RNA-bioinformatics: Tools, Services and Databases for the Analysis of RNA-based Regulation

Rolf Backofen[1,2,*], Jan Engelhardt[3], Anika Erxleben[1], Jörg Fallmann[3], Björn Grüning[1], Uwe Ohler[4], Nikolaus Rajewsky[4] and Peter F. Stadler[3,5,6,7]

[1]Bioinformatics, University of Freiburg, Georges-Köhler-Allee 106, D-79110 Freiburg, Germany
[2]BIOSS Centre for Biological Signaling Studies, University of Freiburg, Schänzlestr. 18, 79104 Freiburg, Germany
[3]Bioinformatics Group, Department of Computer Science; and Interdisciplinary Center for Bioinformatics, University of Leipzig, Härtelstraße 16-18, D-04107 Leipzig, Germany
[4]Max-Delbrück-Centrum (MDC), Robert-Rössle-Str. 10, D-13092 Berlin, Germany
[5]Institute for Theoretical Chemistry, University of Vienna, Währingerstraße 17, A-1090 Wien, Austria
[6]RNomics Group, Fraunhofer Institute for Cell Therapy and Immunology, Perlickstraße 1, D-04103, Leipzig, Germany
[7]Santa Fe Institute, , 1399 Hyde Park Rd., Santa Fe, NM 87501, USA
[*]Corresponding author

May 19, 2017

Abstract

The importance of RNA-based regulation is becoming more and more evident. Genome-wide sequencing efforts have shown that the majority of the DNA in eukaryotic genomes is transcribed. Advanced high-throughput techniques like CLIP for the genome-wide detection of RNA-protein interactions have shown that post-transcriptional regulation by RNA-binding proteins matches the complexity of transcriptional regulation. The need for a specialized and integrated analysis of RNA-based data has led to the foundation of the RNA Bioinformatics Center (RBC) within the German Network of Bioinformatics Infrastructure (de.NBI). This paper describes the tools, services and databases provided by the RBC, and shows example applications. Furthermore, we have setup an RNA workbench within the Galaxy framework. For an easy dissemination, we offer a virtualized version of Galaxy (via Galaxy Docker) enabling other groups to use our RNA workbench in a very simple way.

1 Motivation

Genome-wide sequencing efforts have revealed that a majority of DNA in eukaryotic genomes is pervasively transcribed. Non-coding RNAs and RNA-protein interactions are important parts of cellular regulation that were ignored at first but have received an increasing level of attention over the past decade. While the exact numbers, and even the magnitude, of functional transcripts, regulators and

interactions are a matter of ongoing discussion, they reflect the current challenge for the analysis of whole transcriptome data.

The identification of new classes of regulatory RNAs such as microRNAs (miRNAs), or the genome-wide identification of RNA-protein interactions, which has been enabled by the development of new technologies such as cross-linking and immunoprecipitation (CLIP) methods, suggests that the complexity of post-transcriptional gene regulation is comparable to transcriptional gene regulation. The human genome encodes hundreds to thousands of miRNAs more than 1,000 RNA binding proteins [1–6]. Along with such profiling efforts, a picture has emerged that many human diseases are caused or linked to post-transcriptional gene regulation. Examples include not only rare genetic disorders but cover the entire spectrum of cardio-vascular diseases, cancer, and neurodegenerative disorders (for recent reviews see [7–12]. With increasing evidence that non-coding RNAs are also involved in epigenetic regulatory control, it is clear that RNA biology is of vital, newly emerging importance for research not only in basic molecular biology but also for medical and disease research. Consequently, many of the existing or newly founded centres for common diseases have great need to develop or get access to computational tools and databases that capture and predict regulation by RNA or RNA-protein interactions.

2 Overview over the RNA Bioinformatics Center

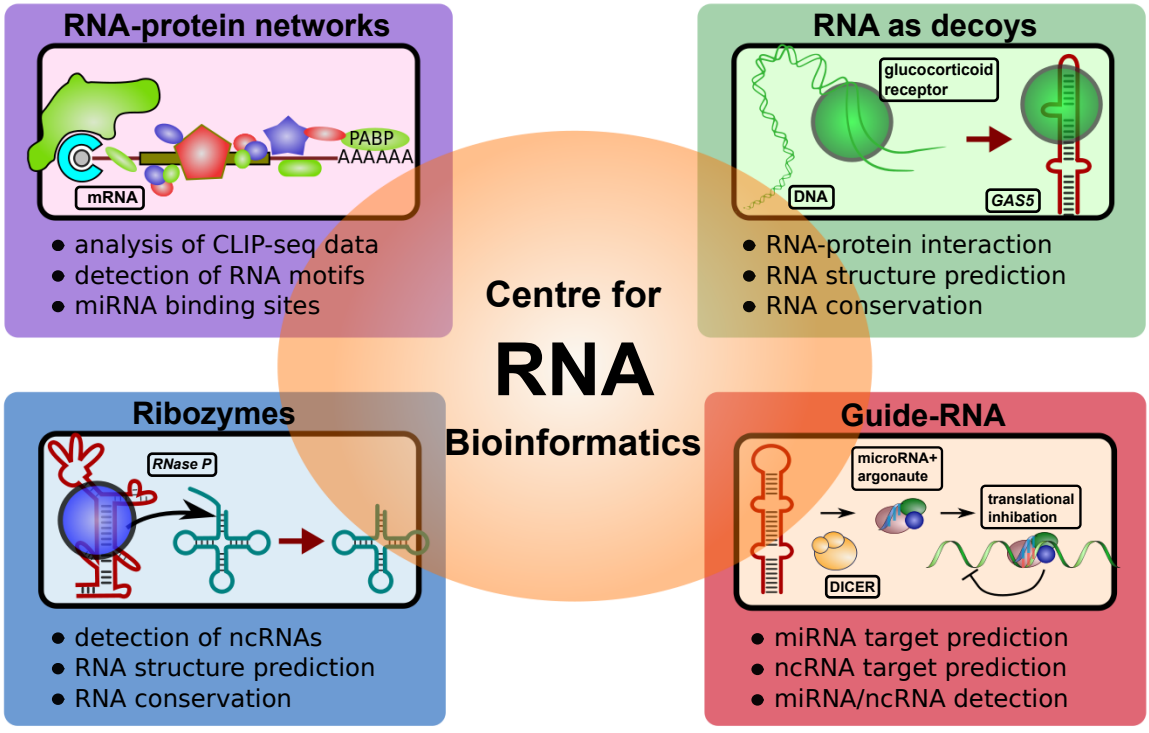

Figure 1: Example functions of RNA and associated bioinformatics services. A comprehensive analysis and annotation of RNA function requires the integration of many different services that are provided by our centre.

Non-coding regulatory RNAs have many possible functions, which require specialized approaches for their detection and analysis (see Figure 1 for an overview of functions and associated bioinformatics services). To name only a few, they can regulate imprinting by modulating chromatin structures, act as guide RNAs for protein complexes, form scaffolds for protein-RNA complexes, regulate other RNAs by

RNA-RNA-interaction [13, 14], function as decoys for proteins and other non-coding RNAs [15] or act as cis-regulatory elements such as riboswitches [16]. MiRNAs [17] are an abundant class of small RNAs, each of which can regulate up to hundreds of transcripts. In total, it is estimated that 60% of all human proteins are regulated by miRNAs. With the advances in high-throughput approaches to detect binding sites of RNA-binding proteins (RBP) such as CLIP-Seq, a plethora of new RNA regulatory mechanisms has been detected when analysing the RBPome (i.e., the network of protein-RNA interactions) [18]. Finally, ribozymes are an important class of ncRNAs that are often involved in the maturation of other RNA or DNA molecules.

With all these potential roles, it has become clear that the analysis of epigenetic and expression data is incomplete if RNA-based regulation is not taken into account. As consequence, the analysis of RNA has to be integrative, combining sequencing datasets with sequence and structure analysis of RNA elements, and allowing for integration with other regulatory mechanisms such as transcription. High-throughput techniques to analyse RNA-based regulation are rapidly evolving, which give rise to a large amount of information but also to the need to constantly adapt databases, annotations and tools.

