COMPUTATIONAL MODELS TO INFER REGULATORS OF GENE EXPRESSION FOR HIGH-THROUGHPUT DATA

Inauguraldissertation
zur
Erlangung der Würde eines Doktors der Philosophie
vorgelegt der
Philosophisch-Naturwissenschaftlichen Fakultät
der Universität Basel

von
MARIA KATSANTONI

Basel, 2023

Genehmigt von der Philosophisch-Naturwissenschaftlichen Fakultät
auf Antrag von

Erstbetreuerin: Prof. Dr. Mihaela Zavolan

Zweitbetreuer: Prof. Dr. Erik van Nimwegen

externer Experte: Dr. Julian König

Basel, den 13.12.2022

Prof. Dr. Marcel Mayor
Dekan

— The real meaning of enlightenment is to gaze with undimmed eyes on all darkness.
- Ascesis: Salvatores dei, Nikos Kazantzakis

Dedicated to
Foivos

ABSTRACT

Ever since the formulation of the central dogma of biology, the focus was shifted into how the various steps up to the protein formation are regulated. RNA-binding proteins (RBPs) have been shown to be instrumental in a vast number of post-transcriptional processes. Crosslinking immunoprecipitation (CLIP) is a mainstay in the experimental approaches used for detecting the binding partners of the RBPs. Many variations of these protocols have been developed, along with multiple software solutions for their analysis. However, most of the existing methods do not include efficient pre-processing in their processing and their statistical models still do not efficiently remove all sources of noise. To deal with these issues, we developed RCRUNCH. RCRUNCH is an automated workflow, that deals with pre-processing, has its own statistical approach to efficiently detect significant binding events and also reliably infers motifs. Additional features are inclusion of multi-mappers, selective removal of ncRNAs and a transcriptomic approach for considering binding events spanning splice junctions. RCRUNCH was shown to have a reliable performance in comparison to the most wide used methods in a number of metrics. ENCODE eCLIP data were analysed using RCRUNCH and many known interactions and motifs were successfully reproduced, along with some new interesting findings. This interest in high-throughput data analysis led to a more collaborative project called ZARP, used for high-throughput analysis of RNA-seq data, developed in the Zavolan group with the FAIR principles in mind, and a good roadmap on how to apply best practices in the specific context of a bioinformatics analysis. In this project we focused more on flexibility and usability of the workflow even with minimal bioinformatics expertise.

ACKNOWLEDGEMENTS

I would like to thank the members of the committee group Julian Koenig, Erik van Nimwegen and Mihaela Zavolan for all their useful comments and suggestions regarding my project all these years, as well as the thesis. Special thanks to Mihaela. For always being there, guiding me with her knowledge and deep understanding of science, but also being receptive to opinions and new ideas, a rare and appreciated quality in a leader. Also, thanks to Erik and his group, for the many useful discussions in the time I spent working on a collaborative project. His undeniable passion for science and insights on approaching scientific questions helped me greatly. I am also grateful to Alex, Christina, Jeremie, Ralf, Andrea, Joao, Maciek, Arka, Nitish, Souvik, Shreemoyeee, and all members of the Zavolan group for their help in the beginning, but also the nice atmosphere, discussions and group activities. Being a first generation PhD student in a different country as well, I want to thank my family for giving me the strength to follow my own path even though it was a different one. Thanks to my friends back in Greece for all the love and support all these years. Lastly, thanks Foivos for the support, understanding and love required since my PhD started when yours was ending. Lots of years spent learning in our home and hopefully more to come!

CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 INTRODUCTION

1.1 GENE EXPRESSION AND ITS REGULATION

The central dogma of biology postulates that the flow of genetic information starts from the DNA, is transferred via transcription to the RNA and then via translation to proteins [1]. The direction of this flow according to the central dogma is irreversible. This basic view has evolved over the years to include a vast array of regulatory processes. Transcription is in itself not as straightforward, with epigenetic marks such as histone modifications and DNA methylation, and regulatory processes such as chromatin remodeling having a great effect on the final outcome. In a similar manner the information transfer from DNA to RNA involves much more than the mere act of transcription. The composition and function of mRNAs further depends on RNA modifications, splicing, polyadenylation, traffic of the mRNA to the correct location to be translated, its stability and susceptibility to degradation. Moreover, transcription does not only produce protein-coding messenger RNAs but also non-coding RNAs (ncRNAs), some of which have regulatory functions in translation (rRNAs, tRNAs), RNA decay and other processes. Translation, as indicated already, requires rRNAs that participate in the formation of the ribosome, and its output is affected by the stability of the mRNA and its localisation. Features of the mRNA, like modifications and length of the untranslated regions (3′ UTRs) have been shown to affect the translation efficiency. RNA binding proteins (RBPs) are important players in the regulation and fine-tuning of the genetic flow of information to the final protein product.

Proteins have been associated with specific functions in these regulatory processes. Transcription factors are essential to the epigenetic regulation [2]. The first indication was the detection of short genomic regions that shared common qualities, like susceptibility to transcription induction after exposure to increased temperatures. Conversely, these small regions, termed heat shock elements (HSE), were absent from genes non-responsive to increased temperatures [3]. To further consolidate the importance of HSEs, they were transferred from the heat inducible gene (hsp70) to the non-inducible gene thymidine kinase [4]. The expression of the latter was increased upon temperature increase, indicating that the HSE does indeed confer heat-inducible transcription. These short DNA sequences were shown to act via binding of specific proteins, called transcription factors (TF), which could affect transcription in a positive or negative manner.

At the time, it was thought that gene expression is determined at the step of transcription, while post-transcriptional processes were largely ignored, perhaps due to the prokaryotic paradigm of regulation [2]. However, it was beginning to become apparent that eukary-

otic cells have a wide range of post-transcriptional processes in place, affecting gene expression and that there are signals and factors (RNA-binding proteins, RBPs) participating in these processes, bringing the RBPs to the forefront, as gene expression regulators [5] [6]. The detection of highly conserved domains in RBPs [7], [8] contributed further to the elucidation of their function, first in the context of development [5, 8]. Since then, many RBPs have been characterized along with their regulation and novel functions in a multitude of different contexts apart from development [9], [10].

Once a protein is produced, it can undergo further interactions or modifications that affect its function. The modifications and interaction partners can also be modulated by small molecules, protein phosphorylation or protein-protein interactions [11].

As we can see, the central dogma of gene expression has been enriched since its conception with a vast number of diverse regulators affecting each and every step of the process, controlling a protein's life, even before its formation at the transcript level and all the way until its final degradation. The complexity and sheer number of regulatory processes governing gene expression, along with the expected stochasticity of these processes which typically involve a small number of molecules, render the exact prediction of the final outcome in a specific context rather inaccurate and complicated.

1.2 RNA-BINDING PROTEINS

1.2.1 *RNA-binding proteins: important regulators of expression*

RNA binding proteins (RBPs) are a major component of gene regulation. They bind a multitude of distinct RNA species at different stages of their life cycle. Since the proposal of RBPs acting as expression regulators [12] [5], more than 1542 RBPs have been uncovered in humans, which is around 7.5% of the protein-coding genome [10, 13]. The plethora of these proteins, their deep evolutionary conservation [10], and their involvement in genetic disorders [14] [15] and human disease in general [16], render these proteins of high interest. In particular, understanding their mode of action as gene regulators will open new avenues for therapeutic interventions.

1.2.2 *RBP discovery*

Despite the fact the first RBPs and their domains were already characterized by the late eighties, the catalog of RBPs, either identified experimentally or predicted, keeps getting longer. Although at first RBPs were studied one at the time, the interest rose for large scale approaches to uncover all RBPs in a living cell. The main high-throughput method for detecting proteins with RNA-binding activity *in vivo* relies on RNA interactome capture (RIC [17] [18] [19]). In this approach, crosslinking is used to stabilize any RBP-RNA interactions taking place in a cell at a given time. Then, the ribonucleoprotein (RNP)

complexes are captured with the help of oligo(dT) beads and the proteins that are bound to RNAs, i.e. RBPs, are identified by mass spectrometry. Modifications of this protocol have enabled not only higher specificity, but also the determination of the exact regions of the RBP that interact with the RNA [17, 19] [20]. Additional fractionation might give more insight into e.g nuclear RBPs and their function [17].

Additional curation to more reliably annotate the RBP repertoire was performed in [10]. Specifically, experimental information on whether proteins interact with RNA directly or indirectly was used, based on which shared domains (so-called Pfam domains [21], represented as position-specific weight matrices) were gathered and used for training hidden Markov models, which were then applied on the whole human genome. This led to the doubling of the number of putative RBPs, relative to the number of RBPs detected based on the experimental methods mentioned above. Furthermore, based on the observation that proteins that often interact with other RBPs in an RNA-dependent manner are usually RBPs themselves, a computational method was developed, which takes advantage of mass spectrometry-determined protein-protein interactomes to predict RNA binding activity of uncharacterised proteins [17, 22]. All prediction methods rely on particular assumptions (e.g that the presence of an RNA-binding domain (RBD) always leads to an RNA-related function) that do not always hold (participation in RNP complex does not always equal RNA interaction) and therefore these predictions generally include many false positives. Additionally, for these models to be trained properly there is the prerequisite of clean training data, which then relies on the accuracy of the experiments used to derive them, which in many cases might be an issue.

1.2.3 *RBPs contain highly conserved RNA-binding domains*

The elucidation of various RBPs in terms of structure and binding preferences along with conservation studies, led to the identification of highly conserved regions of sizes ranging from 60-90 amino acids with common features, called RNA-binding domains (RBD). One of the first RBDs described was the RNP consensus sequence (RNP-CS) [7]. Experimental approaches to identify proteins that bind RNAs and multiple alignments of their conserved regions, uncovered a pattern of 90 amino acid sequences among some RBPs that was additionally found to be conserved all the way from yeast to human [5]. Octapeptides within these regions were found to be further conserved and were hypothesized to mediate the RBP-RNA binding, even though later on more peptides were found to be deeply conserved (RNP1, RNP2).

Some of the main RNA-binding domains described so far are: RNA recognition motif (RRM) [5, 23]: one of the most frequently observed RBDs in human RBPs with a frequency of 0.5-1% across human genes. Its characteristic secondary structure leads to recognition of

between four and eight nucleotides-long motifs. Usually more than one such RBD is required for more specific binding.

K-homology domain (KH): this type of domain has some variability in terms of observed topologies. In this case the RNA binding relies mostly on shape complementarity and hydrogen bonds. This type of domain can also bind single-stranded DNA (ssDNA).

Zinc finger (ZnF): this domain binds DNA as well as RNA. There are further family distinctions within this type, and the specific binding depends on the structure of the DNA, RNA and of the RBP that contains the ZnF domain.

Ribosomal S1-like (in short S1) domain: this domain binds to specific nucleotides via surface complementarity and secondary structure, similarly to RRM.

PAZ and PIWI domains: these domains are found in RBPs participating in the processing of small regulatory RNAs, microRNAs (miRNAs) and small interfering RNAs (siRNAs).

A more extensive catalogue of annotated domains can be found in [24]. RBDs as individual entities offer limited specificity in terms of target specificity. It is the combination of multiple such RBDs and the varying secondary structures of the RNAs that render the RBPs so uniquely specific to certain RNAs during specific stages of gene expression regulation.

The interdomain sequences between specific RBDs have also been shown to significantly affect the affinity and specificity of RBPs towards specific targets. There are also models to calculate the final affinity of either individual RBD and their combinatorial binding affinity [25], that show that shorter linkers lead to increased affinity as opposed to longer linkers (>50-60 residues) which lead to affinity that is close to what it would be if the two domains were acting independently. Furthermore, protein-protein interactions and the varying combinations of proteins into complexes (e.g spliceosome) offer a multitude of different surfaces available for binding [26] [27]. The conformational rigidity of both the RBP interfaces and the target RNA has been correlated to higher specificity of binding while flexibility seems to favor binding affinity [28].

Apart from the RBDs mentioned above, disordered regions have also been shown to maintain important roles in RBP functionality [29]. They mediate RNA-binding, governed by specific motifs, e.g RGG/RG [30], as well as regulate transcription via interaction with the RNA polymerase II [31]. The role of disordered regions in RBP-RNA as well as protein-protein interactions emphasize their importance for gene regulation.

It is evident that although RBPs have distinct RBDs that mediate the RNA binding in a sequence-specific manner, the interactions and localisation of both RBPs and RNAs at a given time could further lead to different outcomes of the RBP-RNA interaction. Therefore, RNA regulation via RBP binding does not seem to be a linear process, but rather the result of intricate interactions of more than one regulators. It is therefore of important to understand the role of individual RBPs

in regulatory processes to be able to decipher their joint functions in determining gene expression.

1.2.4 *RBPs : localisation, binding partners, functions*

RBPs have been observed across all subcellular locations, where they mediate a plethora of regulatory processes. The combination of CLIP methods (see next section) with cellular fractionation aims to shed light into the localisation-dependent functions of RBPs [32]. In recent years, the discovery of more and more membraneless organelles that compartmentalise various processes in eukaryotic cells [33, 34] (e.g Cajal bodies, paraspeckles, nuclear speckles, P-bodies, stress granules), has greatly increased the interest in the mechanisms underlying the formation and function of these structures. Liquid-liquid phase separation (LLPS) is a key mechanism describing the formation of such organelles [35] and RNA-binding proteins have been been shown to provide building blocks for the formation of these condensates. Intrinsically-disordered regions (IDRs) and prion-like domains contribute to the phase separation [34, 36], while defects caused for e.g. by mutations in RBPs have emerged as culprits of various neurodegenerative diseases [37].

The versatility of RBPs continue to surprise scientists. For e.g. recent studies have shown that some RBPs also act as metabolic enzymes, one theory being that RNA-binding might regulate their availability to act as enzymes and the other way round [9, 38]. Another example of dual functionality is chromatin association [39]. A method called SPACE (Silica Particle Assisted Chromatin Enrichment) was developed with the aim of identifying all chromatin-associated proteins and the outcome was that 48% out of a total of 1459 these proteins were annotated as RBPs [39]. This binding was again associated with the IDRs of these RBPs. [40]

The function of RBPs as enhancers or suppressors of miRNA targeting (MT) has also been established and these have been studied in the context of cancer [41]. Since then, there have been approaches to detect more of the MT events across the genome [41, 42].

1.2.5 *Methods to determine RBP-RNA interactions*

As already mentioned, RBPs interact with RNAs via specific conserved domains of 60-90 amino acids called RNA-binding domains or RBDs. Several RBPs are typically bound to an RNA, thus forming ribonucleoprotein complexes or RNPs. The RNP composition is highly dynamic in space, time and cell type or state and it is this dynamics that leads to the expression of different subsets of genes in different cell types. Therefore, mapping these dynamic interactions is of great interest. Many approaches are currently available identifying RNA targets and sites of individual RBPs (protein-centric approaches). Conversely, methods are also available for elucidating

the components of RNP complexes containing a specific RNA (RNA-centric approach).

One of the first methods for isolation of RNP complexes is RNA immunoprecipitation (RIP) [43] [44]. As the name implies, this method relies on selective capture of RNPs containing a specific protein and consequently all the RNAs that are bound to it. The capture step can be done in native conditions [45] or after formaldehyde treatment to stabilize the bound molecules via crosslinking [46]. This method was found to be error-prone, mainly due to random RNP formations after cell lysis [47].

The methodology was superseded by crosslink and immunoprecipitation (CLIP) [48]. The main advantage of this method is the non-reversible crosslinking of the RBP to the UV-irradiated ribonucleotides thanks to a covalent bond formation. The UV crosslinking is shown to lead to higher signal:noise since it does not induce chemical bridges or protein-protein interactions as does formaldehyde crosslinking [49]. Selective immunoprecipitation of a specific RBP and treatment with RNAse allows for detection of RNA-targets with the resolution of individual binding sites, because these are protected by the RBP during sample preparation and subsequently sequenced. Due to insufficient proteinase K treatment it has been shown that reverse transcriptase can be blocked by remaining aminoacid in the position of the covalent bond and therefore lead to truncated reads [50, 51]. However, this was not universally accepted, with an alternative hypothesis being that there is an increased rate of nucleotide misincorporation at the crosslink site during reverse transcription [52, 53]. Of course this depends on the type of the reverse transcriptase and the specific conditions [54].

Various adaptations of this method have emerged over the years, with the aim to provide enhanced accuracy in terms of the targets and nucleotide level detection of the binding sites. An example of such an enhancement is iCLIP [55]. In this method, the incorporation of a cDNA self-circularisation step ensures that cDNAs that are truncated at the crosslink site are still amplified and retained in the sample, in spite of the 5′ adaptor not being reached during reverse transcription. This is estimated to yield higher, nucleotide-level accuracy of the crosslink site. In this method random barcodes were also introduced to deal with PCR amplification artifacts. A latest update to the protocol though has replaced the circularisation step with T4 RNA ligase 1 to ligate a DNA adapter to the 3′ end of the cDNA [56]. eCLIP [57] was designed as a further improvement of iCLIP in terms of yield and performance. It incorporates some of iCLIP's steps concerning library preparation of RNA fragments, but the step of circularisation, deemed inefficient, is replaced by a two step ligation process, which is further improved by optimizing the T4 RNA ligase protocol. The radioactive labeling of the RNA is also skipped to shorten the time required for sample preparation (4 days). The addition of a size-matched input (SMI) sample, which goes through all the steps of sample preparation except for the immunoprecipitation, gives improved background signal. Another modification to

the iCLIP protocol, irCLIP [58], takes advantage of an infrared probe-labeled adaptor, which is as efficient as a standard adaptor, yet reduces the time required for protein–RNA complex visualization 10- to 100-fold. Also, the higher sensitivity of irCLIP adaptor detection leads to faster quality controls during the protocol implementation.

PAR-CLIP [59] is another CLIP variant whose novelty lies in the incorporation of 4-thiouridine (4SU) in cells prior to crosslinking. This allows milder conditions for crosslinking, at >310 nm of UV instead of the 254 nm, leading to higher specificity and efficiency of crosslinking, as well as to a specific signature of crosslinked sites, where T-C mutations occur with much higher frequency than in the background.

In some cases, it is of higher interest to detect the proteins binding to a specific RNA. Methods like RNA affinity protein capture, which enables purification of a specific RNA together with its interactor RBPs, lead to identification of these bound proteins via mass-spectrometry (MS), which might further lead to identification of new proteins acting as RBPs. Another method, RNA-directed proximity-based proteome labeling [60] relies on a specific labeling enzyme which covalently modifies all proteins that are located in the proximity to a specific RNA. This better captures transient interactions and can also be used for specific cell localisations.

A method focusing on a specific RBP bound to a specific target, thus enabling parallel specification of the interacting amino-acids of the RBP and the crosslinked nucleotides, is crosslinking of segmentally isotope-labeled RNA and tandem mass spectrometry (CLIR-MS/MS) that uncovers the structure of the interacting interface and allows for atomic resolution structures of the RNPs formed [61].

Another step in the evolution of these methods is the drop of the requirement for an RBP specific antibody. This approach called TRIBE [62], relies on the fusion of the RBP of interest with a domain from the ADAR family which catalyzes the adenosine-to-inosine changes in the RNAs located in the proximity of the enzyme. The limitation of this method is that a double-stranded RNA where ADAR can bind needs to be in close proximity to the RBP-binding site, which does not come across that often. APOBEC1 was used to circumvent this issue, since it can catalyze a C to U conversion even on single-stranded RNA [63]. Additionally, a domain that recognizes m6A modifications (m6A-binding YTH domain) can be fused to APOBEC1 in order to detect C-U conversions close to m6A sites (DART-seq, [64]). An attempt to apply the DART-seq method at the single-cell level and together with long read sequencing is the method called STAMP (Surveying Targets by APOBEC-Mediated Profiling) [65]. Although these methods successfully eliminate the reliance on a specific RBP antibody with the accompanying sources of noise (antibody specificity, sensitivity), they are still to be evaluated regarding the accuracy of the editing and target identification. Current evaluations rely mostly on one or a small number of RBPs and the absence of ground truth or comparison with other equally noisy methods, as well as limitations on technical aspects render their utility uncertain. Still their importance lies in providing a means to obtain cell and isoform specific

insights on RBP regulation along with plans to transfer this to *in vivo* settings, moving the field forward.

1.2.6 *Software for identifying RBP binding sites from CLIP data*

CLIP is an established method widely used for the discovery of RBP targets and binding sites *in vivo*. However, despite various optimisations that have been implemented, the data still is quite noisy. The issue of PCR duplicates has been dealt with by the introduction of UMIs. The variable specificity of antibodies, variable efficiency of purification and variable extent and sequence-dependence of fragmentation all affect which RNA fragments are captured in a CLIP experiment. Moreover, highly expressed RNAs make their way into the sample despite no binding to the RBP of interest, thus introducing further noise that dampens the true binding signal [66]. Apart from technical sources of error, there is variability in the observed signal to be expected as well, due to biological factors, such as the stochasticity of gene expression, the fact that cells are pooled together and so the signal is averaged across possibly diverse cell states etc. For all these reasons, there have been numerous attempts to provide reliable estimates and model the noise and further rank the detected binding peaks by a measure that reflects the affinity of the RNA-RBP interaction. "Background" samples (coming from RNA-seq, SMI sequencing, or random samples of reads from the CLIP experiment) are used to model the expected noise, while crosslinking-diagnostic events (truncation of the RNA at the crosslinking site, nucleotide substitutions or deletions introduced during reverse transcription) are further used to validate true binding events. Over the years, there has been quite an expansion in the available methods in this field.

iCount [67] is one of the first approaches to the analysis of data produced with the iCLIP method. The crosslink positions inferred by iCLIP are the nucleotides immediately upstream of starts of the sequenced cDNAs, where the reverse transcriptase is expected to fall off, due to a 'defect' in the template caused by the "stub" left by the degradation of the protein at the crosslinked site. The method does not rely on a quantitative model of read sampling, but rather relies on extensive quality-filtering of the reads. The method has more recently evolved to an automatic workflow, including a command line interface [68].

Piranha [69] uses a zero truncated negative binomial distribution (ZTNB) to model the observed read coverage, based on the observation that the distribution of read counts appears to be an overdispersed Poisson, prevalent in high-throughput immunoprecipitation experiments. In this model, it is assumed that most of the regions with some read coverage stem from noise and not real binding signal and therefore, a background model is fitted to the bulk of the data, while the regions deviating from this distribution are considered to have true signal. If additional covariates are provided, then a zero-truncated negative binomial regression (ZTNBR) model is applied.

