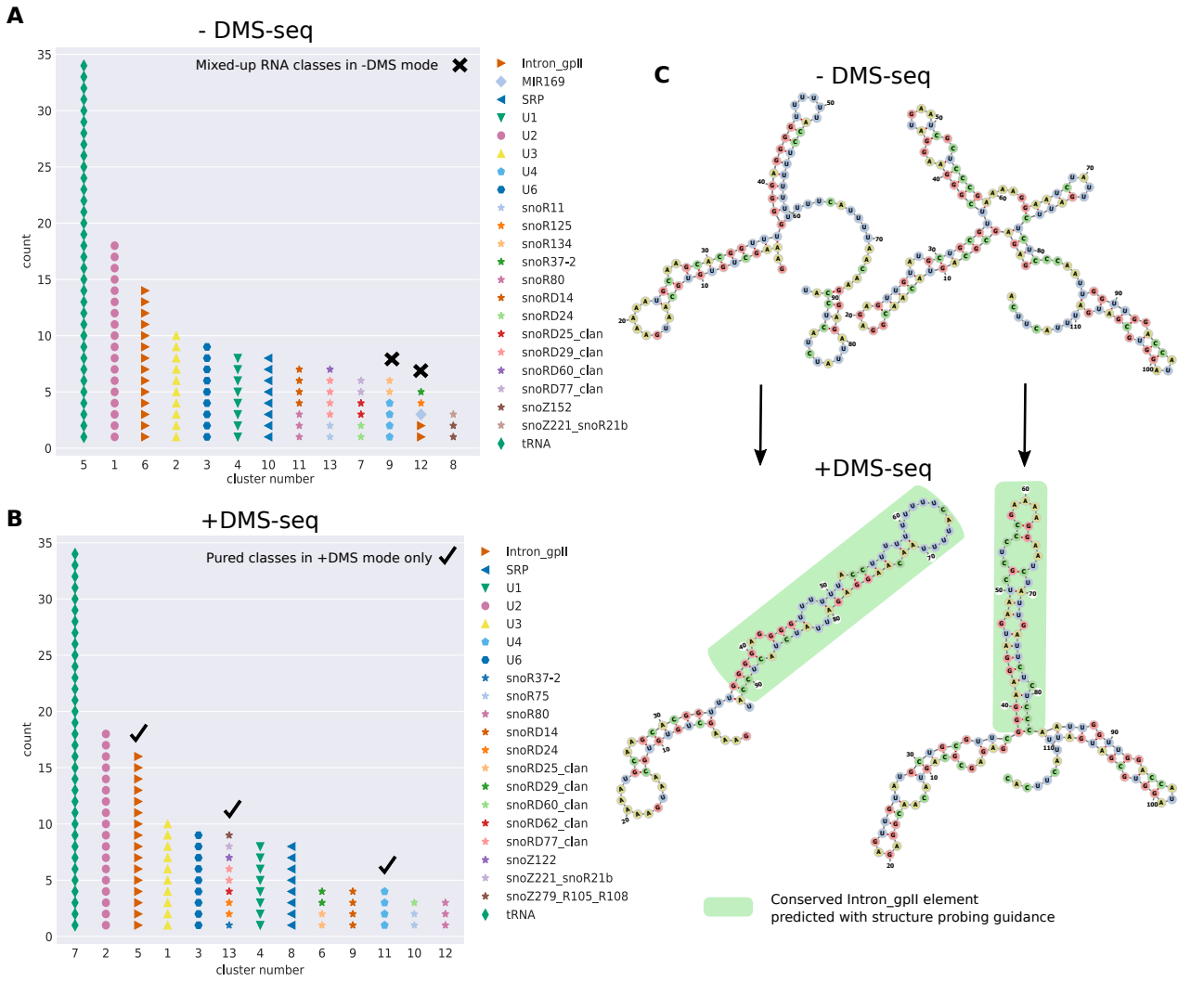

Figure 3. Clustering ncRNAs from *Arabidopsis thaliana* with and without incorporating in vivo DMS-seq structure probing data [49]. (**A**) The predicted clusters without probing data (-DMS-seq) are depicted and the reference family labels are superimposed. Clusters 9 and 12 contain mixed RNA classes. Here, the Group II introns RNA family is split between clusters 6 and 12. (**B**) Similar to A, for +DMS-seq where experimental structure probing data has been used to guide the structure prediction of the generated graphs with pseudo-energy terms. In +DMS-seq mode, only clusters with members from single RNA classes are produced. (**C**) We inspected the predicted structure in more detail. The two transcripts of Intron_gpII family are shown that exhibit substantial structure deviations between their MFE structures (-DMS-seq) and the structures guided by the probing data (+DMS-seq). The structures are predicted using Vienna RNAfold and drawn with forna [90]. The highlighted branches correspond to the conserved references structure from Rfam that are correctly predicted only when the DMS-seq reactivities are incorporated.