To overcome these problems and limitiations, the RNA Bioinformatics Centre (RBC) was founded within the German Network for Bioinformatics Infrastructure (de.NBI) with the following priorities:

1. To establish an integrated, easily accessible RNA analysis workbench which can be used on our own cluster or downloaded and installed on every HPC environment.

2. To work with other Bioinformatics Centers and relevant scientific communities to allow for maximal usefulness, interconnectivity, and added value of the developed infrastructure.

3. to use this infrastructure as foundation for a learning and teaching environment that fosters an awareness for the importance of RNA analysis.

In consequence, our goal in RBC is to serve as contact point for all RNA bioinformatic questions in Germany, ranging from initial study design, over providing protocols and infrastructure, up to developing specialised solutions for individual problems. In addition, the RBC provides specialized curated RNA-related information resources such as databases for protein-RNA interaction or tRNAs, which will be fully integrated into our workbench. Across the three locations Berlin, Freiburg and Leipzig, the joint expertise covers many if not all aspects of RNA biology of current interest, ranging from structure prediction and genome-wide annotations of conserved secondary structures via the detection of members of specific classes of regulatory RNAs, and the interaction of RNA binding proteins and regulatory RNAs with their targets.

3 Individual Tools provided and maintained by the RBC

In this section, we will give an overview of different services and tools that are required to analyse RNA-related data. We will take our emphasis on tools that are provided by the RBC and only shortly mention other related tools. Tools and databases maintained by the RBC will be written in italics. The complete list of tools can be found under `https://github.com/bgruening/galaxy-rna-workbench`.

3.1 Prediction of RNA structure and detection of conserved RNA structure

Many functional RNAs require a specific structure to be formed. Very often, the so-called secondary structure (i.e., the set of Watson-Crick and G—U bonds) is well-conserved and characteristic for the

function of the RNA. Prediction of the secondary structure is a well-established area in RNA-bioinformatics. The *ViennaRNA Package* consists of a C code library and several stand-alone programs for the prediction and comparison of RNA secondary structures. It is also the de-facto standard library for the development of RNA based methods [19].

However, the prediction of the secondary structure is usually only a first step in a whole pipeline for the analysis of RNA-related data. Often it is required to determine the conserved secondary structure, or whether a structure is conserved at all. *MARNA* [20] is an early approach that solved the problem of generating multiple alignments using well-defined pairwise RNA-alignment approaches by using TCOFFEE [21] to combine the pairwise alignments. It computes multiple sequence-structure alignments considering a single fixed structure for each sequence only. *ExpaRNA* is a fast, motif-based comparison and alignment tool for RNA molecules. Instead of computing a full sequence-structure alignment, it computes the best arrangement of sequence-structure motifs common to two RNAs [22]. The gold standard here is the Sankoff approach [23] (and its variants) of performing a sequence-structure alignment of RNAs. One approach that is provided by the RBC is *LocARNA* [24, 25]. It is an efficient variant of the Sankoff approach that computes multiple alignments of RNAs based on their sequence and structure similarity and considers the whole ensemble of secondary structures for each RNA. Thus, LocARNA aligns RNAs with unknown structure and predicts a consensus secondary structure for a set of unaligned RNAs. LocARNA is best suited to compare several (up to about 20) structural RNAs, in particular, of low sequence similarity. *Carna* [26] is a tool for multiple alignment of RNA molecules based on their full ensembles of structures. Carna computes the alignment that fits best to all likely structures simultaneously. Hence, Carna is in particular useful to align RNAs with more than one stable structure, as for example riboswitches, and is able to align arbitrary pseudoknots. The above tools detect whether there is a conserved structure, however, they do not decide whether the structure is *significantly* conserved to indicate a structural RNA. This is solved by *RNAz* [27], which is a program for predicting structurally conserved and thermodynamically stable RNA secondary structures in multiple sequence alignments. It can be used in genome wide screens to detect functional RNA structures, as found in noncoding RNAs and cis-acting regulatory elements of mRNAs.

Genomic screens produce a large set of putative RNAs, however, annotation of these approaches is a critical task. One successful approach is to cluster these RNAs in order to detect RNA classes. These are RNAs that are structurally similar but do not stem from a common anchestor; a prominent example is the class of miRNAs. *GraphClust* [28] is a method based on graph kernels for an alignment-free clustering of ncRNAs. It can be used to detect new ncRNA classes as well as for detecting members of known classes. It is currently the only approach capable of clustering hundred of thousands RNA according to sequence **and** structure. Another approach that uses clustering as means for annotation of ncRNA in a specialized RNA, namely the annotation of CRISPR repeats is *CRISPRmap* [29], which provides a quick and detailed insight into repeat conservation and diversity of both bacterial and archaeal systems. It comprises the largest dataset of CRISPRs to date and enables comprehensive independent clustering analyses to determine conserved sequence families, potential structure motifs for endoribonucleases, and evolutionary relationships.

Finally, one does not only want to detect new natural ncRNAs. For many applications one wants to design, either computationally or even biotechnologically, new synthetic RNAs that are putative members of an RNA class or family. *RNAdesign*, which is part of the Vienna RNA package, is one of the earliest yet still most widely used programs for the design of RNA sequences that fold into a given pseudo-knot free RNA secondary structure. Another successful example is *INFO-RNA* [30], which is a web-server providing RNA designs. *ANTARNA* [31] is an improved design approach based on ant

colony optimization that can control the GC-content.

3.1.1 Identification of specific regulatory non-coding RNA classes

The approaches listed above are general tools that, in theory, can be used for all RNA classes. However, optimized tools exists for specific classes such as miRNAs and snoRNAs that can make use of additional biological knowledge as well as of additional RNA-seq data. The first example is *miRDeep* [32], which is a probabilistic model that detects the presence of expressed animal microRNAs in deep sequencing data. It does this via a set of features that reflect its processing from primary transcript to mature short 22nt sequence, such as the relative frequency of reads aligning to the mature RNA compared to other parts of the precursor. *PiPMir* [33] follows a similar idea to detect new plant miRNAs. Plant precursors can be much longer compared to animals and contain multiple mature miRNAs; *PiPMir* addresses these differences in the pathways of miRNA maturation in plants, for instance via extensive predictions of local secondary structure for precursors up to several hundred nucleotides. *DARIO* [34] is a webservice providing functionalities complementary to *miRDeep*, allowing not only the recognition of novel microRNAs but also small RNAs derived from other types of parental RNAs such as snoRNAs and tRNAs. The tool is being made available also for plant genomes.

As stated above, one of the important ideas behind the above described tools is to combine computational RNA analysis with sequencing data. *NASTI-seq* [35] extends the formalism behind popular differential expression RNAseq tools to strand-specific protocols. It uses an explicit likelihood ratio test to identify the significant presence of overlapping antisense transcription, and consequently, candidate loci for the generation of cis-natural antisense siRNAs.

3.1.2 Identification of targets of regulatory RNAs based on sequence

The assignment of a ncRNA to a specific class is the first annotation task. Many small ncRNAs such as miRNA serve as guide RNA or are directly acting on their target via RNA-RNA interactions. For that reason, target prediction relies on features extracted from RNA-RNA interactions between small ncRNA and its target, possibly combined with additional features. *PicTar* [36, 37] is one of the most established and successful miRNA target predictors based on sequence features of functional miRNA-target interactions.

PicTar is specifically designed for miRNAs. However, other small ncRNAs also act via RNA-RNA interaction. In this case, general approaches for predicting RNA-RNA interactions can be used. *RNAcofold* [38] is part of the *Vienna RNA package* and can predict joint structure of two RNAs, provided that the structure is nested. However, this excludes common interactions such as kissing hairpin loops. These types of interaction can be determined by accessibility-based approaches, which combine the calculation of a duplex energy with a penalty that measures the energy required to make the interaction sides accessible in the two interacting RNAs. *RNAup* [39] was one of the first accessibility-based interaction prediction tools and allows reliable predictions of RNA-RNA binding energies using an approach that is based on the ensemble of RNA-structures of a sequence. It combines the energy for making an interaction site accessible with the energy of duplex-formation. However, it has a complexity of $O(n^2w^2)$ where n is the sequence length, and w is the maximal width for the interaction sites. *IntaRNA* [13, 40] is a fast accessible interaction approach that reduces the complexity by applying a heuristic approach while maintaining a high prediction quality due to the use of a seed-interaction. It has been designed to predict mRNA target sites for given non-coding RNAs (ncRNAs) like eukaryotic microRNAs (miRNAs) or bacterial small RNAs (sRNAs), but it can also be used to predict other types of RNA-RNA interactions.