ASPeak [70] was developed for the analysis of RIP-seq data, but is also expected to support CLIP-seq. Based on the assumption that RNA-seq gives a good proxy of expression for individual transcripts, the latter is used as a background. The genome is split in functionally meaningful intervals (e.g UTRs, introns) and for each nucleotide position the number of read centers is calculated. A negative binomial (NB) distribution is calculated for each interval and if the P-value is less than 0.01, then the position is added to the estimated peak. The parameters p and r for the NB distribution are calculated by taking 3-nts-long windows at the regions of interest and maximizing the expected values of the distribution.

CLIPper [71] is the method also used within the ENCODE consortium [71, 72], based on which the analysis of all available eCLIP data is also provided. The peaks are interpolated with cubic splines. A Poisson distribution is used from the coverage across the pre-mRNA based on which an expected coverage is calculated.

PIPE-CLIP [73] has incorporated in its model substitutions as well as deletions and insertions, so that all types of CLIP experiments can be analyzed. The number of mutations/truncations is calculated with a binomial distribution, where the size is the total number of reads and the success rate, which is the sum of all mapped reads divided by the genome size. P-values are given for each position's mutation rate, which through the Benjamini-Hochberg method are transformed to FDR values. Fisher's method is used to cluster individual crosslinking positions. The motif-search tool MEME [74] is also incorporated in the analysis.

wavClusteR [75] is used for analysis of PAR-CLIP data and relies on a Bayesian network representation of a mixture model that has a chain of three variables, namely the source of a transition (UV-induced or not), the substitution frequency and the observed transitions. An algorithm termed mini-rank norm (MRN) is used to identify clusters of events by fitting the peaks to a rectangle and choosing the clustering that fits best. Local backgrounds are also considered, by sampling positions close to high-confidence transitions and modelling a distribution of fluctuations using a mixture of two Gaussian components of unequal variance, to cover both noisy fluctuations and abrupt jumps.

pyCRAC [76] uses an FDR calculation based on a negative sample (background) or a background constructed from random sampling within genes. What sets this method apart is that it was one of the first attempts at handling complete analysis from pre-processing to motif prediction based on CLIP data.

BMix [70, 77] was designed to specifically analyze PAR-CLIP data. For each position there is a probability for a T-C substitution as well as a G-C or A-C. These can be either due to sequencing error, a single-nucleotide polymorphisms (SNP) or due to the crosslink. Thus, a mixture of binomials is calculated for each of the observations based on these three scenarios. If no background samples are available, the entire data set is used to estimate the A-C and G-C substitutions, as these are not expected to originate from the crosslinking events.

CLAM [78] was designed to take into account the multi-mappers that most models were excluding from the RIP and CLIP experiments. A graph of distinct regions connected by the multi mappers is created and converted to a matrix, where column is the region and row is the read. An EM framework is used to assign each multi mapper to the optimal position, based on the observations and coverage around these possible positions. Peak calling is performed separately for each gene. A background is constructed by randomly assigning each of the observed reads in a gene across this gene 1000 times. In each gene, multiple testing corrected by the Benjamini–Hochberg False Discovery Rate (FDR) and FDR < 0.001 means a locus is significant. This method was shown to also work with m6A RIP-seq data.

CTK [79] is a second generation version of a method developed in the same group [79, 80]. Reads are clustered together into peaks which are ranked by their height and also evaluated by their presence in independent replicates, if available. Crosslinks are determined as transcriptase errors in these positions, an observation based on findings from CLIP samples obtained for the Nova protein [79–81]. Apart from deletions that were observed in the Nova samples, substitutions were also observed at crosslink sites. FDR values on mutation sites are calculated based on real and randomisations.

PureCLIP [82] relies on a Hidden Markov Model (HMM), where each position on the genome is characterized as enriched or not in the CLIP sample and crosslinked or not. The combination of these criteria lead to four hidden states. A Gaussian kernel density estimate helps with estimating an appropriate size for a region when testing if it is enriched or not. The read starts are assumed to be more frequently observed at the crosslink sites. Another added advantage of this method is the possible incorporation of biases (high-expression, overrepresented motifs) via incorporation of position specific external data to the HMM framework.

YODEL [83] was developed with the aim to enable modifying the parameters of the peak specification. There is no specific statistical model governing the peak finding step, the model rather relies on finding maximum values of coverage and looking for consistent decrease to accept clusters of positions as peaks. Thresholds for various variables can be manually set. Despite the usefulness of the tool at the time it was created, there are now other solutions that can automatically fit the observations to specific statistical models.

CLIPick [84] focuses on the inhomogeneity of the background noise due to different expression levels of transcripts interacting with RBPs. Reads are clustered together into peaks, and the coverage is then interpolated using cubic splines. *In silico* randomisation CLIP is used to evaluate the probability that a peak of a particular height could occur by chance, based on a background sample. This randomisation is thought to provide a more accurate peak detection. However, building an expectation from a background sample has already been implemented a number of times as observed in this section where all methods are outlined, but so far there is no agreement of what constitutes the most appropriate background.

omniCLIP [85] relies on a Non-Homogeneous Hidden Markov Model (NHMM) for peak detection, including four states, one for true peaks, another two for either same or higher background signal and one where there is signal neither in the foreground nor in the background. The positions that have the peak state as the most likely are merged into peaks. The scoring of the peaks is the log-likelihood ratio of the peak state against the rest of the states. Coverage derived mean and variance is modeled with a negative binomial based generalized linear approach, based on which a p-value is calculated, while additional diagnostic events (T-C mutations) are modeled using a Dirichlet-multinomial mixture model. As it includes both coverage information as well as diagnostic events, this method is expected to work with most of the CLIP variants.

1.2.7 *Software for motif analysis*

Most of the previous studies relied on additional software to identify enriched motifs in the peak sequences. Most broadly used is MEME [74], an expectation-maximization based method for identifying recurrent sequence motifs in a data set. More recently however, additional approaches to identify motifs in CLIP peaks have started to emerge.

One of these approaches is mCross [86], which starts from the premise that algorithms like MEME were developed with transcription factors in mind, proteins that bind more rigid DNA structures and longer sequence elements than RBPs, and thus these algorithms do not work as efficiently for RBPs that tend to bind lower complexity motifs with lower information content [27, 87]. In mCross positional weight matrices (PWMs) are inferred, similar to other methods, but they additionally consider crosslink-induced mutation sites (CIMS) and crosslink-induced truncation sites (CITS) to direct the motif search to the crosslink positions. A likelihood function is used to jointly model the sequence specificity and the crosslink information. As a way to verify the motifs, heterozygous sites are used, assuming that the proportion of reads mapping to a specific allele is proportional to the score of the site, different between alleles. Based on the alleles that give consistent and inconsistent information and using appropriate statistics, the motifs can be further validated without the need for additional experiments. While for some RBPs this approach gives good results, the issue is that there are not many allelic variations that coincide with the binding sites and also, there are cases of RBPs where the number of alleles yielding consistent and inconsistent information are equal or too small, rendering the generalization of such an approach uncertain.

Another recent approach to infer specific motifs relies on kmer enrichment around crosslink sites [68]. The main difference from the previous approach is that in this model, only the truncations are considered meaningful for specifying the crosslink site, while CIMS [86] are considered uninformative in terms of motif discovery. In this

method normalization of the kmer signal takes place, by considering regions close to peaks but around low-scoring crosslink sites, with the aim of eliminating biases due to UV crosslinking biases or sequence biases of cDNA ligation.

In the context of RCRUNCH, the method presented in this thesis, the PhyloGibbs tool was incorporated for the motif prediction step. PhyloGibbs searches for the optimal assignment of an arbitrary number of binding sites, for an arbitrary number of binding factors (in this case RBPs). In Phylogibbs, phylogenetic information and multiple alignments of regions can be provided to inform the motif identification. However, this step was omitted in RCRUNCH, as RBPs bind to sites across transcript regions with different levels of conservation, which complicates the treatment of the phylogenetic signal.

1.3 Reproducibility in the context of high-throughput analysis

A large problem observed during the work on this thesis, regarding methods for CLIP but also other high-throughput data analyses, is their reproducibility and accuracy. More often than not, these methods start from pre-processed data to construct expected distributions, but this preprocessing is performed in an unclear manner, or with unspecified versions of available tools and in many cases filtering of reads in an arbitrary manner takes place. This very often makes it exceptionally hard to reproduce the analysis and the results, as there are various points of inconsistencies or vagueness. This problem complicated, of course, the benchmarking of these methods during the development of RCRUNCH. Traditional pipelines relying on custom scripts or Make files can reduce the number of manual steps required. Some drawbacks of this approach remain the parameter tracking, the tool versioning, resuming a failed run and the over-reliance on specific infrastructure setups. Maintenance and reproducibility of computational analysis tools therefore remain troublesome [88]. Development of a plethora of workflow managers has led to strides in the pipeline development, as it offers tracing of data provenance, portability, scalability, and re-entrancy [88]. One such workflow manager, used in the context of this thesis, is Snakemake [89]. Snakemake is a bioinformatician-targeted domain-specific language (DSL) [90], with a language similar to Python and a 'backward' logic where starting form the final output files the steps for creating them are reconstructed. A big advantage of this workflow manager is the modularity of the steps, which allows easy maintenance and expansion, thus promoting FAIR (findable, accessible, interoperable, reusable) computational analysis [91, 92]. Containerization software is also a very important part of these workflow managers as they allow direct tool usage, without manual installation step or worry about specific requirements. This approach has led to open-source projects with the aim to support multiple bioinformatics software and contribute further to the FAIR principles [93].

2

ZARP: AN AUTOMATED WORKFLOW FOR PROCESSING OF RNA-SEQ DATA

2.1 ABSTRACT

RNA sequencing (RNA-seq) is a crucial technique for many scientific studies and multiple models, and software packages have been developed for the processing and analysis of such data. Given the plethora of available tools, choosing the most appropriate ones is a time-consuming process that requires an in-depth understanding of the data, as well as of the principles and parameters of each tool. In addition, packages designed for individual tasks are developed in different programming languages and have dependencies of various degrees of complexity, which renders their installation and execution challenging for users with limited computational expertise. The use of workflow languages and execution engines with support for virtualization and encapsulation options such as containers and Conda environments facilitates these tasks considerably. Computational workflows defined in those languages can be reliably shared with the scientific community, enhancing reusability, while improving reproducibility of results by making individual analysis steps more transparent.

Here we present ZARP, a general purpose RNA-seq analysis workflow which builds on state-of-the-art software in the field to facilitate the analysis of RNA-seq data sets. ZARP is developed in the Snakemake workflow language using best software development practices. It can run locally or in a cluster environment, generating extensive reports not only of the data but also of the options utilized. It is built using modern technologies with the ultimate goal to reduce the hands-on time for bioinformaticians and non-expert users. ZARP is available under a permissive Open Source license and open to contributions by the scientific community.

2.2 INTRODUCTION

Recent years have seen an exponential growth in bioinformatics tools [94], a large proportion of which are dedicated to High Throughput Sequencing (HTS) data analysis. For example, for transcript-level analyses there are tools to quantify the expression level of transcripts and genes from RNA-seq data [95], identify RNA-binding protein (RBP) binding sites from crosslinking and immunoprecipitation (CLIP) data [96, 97], improve transcript annotation with the help of RNA 3'end-sequencing data [98, 99], estimate gene expression at the single cell level [100] or improve the annotation of transcripts and quantification of splicing events based on long read sequencing (e.g., on the

Oxford Nanopore platform) [101, 102]. Such tools are written in different programming languages (e.g., Python, R, C, Rust) and have distinct library requirements and dependencies. In most cases, the tools expect the input to be in one of the widely accepted file formats (e.g., FASTQ [103], BAM [104]), but custom formats are also frequently used. In addition, the variations in protocols or instruments across experiments may make it necessary to use different parameterization for every sample, rendering a joint analysis of samples from multiple studies challenging. Combining tools into an analysis protocol is a time-consuming and error-prone process. As these tasks have become so common, and as the data sets and analyses continue to increase in size and complexity, there is an urgent need for expertly curated, well-tested, maintained and easy-to-use reusable computational workflows.

A number of feature-rich, modern workflow specification languages and corresponding management systems [105, 106] like Snakemake [107, 108], Nextflow [109] and CWL [110] are now gaining widespread popularity in life sciences, as they make it easier for such workflows to be developed, tested, shared and executed. This leads to more reusable code and reproducible results, while fostering scientific collaborations and Open Source Software along the way. In addition, to facilitate the installation and execution of these workflows across different hardware architectures and host operating systems, modern workflow management systems make use of virtualization and encapsulation techniques relying on containers (e.g., Docker [111] and Singularity [112]) and/or package managers (e.g., Conda [113] and Bioconda [114]). An added advantage of using workflows is the metadata stored along with the expected results. These can be invaluable for re-analyzing the data but may also provide additional insights into the results and cost analyses (e.g., runtimes, resources usage).s

The aim of the presented work is the development of a flexible, easy-to use workflow for bulk RNA-seq data processing. The inclusion of the most widely used and best performing tools for the various processing steps minimizes time spent by users on making tool choices. Use of a workflow language for the development ensures the reproducibility and reliable execution of each analysis and it facilitates (meta)data management and reporting.

2.3 METHODS/RESULTS

ZARP (Zavolan-Lab Automated RNA-seq Pipeline) is a general purpose RNA-seq analysis workflow that allows users to carry out the most general steps in the analysis of Illumina short-read sequencing libraries with minimum effort. The workflow is developed in Snakemake [107, 108], a widely used workflow language [105]. It relies on publicly available bioinformatics tools that follow best practices in the field [115], and handles bulk, stranded RNA-seq data, single or paired-end.

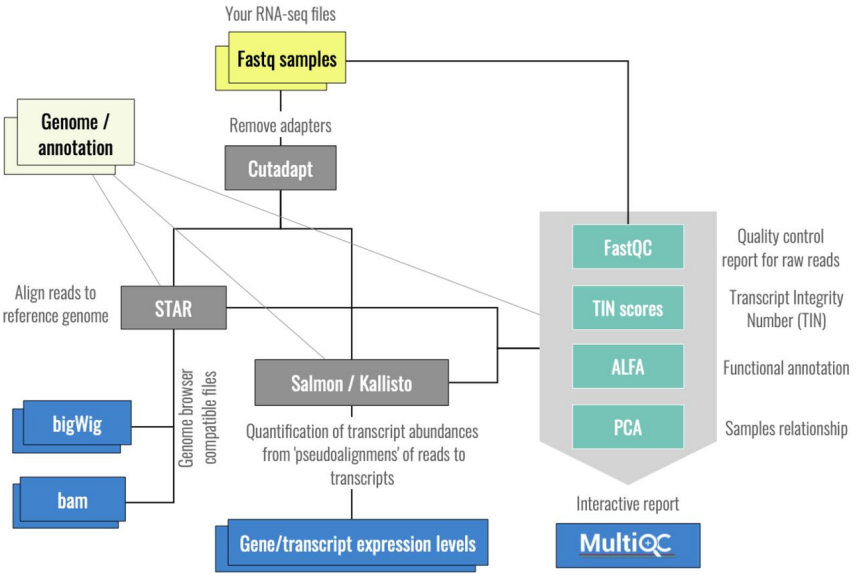


Figure 2.1: Schematic overview of the ZARP workflow.

2.3.1 *Workflow inputs*

ZARP requires two distinct input files: (1) A tab-delimited file with sample-specific information, such as paths to the sequencing data (FASTQ format), reference genome sequence (FASTA format), transcriptome annotation (GTF format) and additional experiment protocol- and library-preparation specifications like adapter sequences or fragment size. (2) A configuration file in YAML format containing workflow-related parameters, such as results and log directory paths and user-related information. Advanced users can take advantage of ZARP's flexible design to provide tool-specific configuration parameters via an optional third input file, which allows adjusting the behaviour of the workflow to their specific needs. More information on the input files can be found in ZARP's documentation [116] .

2.3.2 *Analysis steps*

A general schema of the workflow in its current version (0.3.0) is presented in Figure 2.1 (see Supplementary Figure A.1 for a more technical representation of the entire workflow, including all of its steps). Table 2.1 below lists the main tools/functionalities of ZARP:

Calculation of per-sample quality statistics by applying FastQC [117] directly on the input files (FASTQ) provides a quick assessment of the overall quality of the samples. These consist of a considerable range of metrics, including, for example, GC content, overrepresented sequences and adapter content. An excessive bias in GC content may affect downstream analyses and may have to be corrected for [126]. Overrepresented sequences may be the result of PCR duplication, which, if excessive, may skew expression estimates and other downstream analyses. Information about adapter content may be used to cross-check whether it matches with whatever the user has selected

Tool	Description	Reference
FastQC	Generates various quality control metrics based on raw FASTQ data.	[117]
Cutadapt	Trims sequence fragments of non-biological origin or low information content.	[118]
STAR	Aligns reads to reference genome.	[119]
tin-score-calculation	Calculates a Transcript INtegrity score (TIN) on aligned reads that reflects the state of RNA degradation of a sample.	[120]
ALFA	Annotates read alignments based on gene/transcript annotations.	[121]
kallisto	Estimates gene/transcript expression levels.	[122]
Salmon	Estimates gene/transcript expression levels.	[123]
zpca	Performs principal component analyses of gene/transcript expression level estimates across samples in a given workflow run.	[124]
MultiQC	Aggregates tool results and generates interactive reports.	[125]

Table 2.1: Core tools/functionalities included in ZARP. See main text for more information on use cases for each tool and why we chose those tools to be included in ZARP.

to trim. For more information on the metrics that FastQC reports and how they can be interpreted, please refer to [117].

Trimming of any 5′ and/or 3′ adapters as well as poly(A/T) stretches using Cutadapt [118] ensures a more reliable alignment as well as removal of contaminant adapter sequences. The adapters and poly(A/T) stretches to be removed are indicated by the user.

Alignment against a given set of genome resources (either only the genome or the genome and a set of corresponding gene annotations) is the step where each read is assigned to the genomic region from which it originated. Even though there are many available aligners, STAR has been chosen [119][127]. BAM-formatted files that are then sorted (based on coordinates) and indexed using SAMtools [104]. These steps enable faster random access and visualisation by tools such as genome browsers. The sorted, indexed BAM files are further converted into the BigWig (BedGraphtoBigWig from UCSC tools [128]) format, which allows for library normalisation, and is thus convenient for visualising or comparing coverages across multiple samples.

The aligned reads are also used to calculate per-transcript Transcript Integrity Numbers (TIN scores) [120], a metric to assess the degree of RNA degradation in the sample. This is done with tin-score-calculation [129], which is based on a script originally included in the RSeQC package [130] but modified by us to enable multiprocessing for increased performance.

To provide a high-level topographical/functional annotation of which gene segments (e.g., CDS, 3'UTR, intergenic) and biotypes (e.g., protein coding genes, rRNA) are represented by the reads in a given sample, ZARP includes ALFA [121].

Salmon [123] and kallisto [122] along with a transcriptome are used to infer transcript and gene expression estimates. Since both of these tools have been shown to be equally fast, memory efficient and accurate [131], they are both included in ZARP. The main output metrics provided by either tool are estimates of normalized gene/transcript

expression, in Transcripts Per Million (TPM) [132], as well as raw read counts per gene/transcript.

Within ZARP, TPM estimates are essential for performing principal component analyses (PCA) [133] with the help of zpca [124], a tool created by us for the use in ZARP, but packaged separately so that it can be easily used on its own or as part of other workflows. PCAs on gene/transcript expression levels can help users understand whether differences in gene/transcript expression levels across different sample groups are sufficiently high that meaningful results in downstream analyses may be expected.

TPM and raw count estimates can be further used in downstream analyses, e.g., for differential gene/transcript expression, differential transcript usage or gene set enrichment analyses. Given that such analyses require an experiment design table and are difficult to configure generically for a wide range of experiments, we chose not to include these in ZARP. However, to facilitate downstream analyses, gene/transcript estimates are aggregated for all samples with the aid of Salmon and merge_kallisto [134], which generate summary tables that can be plugged into a variety of available tools.

ZARP produces two user-friendly, web-based, interactive reports: one with a summary of sample-related information generated by MultiQC [125], the other with estimates of utilized computational resources generated by Snakemake itself. Note that both for tin-score-calculation and ALFA, we have created plugins so that the respective results can be explored interactively through MultiQC.

2.3.3 *Reproducibility and reusability*

To enhance reproducibility of results and reusability of the workflow, each step (referred to as "rule" in Snakemake) of the workflow definition relies either on Conda environments mostly hosted in the Bioconda channel [114] or on Docker images. The latter are converted by Snakemake to Singularity images [112] on the fly where needed, enabling seamless execution of the workflow in environments with limited privileges (e.g., HPC clusters). Users can choose between Conda- and container-based execution by selecting or preparing an appropriate profile when/before running a workflow. At the moment, we include profiles for the Slurm job scheduler and we plan to add new profiles over time. For that, we encourage users to feed their own profiles back to the original ZARP repository so that the entire community can benefit.

2.3.4 *Output and documentation*

In addition to the transcript/gene expression tables, ZARP collects log files and metadata for downstream analyses. Intermediate files can be optionally cleaned up by ZARP to minimize disk space usage. The workflow is hosted in its own GitHub repository, and each ZARP

version released is accompanied by an up-to-date workflow-oriented description. Continuous Integration and Testing

To facilitate collaborative development of the workflow and associated software and to reduce the chance of the codebase regressing with ongoing changes, ZARP is making use of a GitHub Actions-based workflow for Continuous Integration and Delivery (CI/CD). Each modification to the remote repository triggers a variety of integration tests (Conda environments test, Snakemake graph test, dry run, minimal-example based run) to guarantee ZARP's correct execution throughout the development cycle as the source code is refactored and new features are added.

2.4 USE CASES

Apart from quickly gaining insights into individual samples or smaller sets of samples, ZARP is very well suited to analyze large RNA-Seq experiments or even run meta-analyses across multiple different experiments.

To demonstrate how ZARP can be used to gain meaningful insights into typical RNA-seq experiments, we tested it on an RNA-seq dataset that was generated by Ham et al. (GEO [135] accession number GSE139213) while analyzing the role of mTORC1 signalling in the age-related loss of muscle mass and function in mice [136]. The dataset consists of 20 single-ended RNA-seq libraries (read length: 101 nt, gzipped FASTQ file sizes ranging from 0.8 to 3.2 Gb, library sizes ranging from 18.5 to 75.3 reads), corresponding to four cohorts of 3-months old mice (with five biological replicates per cohort): (1) wild-type, (2) rapamycin-treated, (3) tuberous sclerosis complex 1 (TSC1) knockout and (4) rapamycin-treated TSC1 knockout. The samples were mapped against ENSEMBL's [137] GRCm38 genome primary assembly and corresponding gene annotations (release: 99) for standard human chromosomes. Other parameters for populating ZARP's samples table were obtained from the GEO accession entries of the respective samples. Sample tables and results for the test run are publicly available [138].