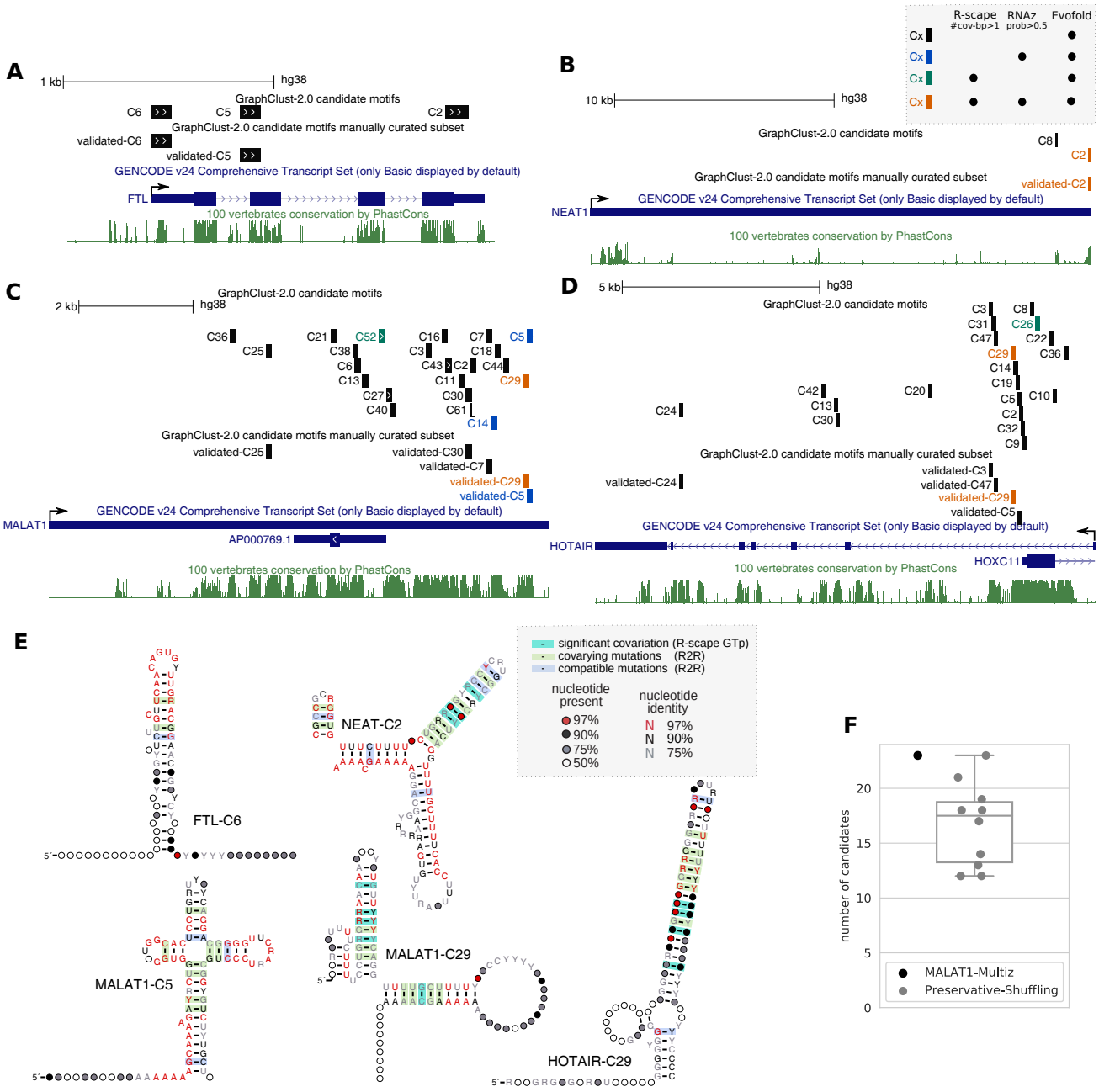

Figure 4. Locally conserved structured elements predicted in FTL mRNA and lncRNAs NEAT1, MALAT1 and HOTAIR. (**A–D**) Locations of the predicted clusters relative to the transcript on the human genome. The clusters under the manually curated subset track, labeled as validated, have passed a qualitative manual screening to exclude unreliable structural alignments (see Results and Discussion). (**E**) Consensus secondary structure for some of the clusters with reliable sequence-structure alignments. Secondary structures are visualized with R2R [89], statistically significant covariation are computed by R-scape and manually overlaid on the R2R visualizations. The alignments are visualized in the Supplementary Figures S6–S11. (**F**) Comparison between the number of predicted candidate motifs of MALAT1 versus ten times Multiperm's preservative shufflings of the same genomic alignment.