It combines the accessibility of interaction-sites with duplex energy and is efficient enough to be used on a genome-wide scale. *RNApredator* [41] combines pre-computed accessibilites for the target genomes with a simplified energy model for the RNA-RNA interaction to speed up genome-wide predictions.

Albeit both approaches are quite successful for predicting targets of small ncRNAs, they still have a quite high false positive rate when applied genome-wide. For that reason, *CopraRNA* [42] computes whole genome predictions by combination of whole genome *IntaRNA* predictions using homologous sRNA sequences from distinct organisms, thus greatly reducing the false positive rate.

3.1.3 Prediction of in vivo RNA-binding protein (RBP) interactions

Hundreds of RBPs have been shown to play a role in virtually all aspects of (post-transcriptional) gene expression regulation, ranging from transcript processing, export and localization, stability to translation (see *e. g.* [43–45]). A manually curated collection of over 1.500 RBPs in human as presented in [3] highlights their vast number and interaction and regulation potential. For many RBPs, their direct interaction with target RNA requires more or less specific sequence motifs [46] and accessible binding sites. So far, most investigated RBPs have been shown to prefer single stranded binding regions, although some interact with structured RNA regions [47] or prefer a structural context such as the location within a loop.

RNA-protein interactions play a key role in the complex interactome of higher organisms rendering their interplay and underlying mechanisms an investigative challenge. Distinguishing true binding sites from sites sharing sequence and/or structure features by chance is a non-trivial task, that becomes even harder as interaction is not necessarily functional. Proteins can, besides specific binding, interact with their targets in a probing manner known as diffusional search [48], which further complicates interaction analysis.

Experimental investigation of RNA-protein interactions requires some knowledge of at least one of the interacting partners, be it to generate specific probes, antibodies, cell-types or substrates. Using RNA-centric methods, an RNA of interest is purified and interacting proteins or protein complexes can be identified via methods like mass spectrometry. Although this allows identification of novel RBPs, or RBPs for which antibodies are hard to come by, RNA-centric methods require the purification of enough protein mass, which means a high amount of starting material [49]. Purified protein can, in contrast to nucleic acids, not be amplified, which makes RNA-centric methods challenging for low abundancy RNAs and proteins.

`In vivo` protein-centric methods are based on specific purification methods for the protein of interest. Antibodies which allow immunoprecipitation (IP) of the latter are most common, however, the quality and specificity of the antibody impacts the quality of the results. To identify interaction partners, co-immunoprecipitated RNA is then reverse transcribed into cDNA, PCR amplified and sequenced. PCR amplification allows to start from low amounts of starting material in contrast to RNA-centric methods. In general, native and denaturating purification methods are available. RNA immunoprecipitation (RIP), preserves physiological conditions and native RNA-protein and protein-protein complexes during native purification . However, the protein of interest can during purification interact with RNAs not natively present in the same cell compartment or interact unspecific with highly abundant RNAs, *e. g.* rRNAs, which can interfere and mask specific interactions with low-abundancy targets. This can be prevented applying denaturing methods, *i. e.* crosslinking the protein of interest to its target RNA. Such a snapshot of interactions at the time of crosslinking prevents non-native interactions in later steps of purification. Short wavelength UV light crosslinking creates covalent bonds between aromatic amino acids of the

protein and RNA nucleotides in close proximity without crosslinking proteins with other proteins. CLIP (crosslink and immunoprecipitation) [50] is an `in vivo` method utilizing UV crosslinking, followed by antibody-purification.

Several types of CLIP procedures have been proposed, *e. g.* HITS-CLIP (HIgh-Throughput Sequencing of RNA isolated by CrossLinking ImmunoPrecipitation) [51], iCLIP (Individual-nucleotide resolution CLIP) [52] and PAR-CLIP (PhotoActivatable-Ribonucleoside-enhanced CrossLinking and ImmunoPrecipitation) [53]. Together with recent methods like eCLIP (enhanced CLIP) [54], irCLIP (infrared CLIP) [55] or hiCLIP (RNA hybrid and individual-nucleotide resolution ultraviolet crosslinking and immunoprecipitation) [56] a bandwidth of experimental designs rely on the same principle, crosslinking protein residues and adjacent nucleotides with UV light, with varying details that affect the specific outcome. PAR-CLIP, for example, makes use of nucleotide analogs like thio-uridine or thio-guanine, which are introduced into the cell as crosslinking agents. These nucleotide analogs can be crosslinked with long-wave UV light (365nm), which helps to circumvent the otherwise low efficiency of UV-crosslinking at 254nm, but works only with cultured cells which readily utilize the nucleotide analogs. The biochemical details behind UV-crosslinking are not yet fully investigated, so that it remains hard to predict how many interactions might be missed completely. However, it is known that reverse transcriptase (RT) misreads crosslinked nucleotides or drops off completely, which is exploited by PAR-CLIP. The introduced nucleotide analogs are in case of thio-uridine misread by RT as guanines, consequentially introducing T-to-C transitions in the resulting sequencing reads. These transitions can then be used to pinpoint interaction sites. iCLIP, as another example, relies on the fact that an amino acid tag left at the crosslink site after proteinase digestion causes termination of reverse transcription to pinpoint the interaction site with nucleotide resolution.

Depending on the CLIP technique used (iCLIP, HITS-CLIP, PAR-CLIP etc.), downstream analysis requires specific algorithms to filter signal from noise. In general, the goal is to filter spurious and unspecific binding to identify true binding sites. A major challenge of many CLIP data sets is the lack of negative control. Without the latter, a measure to distinguish true binding from background binding has to be defined. In this regard, the RBC provides software for the analysis of RBP binding sites, to be specific, *PARalyzer* [57], and *microMUMMIE* [58]. *PARalyzer* is a principled quantitative approach to detect RBP target sites based on a local excess of the diagnostic T-to-C transitions observed at PAR-CLIP derived interaction sites. It computes local kernel density estimates for background and binding sites to distinguish signal from noise, while simultaneously accounting for different RNA expression levels and sequencing depths. This leads to a tighter definition of locations compared to heuristics . *microMUMMIE* was the first approach to directly utilize PAR-CLIP data to identify `in vivo` targets of expressed miRNAs. It integrates PAR-CLIP data profiling RISC protein binding locations with sequence features in a multivariate hidden Markov model to predict which microRNA targeted which of the observed in vivo target sites. Both tools provide the user with CLIP derived binding sites that can readily be used for downstream analysis.

Binding motif prediction After binding sites are defined, the next step is usually the search for binding preferences of the protein of interest. Determination of preferred binding motifs is a routine task with CLIP data, identification of such a motif is, however, non-trivial.

The problem of discovering motifs without any prior knowledge of how the motifs look is described in standard bioinformatics textbooks (see e.g. [59]). The task is to find subsequences that occur more often than expected, *i. e.* they are over-represented from a given set of sequences. The motif of interest can in principle be found by aligning the input sequences and searching for conserved regions, given

that it should occur in many sequences. However, motifs can consists of sub-motifs themselves and do not have to be fully conserved as they can show some variability in their nucleotide content. Position Weight Matrices (PWM), which assign each position in a sequence a probability for containing a certain nucleotide can be generated from alignments. From there, the frequency of each motif in the input can be calculated and compared to the background frequency (*e. g.* number of corresponding motifs in genes), to derive a measure for over-representation. Many algorithms based on this or equal strategies exist, among which MEME [60] is the most widely used. It applies an expectation maximization (EM) algorithm to find the most over-represented motifs in a set of sequences and was successfully used to predicted binding motifs for a set of RBPs from HTS data.