In Figure 2.2 , we are presenting a subset of the outputs that ZARP generated for this dataset. We can see that the GC content of reads (Figure 2.2A) is slightly skewed towards being more AU-rich, yet all samples pass the FastQC-defined threshold for GC bias. Moreover, GC content does not exhibit a strong bias across samples. There is no evidence of extensive sequencing of residual adapters ("adapter contamination") (Figure 2.2B; black), as less than 1% of reads have been discarded in each sample because of insufficient length after adapter trimming. Transcript integrity across samples is also uniform and high (Figure 2.2C), with the highest density of expressed transcripts at TIN scores of 75 to 85. Similarly, alignment statistics as reported by STAR are also consistently high (Figure 2.2D), with rates of reads mapped uniquely against the mouse genome of more than 72% across all samples (<4% unmapped), irrespective of sequencing

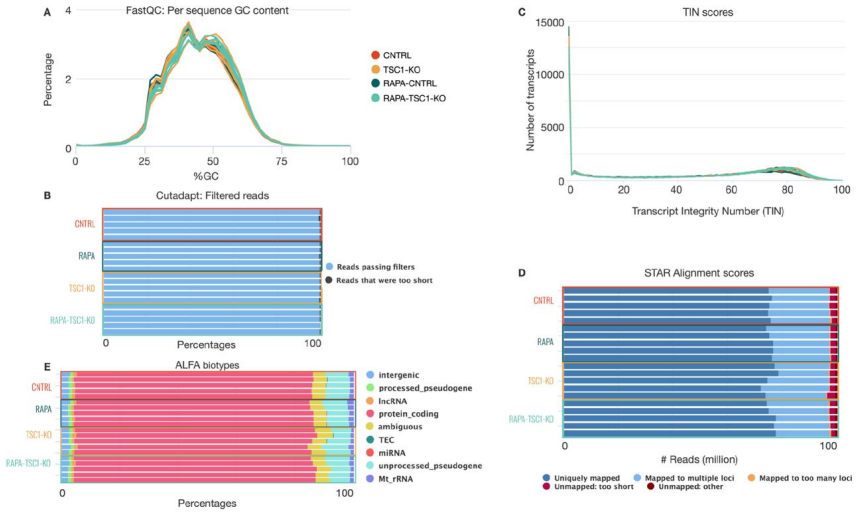


Figure 2.2: Selection of metrics reported by ZARP. Shown are (A) GC content, (B) adapter removal report, (C) Transcript Integrity Number (TIN) score, (D) STAR alignment statistics, and (E) ALFA biotypes for the test run described in the main text. Figures have been edited for visibility purposes in order to group samples according to cohorts. Additionally, some biotypes have been omitted from (E) as they are not meaningfully represented. Note that in (C), transcripts that are not expressed are assigned a TIN score of 0. The complete raw html report can be found at [138].

depth. As expected, ALFA analysis of transcript categories shows that uniquely mapped reads overwhelmingly originated from protein coding genes (over 86% for all samples) (Figure 2.2E). Taken together, these metrics indicate that all samples are of sufficiently high quality for downstream analyses.

In addition to sample-specific metrics, ZARP also provides tooling to compute principal component analyses across samples (Figure 2.3). For the test run, the distribution of samples in the space of the first two principal components shows a clustering by condition, with a clear separation between knockout and wild type, as well as between the untreated and rapamycin-treated TSC1 knockout mice. This separation is more pronounced at the gene expression level (Figure 2.3A), but is also present at the transcript level (Figure 2.3B). This shows that the differences across conditions are more pronounced than any replicate biases (multiplicative noise, sequencing errors), i.e., the signal-to-noise ratio is favorable, which strongly increases the likelihood that any subsequent analyses (e.g., differential gene/transcript expression analysis) will provide targets of biological importance.

The total wall clock time to execute the entire test run was just over one hour (1.01h) for all 20 samples on our Slurm-managed HPC cluster [139], where we could make heavy use of ZARP's parallelization capabilities. This translates to a total CPU time of 68.79 h, out of which 6.68h were run-specific, i.e., jobs that had to be executed only once for all samples. The accumulated sample-specific CPU time used for each sample varied between 2.75h and 8.44h. While the actual runtime may differ considerably across different compute environments, we project that most users would be able to run even large-

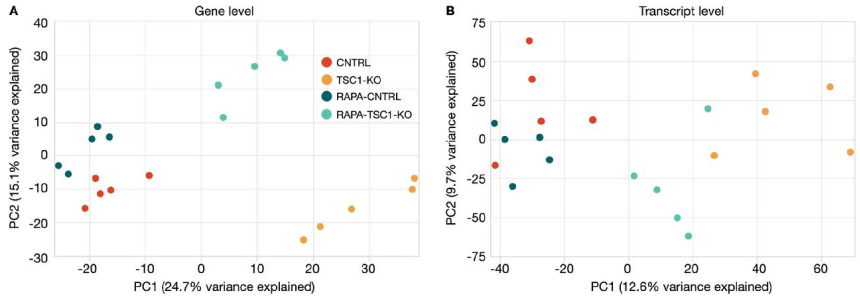


Figure 2.3: Principal component analysis. Principal component analysis (PCA) at the (A) gene and (B) transcript level. PC1 and PC2 correspond to the first and second principal components, respectively. Variances explained by each of them are stated in the parentheses of the corresponding axes labels. Expression levels used in this figure are those reported by kallisto, but ZARP also generates corresponding PCA plots for Salmon-based quantifications.

scale analyses with dozens to hundreds of samples in less than a day on an HPC cluster, with very little hands-on time. Maximum memory usage for any of the steps and across all samples was <32 Gb (for STAR indexing and mapping of/against the human genome), indicating that ZARP is suitable for execution on state of the art computers, albeit at considerably higher runtimes due to limited parallelization capabilities, particularly for large sample groups. None of the jobs took longer than 20 min (wall clock time) for any of the samples (Figure 2.4). Among the most time-consuming steps are the creation of indices (STAR, Salmon, kallisto), which however have to be performed only once per set of genome resources. Among the sample-specific steps, the calculation of the Transcript INtegrity (TIN) score was the most time-consuming. However, we had already considerably reduced its runtime by adding parallelization capabilities to the original script (see subsection "Analysis steps" in section "Methods/Results" for details).

In summary, our test case demonstrates how ZARP can be used to quickly gain informative insights (Figures 2.2 & 2.3) into a non-trivial real-world RNA-seq analysis in a reasonable timeframe (Figure 2.4).

2.5 DISCUSSION/CONCLUSIONS

ZARP is a general purpose, easy-to-use, reliable and efficient RNA-seq processing workflow that can be used by molecular biologists with minimal programming experience. Scientists with access to a UNIX-based computer (ideally a Linux machine with enough memory to align sequencing reads) or a computing cluster can run the workflow to get an initial view of their data on a relatively short time scale. ZARP has been specifically fine-tuned to process bulk RNA-seq datasets, allowing users to run it out of the box with default parameters. At the same time, ZARP allows advanced users to customize workflow behavior, thereby making it a helpful and flexible tool for edge cases, where a more generic analysis with default settings is unsuitable. The outputs that ZARP provides can serve as

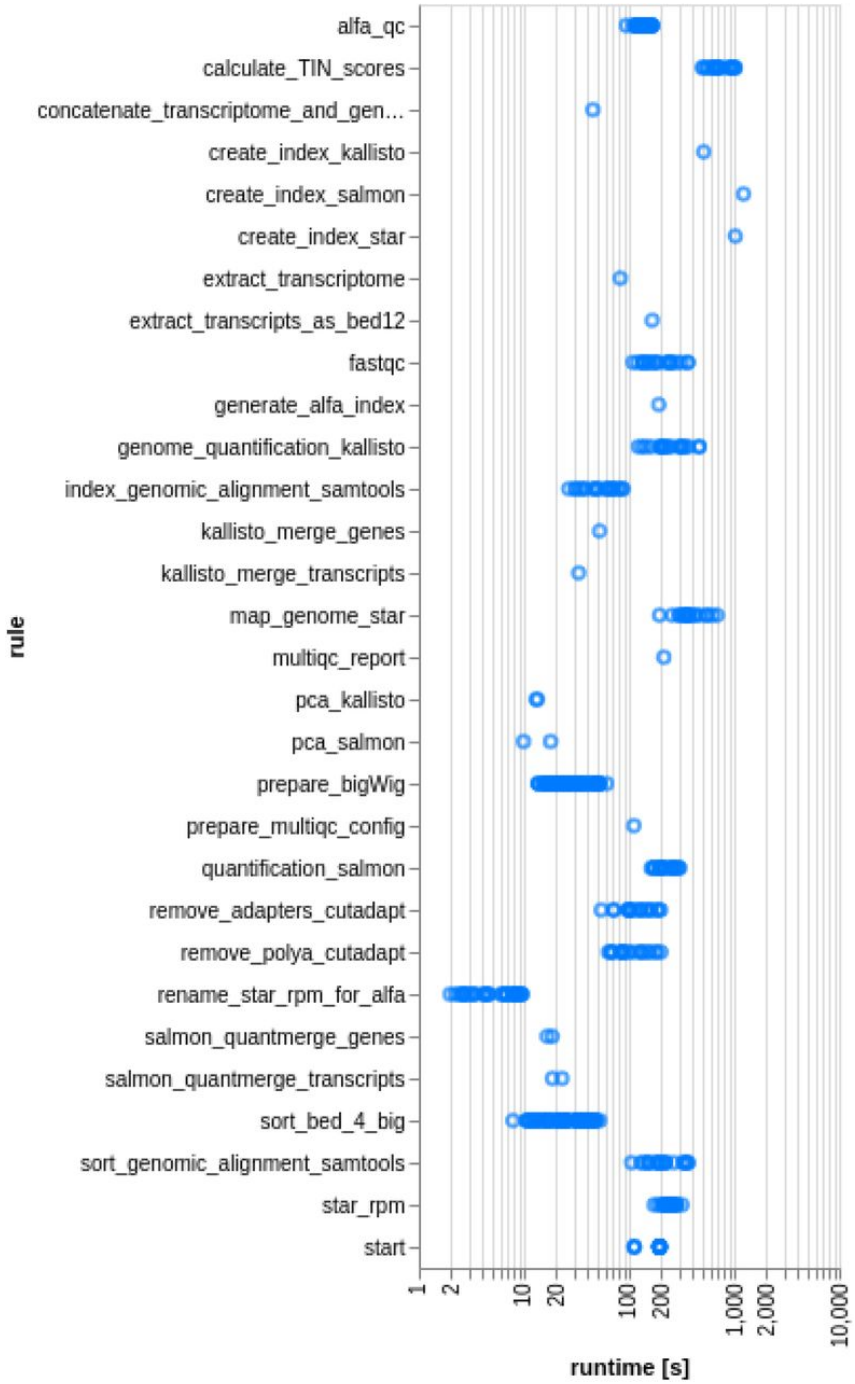


Figure 2.4: Runtime statistics. Runtime (in seconds; wall clock time) of the different steps ("rules") of the workflow run are depicted for each sample. The workflow was executed in an HPC cluster managed by the Slurm job scheduler, so the reported runtimes include the time that jobs spent queuing. Additional variation in runtimes may result from individual jobs being executed on cluster nodes with different specifications.

entry points for other project-specific analyses, such as differential gene and transcript expression analyses. ZARP is publicly available and open source (Apache License, Version 2.0), and contributions from the bioinformatics community are welcome. Please address all development-related inquiries as issues at the official GitHub repository [140].

2.6 DATA AND SOFTWARE AVAILABILITY

2.6.1 *Data*

Raw data analysed in section "Use Cases" are publicly available for anyone to download from the NCBI:GEO server, accession number GSE139213.

2.6.2 *Software*

The ZARP code is available on GitHub at [140] and is published under Apache License, Version 2.0. A snapshot of the ZARP version described in this manuscript (0.3.0) has been additionally uploaded to Zenodo for long-term storage [116]. Both services are public and allow anyone to download the software without prior registration.

2.6.3 *Results*

Analysis results presented in section "Use Cases" are publicly available for anyone to download from Zenodo.

2.6.4 *Acknowledgements*

This work would not have been possible without the sciCORE Team [139] at the University of Basel. We are thankful for their support regarding the computational infrastructure as well as dedicated time and effort to aid us in this project. We would also like to express our deepest gratitude towards all members of the Zavolan Lab who contributed to this work with numerous pieces of advice, during the initial development as well as by testing the workflow in later stages. We would also like to thank the bioconda community for helping us package and distribute some of the custom tools we developed.

3

IMPROVED ANALYSIS OF (E)CLIP DATA WITH RCRUNCH YIELDS A COMPENDIUM OF RNA-BINDING PROTEIN BINDING SITES AND MOTIFS

3.1 ABSTRACT

Crosslinking and immunoprecipitation (CLIP) is used to determine the transcriptome-wide binding sites of RNA-binding proteins (RBPs). Here we present RCRUNCH, an end-to-end solution to CLIP data analysis that enables the reproducible identification of binding sites as well as the inference of RBP sequence specificity. RCRUNCH can analyze not only reads that map uniquely to the genome, but also those that map to multiple genome locations or across splice boundaries. Furthermore, RCRUNCH can consider various types of background in the estimation of read enrichment. By applying RCRUNCH to the eCLIP data from the ENCODE project, we have constructed a comprehensive and homogeneous resource of in vivo-bound RBP sequence motifs. RCRUNCH automates the reproducible analysis of CLIP data, enabling studies of post-transcriptional control of gene expression. RCRUNCH is available at:
https://github.com/zavolanlab/RCRUNCH.

3.2 BACKGROUND

Throughout their life cycle, from transcription to maturation, function and decay, RNAs associate with RNA-binding proteins (RBPs) to form ribonucleoprotein complexes (RNPs) or higher-order RNA granules (e.g paraspeckles, Cajal bodies) [141]. RBPs are abundant in prokaryotes as well as eukaryotes, and methods such as RNA interactome capture (RIC) [142] revealed that over a thousand human and mouse proteins have RNA-binding activity. An RBP is typically composed of multiple RNA-binding domains (RBDs) coming from a limited repertoire [143], and binds to a specific sequence motif and/or secondary structure element. The functional diversity of RBPs rests on the number and arrangement of RBDs that they contain [144], though methods like RIC have uncovered proteins that have RNA-binding activity, despite lacking a known RBD.

As RBPs participate in all steps of RNA metabolism, it is not surprising that they have been implicated in many diseases [145]. However, the critical targets in a particular context are often unknown. The method of choice for mapping the binding sites of an RBP in vivo and transcriptome-wide is crosslinking and immunoprecipitation (CLIP). Introduced in the early 2000s [146], CLIP has a number of variants, all exploiting the photoreactivity of nucleic acids and proteins. Briefly,

ultraviolet light is used to crosslink RBPs to RNAs, the regions of the RNAs that are not protected by RBPs are enzymatically digested, the RBP of interest is purified along with the crosslinked RNAs, and finally the purified RNA [146]fragments are reverse-transcribed and sequenced. One of the main differences between CLIP variants is in the nature of the cDNAs that end up being sequenced. These can either be the result of aborted reverse transcription at the crosslinked site [147] - where a bulky adduct remains after protein digestion -, or the result of reverse transcription through the site of crosslink, which often results in characteristic mutations in the cDNAs [53, 148]. Although the general expectation is that extensive purification leads to a relatively pure population of target sites for a given protein, inspection of the genome coverage by sequenced reads indicates substantial non-specific background. Various approaches have been proposed for background correction, but a systematic benchmarking of these approaches is still lacking (discussed in [149]).

CLIP is analogous to chromatin immunoprecipitation (ChIP), a technique that has been used for many years to determine binding sites of DNA-binding factors. To distinguish protein-specific interactions from various types of background, ChIP includes control samples consisting either of the chromatin input or the material resulting from non-specific binding of antibodies to chromatin. Many computational methods have been developed to identify 'peaks' from such data sets [150]. A previous study of peak finding methods developed for ChIP data has underscored the importance of the model describing the obtained data [151].

In contrast to ChIP, background samples are not always generated in CLIP experiments, as it is less clear what an appropriate background should be. While at the DNA level, genes are generally represented in two copies per cell, the relative abundance of different RNAs in the cell varies over many orders of magnitude. Thus, abundant RNAs are likely to contaminate CLIP samples, leading to false positive sites, while binding sites in low-abundance RNAs may be completely missed. An approach to deal with this issue is to take advantage of crosslinking-induced mutations, identifying regions where such mutations have a higher than expected frequency [53, 86, 148, 152]. This is not unproblematic, because mutations are introduced stochastically and the mutation-containing reads could also come from fragments crosslinked to proteins other than the one of interest in the experiment [153]. Another approach is to correct for the abundance of the RNAs based on RNA-seq data. This is also not ideal, first because differences in sample preparation may lead to RNA-seq data not containing all the potential targets of the RBP, and second because the RNA-seq read coverage profile is not uniform, which will influence the quantification of the local background, and consequently the identification of CLIP sites. Finally, in the eCLIP variant of CLIP, the background coverage of transcripts by reads is inferred from a parallel sample that is prepared from the band corresponding roughly to the size of the protein of interest, obtained by omitting the immunoprecipitation step of sample preparation. This approach

has the caveat that the size of the targeted protein varies from experiment to experiment, and so will the proteins that are contained in the isolated band. This makes it unclear whether the results obtained for different RBPs are of comparable accuracy.

As mentioned above, most RBPs bind their targets in a sequence-dependent manner, and sometimes in the context of specific structure elements [141, 154]. For many RBPs, binding motifs have been inferred with both low and high-throughput approaches, and at least in some of these cases, there is good agreement between the RBP-binding motifs inferred from in vitro [155, 156] and in vivo data [86, 157]. However, in the most comprehensive database to date, ATtRACT [158], there typically are many motifs for an RBP, of widely variable information content and sometimes unrelated. A comprehensive database of RBP binding motifs determined from a consistent in vivo dataset, similar to those available for transcription factors [159, 160], is still lacking.

A number of methods have been proposed for the identification of RBP binding peaks from CLIP data. Benchmarking of various subsets of these methods has revealed a few good performers, such as clipper, the tool developed for the analysis of above-mentioned eCLIP data, omniCLIP and pureCLIP, two recently published tools that use complex models to take advantage not only of the CLIP read coverage, but also of crosslinking-induced mutations [152, 157, 161, 162]. However, none of these tools provides an easy and robust end-to-end solution to the identification of binding sites and sequence motifs from CLIP data, and it has remained unclear how their accuracies compare.

To fill these gaps, we have developed RCRUNCH, a method that further aims to treat appropriately not only reads that map uniquely and contiguously to the genome, but also reads that map across splice junctions in mature mRNAs, as well as multi-mapping reads. The de novo motif discovery component of RCRUNCH, based on the well-established Motevo tool [163], allows an immediate assessment of the quality of the results, including for the comparison of its genome/transcriptome or unique/multi-mapper based approaches. Using data for proteins with well-characterized sequence specificity, we demonstrate that RCRUNCH enables the reproducible extraction of binding sites, with higher enrichment in the expected motifs compared to the other tools. Application of RCRUNCH to the extensive eCLIP datasets generated in the ENCODE project [157], covering 149 RBPs, led to the construction of a comprehensive resource of in vivo binding sites and binding motifs of RBPs. RCRUNCH is available as an entirely automated tool from [164].

3.3 RESULTS

3.3.1 *Automated CLIP data analysis with RCRUNCH*

RCRUNCH is a workflow (Figure 3.1a) for the automated and reproducible analysis of CLIP data, from reads to binding sites and motifs.

It is written in the Snakemake language [165], observing the FAIR (findable, accessible, interoperable and reusable) principles [166]. The peak-calling module at the core of the workflow builds on the CRUNCH model [151] that was extensively validated on ChIP data. Along with the genome sequence and annotation files, the input to RCRUNCH consists of CLIP (foreground) sequencing reads, obtained from immunoprecipitation of a specific RBP, and background reads, which in most of the analyses reported here come from a size-matched control sample as in eCLIP. RCRUNCH's default analysis mode uses reads that map uniquely to the genome, but multi-mappers and/or reads that map across splice junctions of mature mRNAs can also be included. For the latter case, RCRUNCH constructs a representative transcriptome composed of the isoform of each gene that has the highest abundance in the foreground sample. In a first step of its peak finding module, RCRUNCH identifies broad genomic regions whose coverage by reads is significantly higher in the foreground compared to the background sample (Figure 3.1b, see Methods). A Gaussian mixture model is then applied to each of these regions to identify individual peaks and compute associated read enrichment scores (Figure 3.1c). Peaks sorted by the significance of their read enrichment are then used in various analyses, including for the identification of enriched sequence motifs. The workflow provides extensive outputs such as the coordinates of the peaks, their enrichment scores and associated significance measure, enrichment values of known and de novo identified sequence motifs represented as positional weight matrices (PWM).