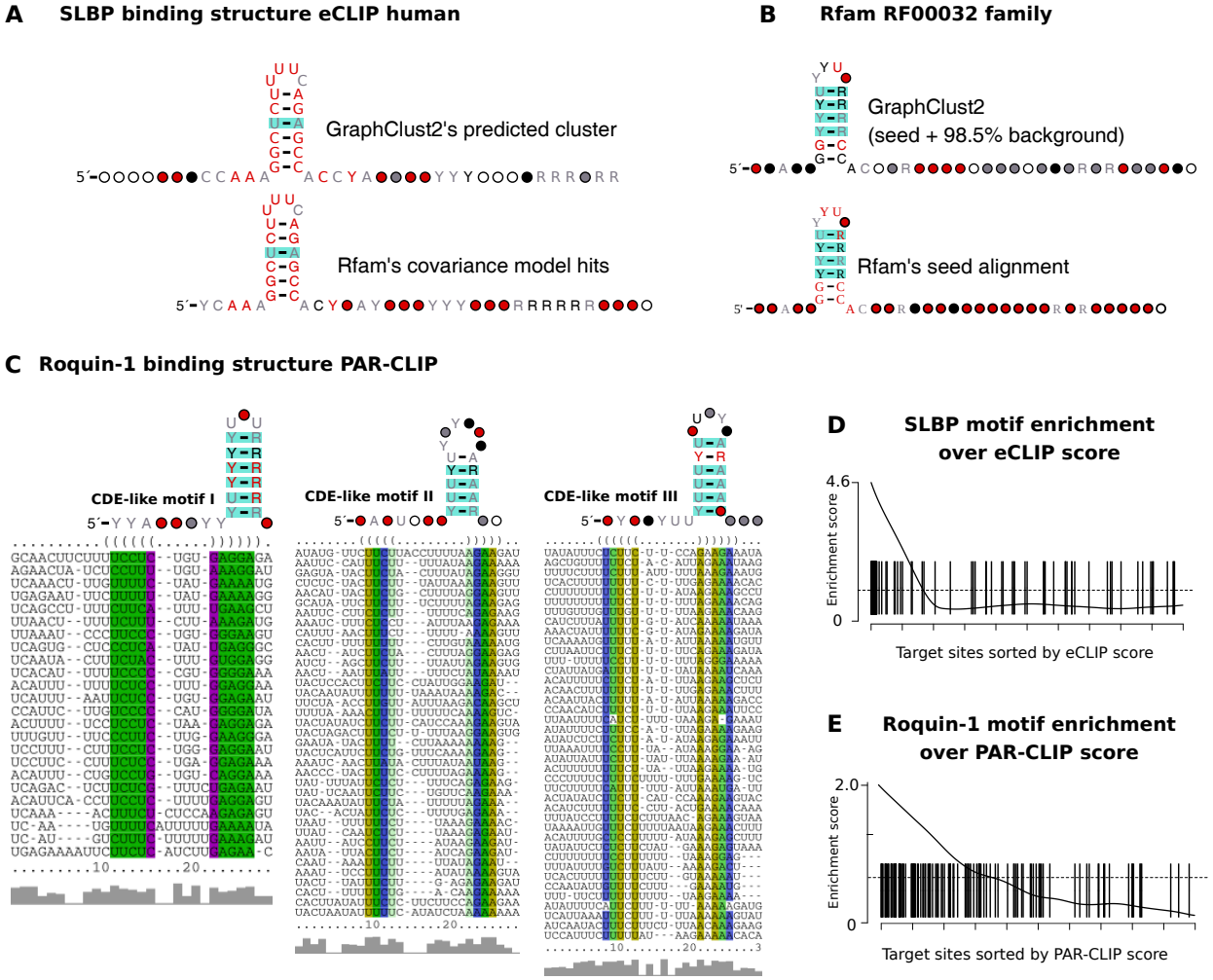

Figure 5. Structured RNA motifs identified by clustering SLBP and Roquin–1 public CLIP data with GraphClust2. (**A**) The consensus secondary structure of the predicted human SLBP motif from eCLIP data versus the consensus structure of cmsearch hits from Rfam's CM for *histone 3'UTR stem-loop* family RF00032. (**B – top**) The consensus secondary structure of the predicted structure motif from clustering 3000 sequences composed of RF00032's 46 seed sequences and 2056 background shuffled full sequences as 98.5% background noise. (**B – bottom**) The Rfam's reference structure for RF00032 seed alignment. (**C**) The consensus secondary structures and alignments of the three clusters with defined consensus structures. The three motifs overlap and have varying loop sizes and uridine content. The structures are akin to the previously validated constitutive decay element (CDE) in TNF–alpha that is a target of Roquin–1. (**C, D**) Gene set enrichment plot of SLBP and Roquin–1 motifs according to the corresponding CLIP scores. SLBP eCLIP has a high enrichment of the stem–loop with strong density in the first hundred target sites. Roquin–1 PAR–CLIP data has a lower enrichment score and low presence in the top 100 target sites. The difference in the enrichment is likely due to the specificity of Roquin–1 that has multiple RNA binding domains and false positive biases of the PAR–CLIP protocol. Scalable clustering assists in overcoming these biases to identify the CDE–like elements. Structures are visualized by R–scape, the color for significant base–pair covariations are adapted to match the legend in Figure 4E. Enrichments are plotted with the Limma R package [91].