CERMIT [61] is a fast sequence motif identification algorithm that utilizes suffix arrays to efficiently find optimal motifs in large sequence sets (such as tens of thousands of sequences identified by chromatin or RNA immunopreciptation experiments). It uses rank-order statistics and accounts for quantitative information for each sequence, and has also been applied to identify the most prominent miRNA seed matches in differentially expressed mRNAs. *miReduce* [62] is another computational algorithm that discovers motifs in mRNAs that explain changes in gene expression, for example upon perturbation of miRNA expression .

In general, RBP binding motifs can be predicted by DNA motif finders that only consider the sequence, or by tools that also consider the RNA secondary structure. For DNA-based motif finders, accessibility in terms of structure is not a factor. The double stranded B-form $\alpha-$helical structure of DNA allows (sequence specific) DNA binding proteins to interact with its major groove. RNA on the other hand is shaped in A-form $\alpha-$helical geometry, which results in a very deep and narrow major groove and a shallow and wide minor groove when double-stranded, rendering it less accessible for proteins. In consequence, most RBPs are thought to prefer single stranded RNA (ssRNA) regions for interaction. It is therefore interesting to include accessibility of binding sites to correctly predict binding motifs for RBPs. *MEMERIS* [63], predicts the probability of being unpaired for sequences and incorporates this single-strandedness information into MEME motif prediction, rendering it more accurate for RBP binding motif prediction.

However, accessibility of the preferred motif is not the only interesting factor to consider, as the structural context of motif embedding regions can of course influence the binding behaviour of RBPs.

GraphProt [64] is an advanced graph kernel-based machine learning algorithm, extracting motifs that were highly predictive for binding from a set of bound and unbound sequences. These motifs can be used to predict binding affinities and `de novo` binding sites that are not present in the experimental output. *GraphProt* is able to use both structural profiles as well as detailed 2D-structures, without the need to decide a priori about the weight of the different structural components. A main advantage is that the full secondary structure information is conserved and not just a structure profile per motif, which decreases the error-rate and can be used to identify structural preferences of RBPs with higher resolution.

3.1.4 Databases

With the ever growing number of experiments detecting new RNAs or targeting RNA-RNA and RNA-RBP interactions, the need for dedicated databases collecting and curating these kind of data emerged. Such databases make it possible to store the results of research projects in a standardized way, fulfilling two very important purposes. They guarantee centralized, long-term and easy access to the results of projects. Keeping data accessible beyond the end of a project is a crucial step for reproducibility and

the advancement of a field, which is however not easy to implement for individual groups with rapidly changing personnel. Specialized databases can maintain a high quality due to manual curation. They can act as a "gold standard" to compare new results to or serve as the initial data set for advanced analysis. Without such databases many datasets could not be used to their full extent. Several of the databases are part of the European `RNAcentral` effort to more tightly integrate all sequence and annotation resources [65].

RBC hosted databases make it possible to compare RNA and RBP targets for shared/unique sequence and structure features, nuclear and mitochondrial tRNA genes as well as special genomic motifs. They build the basis for many downstream analysis tasks.

doRiNA is a database for post-transcriptional regulatory elements, such as RNA:protein interactions obtained via CLIP technologies or computational predictions. Integrating data from different RNA binding proteins, non-coding RNAs, publications and labs is key for understanding combinatorial post-transcriptional gene regulation [66, 67]. doRiNA (`http://dorina.mdc-berlin.de`) curates hundreds of thousands of post-transcriptional regulatory events and is visited by hundreds of researchers world wide.

We have recently proposed circular RNAs as a potentially large class of post-transcriptional regulators [15]. Due to the abundance of circular RNAs across all animals and plants that have been studied so far, we are currently developing *circbase* (`http://www.circbase.org`), where we are curating, storing, and making accessible our own and other circRNA data [68].

Transfer RNA (tRNA) are one of the first known classes of non-coding RNAs and are crucial for proper translation of RNAs to proteins. *tRNAdb* [69] continues Sprinzel's tRNA collection [70] and contains more than 12.000 tRNA genes from 577 species and 623 tRNA sequences from 104 species and is developed in close collaboration with Rfam. Several important features of tRNAs can be extracted from the database, e.g. anticodon, amino acid, position of loop regions as well as the predicted secondary structure.

mitotRNAdb [69] contains more than 30.000 metazoan mitochondrial tRNA genes from more than 1,500 species. Mitochondria are eukaryotic organelles whose main function is the production of adenosine triphosphate (ATP). They are separate regions within a cell and therefore have their own genome and translation machinery. Often mtDNA is the first genome that is sequenced in new organisms. It can already be used for phylogenetic analysis. Mitochondrial tRNAs differ significantly from cytoplasmic ones and are often studied independently, therefore a specialized database was generated (see `http://mttrna.bioinf.uni-leipzig.de/mtDataOutput`).

AREsite2 [71] is a database for the detailed investigation of AU, GU and U-rich elements (ARE, GRE, URE) in the transcriptome of Homo sapiens, Mus musculus, Danio rerio, Caenorhabditis elegans and Drosophila melanogaster. It contains information on genomic location, genic context, RNA secondary structure context and conservation of annotated motifs. Furthermore, it includes data from CLIP-Seq experiments in order to highlight motifs with validated protein interaction. A REST interface for experienced users to interact with the database in an semi-automated manner is available and also part of the RBC RNA-workbench as described in section 5. The database is publicly available at `http://rna.tbi.univie.ac.at/AREsite`.

4 Integrated Service Provided by RBC

Experimental labs now generate data of a complexity that makes computational analyses an absolute necessity, but do often not have the means to employ lab members with advanced practical compu-

tational skills (such as combining tools of different provenance, compiling and installing on different platforms, etc). We aim to close this bottleneck by providing 1) stand-alone platform-independent access to the applications; 2) workflows for standard analyses as well as means to custom adapt them; 3) integration of new data with published relevant datasets, including expansion of existing database resources; 4) training at distinct levels, from effective tool use to deep understanding of the of the algorithms.

For that reason, we aimed at the **integration of tools and databases** in one easy accessible and transparent RNA analysis workbench. The services include both genome annotation tools (e.g., target prediction, RNA structure analysis and comparison) as well as pipelines for the analysis of RNA-related HTS-data. Our integrated systems offer a broad range of different ready-to-use pipelines. As RNA-based tools are only one part in a whole analysis pipeline, we offer different workflows for standard RNA-related HTS-analysis such as analysis of RNA-seq data and the associated determination of differential expression or the analysis of epigenetic-related HTS data such as ChIP-seq. To illustrate how the services provided by the center can be used, we here sketch out a couple of examples. A researcher working on RNA-seq data should be able to easily include expressed ncRNA transcripts using our integrated workbench. For that purpose, he needs to be able to define the corresponding ncRNA transcripts from the RNA-seq data, which is a non-trivial task due to the fact that reads typically do not cover the full ncRNA. For the functional annotation of the found transcripts, he also needs to understand the RNA structure. Only an integrated analysis of binding sites of RBPs, microRNA binding, HTS-structure probing and RNA-structure prediction will allow a comprehensive understanding of function associated with the RNA. A further functional analysis would contain also the assignment of the transcript to ncRNA classes, determination of homologs and the prediction of putative targets. To give another example, a scientist working on disease related synonymous SNPs will be enabled to answer the following questions: 1.) Are there binding sites of microRNAs or RNA-binding proteins that are affected by the SNP? 2.) Does this enhance or decrease the affinity of binding? 3.) If there are no direct binding sites, does the SNP change the secondary structure and thus influence some other binding sites? 4.) Are other regulatory RNA-elements affected? Currently, these kind of questions need a lot of manual work and a very specific expertise in RNA-bioinformatics, and thus cannot be solved by a normal lab person with a side interest in bioinformatics analysis of high-throughput sequencing data.

One of our main goals is to strengthen the awareness of ncRNA importance during analysis of biological data like differential gene expression or transcriptomics data and the impact of non-coding sequence variation. Combining different data sources is already becoming a standard approach in other areas. To give an example, whole genome methylation data has gained attraction recently and is often being combined with ChIP-Seq data [72]. However, ncRNA data is not taken into account, leading to a systematic knowledge gap.