3.3.2 *Comparative evaluation of CLIP peak finding methods*

By automating the analysis of CLIP data from reads to binding sites and motifs, RCRUNCH facilitates studies that rely on such data to a much larger extent than it was possible so far. To demonstrate its performance, we compared RCRUNCH with a few recently published and more broadly used tools for CLIP site identification. These were: clipper, the method used in the ENCODE project that generated the eCLIP data [157], already shown to supersede a few other methods [161], PURE-CLIP [162] and omniCLIP [152]. The latter two can use both the CLIP read density as well as the type and frequency of mutations in cDNAs to identify binding sites. As peak calling is not fully automated in clipper, for this tool we used the peaks provided by the ENCODE consortium [167, 168]. The other tools were provided with genome-mapped reads obtained with the pre-processing module of RCRUNCH. As done in previous studies [152, 162], in the comparative evaluation we used proteins for which the binding patterns and motifs have been extensively studied and are thus well understood [152, 157, 162]. These are hnRNPC, a splicing regulator that binds (U)5 sequences [169], PTBP1, another splicing regulator with a CU-rich binding motif [170], PUM2, a post-transcriptional regulator binding to UGUANAUA sequence elements [171, 172] and RBFOX2,

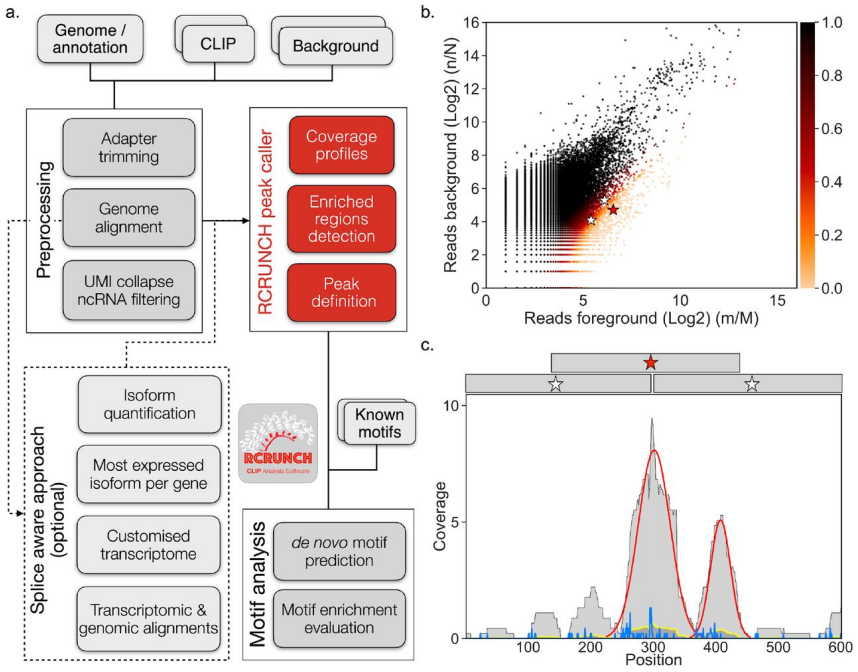


Figure 3.1: Schematic representation of RCRUNCH. a. Overview of the workflow. b. Scatterplot of the proportion of reads (log2) in individual genomic regions in a foreground (CLIP) sample generated for the PUM2 protein (replicate 2 of dataset ENCSR661ICQ from the ENCODE project, see Methods) and a corresponding background sample (size-matched input control, SMI). Each dot corresponds to a 300 nucleotides-long genomic region. Marked in color are regions that are enriched in reads in the foreground relative to the background sample (FDR < 0.1, see color legend). Three genomic regions (zoom-ins in panel c) with various enrichment scores are marked with stars. c. Coverage of three overlapping genomic regions (highlighted in panel b and indicated at the top of the panel) by CLIP reads (light gray) and by background reads (dark gray) from the SMI sample. The yellow line represents the envelope of the SMI read coverage profile. The significant peaks identified in the genomic region spanned by the three windows are shown in red. The blue line shows the number of CLIP read starts (5′ end of mate 2 reads) in the genomic region. Read starts indicate crosslinked nucleotides, where the reverse transcriptase falls off during sample preparation.

a splicing regulator known to recognize the sequence (U)GCAUG(U) [167]. For each of these proteins we applied RCRUNCH to the corresponding eCLIP samples in ENCODE, typically 2 replicates in each of two cell lines, and extracted the 1000 highest scoring sites from each sample.

The reproducibility of results obtained from replicate experiments is an important indicator of a method's accuracy [149]. To estimate the agreement between sets of peaks, obtained from either replicate experiments or by different methods applied to the same dataset, we used the Jaccard similarity index (see Methods). The agreement of the top peaks inferred from two replicate experiments for the same RBP in a given cell line (Figure 3.2a) was 20-40%, in the range reported for CLIP samples before [148]. RCRUNCH consistently provided values at the top of this range: for 5 of the 7 datasets RCRUNCH gave the highest agreement, and in the 2 cases when it did not, it was still a close second performer (Figure 3.2b). We also asked how large is the overlap between the peaks reported by different methods. Although this was generally lower than the overlap of peaks identified by one method from replicate experiments, RCRUNCH had the overall highest agreement with the other methods (Figure 3.2c). Finally, as the computational cost incurred by a tool is also an important factor in its adoption, we recorded the clock time for the peak calling step of all methods on the benchmarking data sets. For RCRUNCH the clock time was up to 3 hours (Figure 3.2d), while PURE-CLIP needed up to 6 hours and omniCLIP up to 9 hours for an individual dataset/RBP.

Thus, RCRUNCH outperforms currently used methods both in terms of peak reproducibility between replicate experiments and in terms of running time. Moreover, RCRUNCH has, on average, the highest agreement with other peak finding methods, indicating that it capitalizes on some of the same information, while diminishing some of the biases of these other methods.

To further evaluate the quality of the detected peaks, we determined their enrichment in the motif known to be bound by the RBP targeted in each CLIP experiment. We extracted the literature-supported motifs for each RBP from the ATtRACT database [158] and calculated their enrichment in the 1000 top peaks predicted by each method relative to random genomic regions, unlikely to be bound by the RBP (see Methods). As shown in Figure 3.3a, the known motifs were indeed enriched in the peaks relative to background sequences, up to 5-fold, as observed before [157, 173]. RCRUNCH peaks showed enrichment values at the high end of the achieved range (Figure 3.3a) for all proteins/samples except hnRNPC. To verify that the peaks were indeed most enriched in the motifs known to be bound by the RBP (as opposed to any other), we further applied the Phylogibbs algorithm [174], to discover de novo the motif that is most overrepresented in the top peaks. Some of the de novo motifs were indeed similar to the expected ones, but they tended to be less polarized and more enriched (Figure 3.3b-d). Strikingly, while Phylogibbs identified de novo motifs that were very strongly enriched in the hnRNPC peaks, these motifs did not have any resemblance to the expected

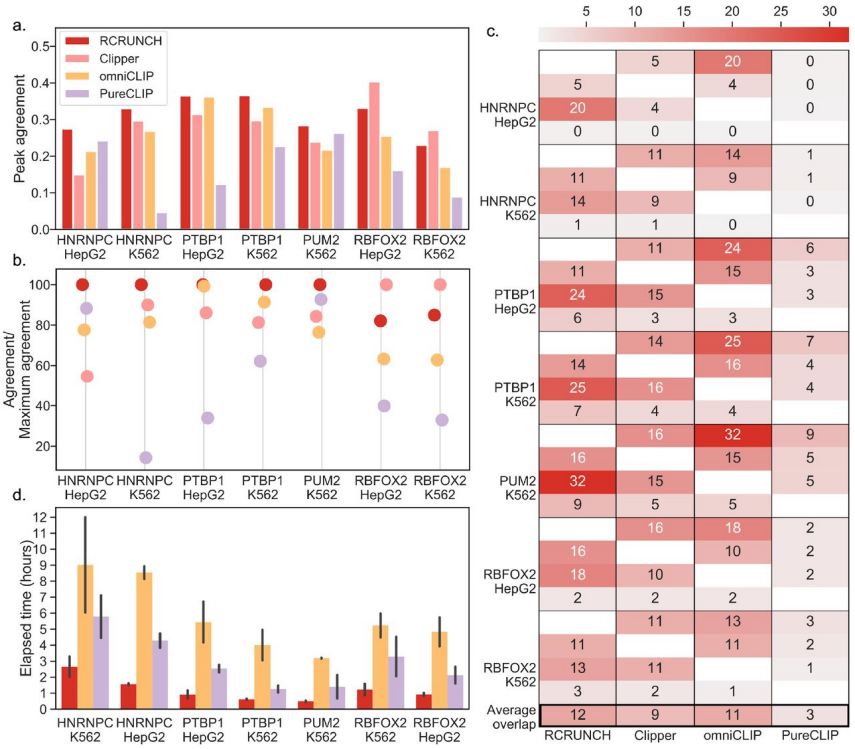


Figure 3.2: Comparison of CLIP peak calling methods. a. Barplot showing the Jaccard similarity index of the peaks identified by each computational method (shown in the legend) from replicate experiments. For all RBPs, except PUM2, data were available from two distinct cell lines, HepG2 and K562. b. Replicate agreement, calculated based on the data in panel a, as a percent of the maximum obtained by any method on each individual dataset. c. Heatmap showing the Jaccard similarity index of the peaks identified by two distinct methods for a given protein in a given cell line (in percentages, averaged over two replicate experiments). The methods are shown in order in the x-axis and the same order is used in each block (corresponding to one protein and cell line) on the y-axis. The average similarity of a method with any other method on all datasets is shown at the bottom. d. Barplot showing the running times of peak calling steps for RCRUNCH, omniCLIP, and PureCLIP. Error bars show the standard deviations from the 2 replicate runs.

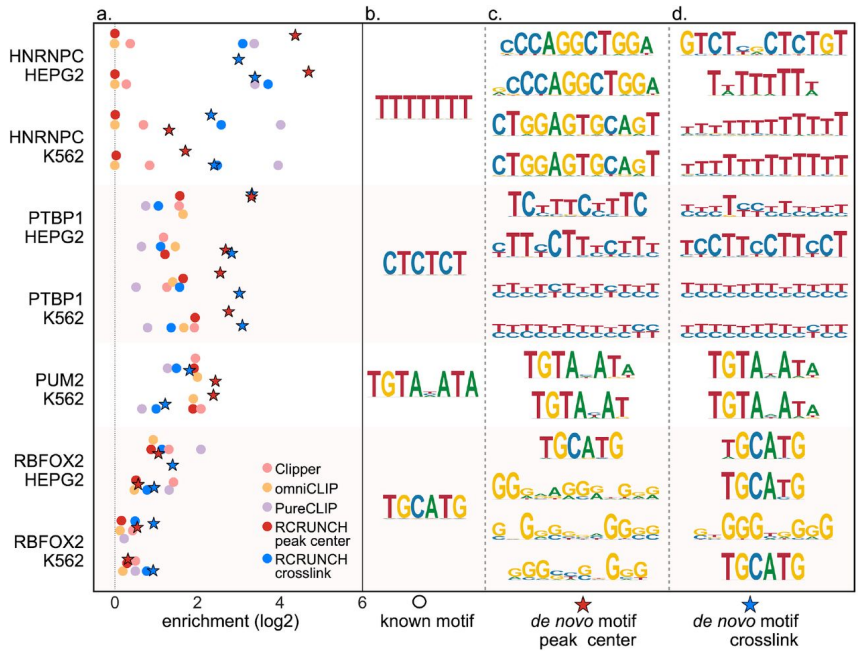


Figure 3.3: Enrichment of known and de novo sequence motifs in the CLIP peaks of individual RBPs. a. Enrichment scores computed by comparing the frequency of known motifs among the top CLIP peaks identified by the indicated method with the frequency in background regions (random subsets of regions that were least enriched in CLIP reads, see Methods). Each peak finder is shown in a different color. The enrichments of the de novo motifs are indicated by stars, red for RCRUNCH peak center and blue for RCRUNCH crosslink. b. The known RBP-specific motifs from ATtRACT [158] that were used in the analysis. c. Most significantly enriched de novo motif predicted by Phylogibbs [174] in the "RCRUNCH peak center" sites from each sample. d. Most significantly enriched de novo motif predicted by Phylogibbs in the "RCRUNCH crosslink" sites from each sample.

(U)5 motif corresponding to this protein. To determine whether "truncated" reads, resulting from the reverse transcriptase falling off at the crosslinked nucleotide, allow a more accurate identification of RBP-binding motifs than the peaks in read coverage, we implemented the "RCRUNCH crosslink" variant, in which RBP binding sites are extracted from around the most crosslinked position within each peak (position where most reads start), in contrast to the "RCRUNCH peak center" discussed so far, in which sites are extracted relative to coverage peak centers. RCRUNCH crosslink clearly recovered the hnRNPC-specific (U)5 motif (Figure 3.4d) and further improved the identification of the RBFOX2-specific UGCAUG motif, while the recovery of the PUM2 and PTBP1-specific motifs was unaffected.

These results demonstrate that the peaks predicted by different methods are enriched to fairly similar extents in the expected motifs, though RCRUNCH has the most reliable high enrichments. Furthermore, the de novo motifs identified by RCRUNCH are more enriched in the peaks than the known motifs, even when they appear quite similar. For some proteins, specifically hnRNPC and RBFOX2, the read starts enable a more precise identification of RBP-specific bind-

ing motifs, while for others, like PTBP1 and PUM2, the coverage peak centers are equally informative.

3.3.3 *RCRUNCH helps elucidate how RBPs interact with and crosslink to RNAs*

To better understand why the sites extracted from around coverage peak centers contain the RBP-binding motif for some proteins but not for others, we carried out the following analysis. Within each of the 1000 most significant peaks based on the enrichment in reads we identified both the location of the highest-scoring match to the RBP-specific motif and the crosslink position, where most reads started (Supplementary Table 1). Then, we anchored the peaks on the center of the motif match (position 0), and constructed the histograms of distances between the crosslinks and motifs, and between coverage peak centers and motifs. As already suggested by the observations from the previous section, the relationship between these two histograms is highly dependent on the studied RBP (Figure 3.4a-d). For PUM2, RBFOX2, and hnRNPC, crosslink positions strongly co-localize with the RBP-binding motif, while for PTBP1 this is not the case. In contrast, the peak centers show weak co-localization with the binding motif of PUM2 and PTBP1, but occur clearly downstream of the binding motifs for RBFOX2 and hnRNPC.

In the case of hnRNPC, highly enriched motifs were recovered around peak centers as well, and these motifs were very different from the expected (U)5 (Figure 3.3c). A literature search revealed that these motifs correspond to the Alu antisense element (AAE) [175], consistent with reported function of hnRNPC in suppressing the exonization of these repetitive elements [147]. Computing the peak center - motif and crosslink - motif histograms relative to the AAE showed that hnRNPC binding sites containing AAEs have a very specific configuration, crosslinking occurring upstream of the AAE, within the U-rich motif, leading to CLIP read starts in this region, while the peak in read coverage is on the AAE (Figure 3.4e,f). To better understand how these patterns relate to the specific interaction of an RBP with RNAs, we carried out a simulation of an RBP binding to its cognate motif, crosslinking to the RNA and protecting an extended region of the target from digestion. As the efficiency of crosslinking depends on the identity of both the nucleotides and the amino acids that participate in the RNA-RBP interaction [176], we simulated three scenarios, corresponding to the nucleotide with the highest efficiency of crosslinking being located upstream, within or downstream of the RBP-specific binding motif. For each of these cases we simulated scenarios where the probability of reverse transcriptase reading through the crosslinked nucleotide is very low, intermediate or very high (Supplementary Figure B.1). We found that when the readthrough probability is low, the crosslink is a better indicator of the binding motif than the peak center (Figure 3.4g). In contrast, when the readthrough probability is moderate to high, the crosslink

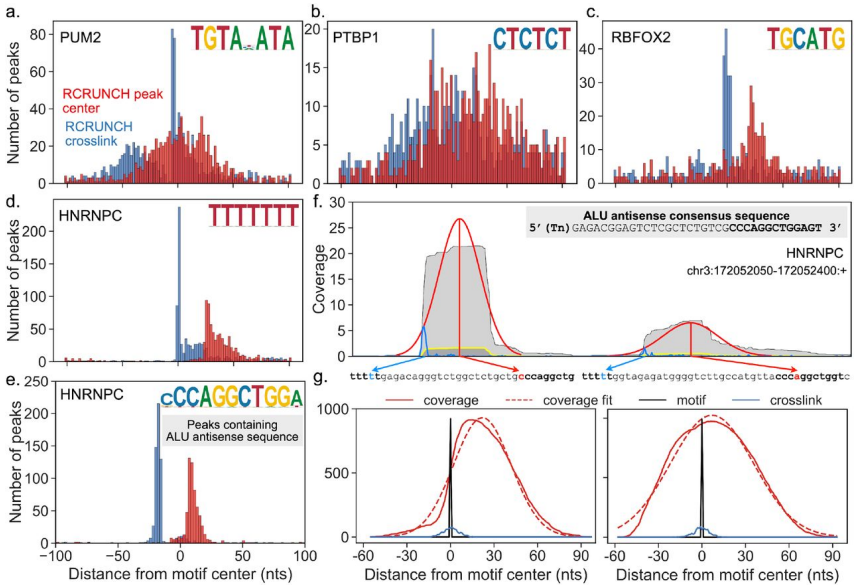


Figure 3.4: Configuration of binding and crosslinking differ across RBPs. a-d. Histograms of distances between the cognate motifs of RBPs and the centers of the read coverage peaks (RCRUNCH peak center, in red) or between the motifs and the most frequent read start position in a given peak (RCRUNCH crosslink, in blue). The top 1000 peaks (in the order of their z-score) from one of the available samples for PUM2 (a), PTBP1 (b), RBFOX2 (c) and hnRNPC (d) were extracted. The cognate motif with the highest posterior probability given the RBP's PWM was determined and the peak was retained only when the posterior was at least 0.3. e. For hnRNPC we carried out the same analysis relative to the Alu-related motif. f. Example of two hnRNPC binding peaks located on Alu antisense elements. The library-size-normalized read coverage in the eCLIP sample is shown in gray, while the coverage in the background SMI sample is shown in dark gray with a yellow outline. The fitted Gaussian peaks predicted by RCRUNCH are shown in red, while the distribution of reads starts in this region is depicted with the blue line. The most frequent read start within a coverage peak is chosen by RCRUNCH crosslink as the crosslink position and the corresponding nucleotide is indicated here by the blue arrow. The red arrow links the peak center to the corresponding nucleotide within the Alu antisense element. g. Results of CLIP experiment simulations, showing the read coverage profile (full red line), corresponding Gaussian fit (dashed red line), and frequency of crosslinks (blue) with respect to the RBP-specific motif (black, centered on position 0), for low (0.1, left) and high (0.9, right) probability of reverse transcriptase readthrough.

position and the coverage peak center are located at comparable distances from the RBP-binding motif, so that either could be used to identify the RBP binding site (Figure 3.4g). Thus, the motif-crosslink and motif-peak center distance relationships that we observed for the selected proteins indicate that the probability that the reverse transcriptase reads through the RBP-RNA crosslink is much lower in the case of RBFOX2 and hnRNPC compared PUM2 and PTBP1, leading to a more precise identification of the binding motif when binding sites are centered on the crosslink position.

These results suggest that model-driven analyses of CLIP data, taking into account the architecture of protein-RNA interactions, could further improve the identification of binding sites and the interpretation of the observed binding patterns. Furthermore, the variant RCRUNCH workflows provide a flexible platform to explore the architecture of RBP-RNA interaction sites.

3.3.4 *RCRUNCH variants enable detection of specific classes of RBP targets*

Tools for CLIP data analysis focus almost exclusively on reads that map uniquely to the genome, leaving out multi-mapping reads or reads that map across splice junctions, which are more challenging to map and quantify correctly. To provide users with the opportunity to investigate RBPs that specifically bind to repetitive elements or mature mRNAs, we have implemented and evaluated a few variations of the RCRUNCH workflow. Specifically, we have implemented the option of identifying binding sites that are located in the immediate vicinity of exon-exon junctions in mature mRNAs, as well as the option of using reads that map to multiple genomic locations (multi-mappers). In the first situation, some reads end up mapping to the genome in a split manner, partly to the 5' exon and partly to the 3' exon. This in turn can lead to multiple distant peaks, with the RBP-binding motif being present at only one or perhaps neither of these peaks. Finally, the question of an appropriate "background" for estimating the enrichment of reads in CLIP samples is still open [149]. Aside from the size-matched control used here, the relative abundance of mRNAs (estimated based on RNA-seq data) is sometimes taken into account [177]. By providing an appropriate file with sequenced reads, RCRUNCH allows an easy incorporation of different types of background. Here we used the mRNA-seq data generated for the specific cell lines included in the ENCODE project. We benchmarked the performance of RCRUNCH variants on the proteins chosen at the beginning of our study. In all cases we extracted sites anchored at the crosslink position within each peak and compared the "standard" RCRUNCH crosslink with individual variants across a few different measures. These measures were: the number of significant sites (at FDR = 0.1) identified in a sample (Figure 3.5a-c), the enrichment of the known motif in the top 1000 sites identified in each sample (Figure 3.5d-f), and the similarity of the known motif

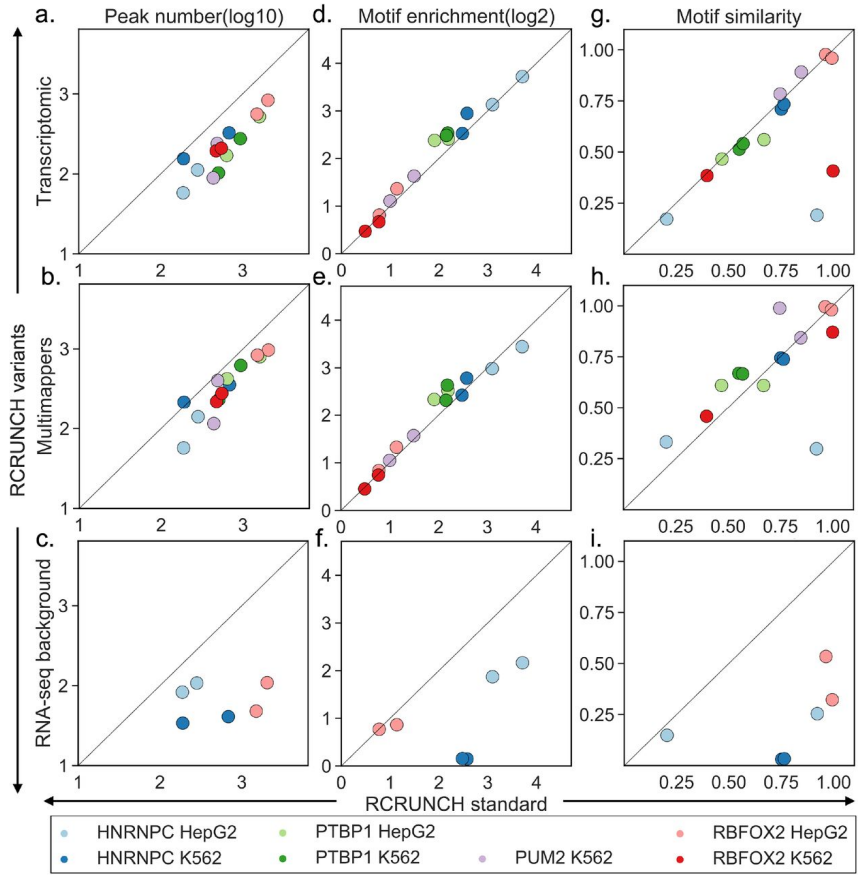


Figure 3.5: Performance evaluation of RCRUNCH variants. "RCRUNCH standard" refers to peaks identified based on reads that map uniquely to the genome, and binding sites that are always centered on the most frequent crosslink position in each peak. Rows correspond to variants of the RCRUNCH workflow, while columns show different metrics used to compare each of the variants with the "standard" RCRUNCH: a-c. Number of significant peaks in a sample (at FDR < 0.1); d-f. Enrichment of the known motif of the RBP that was assayed in a particular sample; g-i. Similarity of the motif identified de novo from the top peaks of a given sample and the known motif for the respective RBP. In the case of the RNA-seq background variant, some samples did not yield any significant peaks (FDR = 0.1). These samples are therefore not represented in the plots.

of the RBP to the de novo motif identified from the top 1000 peaks of a given sample (Figure 3.5g-i). While for the RCRUNCH transcriptomic and multi-mapper approaches the results were very comparable with those of the standard RCRUNCH, the choice of RNA-seq as background results in a strong decrease in performance. Including multi-mappers or splice junction reads led to the recovery of somewhat fewer sites, but the quality of the peaks, measured in terms of their enrichment in motifs, was not affected.