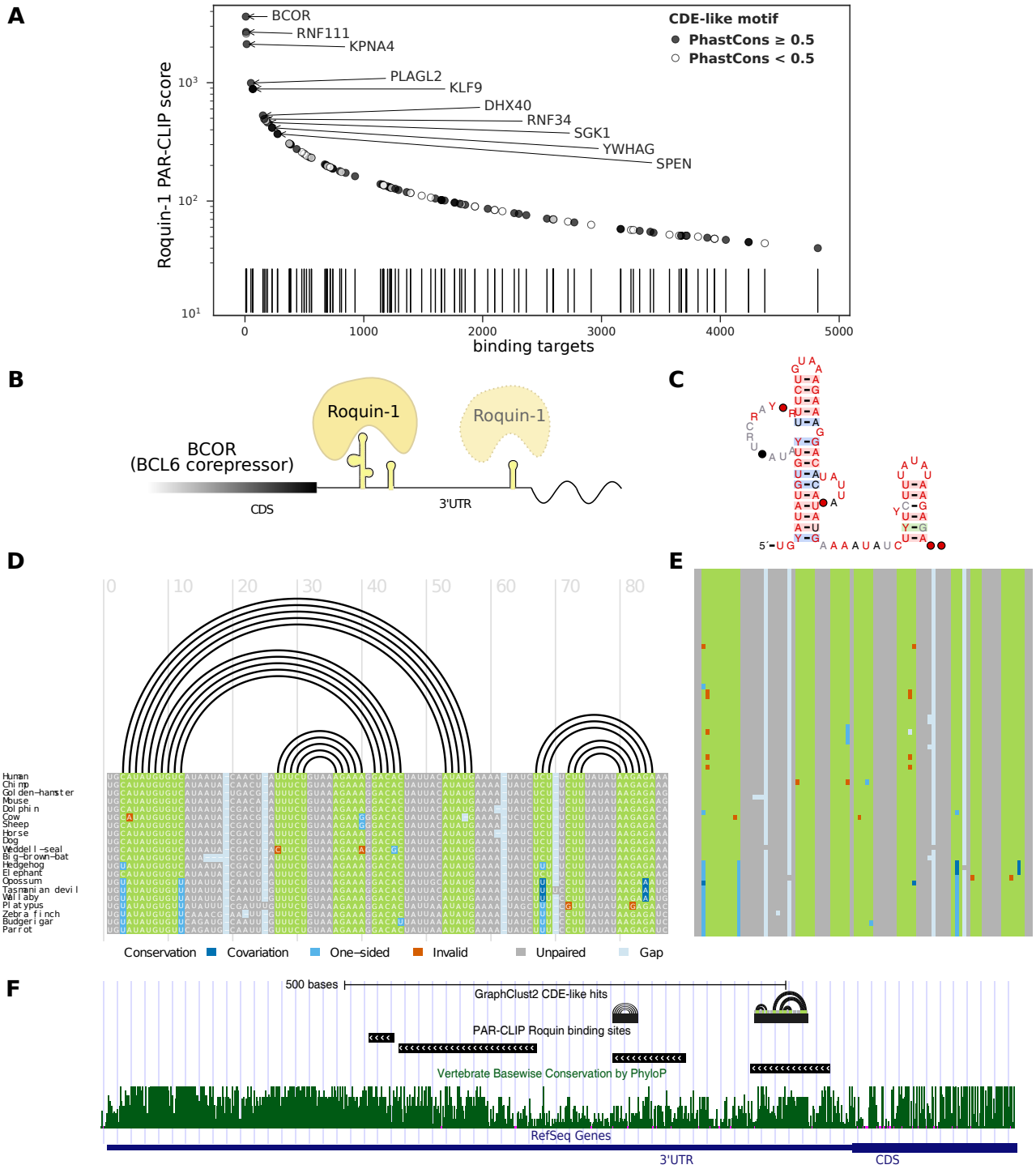

Figure 6. (**A**) The distribution of the Roquin-1 target sites bearing the CDE-like motifs on the 3'UTRs of genes, according to the binding affinity scores. Top 10 highest binding sites with a conserved CDE-like motif are labeled with the associated gene names. (**B**) Roquin-1 binds to a highly conserved double stem-loop element on the 3'UTR of BCOR (BCL6 CoRepressor) with very high affinity. Another CDE-like element with lower affinity downstream of the first element is also spotted. (**C**) R2R visualization for the RNAalifold consensus secondary structure of the conserved double stem-loop element from the vertebrate Multiz alignment. (**D**) Genomic alignment of 20 selected species that is annotated with the consensus structure and the base-pair covariations information. Alignments and compensatory mutations are visualized with R-chie package [92]. (**E**) Genomic alignment overview for the available species extracted from the 100way Multiz alignment. (**F**) Conservation track of BCOR 3'UTR end plus the location of the CDE-like motifs on the negative strand of the locus on the human X chromosome.

Supplementary Table T1

Click here to access/download

Supplementary Material

Supplementary_T1_longRNAs_candidates_metrics.xlsx