Figure 2 puts the core tasks of RNA bioinformatics analysis into a broader context. Some of these interconnections already have been described above, such as the use of services related to RNA-structure and RNA-protein-interactions for genome-wide association studies, or the use of RNA-target prediction and RNA-gene detection for the analysis of RNA-seq data. The investigation of RNA-protein interactions clearly needs proteomics for exact quantification of proteins. On the other side, proteomics also needs RNA-target prediction and RNA-protein interaction to answer questions that are related to translation such as mRNA stability and translational efficiency. ChIP-seq data is e.g. used to investigate transcriptional regulation, and provides information that might be used to improve the prediction of RNA-genes. Conversely, epigenetic modifications that are investigated by using ChIP-seq or by analysing genome-wide methylation often show effects on long non-coding RNAs, which than can be investigated

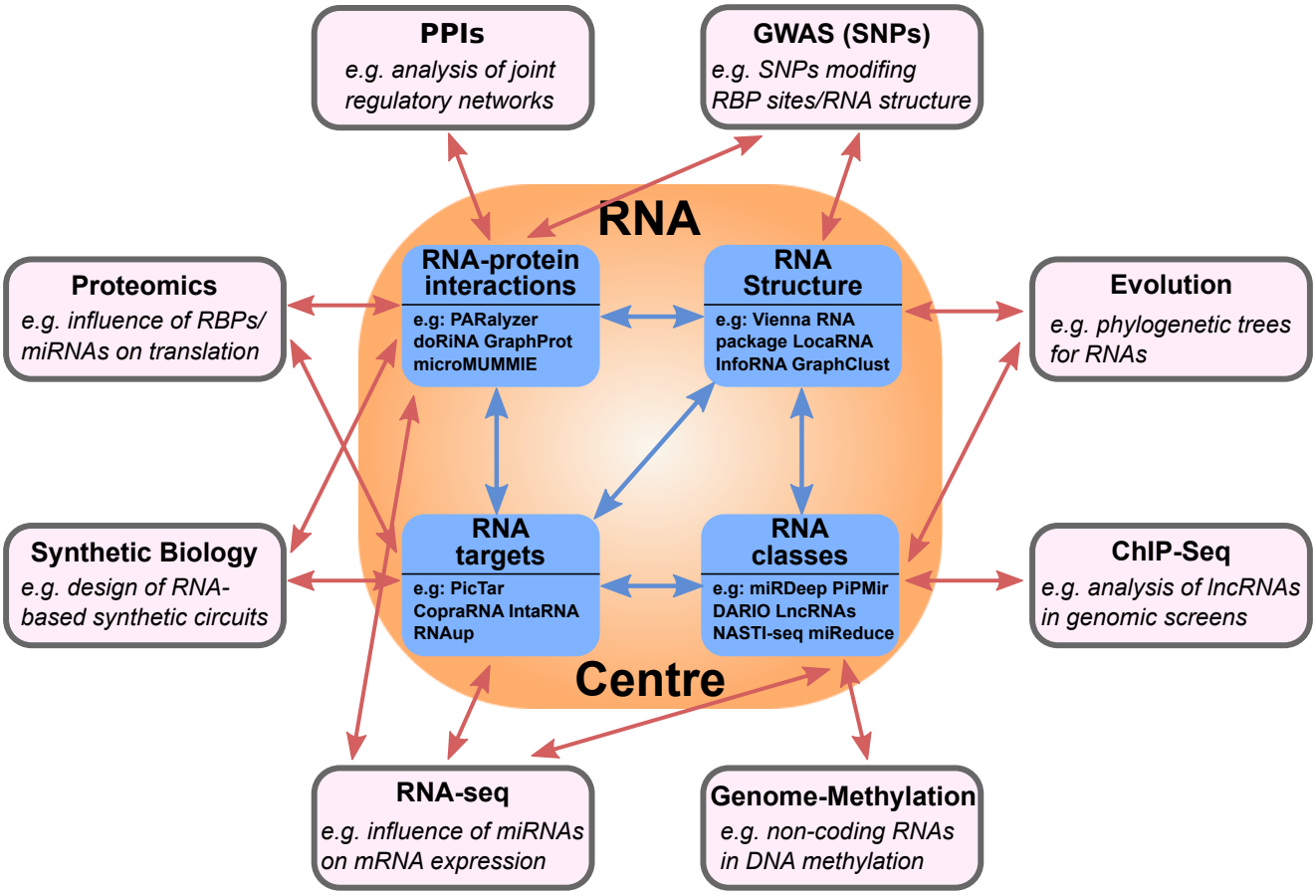

Figure 2: Possible general bioinformatic tasks (outer circle, with example problems) that are related to services provided by our RNA-Bioinformatics Centre (inner circle). Arrows indicate possible interactions. See text for detailed explanations of some of the interactions.

using our services. Finally, the use of RNA-target prediction and RNA-protein interactions have already been established in synthetic biology.

5 The Freiburg Galaxy Server

For the dissemination of our RNA workbench we have chosen the Galaxy platform [73], because it allows to set up appropriate advisory and team-based structures to allow for an effective integration of tools across topics, and to avoid duplication of efforts. Galaxy is a highly modular, flexible and extensible system, that focuses on easy accessible and reproducible research. As part of this endeavor, we have invested massively in the definition and sharing of analysis workflows. Since the workbench is intended to be usable in a standard laboratory setting, we did not restrict ourself to RNA-based tools only but also integrated workflows for related task such as the analysis of RNA-seq data or pipelines for epigenetic research.

Since February 2013, RBC has been running a Galaxy server for high-throughput sequencing (HTS) data analysis. Since the start of this Freiburg Galaxy server we gained more than 500 users from a diversity of scientific disciplines, who use this service on a regular basis with about 1 Million jobs in 2016. Reproducible research is taken seriously: every version of an application, the raw data and the executed workflows and analysis histories are stored in a dedicated file server with a capacity of more than 100 TB and an extensive backup strategy. RBC and the Freiburg Galaxy Server is thus well positioned to deal with the challenges of big data. Moreover, group members of the RBC are experts in

Galaxy development, active community members and part of a commission that ensures the functional correctness of Galaxy applications and fulfill the strict rules to enable reproducibility.

The Freiburg Galaxy server offers data analysis tools in an easy accessible user-friendly way without any required knowledge in programming. Beside text manipulation and format converters, the Freiburg Galaxy instance offers tools for the analysis of HTS data from e.g. ChIP-seq, CLIP-seq, Exome-seq, genome annotation, and MethylC-seq experiments in addition to RNA-based pipelines. New tools are continuously developed and integrated with existing tools and databases as well as standard pipelines for the analysis of high-throughput sequencing data.

While the standard pipelines cover many aspects of the HTS analysis and are sufficient for large group of users, there are many additional analysis steps and visualization techniques that can be performed on an individual level. This includes comparison with publically available data or gene and pathway enrichment analysis. Even more, there are many individual experiments that require a deviation from the standard protocol. This can be due to the type of experiments (e.g. CLIP-seq or MethylC-seq), the quality of data (e.g., low coverage or different biases) or just an unusual use case (e.g. sequencing of compartments with low RNA expression).

In the definition of standard workflows we try to be as comprehensive as possible. To give an example, various tools for the analysis of RNAseq data are available on the Freiburg Galaxy server. First the raw data files are checked for their quality by using the tool FastQC. Preprocessing of the fastqsanger files, e.g. by trimming using Trim Galore!, is followed by mapping of reads to a reference genome. In our Galaxy server, several mappers are included, such as HISAT, TopHat2, and STAR. After read counting (htseq-count, feature counts), differentially expressed genes are calculated by the tool DESeq2 or edgeR. The output then needs to be filtered by e.g. p-value and sorted by fold change. In Galaxy, the results can be visualized by various bar charts, diagrams and heatmaps. All tools in Galaxy can be combined into shareable workflows, where all parameter settings and tool versions are saved. The standard RNAseq analysis workflow (Fig 3) is published on Galaxy (http://galaxyproject.github.io/training-material).