These results demonstrate that RCRUNCH is a flexible and performant method for CLIP data analysis. The choice of background is important, and in the case of eCLIP, the size-matched control samples provide a more appropriate background for estimating read enrichment in binding sites than the mRNA-samples.

3.3.5 *A compendium of RBP binding motifs inferred from eCLIP data*

Although various analyses of the ENCODE eCLIP datasets have been carried out, a consolidated compendium of binding motifs inferred for individual proteins from these data is not available. To fill this gap, we have applied RCRUNCH (both peak center and crosslink variants) to all available eCLIP samples, determined peaks that are enriched in reads relative to the SMI background, inferred the most significantly enriched sequence motifs and finally, for each RBP, identified the motif with the highest average enrichment across all samples corresponding to the RBP (Supplementary Table 2). The distribution of the number of binding sites per RBP, as shown in Figure 3.6a, indicates that two thirds of the samples yielded more than 100 binding sites, with few samples (for the HNRNPL, AGGF1, DDX3, TARDBP proteins) yielding thousands of sites. As may have been expected, a known binding motif in ATtRACT is indicative of the protein having a high number of binding sites (Figure 3.6a). We next calculated the average Jaccard similarity index of peaks identified from pairs of samples, either corresponding to the same protein, or to different proteins (Figure 3.3b-c and Supplementary Figure B.2). We also carried out an analysis of the motifs enriched in individual samples (Supplementary Figure B.4), ultimately identifying the motif that best explains the entire data obtained for a given protein (highest sum of log-likelihood ratios across all samples, Figure 3.6c-d). Of 149 proteins, 86 yielded an enriched motif in our analysis, and 26 of these already had a specific motif in the ATtRACT database. 21 of the proteins for which we could not identify an enriched motif in this study were covered by the ATtRACT database (Figure 3.6b). The heatmap of peak overlaps shows good consistency among different experiments involving the same protein (Figure 3.6c, diagonal), and also highlights interesting cases of proteins that bind to similar regions, in many cases because the proteins take part in the same multi-molecular complex. For example, we found high overlaps between the sites of splicing factors U2AF1 and U2AF2 [178, 179] , of the DXH30 and FASTKD2 proteins involved in the ribosomes biogenesis in the mitochondria (Supplementary Figure B.2) [180], of the DGCR8 and DROSHA components of the miRNA biogenesis complex [181] and a few others (Figure 3.6c). These results lend further support to the notion that our method recovers expected signals in the eCLIP data. Additional per sample analyses are shown in Supplementary Figure B.2. In brief, we identified enriched motifs in 90% of the samples, similar between the RCRUNCH crosslink and RCRUNCH peak center approaches (Supplementary Figure B.4c), but in many cases the enrichments were small. As we have seen for the benchmarked proteins, the enrichment of the de novo identified motif was higher than the enrichment of the known motif for the studied protein (Supplementary Figure B.4b). We further calculated the similarity (see Methods) between the known and de novo motifs, the latter obtained either from RCRUNCH crosslink or RCRUNCH peak center-predicted binding sites. We found that around 80% of the samples yielded motifs with at least 0.4 similarity to the known motif, which is

a much higher proportion than when comparing random pairs of motifs. The similarity was slightly higher when sites were identified by RCRUNCH crosslink (Supplementary Figure B.4d). Given the large number of motifs identified for RBPs that are not represented in ATtRACT, we asked whether these motifs are reproduced between replicate samples of an RBP. Indeed, the similarity of motifs obtained from replicate samples was similar for proteins with and without a known motif, and it correlated with the number of sites inferred from the samples (Supplementary Figure B.4e).

These data indicate that the motifs identified from RCRUNCH-extracted CLIP peaks are reliable, conform with prior knowledge, and explain the binding data better than the motifs that are currently available in databases. Altogether, the compendium that we have constructed (Supplementary Table 2), provides sequence specificity data for 86 RBPs, thus being, to our knowledge, the most extensive collection of in vivo, reproducibly-identified, consensus RBP binding motifs.

3.4 DISCUSSION

Within over a decade of development and application, CLIP has provided a wealth of insight into the RNA-binding protein-dependent regulation of cellular processes such as RNA maturation, turnover, localization, and translation. The large collection of RBP-centric high-throughput datasets that has been generated as part of the ENCODE project [182] is broadly used to unravel the functions of RBPs, many of which were only recently found to bind RNAs. As other types of high-throughput data, CLIP also requires dedicated computational analysis methods. In this context, RCRUNCH makes the following main contributions.

First, it is the first completely automated solution to CLIP data analysis, from reads to binding sites and sequence motifs, with accuracy and running times that compare favorably to those of the most broadly used tools to date. Key to this performance are the enrichment-based prioritization of genomic regions likely to contain RBP binding sites, and the model for evaluating this enrichment in individual sites.

Second, to accommodate the variability of target types across RBPs, RCRUNCH goes beyond the typical approach of using uniquely genome-mapped reads, allowing the inclusion of multi-mappers and/or of reads that map across splice junctions. Both of these situations make it difficult to determine the locus of origin of the reads and may lead to a decreased accuracy of binding site inference. Nevertheless, for RBPs that specifically bind repeat elements, or in the vicinity of splice junctions, taking into account such reads and appropriately defining the peaks in read coverage is a must. The variant RCRUNCH workflows fulfill this need. RCRUNCH multi-mapper considers reads that map to a maximum number of genomic locations (specified by the user), distributing the reads equally among the loci with maximum

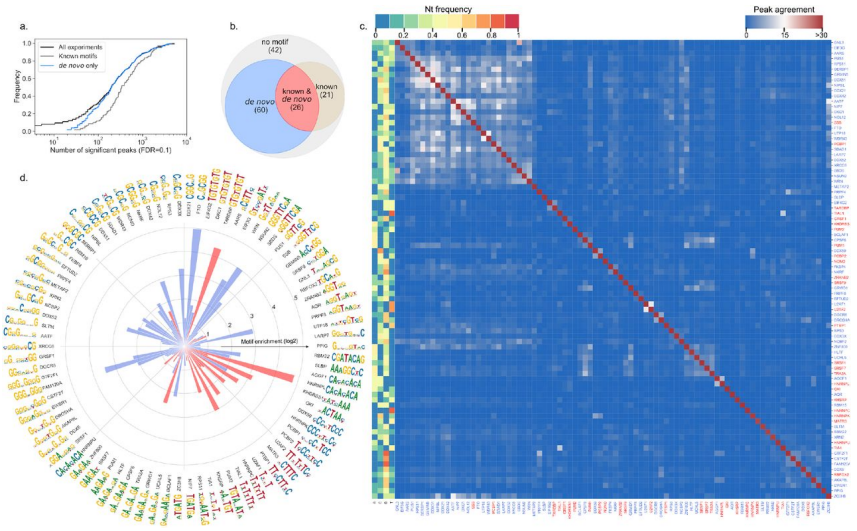


Figure 3.6: RCRUNCH results for all ENCODE eCLIP data currently available. a. Cumulative distribution of the number of significant binding sites detected per experiment (FDR threshold=0.1, in black). Cumulative distributions are also shown separately for samples corresponding to proteins with a known binding motif (gray) and to proteins for which no known motif is available in ATtRACT, but one was found by RCRUNCH (blue). b. Venn diagram summarizing the motif inference in the identified peaks. We distinguished four categories of proteins: for which (1) no motif is known and also no enriched motif was identified in this study (grey), (2) a de novo motif was found for a protein for which no motif is given in ATtRACT (blue) (3) a de novo motif was found for a protein with a known motif in ATtRACT (coral) and (4) a motif is known, but none was identified de novo. c. Heatmap of mean peak agreement across RBPs (only the RBPs. The agreement is calculated as the Jaccard index of the nucleotides in the peaks, where the intersection of two sets of peaks is the number of nts covered in both sets, while the union is the number of nts covered in at least one of the two sets. The color range is capped at a similarity of 0.4 to make the clusters more easily distinguishable. The top peaks are taken according to the FDR threshold (0.1), extending by 20 nts upstream and downstream from the crosslink site. Since there are more than 1 replicates per RBP, the mean of these agreement calculations is used here. The colors on the left indicate the relative frequency of each nucleotide type averaged over all positions of the PWM. d. Polar projection of the enrichment of de novo motifs inferred for individual RBPs from all peaks with FDR >0.1 extracted from the ENCODE samples. Only RBPs for which an enriched motif (motif enrichment higher than 1) are included. The color of the bars indicate whether the respective RBP already has a known motif in ATtRACT (coral) or not (blue).

alignment score. Compared to the run that only allowed uniquely mapped reads, this approach yields binding sites with similar enrichments in the expected sequence motifs of the benchmarked proteins, including those with repetitive binding motifs like PTBP1. More sophisticated models, e.g. applying an expectation/maximization approach to read-to-locus assignment, have been proposed before [78], but they come at the cost of increased running times, and have not been broadly adopted. The RCRUNCH transcriptome variant was designed to address another special case, namely that of RBPs that bind predominantly mature mRNAs. Many RBPs are located in the cytoplasm where they orchestrate RNA traffic and localization [173]. As internal exons of human transcripts are relatively short, 147 nucleotides [183], it is likely that CLIP reads for RBPs that bind to these exons will cover exon-exon junctions [184]. Recent generation alignment programs such as STAR [185] can carry out spliced alignment. However, reads that span splice junctions will give rise to multiple peaks, in the 5′ and the 3′ exons that flank the splice junction. This will affect the accuracy of binding site and motif identification. A way to circumvent the issue is to map the reads to transcripts instead of genome sequences. This is not without drawbacks. First, given that it is not generally known a priori whether the RBP of interest binds mature RNAs or other classes of transcripts, a hybrid strategy will need to be adopted, to allow the identification of binding sites in introns as well as in mature mRNAs. Second, given the large number of possible isoforms per gene, accurate assignment of reads to isoforms and peak identification in multiple isoforms are not trivial. In RCRUNCH transcriptome we have implemented a general hybrid strategy to capture binding sites across all types of types of transcripts, including pre-mRNAs and mature mRNAs, without incurring large computational costs. Namely, taking advantage of the observation that individual cell types express predominantly one isoform from a given gene [186], we first determine which of the known isoforms of each gene has the highest expression level in the CLIP sample. We then use these isoforms as pseudo-chromosomes, assigning reads to the best-scoring loci, but with priority given to the spliced isoforms over genomic loci. This approach gave good results for the benchmarked proteins, including PUM2, which is known to bind to mature mRNAs [187], but overall, the number of sites that spanned splice junctions was small for the benchmarked proteins. On the other hand, for splicing factors we identified many sites in the vicinity of splice sites, as expected, indicating that these data can be studied further to determine to what extent these splicing factors remain associated with the mature RNAs (Supplementary Figure B.5). Nevertheless, our exploration of ways to handle reads that originate in various categories of targets was by no means exhaustive and this could be a direction for further development of the workflow.

Third, our analysis of the ENCODE eCLIP data yielded enriched sequence motifs for 86 RBPs. These were selected using a uniform procedure, based on the maximum enrichment across all samples available for a given RBP, in contrast to resources such as ATtRACT, which con-

tain multiple, heterogeneous RBP-specific sequence motifs obtained with a wide range of techniques. While the eCLIP datasets were the focus of various previous studies (e.g. [86, 157, 173]), a compendium of reproducible binding motifs inferred from these data sets is not available. Moreover, although RBPs typically bind to a defined sequence (or sometimes structure) motif, the binding specificity of RBPs inferred from eCLIP data has been described in terms of collections of short motifs [86, 173], for reasons that remain unclear. The main aim of the work presented here was to provide a uniform procedure for inferring RBP sequence specificity from binding data, and a resource of RBP-specific motifs similar to those available for transcription factors (e.g. [188]). For RBPs that have been extensively studied, the motifs that we identified de novo from the CLIP peaks conform with prior knowledge, though they differ in quantitative detail. Moreover, the de novo motifs have higher enrichment in the peaks compared to the known motifs, which may indicate context-specific contributions to the binding affinity. Overall, we identified enriched sequence motifs for 86 proteins, 60 of which are not represented in the ATtRACT database. In some cases, the most enriched motif in a given sample was not the one known to be bound by the corresponding protein, as observed before [173]. Repetitive motifs (G/G&C/C-rich) were occasionally found to be enriched in various samples, and this enrichment was also reproduced in replicate samples for the same protein. This raises the question of whether these motifs represent some sort of non-specific background in CLIP samples [162]. However, we did not find a larger overlap among the binding sites containing such motifs relative to binding sites of randomly chosen pairs of proteins (Figure 3.6c). In fact, overlaying the pairwise overlap data with data on protein complex composition revealed compelling cases of high overlap for proteins of the same complex such as the spliceosome, the pre-rRNA processing complex, a paraspeckle-related complex and others (Supplementary Figure B.2). Thus, our analysis does not support the concept that general non-specific background in eCLIP leads to similar motifs for unrelated RBPs, though it will be interesting to investigate further the functional significance of the identified motifs.

Finally, our analysis of the motif-crosslink and motif-peak center distances revealed distinct RBP-dependent patterns. Most striking was the peak in coverage observed over the AAEs in the hnRNPC eCLIP. HnRNPC binds (U)5 elements [147], while the read starts, indicative of crosslink positions, were located in a U-rich region upstream of the AAE. Our simulation of a CLIP experiment suggests that the hnRNPC data is quite unusual relative to data for other RBPs. HnRNPC has a large number of binding sites in Alu antisense elements that have a conserved consensus extending much beyond the U-rich element. Motif finding methods will identify this consensus as extremely enriched, more so than the much shorter (U)5 motif. The strong colocalization of the most frequent crosslink position within a peak and the RBP-specific motif supports the notion that the RT has a high propensity to stop extending the cDNA when it encounters the RNA-RBP crosslink [50]. However, our analysis also suggests

that the readthough probability varies substantially between RBPs, being very low for hnRNPC, and relatively high for other proteins like PTBP1 and PUM2. This highlights the importance of a flexible but principled approach to binding site and motif identification, using a general measure of performance such as the motif enrichment score, as done here. This is because the configuration of RBP-RNA interactions varies across RBPs, influencing the nature of the reads that are captured in the CLIP experiment. Although beyond the scope of our present work, exploration of the ENCODE data set with a simulation-driven approach may yield further insight into the interactions of individual RBPs with their binding sites.

3.5 CONCLUSIONS

Our study provides a general, end-to-end solution for CLIP data analysis, starting from sequenced reads and ending in binding sites and RBP-specific sequence motifs. The tool compares favorably with the most broadly used tools to date, and further extends the type of reads that can be analyzed, to multi-mapping and split-mapping reads. By applying RCRUNCH to the entire ENCODE set of samples available to date, we provide a compendium of reproducibly enriched sequenced motifs for 86 RBPs, of which only 26 are represented in extensive databases available today, such as ATtRACT. Finally, our simulations suggest that the architecture of RBP-RNA interactions imposes strong variation in the probability of identifying the precise position of crosslinking from CLIP data.

3.6 METHODS

3.6.1 *Inputs to RCRUNCH*

RCRUNCH performs its analysis on at least one paired-end, stranded CLIP sample, and a corresponding background sample (which could be size-matched control (SMI) from eCLIP experiments or RNA-seq) both provided in fastq format. All necessary parameters for the run such as sample file names, adapters, fragment size, presence of UMIs etc. should be provided in a config file, a template of which can be found in the repository along with a test case. The tool also requires the genome sequence fasta file and Ensembl [189] gtf annotation of the corresponding organism, also given in the config file. RCRUNCH can additionally perform some optional analyses, one being the filtering out of sequences that correspond to specific non-coding RNA biotypes and the other being the enrichment analysis for known sequence motifs. If these options are chosen, paths to corresponding files should be provided in the config file, according to the instructions in the README file accompanying the software.

3.6.2 *Read preprocessing*

3.6.2.1 *Adapter removal*

3' and 5' adapters for read1 and/or read2 specified in the config file are trimmed with Cutadapt [190].

3.6.2.2 *Alignment of reads to reference genome*

The alignment of reads to the reference genome is done with STAR [185], disabling the soft-clipping option. Some of the options to STAR differ from the standard value, to allow the alignment of short reads with only few mismatches (outFilterScoreMinOverLread 0.2, –outFilterMatchNminOverLrea 0.2, outFilterMismatchNoverLmax 0.1). Multi-mapper reads (that map equally well - same number of errors - to multiple regions in the genome) can be included in the analysis by setting the 'multimappers' field in the config file to the desired number of equivalent mappings to consider for a read. In this case, reads that map to at most 'multimappers' locations in the genome are counted towards each of these locations with a weight of 1/'multimappers'.

3.6.2.3 *Removal of reads from abundant non-coding RNAs*

Reads derived from some non-coding RNAs (e.g ribosomal (rRNAs), transfer (tRNAs) and small nuclear RNAs (snRNAs)) are abundant in many CLIP samples and thus believed to be largely contaminants [157]. Frequently, these abundant RNAs are also encoded in highly repetitive genomic loci. For these reasons RCRUNCH allows the option of selective removal of reads mapping to ncRNAs, based on the annotation from RNAcentral [191]. For this, the user will need to provide a gff3-formatted file for the appropriate species, which can be downloaded from RNAcentral. Specific biotypes of ncRNAs can be selectively removed by filling out the 'ncRNA_biotypes' option in the config. The names of reads that overlap in the genome with any of the selected ncRNAs specified by the user, are saved in a list. This is then used as input in the FilterSamReads function of the Picard software [192] to remove the reads from the alignment file that passed to downstream analysis.

3.6.2.4 *Removal of PCR duplicates*

PCR amplification is a well-established source of error in the estimation of transcript counts [193]. However, different CLIP protocols differ in whether and how they deal with this issue. Accordingly, RCRUNCH offers multiple options. The default is to not carry out any PCR duplicate removal, which can be specified by choosing 'standard' as the value for the 'dup_type' fields in the config file. Alternatively, RCRUNCH can take advantage of Unique Molecular Identifiers (UMIs), which are introduced by ligation of a DNA adapter containing a random oligonucleotide (the UMI or randomer) to the cDNA fragments, as done in eCLIP [157]. As the UMI is preserved

during PCR amplification, it can be used to identify reads that are copies of the same initial fragment. To remove PCR duplicates we use UMI-tools [194], which assumes that the UMI sequences are suffices of the read names. However, data from the ENCODE project has the UMIs as prefixes to the read names. Thus, we use a specific rule to make this transformation, which is controlled by the field 'format' in the config file, and can be either 'encode' or 'standard'. If standard is chosen, no reformatting occurs and it is up to the user to make sure the format of the fastq files they provide is compatible with UMI-tools processing. Finally, if the sample preparation did not include the addition of UMIs, RCRUNCH can still attempt this removal via the deduplication function of STAR [185] via filling out the 'dup_type' option with 'duplicates'. If no duplicate removal is desired then the 'dup_type' can take the option 'with_duplicates'.

3.6.2.5 *Additional preprocessing steps for the 'RCRUNCH transcriptome' approach*

If the user chooses the transcriptomic mode of RCRUNCH ('method_types' as 'TR' in the config), a few additional steps are needed to identify reads that map across splice junctions. First, reads are aligned to the genome (as described above), and the alignments are used to remove PCR duplicates and possibly ncRNAs. The remaining read alignments for the foreground sample are used by the Salmon software [185, 192, 195] to select the most expressed transcript isoform for each gene and construct a dataset-specific transcriptome. The reads that were selected in the first step are aligned to the transcriptome, after which the genome and transcriptome alignment files are jointly analyzed to identify the highest scoring alignment (AS score in the bamfiles) for each read. If the AS score for the transcriptome alignment is greater than the AS score of the genome alignment -3, the alignment to the transcriptome is selected. We chose this criterion rather than requiring the transcriptome alignment score to be strictly better than the genome alignment score to conservatively assign the reads preferentially to the transcriptome. Peaks are then detected either on the genome or the transcriptome, treating individual transcripts as we treat chromosomes. This approach allows us to detect and properly quantify RBP binding sites in the vicinity or even spanning splice junctions.

3.6.3 *The RCRUNCH model for the detection of RBP binding regions*

Genome/transcriptome-wide identification of peaks corresponding to individual binding sites for an RBP is time-consuming. For this reason RCRUNCH implements a two-step process, as previously done for analyzing chromatin immunoprecipitation data [151]. That is, broader genomic regions that are enriched in reads in the foreground (CLIP) sample compared to the background are first identified, and then individual peaks are fitted to the CLIP read coverage profiles within the selected windows. More specifically, we tabulate the number of

fragments that map to sliding windows of a specific size (e.g 300 nucleotides) both in the foreground (CLIP) and the background (e.g. SMI) sample. For windows that are not enriched in binding, fluctuations in the number of reads across replicate samples have been found to be well-described by the convolution of a log-normal distribution due to multiplicative noise in the sample preparation and Poisson sampling noise [151]. The frequency of reads in windows with no RBP binding should in principle be the same between the foreground and the background sample. However, as in the foreground sample reads are expected to come largely from bound regions, the unbound regions will be somewhat depleted of reads. Including a correction term μ to account for this depletion, the probability of the data for an unbound region can be modeled as

$$P_u(n|N, m, M, \sigma, \mu) = \frac{1}{\sqrt{2\pi(2\sigma^2 + \frac{1}{n} + \frac{1}{m})}} exp(-\frac{(log(\frac{n}{N}) - log(\frac{m}{M}) - \mu)^2}{2(2\sigma^2 + \frac{1}{n} + \frac{1}{m})})$$

where n is the number of reads in a given window (of a total of N reads) in the foreground sample, m is the number of reads (of a total of M) in the background sample, 22 is the variance due to multiplicative noise in the two samples, and $1/n$ and $1/m$ are the variances due to the Poisson noise. For the probability of read counts due to binding, we assume a uniform distribution over the range corresponding to the maximum and minimum difference in read frequency between foreground and background samples across all windows:

$$P_b(n|N, m, M) = \frac{1}{\delta_{max} - \delta_{min}}$$

where $\delta = \frac{n}{N} - \frac{m}{M}$

Finally, the probability of observing the n reads is given by a mixture model, of the window representing a background region(with probability) or a region of RBP binding (with probability $1 - \rho$):

$$P(n|N, m, M, \sigma, \mu, \rho) = \rho P_u(n|N, m, M, \sigma, \mu) + (1 - \rho)P_b(n|N, m, M)$$

We fit the parameters ρ, σ, μ by expectation-maximization (Detailed method described in [151], Supplemental material, section 1.9). Regions that are found to be enriched in the foreground sample (based on an FDR threshold and a corresponding z-score threshold) are analyzed individually with a second mixture model, to fit peaks corresponding to the individual binding events. The z-scores of these peaks are recalculated and those that still have a high enough z-score are kept as significant (Supplementary, RCRUNCH model). The regions that are considered for the peak fitting might also have slightly lower z-scores. This aims to be more inclusive and reduce false negatives (cases where the region has significant peaks close to high background regions).