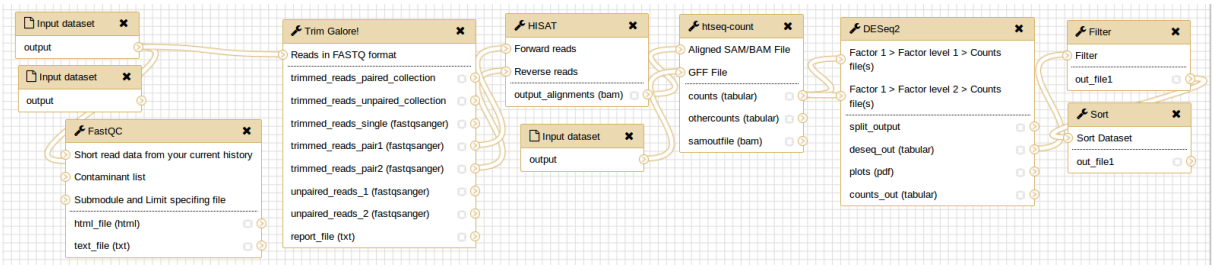

Figure 3: Standard RNAseq data analysis workflow in Galaxy. After raw data preprocessing, reads are mapped on a reference genome and counted. Differentially expressed genes can be calculated by e.g. DESeq2

Moreover, a virtualized version of Galaxy (via Galaxy Docker) enable other groups to use our RNA workbench and to do data analysis behind firewalls for e.g. sensitive data. Docker allows to wrap a complete Galaxy instance into a container that contains everything to run the instance. The only requirement is the basic installation and maintenance of the machine plus the installation of the Docker software. We have already developed several docker containers for a basic Galaxy instance as well as for several extensions, which allows to build up a Galaxy instance of a specific “flavour”, i.e., a Galaxy instance containing all necessary tools to handle some types of experiments such as RNA-seq or ChIP-seq experiments. The Freiburg Galaxy group also offers a comprehensive set of training material online

for self-study and invites the community to contribute to it (https://github.com/galaxyproject/training-material)

6 Concluding Remarks

With the recent advent of high-throughput RNA-based methods such as CLIP-seq, RIP-seq, ChIRP-Seq or Shape-Seq, the investigation of RNA-based regulation has become a central topic in molecular biology research. The RNA workbench curated by the RNA Bioinformatics center provides a comprehensive set of analysis tools and workflows for the analysis of this type of data. The integration of these tools in the Galaxy framework allows easy access to our RNA workbench. However, the technology in this field is rapidly developing and novel protocols are established in an increasing rate. Our current developments therefore include development of RNA-centric annotation efforts that take RNA processing steps into account [74], sequence/structure integrated motif finding, extending RBP peak callers to new protocols, or detecting RNA modifications. As a specific example, profiling the mRNA portions that are covered by translating ribosomes, so-called RiboSeq, is rapidly gaining in popularity and therefore motivated us to develop a new dedicated computational approach [75]. We also cope with the need for new approaches to integrate RNA secondary structure information into RBP binding site prediction [76]. Ultimately, an important task for the future will be the design and development of novel analysis tools and integration in our workbench to accommodate the technological progress.

Acknowledgements

This work was supported by the BMBF-funded project “Center for RNA-Bioinformatics” (Förderkennzeichen 031A538A (de.NBI-RBC) and 031L0101C (de.NBI-epi)) within the German Network for Bioinformatics Infrastructure (de.NBI).

References

[1] J. Medenbach, M. Seiler, and M. W. Hentze, “Translational control via protein-regulated upstream open reading frames,” *Cell*, vol. 145, no. 6, pp. 902–13, 2011.

[2] A. G. Baltz, M. Munschauer, B. Schwanhausser, A. Vasile, Y. Murakawa, M. Schueler, N. Youngs, D. Penfold-Brown, K. Drew, M. Milek, E. Wyler, R. Bonneau, M. Selbach, C. Dieterich, and M. Landthaler, “The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts,” *Mol Cell*, vol. 46, no. 5, pp. 674–90, 2012.

[3] S. Gerstberger, M. Hafner, and T. Tuschl, “A census of human RNA-binding proteins,” *Nature Reviews Genetics*, vol. 15, pp. 829–845, Nov. 2014.

[4] K. W. Brannan, W. Jin, S. C. Huelga, C. A. Banks, J. M. Gilmore, L. Florens, M. P. Washburn, E. L. V. Nostrand, G. A. Pratt, M. K. Schwinn, D. L. Daniels, and G. W. Yeo, “SONAR discovers RNA-binding proteins from analysis of large-scale protein-protein interactomes,” *Molecular Cell*, vol. 64, pp. 282–293, oct 2016.

[5] C. He, S. Sidoli, R. Warneford-Thomson, D. C. Tatomer, J. E. Wilusz, B. A. Garcia, and R. Bonasio, “High-Resolution Mapping of RNA-Binding Regions in the Nuclear Proteome of Embryonic Stem Cells,” *Mol. Cell*, vol. 64, pp. 416–430, Oct 2016.

[6] A. Castello, B. Fischer, C. K. Frese, R. Horos, A. M. Alleaume, S. Foehr, T. Curk, J. Krijgsveld, and M. W. Hentze, "Comprehensive Identification of RNA-Binding Domains in Human Cells," *Mol. Cell*, vol. 63, pp. 696–710, Aug 2016.

[7] K. Kapeli and G. W. Yeo, "Genome-wide approaches to dissect the roles of RNA binding proteins in translational control: implications for neurological diseases," *Front Neurosci*, vol. 6, p. 144, 2012.

[8] J. C. Darnell and J. D. Richter, "Cytoplasmic RNA-binding proteins and the control of complex brain function," *Cold Spring Harb Perspect Biol*, vol. 4, no. 8, p. a012344, 2012.

[9] J. Kong and P. Lasko, "Translational control in cellular and developmental processes," *Nat Rev Genet*, vol. 13, no. 6, pp. 383–94, 2012.

[10] F. Ibrahim, T. Nakaya, and Z. Mourelatos, "RNA dysregulation in diseases of motor neurons," *Annu Rev Pathol*, vol. 7, pp. 323–52, 2012.

[11] S. Rosina and L. D. Hurst, "Both maintenance and avoidance of RNA-binding protein interactions constrain coding sequence evolution," *Mol. Biol. Evol.*, Jan 2017.

[12] D. Schmiedel, J. Tai, R. Yamin, O. Berhani, Y. Bauman, and O. Mandelboim, "The rna binding protein imp3 facilitates tumor immune escape by downregulating the stress-induced ligands ulpb2 and micb," *eLife*, vol. 5, p. e13426, mar 2016.

[13] A. Busch, A. S. Richter, and R. Backofen, "IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions," *Bioinformatics*, vol. 24, no. 24, pp. 2849–56, 2008.

[14] U. Mückstein, H. Tafer, J. Hackermüller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker, "Thermodynamics of RNA-RNA binding," *Bioinformatics*, vol. 22, no. 10, pp. 1177–82, 2006.

[15] S. Memczak, M. Jens, A. Elefsinioti, F. Torti, J. Krueger, A. Rybak, L. Maier, S. D. Mackowiak, L. H. Gregersen, M. Munschauer, A. Loewer, U. Ziebold, M. Landthaler, C. Kocks, F. le Noble, and N. Rajewsky, "Circular RNAs are a large class of animal RNAs with regulatory potency," *Nature*, vol. 495, no. 7441, pp. 333–8, 2013.

[16] M. Wachsmuth, S. Findeiss, N. Weissheimer, P. F. Stadler, and M. Morl, "De novo design of a synthetic riboswitch that regulates transcription termination," *Nucleic Acids Res*, vol. 41, no. 4, pp. 2541–51, 2013.

[17] N. Rajewsky, "microRNA target predictions in animals," *Nat Genet*, vol. 38 Suppl, pp. S8–13, 2006.

[18] J. L. Rinn and J. Ule, "Oming in on RNA–protein interactions," *Genome Biology*, vol. 15, no. 1, p. 401, 2014.

[19] R. Lorenz, S. H. Bernhart, C. Höner Zu Siederdissen, H. Tafer, C. Flamm, P. F. Stadler, and I. L. Hofacker, "ViennaRNA Package 2.0," *Algorithms Mol Biol*, vol. 6, p. 26, 2011.

[20] S. Siebert and R. Backofen, "Methods for multiple alignment and consensus structure prediction of RNAs implemented in MARNA," *Methods Mol Biol*, vol. 395, pp. 489–502, 2007.