3.6.4 *De novo motif identification and enrichment calculation*

For each peak we extract the region covering 25 nucleotides (nts) upstream and 25 nts downstream of the peak center in the case of the

RCRUNCH peak center approach (option 'peak_center' in the config). For RCRUNCH crosslink we extract the same type of window, centered not on the peak center but rather on the position where most read starts within the peak are located. To avoid double-counting of motifs, we merge overlapping peaks, obtaining thus a set of non-redundant peaks. To estimate the enrichment of known and de novo motifs we used the MotEvo software [163] as described in [151]. Namely, a subset of the peaks is used to train a prior probability of non-specific binding, and the motif enrichment is then estimated from the test set, using this prior. We carried out this procedure 5 times (parameter 'runs', can be modified by the user), to both estimate the mean enrichment for known motifs from the ATtRACT database [158], and to identify de novo motifs (represented as positional weight matrices, PWMs) of various sizes (provided by the user, default values: 6, 10, 14) that are enriched in the foreground peaks relative to unbound sequences using PhyloGibbs [196]. As unbound sequences we use those genomic regions obtained in the CLIP experiment that had the lowest z-scores, meaning that they were depleted in CLIP reads. We sampled 20 background sequences of equal size (50nts) for each foreground peak. For each motif length, we extracted the top two motifs in the order of cross-validated enrichment and trimmed off positions from the boundaries of the motif until the information score became at least 0.5. For both the known motifs, as well as the de novo motifs we report the motif and corresponding enrichment for each of the 5 runs, as the motif can vary to some extent from run to run. The de novo motifs from all these runs are then collected together in an extra step and along with the known motifs, the enrichment over all the significant peaks is estimated (in this step there cannot be any prior estimate).

3.6.5 *Benchmarking peak finder tools*

We compared RCRUNCH with the recently developed and broadly used tools clipper [157], PURE-CLIP [162] and omniCLIP [152]. To reliably and reproducibly perform this analysis we created a separate snakemake workflow. For PURE-CLIP and omniCLIP, docker images [197] were either created or used from existing repositories. As we could not implement clipper in the same type of workflow as the other tools, we relied on the bed files of clipper-predicted binding sites from each sample provided by ENCODE. To benchmark the tools we used eCLIP data generated for some RBPs whose binding motifs are well-known: hnRNPC [198, 199], IGF2BP3 [198, 200], PTBP1 [201, 202], PUM2 ([203]) and RBFOX2 [202, 204, 205]. For PURE-CLIP and omniCLIP the eCLIP data had to be pre-processed separately, as the tools use alignments as input. To facilitate the comparison across methods, we used the pre-processed data from RCRUNCH. The execution of RCRUNCH was done using some specific options as explained in the RCRUNCH workflow description. Firstly, only unique mappers were aligned to the genome. PCR deduplication

was performed, using the UMIs that eCLIP experiments contain. We did not remove any reads mapping to ncRNAs, and we used the RCRUNCH genomic approach. Each of the different peak calling methods was applied, and the top 1000 peaks were extracted for the motif analysis, irrespective of FDR threshold. The motif analyses are included as a post-processing part of RCRUNCH. For RCRUNCH, both the 'crosslink' and 'peak center' positions were used as anchors for extending by 25 nts on either side and obtaining the sequences for the motif analysis. For the other methods, we used the method-predicted crosslink positions as anchors for extracting similar regions of 50 nts in length. For all methods overlapping peaks were merged to ensure non-redundancy in the sequence set. To ensure comparability, we used the same set of sequences with lowest z-scores as background for motif enrichment estimation in the peaks predicted by all samples.

3.6.6 *Calculation of peak agreement between replicate samples and between methods*

To calculate the peak agreement between methods and across replicates we used the jaccard distance metric, defined as:

$$A(nts_1, nts_2) = \frac{|nts_1 \cap nts_2|}{nts_1 + nts_2 - |nts_1 \cap nts_2|}$$

, where nts_1 and nts_2 are the total number of nucleotides contained in the top number of peaks chosen for sample 1 and sample 2, respectively.

3.6.7 *Calculation of motif similarity*

We defined the motif similarity M of two sequence motifs m_1 and m_2, $M(m_1, m_2) = \frac{2S(m_1,m_2)}{S(m_1,m_1)+S(m_2,m_2)}$, where $S(m_1, m_2) = max_1[I(m_1, m_2, d)]$ and $I(m_1, m_2, d) = \sum_i m_1(i)m_2(i-d)$ is the inner product of the motifs with the second motif being at offset d compared to the first motif [151]. This measure allows for the comparison of motifs of different lengths, and takes values between 0 (when the base frequency vectors are orthogonal) and 1 (when the two motifs are identical).

3.6.8 *RCRUNCH analysis of ENCODE eCLIP data*

We applied the RCRUNCH workflow to all of the ENCODE eCLIP datasets, consisting of 220 distinct eCLIP experiments with 143 different RBPs in two cell lines (K562, HepG2). As for the benchmarks, we used reads mapping uniquely to the genome, performed read deduplication based on the UMIs, and did not exclude reads mapping to ncRNAs. To identify the RBP-specific binding motifs, we used the top peaks for each RBP, based on FDR threshold < 0.1. Agreement across replicates and motif agreement with existing knowledge were the main metrics of performance evaluation.

3.6.9 *RCRUNCH variants*

To evaluate the RCRUNCH variants, we used the same samples that were used for benchmarking the computational methods. The pre-processing and post-processing steps (motif analysis) were the same as those implemented in the benchmark analysis, but the selection of reads, regions and peaks differed. Specifically, in the RCRUNCH transcriptome approach we first construct a reference transcriptome composed of the most abundant isoform of each gene in the CLIP data, and use these transcripts as 'pseudo-chromosomes' in the mapping process. We then map reads in an unspliced manner to both this reference transcriptome and to the genome and retain the mappings with the highest score. In cases when the alignments to transcriptome and genome are very close in score (transcriptome - genome scores >= -3 points), we give precedence to the transcriptome mappings and ignore the genomic ones. We then apply the standard RCRUNCH (see section: The RCRUNCH model for the detection of RBP binding regions). In RCRUNCH multi-mappers we consider not only reads that map uniquely to the genome, but also those that have up to 50 of equally good mappings (see section: Alignment of reads to reference genome). For RCRUNCH RNA-seq background we simply used RNA-seq samples that are provided by ENCODE for the cell lines (K562, HepG2) used for CLIP. These were treated the same as the SMI sample.

3.7 ABBREVIATIONS

- AAE: Alu antisense element
- AS: alignment score
- cDNA: complementary DNA
- ChIP: chromatin immunoprecipitation
- CLIP: crosslinking and immunoprecipitation (CLIP)
- DNA: deoxyribonucleic acid
- FAIR principles: findable, accessible, interoperable, reusable
- FDR: False discovery rate
- mRNA: messenger RNA
- ncRNA: non-coding RNA
- PCR: polymerase chain reaction
- PWM: positional weight matrix
- RBDs: RNA-binding domains
- RBPs: RNA-binding proteins

- RIC: RNA-interactome capture
- RNA: ribonucleic acid
- RNA-seq: RNA sequencing
- RNPs: ribonucleoprotein complexes
- UMI: unique molecular identifier
- rRNA: ribosomal RNA
- tRNA: transfer RNA
- snRNA: small nuclear RNA
- SMI: size-matched input

3.8 AVAILABILITY OF DATA AND MATERIALS

- Project name: RCRUNCH
- Project home page: https://github.com/zavolanlab/RCRUNCH [164]
- Programming language: Python
- License: Apache-2.0

3.9 ACKNOWLEDGEMENTS

This work would not have been possible without the sciCORE Team at the University of Basel. We are thankful for their support regarding the computational infrastructure as well as dedicated time and effort to aid us in this project. Special thanks to Mikhail Pachkov who provided guidance and help during the adaptation of parts of the CRUNCH model. Many thanks to the Zavolan Lab who contributed to this work with numerous pieces of advice, during the initial development. Special thanks to Dominik Burri and Christina J. Herrmann for help with testing during development.

4

DISCUSSION

RCRUNCH is a method developed with the goal of providing an end-to-end solution to CLIP data analysis, and especially the preprocessing steps, which are almost always poorly documented. RCRUNCH also incorporates a new model to describe the background noise and thus lead to a reliable detection of true binding sites. *De novo* motif prediction and the evaluation of *de novo* and known motif enrichments are also incorporated. Additional features are the possibility to include multimappers, selectively filter out non-coding RNA categories, as well as a transcriptomic approach that allows for peak detection spanning splice junctions. The workflow has been developed using Snakemake [89] and following the FAIR principles [206], with the aim of better reproducibility and maintenance along the way, as well as room for expansion of its functionality.

A few directions for further RCRUNCH development are as follows. First, RCRUNCH currently builds a background model from suitable experimental samples, e.g. SMI or RNA-seq. A relatively easily-implemented extension would be to build the background model of coverage from broader regions around the putative peaks. This would extend the applicability of RCRUNCH, especially to iCLIP data sets, where background samples are not routinely generated.

Another useful expansion could be identification of RNA structures based on the binding sites detected by RCRUNCH. Most of the analysis papers based on CLIP experiments put a lot of weight in detection of specific sequence motifs as the drivers of binding, but RNA secondary structure information [207] could potentially shed light to other modes of binding. To solve this task, one would have to use methods that infer sequence-structure models of binding sites.

Incorporation of phylogenetic information as well in the motif detection might prove useful and lead to a better motif accuracy, as it has been shown to be the case for transcription factor motif detection [163]. However, one would have to be able to handle inhomogeneous background conservation, as RBP binding sites occur in 5'/3'UTRs, coding regions, as well as introns, all with different types of evolutionary constraints and thereby conservation level.

So far, most methods rely on truncation events or mutations to reliably detect the crosslink sites within the binding sites. In our work, we showed that although truncations are meaningfully correlated with the crosslink sites because they co-localise with motifs known to be bound by RBPs, this does not seem to universally apply to all of the RBPs. Rather, we hypothesized that the extent to which truncations are informative for precisely locating the binding sites/motifs might have to do with the mode of binding of the specific RBP. Late approaches in studying the RBP-RNA interactions from a structural perspective [208], show discrepancies between NMR based observations

of proximity and UV-induced covalent bonds, suggesting that CLIP may provide a biased view of the crosslink sites. We suggested that studying the relationship between the positions where RBP-specific motifs, truncations and coverage peaks occur could provide new information regarding how the RBP contacts the RNA. Conversely, the binding sites within peaks of coverage may be more accurately identifiable with a model that includes the distance relationship between motifs and crosslinking-induced mutations. While this was attempted before, e.g. with mCross [86], the model for the inferred motifs was rather simplistic (kmers).

Using RCRUNCH, we analyzed the eCLIP data available in ENCODE to generate a set of consensus motifs detected for a big number of RBPs. These motifs could then be used as a means to explain the changes in expression of RNAs in terms of activities of motifs. Similar methods exist to explain the transcription factor-dependent changes in promoter activity [209, 210]. Some combinatorial approach that would encapsulate both transcription factor and RBPs would be even more interesting. Moreover, motif predictions genome-wide could be used in conjunction with other features, like topology, expression in specific context, RBDs, to detect combinatorial effects of RBPs (antagonistic-synergistic effects or participation in complexes).

RCRUNCH was benchmarked against a number of the methods available regarding CLIP analysis. The main metrics used for evaluation were the binding peak agreement between replicates across methods and the identity of the predicted motifs for RBPs for which the binding motif was known. We found 30% agreement between the top 1000 peaks obtained in different replicate experiments, while *de novo* predicted motifs were largely in line with current knowledge. A key limitation, though not specific to our study, is the fact that there is no ground truth regarding the exact binding behavior of a specific RBP. We have seen that the pattern of read coverage is affected by the UV-crosslinking technique, the composition of the RBP (e.g. when just the binding domain is considered as opposed to the entire protein), or even by the antibody specificity and the particular RBP isoform present in the specific cells in the context of the *in vivo* experiments. Each of these aspects leads to different answers making this a difficult problem to address. Still, with the vast amount of methods available for analysis of CLIP data, as well as the different CLIP variations, the need of a benchmark with diverse RBPs evaluated is imperative. Perhaps, the main conclusion of this study in spite of RNA-binding be very complex, we still have some way of handling the main biases with a unique model.

ZARP is the result of a group effort in the Zavolan lab to develop an RNA-seq processing pipeline based on a workflow management system while following the FAIR principles. Working in a collaborative manner we organized hackathons and followed best practices, like version control, tracking issues, automating tests via continuous integration (CI), using services that host open source software projects during development. One of the main goals in the beginning was to enable the wet lab people of the group to analyze their own data

without relying on the computational biologists. Snakemake [89] was again the workflow manager of choice and all software used was packaged using conda [89, 211] or containerized using Docker [197]. The inputs required for ZARP are a table with sample information and a config file that contains the parameters for the specific run. This project is planned to be further expanded by development of a command line interface (CLI) as well as a separate workflow that would recognise features of the samples so as to remove the need for specifying sample-specific information (such as the species from which it originated, the adaptors that were used or the type of protocol, stranded or unstranded, that was used for sample preparation. We used the ZARP development process to gain insights into how software development can be optimized in an academic setting, and to generate a useful template for future members of the group. Last but not least, ZARP is publicly available and open to contribution from the community.

A

ZARP SUPPLEMENTS

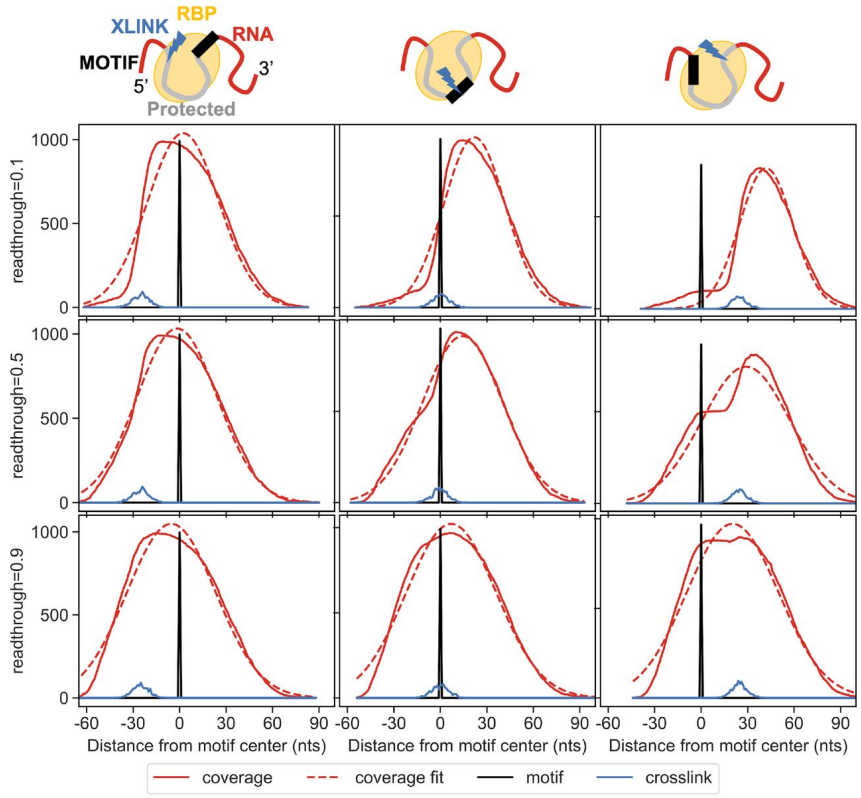
Supplementary material to chapter 2.

Supplementary Figure A.1: ZARP workflow schema. Graph-based representation of ZARP v0.3.0, including all of its steps ("rules"), as produced by running Snakemake with the –rulegraph option. Steps for both the single and the paired end workflows are shown.

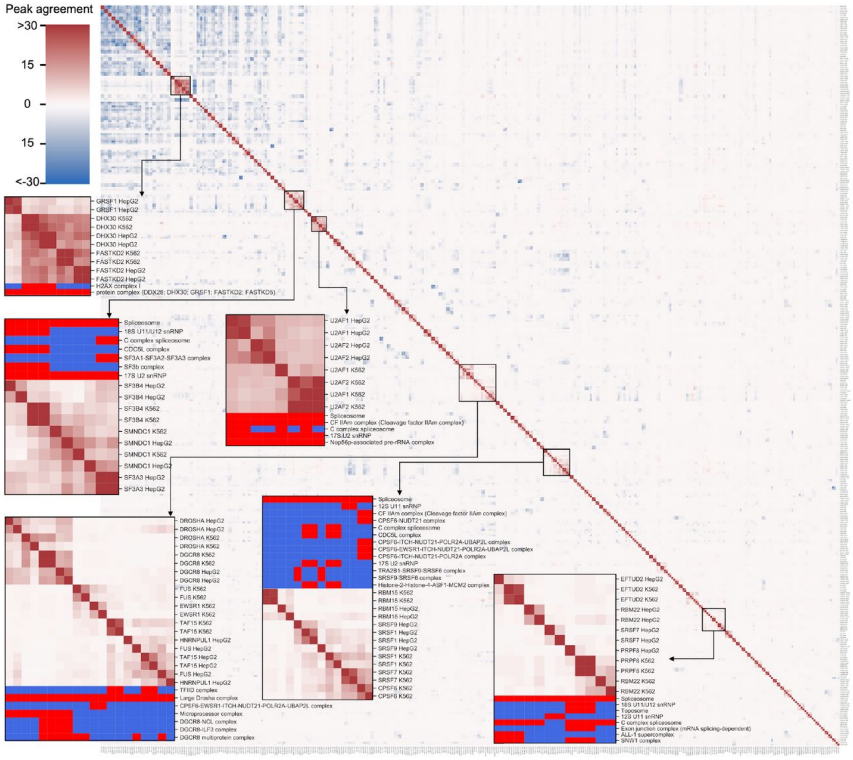
B

RCRUNCH SUPPLEMENTS

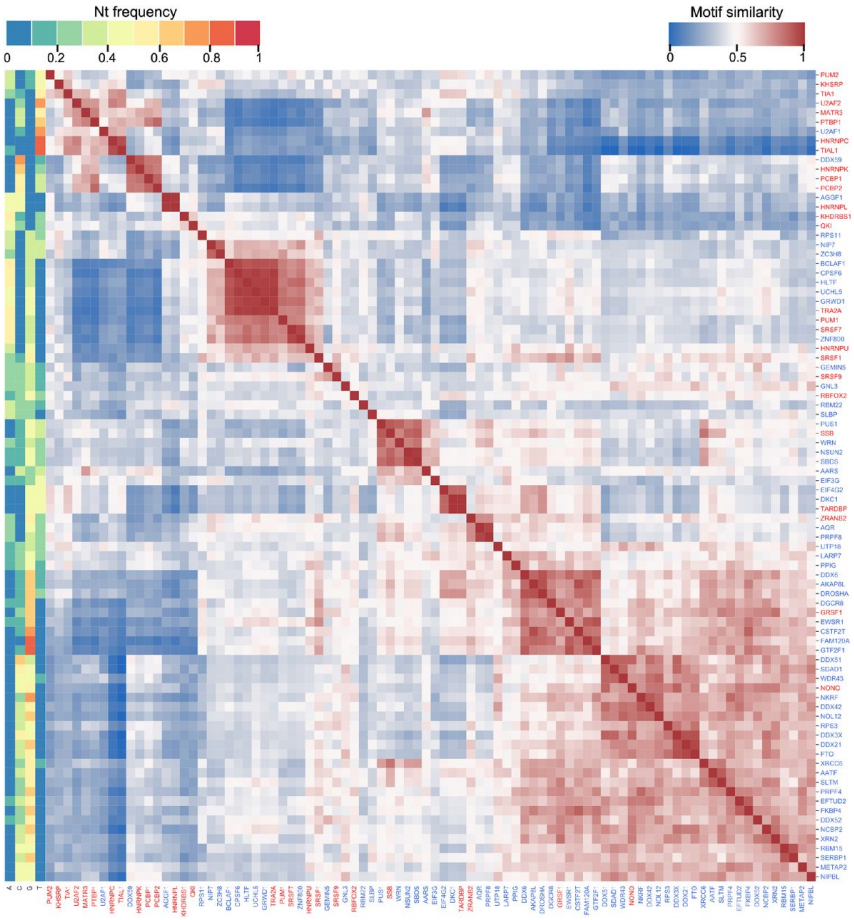
Supplementary material to chapter 3.



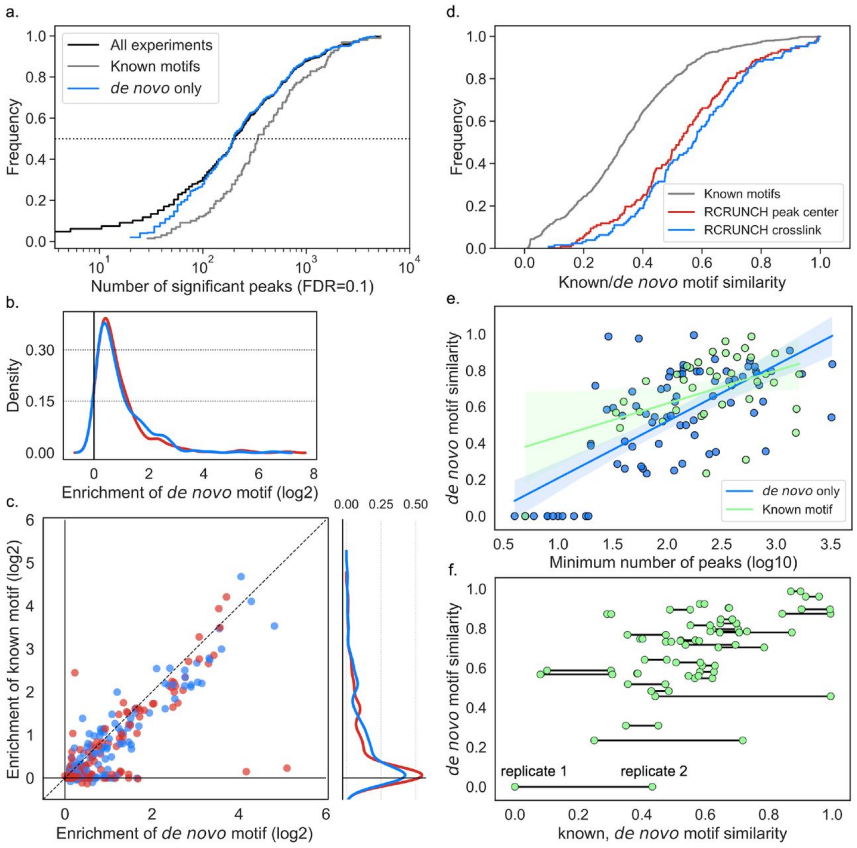
Supplementary Figure B.1: Simulation of a CLIP experiment. Upper schema: an RBP binding to its cognate motif, (black box), crosslinking to the RNA at the site indicated by the blue arrow, and protecting an extended region of the target (shown in gray) from digestion. From right to left, the relative position of the crosslink is changed relative to the motif position. The different columns correspond, as the schema shows, to different positions of the crosslink relative to the motif position. Each row corresponds to different probabilities of readthrough, varying from lower to higher from top to bottom.



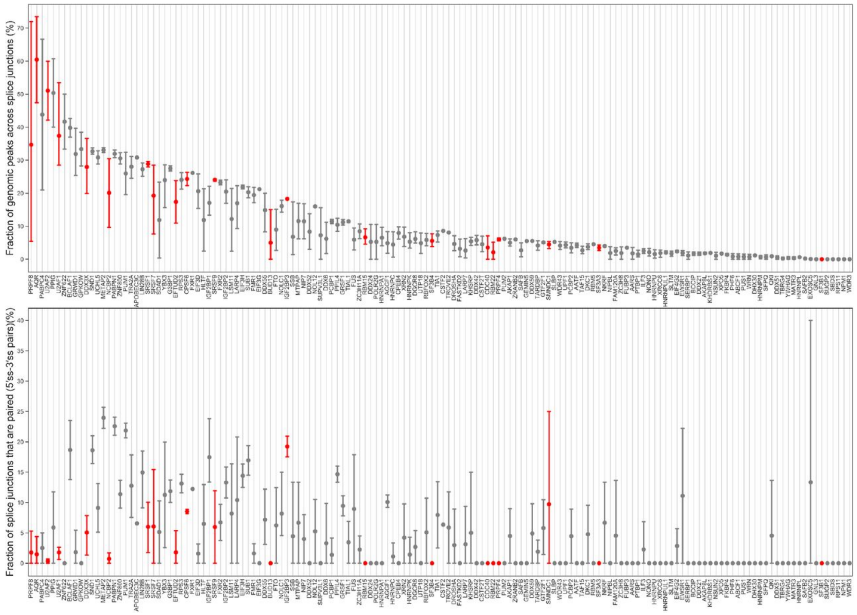
Supplementary Figure B.2: Agreements of peaks identified between individual ENCODE samples. The agreement is calculated as the Jaccard distance of the nucleotides in the peaks, where the intersection of two sets of peaks is the number of nts covered in both sets, while the union is the number of nts covered in at least one of the two sets. The color range is capped at a similarity of 0.4 to make the clusters more easily distinguishable. The top peaks are taken according to the FDR threshold (0.1), extending by 20 nts upstream and downstream from the crosslink site. Only samples with more than 100 peaks are included in this plot. The membership of RBPs in complexes is taken into account (based on CORUM [212]), by multiplying the value of the agreement by 1 if two proteins are known to participate in the same complex, and - 1 otherwise. Resulting negative values are shown in blue, while positive values are shown in red. That is, shown in red are peak agreements of samples that either correspond to the same protein, or to proteins known to interact with each other in complexes. On the right, a few clusters of samples containing proteins that are known to take part in complexes are highlighted.



Supplementary Figure B.3: Similarity of de novo predicted motifs of different RBPs. Clustermap of de novo motif similarity of all RBPs covered in the ENCODE dataset. Similarity is calculated as described in the methods section (Calculation of motif similarity), taking for each RBP the motif that best explains all of the samples corresponding to that RBP. On the left, a representation of the nucleotide composition of the motif is shown, each column corresponding to one nucleotide, while the color indicates the relative frequency of that nucleotide averaged over all positions of the PWM. The RBP names are colored according to whether they have a motif (red) or not (blue) in ATtRACT.



Supplementary Figure B.4: RCRUNCH results for all ENCODE eCLIP data currently available. a. Cumulative distribution of the number of significant binding sites detected per experiment (FDR threshold=0.1, in black). Cumulative distributions are also shown separately for samples corresponding to proteins with a known binding motif (gray) and to proteins for which no known motif is available in ATtRACT, but one was found by RCRUNCH (blue). b. Left: Distribution of enrichment scores for the de novo identified motifs in RCRUNCH crosslink (blue) and RCRUNCH peak center peaks, for RBPs that are not represented in ATtRACT. Right: Venn diagram illustrating the type of motifs identified across samples. We distinguished three categories of samples: for which (1) no significant or fewer than 20 significant peaks were found and thus no motif was searched/reported, (2) only a de novo motif was found and (3) a de novo motif was found and a motif is already known. c. Enrichment of the de novo motif predicted in the RCRUNCH-identified peaks from each sample, versus the enrichment score of the known motif for the protein assayed in the respective experiment. Marginal distributions of the known motif enrichments in RCRUNCH crosslink (blue) and RCRUNCH peak center (red) peaks are also shown. d. Cumulative density function of pairwise similarity scores for random pairs of known RBP-binding motifs (gray), known and de novo motifs identified from RCRUNCH crosslink sites (blue), known and de novo motifs identified for RCRUNCH peak center sites of individual RBPs (red). The same known motif was used for a given protein. e. Relationship between the similarity of de novo motifs inferred from replicate experiments and the minimum number of binding sites identified in these replicates. Experiments (each corresponding to an RBP and cell line) are colored according to whether (green) or not (blue) a motif was found in ATtRACT for the assayed RBP. f. Relationship between the similarity of de novo motifs identified in replicate experiments for a given RBP and the similarities of these de novo motifs and the known motif of the corresponding RBP. Lines connect replicate samples.



Supplementary Figure B.5: Binding events spanning splice junctions. a. Scatterplot of average fraction of peaks predicted by RCRUNCH (genomic approach) that overlap with annotated splice junctions. The samples are from ENCODE and the average is for sample replicates and cell lines. b. Scatterplot showing the fraction of the peaks that overlap a junction (e.g 3'ss) and also have a "pair" peak which overlaps the other side of the same junction (5'ss).

C

PUBLICATIONS AND CONTRIBUTION

This PhD thesis is based on the following publications:

1. **An automated workflow for processing of RNA-seq data**
   Maria Katsantoni, Foivos Gypas, Christina J. Herrmann, Dominik Burri, Maciej Bak, Paula Iborra, Krish Agarwal, Meric Ataman, Anastasiya Börsch, Mihaela Zavolan, Alexander Kanitz
   2021 (in preparation)

2. **Improved analysis of (e)CLIP data with RCRUNCH yields a compendium of RNA-binding protein binding sites and motifs**
   Maria Katsantoni, Erik van Nimwegen, Mihaela Zavolan (under review)

BIBLIOGRAPHY

[1] F Crick. "Central dogma of molecular biology." en. In: *Nature* 227.5258 (Aug. 1970), pp. 561–563.

[2] D S Latchman. "Transcription factors: an overview." en. In: *Int. J. Biochem. Cell Biol.* 29.12 (Dec. 1997), pp. 1305–1312.

[3] A P McMahon, T J Novak, R J Britten, and E H Davidson. "Inducible expression of a cloned heat shock fusion gene in sea urchin embryos." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 81.23 (Dec. 1984), pp. 7490–7494.

[4] H R Pelham and M Bienz. "A synthetic heat-shock promoter element confers heat-inducibility on the herpes simplex virus thymidine kinase gene." en. In: *EMBO J.* 1.11 (1982), pp. 1473–1477.

[5] R J Bandziulis, M S Swanson, and G Dreyfuss. *RNA-binding proteins as developmental regulators*. 1989.

[6] H Siomi and G Dreyfuss. "RNA-binding proteins as regulators of gene expression." en. In: *Curr. Opin. Genet. Dev.* 7.3 (June 1997), pp. 345–353.

[7] S A Adam, T Nakagawa, M S Swanson, T K Woodruff, and G Dreyfuss. "mRNA polyadenylate-binding protein: gene isolation and sequencing and identification of a ribonucleoprotein consensus sequence." en. In: *Mol. Cell. Biol.* 6.8 (Aug. 1986), pp. 2932–2943.

[8] M S Swanson, T Y Nakagawa, K LeVan, and G Dreyfuss. "Primary structure of human nuclear ribonucleoprotein particle C proteins: conservation of sequence and domain structures in heterogeneous nuclear RNA, mRNA, and pre-rRNA-binding proteins." en. In: *Mol. Cell. Biol.* 7.5 (May 1987), pp. 1731–1739.

[9] Matthias W Hentze, Alfredo Castello, Thomas Schwarzl, and Thomas Preiss. "A brave new world of RNA-binding proteins." en. In: *Nat. Rev. Mol. Cell Biol.* 19.5 (May 2018), pp. 327–341.

[10] Stefanie Gerstberger, Markus Hafner, and Thomas Tuschl. "A census of human RNA-binding proteins." en. In: *Nat. Rev. Genet.* 15.12 (Dec. 2014), pp. 829–845.

[11] Geoffrey M Cooper. "Regulation of Protein Function." In: *The Cell: A Molecular Approach. 2nd edition*. Sinauer Associates, 2000.

[12] A Kumar, K R Williams, and W Szer. "Purification and domain structure of core hnRNP proteins A1 and A2 and their relationship to single-stranded DNA-binding proteins." en. In: *J. Biol. Chem.* 261.24 (Aug. 1986), pp. 11266–11273.

[13] Stefanie Gerstberger, Markus Hafner, Manuel Ascano, and Thomas Tuschl. "Evolutionary conservation and expression of human RNA-binding proteins and their role in human genetic disease." en. In: *Adv. Exp. Med. Biol.* 825 (2014), pp. 1–55.

[14] Fátima Gebauer, Thomas Schwarzl, Juan Valcárcel, and Matthias W Hentze. "RNA-binding proteins in human genetic disease." en. In: *Nat. Rev. Genet.* 22.3 (Mar. 2021), pp. 185–198.

[15] Kiven E Lukong, Kai-Wei Chang, Edouard W Khandjian, and Stéphane Richard. "RNA-binding proteins in human genetic disease." en. In: *Trends Genet.* 24.8 (Aug. 2008), pp. 416–425.

[16] Yaseswini Neelamraju, Seyedsasan Hashemikhabir, and Sarath Chandra Janga. "The human RBPome: from genes and proteins to human disease." en. In: *J. Proteomics* 127.Pt A (Sept. 2015), pp. 61–70.

[17] Thomas Conrad, Anne-Susann Albrecht, Veronica Rodrigues de Melo Costa, Sascha Sauer, David Meierhofer, and Ulf Andersson Ørom. "Serial interactome capture of the human cell nucleus." en. In: *Nat. Commun.* 7 (Apr. 2016), p. 11212.

[18] Alexander G Baltz et al. "The mRNA-bound proteome and its global occupancy profile on protein-coding transcripts." en. In: *Mol. Cell* 46.5 (June 2012), pp. 674–690.

[19] Alfredo Castello et al. "Insights into RNA biology from an atlas of mammalian mRNA-binding proteins." en. In: *Cell* 149.6 (June 2012), pp. 1393–1406.

[20] Katharina Kramer, Timo Sachsenberg, Benedikt M Beckmann, Saadia Qamar, Kum-Loong Boon, Matthias W Hentze, Oliver Kohlbacher, and Henning Urlaub. "Photo-cross-linking and high-resolution mass spectrometry for assignment of RNA-binding sites in RNA-binding proteins." en. In: *Nat. Methods* 11.10 (Oct. 2014), pp. 1064–1070.

[21] Jaina Mistry et al. "Pfam: The protein families database in 2021." en. In: *Nucleic Acids Res.* 49.D1 (Jan. 2021), pp. D412–D419.

[22] Kristopher W Brannan et al. "SONAR Discovers RNA-Binding Proteins from Analysis of Large-Scale Protein-Protein Interactomes." en. In: *Mol. Cell* 64.2 (Oct. 2016), pp. 282–293.

[23] Christophe Maris, Cyril Dominguez, and Frédéric H-T Allain. "The RNA recognition motif, a plastic RNA-binding platform to regulate post-transcriptional gene expression." en. In: *FEBS J.* 272.9 (May 2005), pp. 2118–2131.

[24] Meredith Corley, Margaret C Burns, and Gene W Yeo. "How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms." en. In: *Mol. Cell* 78.1 (Apr. 2020), pp. 9–29.

[25] Y Shamoo, N Abdul-Manan, and K R Williams. "Multiple RNA binding domains (RBDs) just don't add up." en. In: *Nucleic Acids Res.* 23.5 (Mar. 1995), pp. 725–728.

[26] Stephanie Helder, Amanda J Blythe, Charles S Bond, and Joel P Mackay. "Determinants of affinity and specificity in RNA-binding proteins." en. In: *Curr. Opin. Struct. Biol.* 38 (June 2016), pp. 83–91.

[27] Bradley M Lunde, Claire Moore, and Gabriele Varani. *RNA-binding proteins: modular design for efficient function*. 2007.

[28] Antoine Cléry, Markus Blatter, and Frédéric H-T Allain. "RNA recognition motifs: boring? Not quite." en. In: *Curr. Opin. Struct. Biol.* 18.3 (June 2008), pp. 290–298.

[29] Sara Calabretta and Stéphane Richard. "Emerging Roles of Disordered Sequences in RNA-Binding Proteins." en. In: *Trends Biochem. Sci.* 40.11 (Nov. 2015), pp. 662–672.

[30] Palaniraja Thandapani, Timothy R O'Connor, Timothy L Bailey, and Stéphane Richard. "Defining the RGG/RG motif." en. In: *Mol. Cell* 50.5 (June 2013), pp. 613–623.

[31] Ilmin Kwon, Masato Kato, Siheng Xiang, Leeju Wu, Pano Theodoropoulos, Hamid Mirzaei, Tina Han, Shanhai Xie, Jeffry L Corden, and Steven L McKnight. "Phosphorylation-regulated binding of RNA polymerase II to fibrous polymers of low-complexity domains." en. In: *Cell* 155.5 (Nov. 2013), pp. 1049–1060.

[32] Daniel Benhalevy, Dimitrios G Anastasakis, and Markus Hafner. "Proximity-CLIP provides a snapshot of protein-occupied RNA elements in subcellular compartments." en. In: *Nat. Methods* 15.12 (Dec. 2018), pp. 1074–1082.

[33] Marina Feric, Nilesh Vaidya, Tyler S Harmon, Diana M Mitrea, Lian Zhu, Tiffany M Richardson, Richard W Kriwacki, Rohit V Pappu, and Clifford P Brangwynne. "Coexisting Liquid Phases Underlie Nucleolar Subcompartments." en. In: *Cell* 165.7 (June 2016), pp. 1686–1697.

[34] Salman F Banani, Hyun O Lee, Anthony A Hyman, and Michael K Rosen. "Biomolecular condensates: organizers of cellular biochemistry." en. In: *Nat. Rev. Mol. Cell Biol.* 18.5 (May 2017), pp. 285–298.

[35] Bin Wang, Lei Zhang, Tong Dai, Ziran Qin, Huasong Lu, Long Zhang, and Fangfang Zhou. "Liquid–liquid phase separation in human health and diseases." en. In: *Signal Transduction and Targeted Therapy* 6.1 (Aug. 2021), pp. 1–16.

[36] Titus M Franzmann and Simon Alberti. "Prion-like low-complexity sequences: Key regulators of protein solubility and phase behavior." en. In: *J. Biol. Chem.* 294.18 (May 2019), pp. 7128–7136.

[37] Yongdae Shin and Clifford P Brangwynne. "Liquid phase condensation in cell physiology and disease." en. In: *Science* 357.6357 (Sept. 2017).

[38] Nicole J Curtis and Constance J Jeffery. "The expanding world of metabolic enzymes moonlighting as RNA binding proteins." en. In: *Biochem. Soc. Trans.* 49.3 (June 2021), pp. 1099–1108.

[39] David G Hendrickson, David R Kelley, Danielle Tenen, Bradley Bernstein, and John L Rinn. "Widespread RNA binding by chromatin-associated proteins." en. In: *Genome Biol.* 17 (Feb. 2016), p. 28.

[40] Mahmoud-Reza Rafiee, Julian A Zagalak, Sviatoslav Sidorov, Sebastian Steinhauser, Karen Davey, Jernej Ule, and Nicholas M Luscombe. "Chromatin-contact atlas reveals disorder-mediated protein interactions and moonlighting chromatin-associated RBPs." en. In: *Nucleic Acids Res.* 49.22 (Dec. 2021), pp. 13092–13107.

[41] Marieke van Kouwenhove, Martijn Kedde, and Reuven Agami. "MicroRNA regulation by RNA-binding proteins and its implications for cancer." en. In: *Nat. Rev. Cancer* 11.9 (Aug. 2011), pp. 644–656.

[42] Sukjun Kim et al. "The regulatory impact of RNA-binding proteins on microRNA targeting." en. In: *Nat. Commun.* 12.1 (Aug. 2021), p. 5057.

[43] P Trifillis, N Day, and M Kiledjian. "Finding the right RNA: identification of cellular mRNA substrates for RNA-binding proteins." en. In: *RNA* 5.8 (Aug. 1999), pp. 1071–1082.

[44] S A Brooks and W F Rigby. "Characterization of the mRNA ligands bound by the RNA binding protein hnRNP A2 utilizing a novel in vivo technique." en. In: *Nucleic Acids Res.* 28.10 (May 2000), E49.

[45] S A Tenenbaum, C C Carson, P J Lager, and J D Keene. "Identifying mRNA subsets in messenger ribonucleoprotein complexes by using cDNA arrays." en. In: *Proc. Natl. Acad. Sci. U. S. A.* 97.26 (Dec. 2000), pp. 14085–14090.

[46] Somashe Niranjanakumari, Erika Lasda, Robert Brazas, and Mariano A Garcia-Blanco. "Reversible cross-linking combined with immunoprecipitation to study RNA-protein interactions in vivo." en. In: *Methods* 26.2 (Feb. 2002), pp. 182–190.

[47] Stavroula Mili and Joan A Steitz. "Evidence for reassociation of RNA-binding proteins after cell lysis: implications for the interpretation of immunoprecipitation analyses." en. In: *RNA* 10.11 (Nov. 2004), pp. 1692–1694.

[48] Jernej Ule, Kirk Jensen, Aldo Mele, and Robert B Darnell. "CLIP: A method for identifying protein–RNA interaction sites in living cells." In: *Methods* 37.4 (Dec. 2005), pp. 376–386.

[49] Elizabeth A Hoffman, Brian L Frey, Lloyd M Smith, and David T Auble. "Formaldehyde crosslinking: a tool for the study of chromatin complexes." en. In: *J. Biol. Chem.* 290.44 (Oct. 2015), pp. 26404–26411.

[50] Yoichiro Sugimoto, Julian König, Shobbir Hussain, Blaž Zupan, Tomaž Curk, Michaela Frye, and Jernej Ule. "Analysis of CLIP and iCLIP methods for nucleotide-resolution studies of protein-RNA interactions." en. In: *Genome Biol.* 13.8 (Aug. 2012), R67.

[51] Henning Urlaub, Klaus Hartmuth, and Reinhard Lührmann. "A two-tracked approach to analyze RNA-protein crosslinking sites in native, nonlabeled small nuclear ribonucleoprotein particles." en. In: *Methods* 26.2 (Feb. 2002), pp. 170–181.

[52] Kirk B Jensen and Robert B Darnell. "CLIP: Crosslinking and ImmunoPrecipitation of In Vivo RNA Targets of RNA-Binding Proteins." en. In: *RNA-Protein Interaction Protocols* (2008), pp. 85–98.

[53] Chaolin Zhang and Robert B Darnell. "Mapping in vivo protein-RNA interactions at single-nucleotide resolution from HITS-CLIP data." en. In: *Nat. Biotechnol.* 29.7 (June 2011), pp. 607–614.

[54] John R Lorsch, David P Bartel, and Jack W Szostak. "Reverse transcriptase reads through a 2–5 linkage and a 2-thiphosphate in a template." en. In: *Nucleic Acids Res.* 23.15 (Aug. 1995), pp. 2811–2814.

[55] Julian Konig, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. "iCLIP–transcriptome-wide mapping of protein-RNA interactions with individual nucleotide resolution." en. In: *J. Vis. Exp.* 50 (Apr. 2011).

[56] Andreas Buchbender, Holger Mutter, F X Reymond Sutandy, Nadine Körtel, Heike Hänel, Anke Busch, Stefanie Ebersberger, and Julian König. "Improved library preparation with the new iCLIP2 protocol." en. In: *Methods* 178 (June 2020), pp. 33–48.

[57] Eric L Van Nostrand et al. "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)." en. In: *Nat. Methods* 13.6 (June 2016), pp. 508–514.

[58] Brian J Zarnegar, Ryan A Flynn, Ying Shen, Brian T Do, Howard Y Chang, and Paul A Khavari. "irCLIP platform for efficient characterization of protein-RNA interactions." en. In: *Nat. Methods* 13.6 (June 2016), pp. 489–492.

[59] Jessica Spitzer et al. "PAR-CLIP (Photoactivatable Ribonucleoside-Enhanced Crosslinking and Immunoprecipitation): a step-by-step protocol to the transcriptome-wide identification of binding sites of RNA-binding proteins." en. In: *Methods Enzymol.* 539 (2014), pp. 113–161.

[60] Ronja Weissinger, Lisa Heinold, Saira Akram, Ralf-Peter Jansen, and Orit Hermesh. "RNA Proximity Labeling: A New Detection Tool for RNA-Protein Interactions." en. In: *Molecules* 26.8 (Apr. 2021).

[61] G Dorn, A Leitner, J Boudet, S Campagne, C von Schroetter, A Moursy, R Aebersold, and F H-T Allain. "Structural modeling of protein-RNA complexes using crosslinking of segmentally isotope-labeled RNA and MS/MS." en. In: *Nat. Methods* 14.5 (May 2017), pp. 487–490.

[62] Aoife C McMahon, Reazur Rahman, Hua Jin, James L Shen, Allegra Fieldsend, Weifei Luo, and Michael Rosbash. "TRIBE: Hijacking an RNA-Editing Enzyme to Identify Cell-Specific Targets of RNA-Binding Proteins." en. In: *Cell* 165.3 (Apr. 2016), pp. 742–753.