[21] C. Notredame, D. G. Higgins, and J. Heringa, "T-Coffee: A novel method for fast and accurate multiple sequence alignment.," vol. 302, no. 1, pp. 205–17, 2000.

[22] C. Smith, S. Heyne, A. S. Richter, S. Will, and R. Backofen, “Freiburg RNA Tools: a web server integrating IntaRNA, ExpaRNA and LocARNA,” *Nucleic Acids Res*, vol. 38 Suppl, pp. W373–7, 2010.

[23] D. Sankoff, “Simultaneous solution of the RNA folding, alignment and protosequence problems.,” *SIAM J. Appl. Math.*, vol. 45, no. 5, pp. 810–825, 1985.

[24] S. Will, K. Reiche, I. L. Hofacker, P. F. Stadler, and R. Backofen, “Inferring non-coding RNA families and classes by means of genome-scale structure-based clustering,” *PLoS Comput Biol*, vol. 3, no. 4, p. e65, 2007.

[25] S. Will, T. Joshi, I. L. Hofacker, P. F. Stadler, and R. Backofen, “LocARNA-P: Accurate boundary prediction and improved detection of structural RNAs,” *RNA*, vol. 18, no. 5, pp. 900–14, 2012.

[26] A. D. Palu, M. Möhl, and S. Will, “Alignnment of RNA with structures of unlimited complexity,” in *Proceedings of the Workshop on Constraint Based Methods for Bioinformatics (WCB 2010)*, p. 7, 2010.

[27] A. R. Gruber, S. Findeiss, S. Washietl, I. L. Hofacker, and P. F. Stadler, “RNAZ 2.0: IMPROVED NONCODING RNA DETECTION,” in *PSB10*, vol. 15, pp. 69–79, 2010.

[28] S. Heyne, F. Costa, D. Rose, and R. Backofen, “GraphClust: alignment-free structural clustering of local RNA secondary structures,” *Bioinformatics*, vol. 28, no. 12, pp. i224–i232, 2012.

[29] S. J. Lange, O. S. Alkhnbashi, D. Rose, S. Will, and R. Backofen, “CRISPRmap: an automated classification of repeat conservation in prokaryotic adaptive immune systems,” *Nucleic Acids Res*, 2013. SJL, OSA, and DR contributed equally to this work.

[30] A. Busch and R. Backofen, “INFO-RNA–a server for fast inverse RNA folding satisfying sequence constraints,” *Nucleic Acids Res*, vol. 35, no. Web Server issue, pp. W310–3, 2007.

[31] R. Kleinkauf, T. Houwaart, R. Backofen, and M. Mann, “antaRNA - multi-objective inverse folding of pseudoknot RNA using ant-colony optimization,” *BMC Bioinformatics*, vol. 16, no. 1, pp. 1–7, 2015.

[32] M. R. Friedlander, W. Chen, C. Adamidi, J. Maaskola, R. Einspanier, S. Knespel, and N. Rajewsky, “Discovering microRNAs from deep sequencing data using miRDeep,” *Nat Biotechnol*, vol. 26, no. 4, pp. 407–15, 2008.

[33] N. W. Breakfield, D. L. Corcoran, J. J. Petricka, J. Shen, J. Sae-Seaw, I. Rubio-Somoza, D. Weigel, U. Ohler, and P. N. Benfey, “High-resolution experimental and computational profiling of tissue-specific known and novel miRNAs in Arabidopsis,” *Genome Res*, vol. 22, no. 1, pp. 163–76, 2012.

[34] M. Fasold, D. Langenberger, H. Binder, P. F. Stadler, and S. Hoffmann, “DARIO: a ncRNA detection and analysis tool for next-generation sequencing experiments,” *Nucleic Acids Res*, vol. 39, no. Web Server issue, pp. W112–7, 2011.

[35] S. Li, L. M. Liberman, N. Mukherjee, P. N. Benfey, and U. Ohler, “Integrated detection of natural antisense transcripts using strand-specific RNA sequencing data,” *Genome Res*, 2013.

[36] A. Krek, D. Grun, M. N. Poy, R. Wolf, L. Rosenberg, E. J. Epstein, P. MacMenamin, I. da Piedade, K. C. Gunsalus, M. Stoffel, and N. Rajewsky, "Combinatorial microRNA target predictions," *Nat Genet*, vol. 37, no. 5, pp. 495–500, 2005.

[37] S. Lall, D. Grun, A. Krek, K. Chen, Y.-L. Wang, C. N. Dewey, P. Sood, T. Colombo, N. Bray, P. Macmenamin, H.-L. Kao, K. C. Gunsalus, L. Pachter, F. Piano, and N. Rajewsky, "A genome-wide map of conserved microRNA targets in C. elegans," *Curr Biol*, vol. 16, no. 5, pp. 460–71, 2006.

[38] S. H. Bernhart, H. Tafer, U. Muckstein, C. Flamm, P. F. Stadler, and I. L. Hofacker, "Partition function and base pairing probabilities of RNA heterodimers," *Algorithms Mol Biol*, vol. 1, no. 1, p. 3, 2006.

[39] U. Muckstein, H. Tafer, J. Hackermuller, S. H. Bernhart, P. F. Stadler, and I. L. Hofacker, "Thermodynamics of RNA-RNA binding," *Bioinformatics*, vol. 22, no. 10, pp. 1177–82, 2006.

[40] P. R. Wright, J. Georg, M. Mann, D. A. Sorescu, A. S. Richter, S. Lott, R. Kleinkauf, W. R. Hess, and R. Backofen, "CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains," vol. 42, no. Web Server issue, pp. W119–23, 2014. PRW, JG and MM contributed equally to this work.

[41] F. Eggenhofer, H. Tafer, P. F. Stadler, and I. L. Hofacker, "RNApredator: Fast accessibility-based prediction of sRNA targets," *Nucleic Acids Res.*, vol. 39, pp. W149–W154, 2011.

[42] P. R. Wright, A. S. Richter, K. Papenfort, M. Mann, J. Vogel, W. R. Hess, R. Backofen, and J. Georg, "Comparative genomics boosts target prediction for bacterial small RNAs," *Proc Natl Acad Sci USA*, 2013.

[43] R. H. Baltz, "Combinatorial biosynthesis of cyclic lipopeptide antibiotics: a model for synthetic biology to accelerate the evolution of secondary metabolite biosynthetic pathways," *ACS Synthetic Biology*, p. 120809123853000, Aug. 2012.

[44] A. Castello, B. Fischer, K. Eichelbaum, R. Horos, B. M. Beckmann, C. Strein, N. E. Davey, D. T. Humphreys, T. Preiss, L. M. Steinmetz, J. Krijgsveld, and M. W. Hentze, "Insights into RNA Biology from an Atlas of Mammalian mRNA-Binding Proteins," *Cell*, vol. 149, pp. 1393–1406, June 2012.

[45] D. Ray, H. Kazan, K. B. Cook, M. T. Weirauch, H. S. Najafabadi, X. Li, S. Gueroussov, M. Albu, H. Zheng, A. Yang, H. Na, M. Irimia, L. H. Matzat, R. K. Dale, S. A. Smith, C. A. Yarosh, S. M. Kelly, B. Nabet, D. Mecenas, W. Li, R. S. Laishram, M. Qiao, H. D. Lipshitz, F. Piano, A. H. Corbett, R. P. Carstens, B. J. Frey, R. A. Anderson, K. W. Lynch, L. O. F. Penalva, E. P. Lei, A. G. Fraser, B. J. Blencowe, Q. D. Morris, and T. R. Hughes, "A compendium of RNA-binding motifs for decoding gene regulation," *Nature*, vol. 499, pp. 172–177, July 2013.

[46] K. B. Cook, H. Kazan, K. Zuberi, Q. Morris, and T. R. Hughes, "RBPDB: a database of RNA-binding specificities.," *Nucleic acids research*, vol. 39, pp. D301–8, Jan. 2011.

[47] S. D. Auweter, F. C. Oberstrass, and F. H.-T. Allain, "Sequence-specific binding of single-stranded RNA: is there a code for recognition?," *Nucleic Acids Research*, vol. 34, pp. 4943–4959, Sept. 2006.

[48] G. V. Mechetin and D. O. Zharkov, "Mechanisms of diffusional search for specific targets by DNA-dependent proteins," *Biochemistry (Moscow)*, vol. 79, pp. 496–505, June 2014.

[49] M. Bantscheff, M. Schirle, G. Sweetman, J. Rick, and B. Kuster, “Quantitative mass spectrometry in proteomics: a critical review,” *Analytical and Bioanalytical Chemistry*, vol. 389, pp. 1017–1031, Sept. 2007.

[50] J. Ule, K. Jensen, A. Mele, and R. B. Darnell, “CLIP: A method for identifying protein-RNA interaction sites in living cells,” *Methods*, vol. 37, pp. 376–386, Dec. 2005.

[51] G. W. Yeo, N. G. Coufal, T. Y. Liang, G. E. Peng, X.-D. Fu, and F. H. Gage, “An RNA code for the FOX2 splicing regulator revealed by mapping RNA-protein interactions in stem cells.,” *Nature structural & molecular biology*, vol. 16, pp. 130–7, Feb. 2009.

[52] J. König, K. Zarnack, G. Rot, T. Curk, M. Kayikci, B. Zupan, D. J. Turner, N. M. Luscombe, and J. Ule, “iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution.,” *Nature structural & molecular biology*, vol. 17, pp. 909–15, July 2010.

[53] M. Hafner, M. Landthaler, L. Burger, M. Khorshid, J. Hausser, P. Berninger, A. Rothballer, M. Ascano, A.-c. Jungkamp, M. Munschauer, A. Ulrich, G. S. Wardle, S. Dewell, M. Zavolan, and T. Tuschl, “Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP.,” *Cell*, vol. 141, pp. 129–41, Apr. 2010.

[54] E. L. Van Nostrand, G. A. Pratt, A. A. Shishkin, C. Gelboin-Burkhart, M. Y. Fang, B. Sundararaman, S. M. Blue, T. B. Nguyen, C. Surka, K. Elkins, R. Stanton, F. Rigo, M. Guttman, and G. W. Yeo, “Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP),” *Nature Methods*, vol. 13, pp. 508–514, Mar. 2016.

[55] B. J. Zarnegar, R. A. Flynn, Y. Shen, B. T. Do, H. Y. Chang, and P. A. Khavari, “irCLIP platform for efficient characterization of protein-RNA interactions,” *Nature Methods*, vol. 13, pp. 489–492, Apr. 2016.

[56] Y. Sugimoto, A. Vigilante, E. Darbo, A. Zirra, C. Militti, A. DÁmbrogio, N. M. Luscombe, and J. Ule, “hiCLIP reveals the in vivo atlas of mRNA secondary structures recognized by Staufen 1,” *Nature*, vol. 519, pp. 491–494, Mar. 2015.

[57] D. L. Corcoran, S. Georgiev, N. Mukherjee, E. Gottwein, R. L. Skalsky, J. D. Keene, and U. Ohler, “PARalyzer: definition of RNA binding sites from PAR-CLIP short-read sequence data,” *Genome Biol*, vol. 12, no. 8, p. R79, 2011.

[58] W. H. Majoros, P. Lekprasert, N. Mukherjee, R. L. Skalsky, D. L. Corcoran, B. R. Cullen, and U. Ohler, “MicroRNA target site identification by integrating sequence and binding information,” *Nat Methods*, vol. 10, no. 7, pp. 630–3, 2013.

[59] N. C. Jones and P. Pevzner, *An introduction to bioinformatics algorithms*. Computational molecular biology, Cambridge, MA: MIT Press, 2004.

[60] T. L. Bailey and C. Elkan, “Fitting a mixture model by expectation maximization to discover motifs in biopolymers.,” *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, vol. 2, pp. 28–36, Jan. 1994.

[61] S. Georgiev, A. P. Boyle, K. Jayasurya, X. Ding, S. Mukherjee, and U. Ohler, “Evidence-ranked motif identification,” *Genome Biology*, vol. 11, no. 2, p. R19, 2010.

[62] P. Sood, A. Krek, M. Zavolan, G. Macino, and N. Rajewsky, “Cell-type-specific signatures of microRNAs on target mRNA expression,” *Proc Natl Acad Sci USA*, vol. 103, no. 8, pp. 2746–51, 2006.

[63] M. Hiller, R. Pudimat, A. Busch, and R. Backofen, “Using RNA secondary structures to guide sequence motif finding towards single-stranded regions.,” *Nucleic acids research*, vol. 34, p. e117, Jan. 2006.

[64] D. Maticzka, S. J. Lange, F. Costa, and R. Backofen, “GraphProt: modeling binding preferences of RNA-binding proteins,” *Genome Biol*, vol. 15, no. 1, p. R17, 2014.

[65] The RNA Central Consortium, “RNAcentral: a comprehensive database of non-coding RNA sequences,” *Nucleic Acids Res.*, vol. 45, pp. D128–D134, 2017.

[66] G. Anders, S. D. Mackowiak, M. Jens, J. Maaskola, A. Kuntzagk, N. Rajewsky, M. Landthaler, and C. Dieterich, “doRiNA: a database of RNA interactions in post-transcriptional regulation,” *Nucleic Acids Res*, vol. 40, no. Database issue, pp. D180–6, 2012.

[67] K. Blin, C. Dieterich, R. Wurmus, N. Rajewsky, M. Landthaler, and A. Akalin, “DoRiNA 2.0–upgrading the doRiNA database of RNA interactions in post-transcriptional regulation,” *Nucleic Acids Research*, vol. 43, pp. D160–D167, nov 2014.

[68] P. Glažar, P. Papavasileiou, and N. Rajewsky, “circBase: a database for circular RNAs,” *RNA*, vol. 20, pp. 1666–1670, sep 2014.

[69] F. Jühling, M. Mörl, R. K. Hartmann, M. Sprinzl, P. F. Stadler, and J. Pütz, “trnadb 2009: compilation of trna sequences and trna genes,” *Nucleic Acids Research*, vol. 37, p. D159, 2008.

[70] M. Sprinzl and K. S. Vassilenko, “Compilation of trna sequences and sequences of trna genes,” *Nucleic Acids Research*, vol. 33, p. D139, 2005.

[71] J. Fallmann, V. Sedlyarov, A. Tanzer, P. Kovarik, and I. L. Hofacker, “AREsite2: an enhanced database for the comprehensive investigation of AU/GU/U-rich elements,” *Nucleic Acids Research*, vol. 44, pp. D90–D95, Jan. 2016.

[72] R. Gilsbach, S. Preissl, B. A. Grüning, T. Schnick, L. Burger, V. Benes, A. Würch, U. Bönisch, S. Günther, R. Backofen, B. K. Fleischmann, D. Schübeler, and L. Hein, “Dynamic DNA methylation orchestrates cardiomyocyte development, maturation and disease,” *Nature Communications*, vol. 5, p. 5288, oct 2014.

[73] E. Afgan, D. Baker, M. van den Beek, D. Blankenberg, D. Bouvier, M. Čech, J. Chilton, D. Clements, N. Coraor, C. Eberhard, B. Grüning, A. Guerler, J. Hillman-Jackson, G. V. Kuster, E. Rasche, N. Soranzo, N. Turaga, J. Taylor, A. Nekrutenko, and J. Goecks, “The galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update,” *Nucleic Acids Research*, vol. 44, pp. W3–W10, may 2016.

[74] N. Mukherjee, L. Calviello, A. Hirsekorn, S. de Pretis, M. Pelizzola, and U. Ohler, “Integrative classification of human coding and noncoding genes through RNA metabolism profiles,” *Nature Structural & Molecular Biology*, vol. 24, pp. 86–96, nov 2016.

[75] L. Calviello, N. Mukherjee, E. Wyler, H. Zauber, A. Hirsekorn, M. Selbach, M. Landthaler, B. Obermayer, and U. Ohler, "Detecting actively translated open reading frames in ribosome profiling data," *Nature Methods*, vol. 13, pp. 165–170, dec 2015.

[76] J. Fallmann, S. Will, J. Engelhardt, B. Grüning, R. Backofen, and P. F. Stadler, "The german center for RNA-bioinformatics (RBC)," *J. Biotech.*, 2017. this issue.