[63] N Navaratnam, J R Morrison, S Bhattacharya, D Patel, T Funahashi, F Giannoni, B B Teng, N O Davidson, and J Scott. "The p27 catalytic subunit of the apolipoprotein B mRNA editing enzyme is a cytidine deaminase." en. In: *J. Biol. Chem.* 268.28 (Oct. 1993), pp. 20709–20712.

[64] Kate D Meyer. "DART-seq: an antibody-free method for global mA detection." en. In: *Nat. Methods* 16.12 (Dec. 2019), pp. 1275–1280.

[65] Kristopher W Brannan et al. "Robust single-cell discovery of RNA targets of RNA-binding proteins and ribosomes." en. In: *Nat. Methods* 18.5 (May 2021), pp. 507–519.

[66] Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. "CLIP and complementary methods." en. In: *Nature Reviews Methods Primers* 1.1 (Mar. 2021), pp. 1–23.

[67] Zhen Wang, Melis Kayikci, Michael Briese, Kathi Zarnack, Nicholas M Luscombe, Gregor Rot, Blaž Zupan, Tomaž Curk, and Jernej Ule. "iCLIP Predicts the Dual Splicing Effects of TIA-RNA Interactions." In: *PLoS Biol.* 8.10 (Oct. 2010), e1000530.

[68] Klara Kuret, Aram Gustav Amalietti, D Marc Jones, Charlotte Capitanchik, and Jernej Ule. "Positional motif analysis reveals the extent of specificity of protein-RNA interactions observed by CLIP." en. In: *Genome Biol.* 23.1 (Sept. 2022), p. 191.

[69] Philip J Uren, Emad Bahrami-Samani, Suzanne C Burns, Mei Qiao, Fedor V Karginov, Emily Hodges, Gregory J Hannon, Jeremy R Sanford, Luiz O F Penalva, and Andrew D Smith. "Site identification in high-throughput RNA-protein interaction data." en. In: *Bioinformatics* 28.23 (Dec. 2012), pp. 3013–3020.

[70] Alper Kucukural, Hakan Özadam, Guramrit Singh, Melissa J Moore, and Can Cenik. "ASPeak: an abundance sensitive peak detection algorithm for RIP-Seq." en. In: *Bioinformatics* 29.19 (Oct. 2013), pp. 2485–2486.

[71] Michael T Lovci et al. "Rbfox proteins regulate alternative mRNA splicing through evolutionarily conserved RNA bridges." en. In: *Nat. Struct. Mol. Biol.* 20.12 (Dec. 2013), pp. 1434–1442.

[72] ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." en. In: *Nature* 489.7414 (Sept. 2012), pp. 57–74.

[73] Beibei Chen, Jonghyun Yun, Min Soo Kim, Joshua T Mendell, and Yang Xie. "PIPE-CLIP: a comprehensive online tool for CLIP-seq data analysis." en. In: *Genome Biol.* 15.1 (Jan. 2014), R18.

[74] Timothy L Bailey, James Johnson, Charles E Grant, and William S Noble. *The MEME Suite*. 2015.

[75] Federico Comoglio, Cem Sievers, and Renato Paro. "Sensitive and highly resolved identification of RNA-protein interaction sites in PAR-CLIP data." en. In: *BMC Bioinformatics* 16 (Feb. 2015), p. 32.

[76] Jai J Tree, Sander Granneman, Sean P McAteer, David Tollervey, and David L Gally. "Identification of bacteriophage-encoded anti-sRNAs in pathogenic Escherichia coli." en. In: *Mol. Cell* 55.2 (July 2014), pp. 199–213.

[77] Monica Golumbeanu, Pejman Mohammadi, and Niko Beerenwinkel. "BMix: probabilistic modeling of occurring substitutions in PAR-CLIP data." en. In: *Bioinformatics* 32.7 (Apr. 2016), pp. 976–983.

[78] Zijun Zhang and Yi Xing. "CLIP-seq analysis of multi-mapped reads discovers novel functional RNA regulatory sites in the human transcriptome." en. In: *Nucleic Acids Res.* 45.16 (Sept. 2017), pp. 9260–9271.

[79] Ankeeta Shah, Yingzhi Qian, Sebastien M Weyn-Vanhentenryck, and Chaolin Zhang. "CLIP Tool Kit (CTK): a flexible and robust pipeline to analyze CLIP sequencing data." en. In: *Bioinformatics* 33.4 (Feb. 2017), pp. 566–567.

[80] Michael J Moore, Chaolin Zhang, Emily Conn Gantman, Aldo Mele, Jennifer C Darnell, and Robert B Darnell. "Mapping Argonaute and conventional RNA-binding protein interactions with RNA at single-nucleotide resolution using HITS-CLIP and CIMS analysis." en. In: *Nat. Protoc.* 9.2 (Feb. 2014), pp. 263–293.

[81] Sander Granneman, Grzegorz Kudla, Elisabeth Petfalski, and David Tollervey. *Identification of protein binding sites on U3 snoRNA and pre-rRNA by UV cross-linking and high-throughput analysis of cDNAs*. 2009.

[82] Sabrina Krakau, Hugues Richard, and Annalisa Marsico. "PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data." en. In: *Genome Biol.* 18.1 (Dec. 2017), p. 240.

[83] Lance E Palmer, Mitchell J Weiss, and Vikram R Paralkar. "YODEL: Peak calling software for HITS-CLIP data." en. In: *F1000Res.* 6.1138 (July 2017), p. 1138.

[84] Sihyung Park, Seung Hyun Ahn, Eun Sol Cho, You Kyung Cho, Eun-Sook Jang, and Sung Wook Chi. "CLIPick: a sensitive peak caller for expression-based deconvolution of HITS-CLIP signals." en. In: *Nucleic Acids Res.* 46.21 (Nov. 2018), pp. 11153–11168.

[85] Philipp Drewe-Boss, Hans-Hermann Wessels, and Uwe Ohler. *omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data*. 2018.

[86] Huijuan Feng, Suying Bao, Mohammad Alinoor Rahman, Sebastien M Weyn-Vanhentenryck, Aziz Khan, Justin Wong, Ankeeta Shah, Elise D Flynn, Adrian R Krainer, and Chaolin Zhang. "Modeling RNA-Binding Protein Specificity In Vivo by Precisely Registering Protein-RNA Crosslink Sites." en. In: *Mol. Cell* 74.6 (June 2019), 1189–1204.e6.

[87] Ravinder Singh and Juan Valcárcel. "Building specificity with nonspecific RNA-binding proteins." en. In: *Nat. Struct. Mol. Biol.* 12.8 (Aug. 2005), pp. 645–653.

[88] Laura Wratten, Andreas Wilm, and Jonathan Göke. "Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers." en. In: *Nat. Methods* 18.10 (Oct. 2021), pp. 1161–1168.

[89] Felix Mölder et al. "Sustainable data analysis with Snakemake." en. In: *F1000Res.* 10 (Jan. 2021), p. 33.

[90] *Website*. https://doi.org/10.1002/cpe.5802. Accessed: NA-NA-NA.

[91] Carole Goble, Sarah Cohen-Boulakia, Stian Soiland-Reyes, Daniel Garijo, Yolanda Gil, Michael R Crusoe, Kristian Peters, and Daniel Schober. "FAIR Computational Workflows." en. In: *Data Intelligence* 2.1-2 (Jan. 2020), pp. 108–121.

[92] Anna-Lena Lamprecht et al. "Towards FAIR principles for research software." In: *Data sci.* 3.1 (June 2020), pp. 37–59.

[93] Jingwen Bai, Chakradhar Bandla, Jiaxin Guo, Roberto Vera Alvarez, Mingze Bai, Juan Antonio Vizcaíno, Pablo Moreno, Björn Grüning, Olivier Sallou, and Yasset Perez-Riverol. "Bio-Containers Registry: Searching Bioinformatics and Proteomics Tools, Packages, and Containers." en. In: *J. Proteome Res.* 20.4 (Apr. 2021), pp. 2056–2061.

[94] Levin Clément, Dynomant Emeric, Gonzalez Bruno J, Mouchard Laurent, Landsman David, Hovig Eivind, and Vlahovicek Kristian. "A data-supported history of bioinformatics tools." In: (July 2018). arXiv: 1807.06808 [cs.DL].

[95] Alexander Kanitz, Foivos Gypas, Andreas J Gruber, Andreas R Gruber, Georges Martin, and Mihaela Zavolan. "Comparative assessment of methods for the computational inference of transcript isoform abundance from RNA-seq data." en. In: *Genome Biol.* 16 (July 2015), p. 150.

[96] Mohsen Khorshid, Christoph Rodak, and Mihaela Zavolan. "CLIPZ: a database and analysis environment for experimentally determined binding sites of RNA-binding proteins." en. In: *Nucleic Acids Res.* 39.Database issue (Jan. 2011), pp. D245–52.

[97] Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. "CLIP and complementary methods." en. In: *Nature Reviews Methods Primers* 1.1 (Mar. 2021), pp. 1–23.

[98] Andreas J Gruber, Foivos Gypas, Andrea Riba, Ralf Schmidt, and Mihaela Zavolan. "Terminal exon characterization with TECtool reveals an abundance of cell-specific isoforms." en. In: *Nat. Methods* (Sept. 2018).

[99] Christina J Herrmann, Ralf Schmidt, Alexander Kanitz, Panu Artimo, Andreas J Gruber, and Mihaela Zavolan. "PolyASite 2.0: a consolidated atlas of polyadenylation sites from 3' end sequencing." en. In: *Nucleic Acids Res.* 48.D1 (Oct. 2019), pp. D174–D179.

[100] Jérémie Breda, Mihaela Zavolan, and Erik van Nimwegen. "Bayesian inference of gene expression states from single-cell RNA-seq data." en. In: *Nat. Biotechnol.* (Apr. 2021).

[101] Charlotte Soneson, Yao Yao, Anna Bratus-Neuenschwander, Andrea Patrignani, Mark D Robinson, and Shobbir Hussain. "A comprehensive examination of Nanopore native RNA sequencing for characterization of complex transcriptomes." en. In: *Nat. Commun.* 10.1 (July 2019), p. 3359.

[102] E D Karousis, F Gypas, M Zavolan, and O Mühlemann. "Nanopore sequencing reveals endogenous NMD-targeted isoforms in human cells." In: *bioRxiv* (2021).

[103] Peter J A Cock, Christopher J Fields, Naohisa Goto, Michael L Heuer, and Peter M Rice. "The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants." In: *Nucleic Acids Res.* 38.6 (Dec. 2009), pp. 1767–1771.

[104] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. "The Sequence Alignment/Map format and SAMtools." In: *Bioinformatics* 25.16 (Aug. 2009), pp. 2078–2079.

[105] Jeffrey M Perkel. "Workflow systems turn raw data into scientific knowledge." en. In: *Nature* 573.7772 (Sept. 2019), pp. 149–150.

[106] Laura Wratten, Andreas Wilm, and Jonathan Goke. "Reproducible, scalable, and shareable analysis pipelines with bioinformatics workflow managers." en. In: *Nat. Methods* (Sept. 2021), pp. 1–8.

[107] J Köster and S Rahmann. "Snakemake—a scalable bioinformatics workflow engine." In: *Bioinformatics* (2012).

[108] Felix Mölder et al. "Sustainable data analysis with Snakemake." en. In: *F1000Res.* 10 (Jan. 2021), p. 33.

[109] Paolo Di Tommaso, Maria Chatzou, Evan W Floden, Pablo Prieto Barja, Emilio Palumbo, and Cedric Notredame. "Nextflow enables reproducible computational workflows." en. In: *Nat. Biotechnol.* 35.4 (Apr. 2017), pp. 316–319.

[110] Peter Amstutz et al. *Common Workflow Language, v1.0*. July 2016.

[111] Dirk Merkel and Others. "Docker: lightweight linux containers for consistent development and deployment." In: *Linux J.* 2014.239 (2014), p. 2.

[112] Gregory M Kurtzer, Vanessa Sochat, and Michael W Bauer. "Singularity: Scientific containers for mobility of compute." en. In: *PLoS One* 12.5 (May 2017), e0177459.

[113] *Anaconda Documentation — Anaconda documentation*. https://docs.anaconda.com. Accessed: 2021-8-23.

[114] Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A Chapman, Jillian Rowe, Christopher H Tomkins-Tinch, Renan Valieris, Johannes Köster, and Bioconda Team. "Bioconda: sustainable and comprehensive software distribution for the life sciences." en. In: *Nat. Methods* 15.7 (July 2018), pp. 475–476.

[115] Ana Conesa et al. "A survey of best practices for RNA-seq data analysis." en. In: *Genome Biol.* 17 (Jan. 2016), p. 13.

[116] Maria Katsantoni et al. *ZARP: An automated workflow for processing of RNA-seq data*. 2021.

[117] Simon Andrews. *FastQC: a quality control tool for high throughput sequence data*. 2010.

[118] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads." In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12.

[119] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. "STAR: ultrafast universal RNA-seq aligner." In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21.

[120] Liguo Wang et al. "Measure transcript integrity using RNA-seq data." en. In: *BMC Bioinformatics* 17 (Feb. 2016), p. 58.

[121] Mathieu Bahin, Benoit F Noël, Valentine Murigneux, Charles Bernard, Leila Bastianelli, Hervé Le Hir, Alice Lebreton, and Auguste Genovesio. "ALFA: annotation landscape for aligned reads." en. In: *BMC Genomics* 20.1 (Mar. 2019), p. 250.

[122] Nicolas L Bray, Harold Pimentel, Páll Melsted, and Lior Pachter. "Near-optimal probabilistic RNA-seq quantification." en. In: *Nat. Biotechnol.* 34.5 (May 2016), pp. 525–527.

[123] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. “Salmon provides fast and bias-aware quantification of transcript expression.” en. In: *Nat. Methods* (Mar. 2017).

[124] *zpca: PCA analysis*.

[125] Philip Ewels, Måns Magnusson, Sverker Lundin, and Max Käller. “MultiQC: summarize analysis results for multiple tools and samples in a single report.” en. In: *Bioinformatics* 32.19 (Oct. 2016), pp. 3047–3048.

[126] Yuval Benjamini and Terence P Speed. “Summarizing and correcting the GC content bias in high-throughput sequencing.” en. In: *Nucleic Acids Res.* 40.10 (May 2012), e72.

[127] Giacomo Baruzzo, Katharina E Hayer, Eun Ji Kim, Barbara Di Camillo, Garret A FitzGerald, and Gregory R Grant. “Simulation-based comprehensive benchmarking of RNA-seq aligners.” en. In: *Nat. Methods* 14.2 (Dec. 2016), pp. 135–139.

[128] Robert M Kuhn, David Haussler, and W James Kent. “The UCSC genome browser and associated tools.” en. In: *Brief. Bioinform.* 14.2 (Mar. 2013), pp. 144–161.

[129] *tin-score-calculation: Given a set of BAM files and a gene annotation BED file, calculates the Transcript Integrity Number (TIN) for each transcript*.

[130] Liguo Wang, Shengqin Wang, and Wei Li. “RSeQC: quality control of RNA-seq experiments.” en. In: *Bioinformatics* 28.16 (Aug. 2012), pp. 2184–2185.

[131] Mingxiang Teng et al. “A benchmark for RNA-seq quantification pipelines.” en. In: *Genome Biol.* 17 (Apr. 2016), p. 74.

[132] Günter P Wagner, Koryu Kin, and Vincent J Lynch. “Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples.” In: *Theory Biosci.* 131.4 (Dec. 2012), pp. 281–285.

[133] Ian Jolliffe. “Principal Component Analysis.” In: *Encyclopedia of Statistics in Behavioral Science*. Chichester, UK: John Wiley & Sons, Ltd, Oct. 2005.

[134] *merge_kallisto: Merge kallisto results from multiple runs*.

[135] Tanya Barrett et al. “NCBI GEO: archive for functional genomics data sets–update.” en. In: *Nucleic Acids Res.* 41.Database issue (Jan. 2013), pp. D991–5.

[136] Daniel J Ham et al. “The neuromuscular junction is a focal point of mTORC1 signaling in sarcopenia.” en. In: *Nat. Commun.* 11.1 (Sept. 2020), p. 4510.

[137] Kevin L Howe et al. “Ensembl 2021.” en. In: *Nucleic Acids Res.* 49.D1 (Jan. 2021), pp. D884–D891.

[138] Meric Ataman, Anastasiya Börsch, and Maciej Bak. *ZARP: Supplementary Materials*. 2021.

[139] *sciCORE*. http://scicore.unibas.ch/. Accessed: 2021-11-15.

[140] zavolanlab. *GitHub - zavolanlab/zarp: Zavolan-Lab Automated RNA-Seq Pipeline*. https://github.com/zavolanlab/zarp. Accessed: 2021-11-15.

[141] María Gabriela Thomas, Mariela Loschi, María Andrea Desbats, and Graciela Lidia Boccaccio. "RNA granules: the good, the bad and the ugly." en. In: *Cell. Signal.* 23.2 (Feb. 2011), pp. 324–334.

[142] Matthias W Hentze, Alfredo Castello, Thomas Schwarzl, and Thomas Preiss. "A brave new world of RNA-binding proteins." en. In: *Nat. Rev. Mol. Cell Biol.* 19.5 (May 2018), pp. 327–341.

[143] Bradley M Lunde, Claire Moore, and Gabriele Varani. "RNA-binding proteins: modular design for efficient function." en. In: *Nat. Rev. Mol. Cell Biol.* 8.6 (June 2007), pp. 479–490.

[144] Richard Stefl, Lenka Skrisovska, and Frédéric H-T Allain. "RNA sequence- and shape-dependent recognition by proteins in the ribonucleoprotein particle." en. In: *EMBO Rep.* 6.1 (Jan. 2005), pp. 33–38.

[145] Kiven E Lukong, Kai-Wei Chang, Edouard W Khandjian, and Stéphane Richard. "RNA-binding proteins in human genetic disease." en. In: *Trends Genet.* 24.8 (Aug. 2008), pp. 416–425.

[146] Jernej Ule, Kirk B Jensen, Matteo Ruggiu, Aldo Mele, Aljaz Ule, and Robert B Darnell. "CLIP identifies Nova-regulated RNA networks in the brain." In: *Science* 302.5648 (Nov. 2003), pp. 1212–1215.

[147] Julian König, Kathi Zarnack, Gregor Rot, Tomaz Curk, Melis Kayikci, Blaz Zupan, Daniel J Turner, Nicholas M Luscombe, and Jernej Ule. "iCLIP reveals the function of hnRNP particles in splicing at individual nucleotide resolution." en. In: *Nat. Struct. Mol. Biol.* 17.7 (July 2010), pp. 909–915.

[148] Shivendra Kishore, Lukasz Jaskiewicz, Lukas Burger, Jean Hausser, Mohsen Khorshid, and Mihaela Zavolan. "A quantitative analysis of CLIP methods for identifying binding sites of RNA-binding proteins." In: *Nat. Methods* 8.7 (July 2011), pp. 559–564.

[149] Markus Hafner, Maria Katsantoni, Tino Köster, James Marks, Joyita Mukherjee, Dorothee Staiger, Jernej Ule, and Mihaela Zavolan. "CLIP and complementary methods." In: *Nature Reviews Methods Primers* 1.1 (Mar. 2021), p. 20.

[150] Hyeongrin Jeon, Hyunji Lee, Byunghee Kang, Insoon Jang, and Tae-Young Roh. "Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis." en. In: *Genomics Inform.* 18.4 (Dec. 2020), e42.

[151] Severin Berger, Mikhail Pachkov, Phil Arnold, Saeed Omidi, Nicholas Kelley, Silvia Salatino, and Erik van Nimwegen. "Crunch: integrated processing and modeling of ChIP-seq data in terms of regulatory motifs." en. In: *Genome Res.* 29.7 (July 2019), pp. 1164–1177.

[152] Philipp Drewe-Boss, Hans-Hermann Wessels, and Uwe Ohler. "omniCLIP: probabilistic identification of protein-RNA interactions from CLIP-seq data." en. In: *Genome Biol.* 19.1 (Nov. 2018), p. 183.

[153] Matthew B Friedersdorf and Jack D Keene. "Advancing the functional utility of PAR-CLIP by quantifying background binding to mRNAs and lncRNAs." en. In: *Genome Biol.* 15.1 (Jan. 2014), R2.

[154] Meredith Corley, Margaret C Burns, and Gene W Yeo. "How RNA-Binding Proteins Interact with RNA: Molecules and Mechanisms." en. In: *Mol. Cell* 78.1 (Apr. 2020), pp. 9–29.

[155] Debashish Ray, Hilal Kazan, Esther T Chan, Lourdes Peña Castillo, Sidharth Chaudhry, Shaheynoor Talukder, Benjamin J Blencowe, Quaid Morris, and Timothy R Hughes. "Rapid and systematic analysis of the RNA recognition specificities of RNA-binding proteins." en. In: *Nat. Biotechnol.* 27.7 (July 2009), pp. 667–670.

[156] Nicole Lambert, Alex Robertson, Mohini Jangi, Sean McGeary, Phillip A Sharp, and Christopher B Burge. "RNA Bind-n-Seq: quantitative assessment of the sequence and structural binding specificity of RNA binding proteins." en. In: *Mol. Cell* 54.5 (June 2014), pp. 887–900.

[157] Eric L Van Nostrand et al. "Robust transcriptome-wide discovery of RNA-binding protein binding sites with enhanced CLIP (eCLIP)." en. In: *Nat. Methods* 13.6 (June 2016), pp. 508–514.

[158] Girolamo Giudice, Fátima Sánchez-Cabo, Carlos Torroja, and Enrique Lara-Pezzi. *ATtRACT—a database of RNA-binding proteins and associated motifs*. 2016.

[159] Anthony Mathelier et al. "JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles." en. In: *Nucleic Acids Res.* 44.D1 (Jan. 2016), pp. D110–5.

[160] Ivan V Kulakovskiy et al. "HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis." en. In: *Nucleic Acids Res.* 46.D1 (Jan. 2018), pp. D252–D259.

[161] Xiaoli Chen, Sarah A Castro, Qiuying Liu, Wenqian Hu, and Shaojie Zhang. "Practical considerations on performing and analyzing CLIP-seq experiments to identify transcriptomic-wide RNA-protein interactions." en. In: *Methods* 155 (Feb. 2019), pp. 49–57.

[162] Sabrina Krakau, Hugues Richard, and Annalisa Marsico. "PureCLIP: capturing target-specific protein-RNA interaction footprints from single-nucleotide CLIP-seq data." en. In: *Genome Biol.* 18.1 (Dec. 2017), p. 240.

[163] Phil Arnold, Ionas Erb, Mikhail Pachkov, Nacho Molina, and Erik van Nimwegen. "MotEvo: integrated Bayesian probabilistic methods for inferring regulatory sites and motifs on multiple alignments of DNA sequences." en. In: *Bioinformatics* 28.4 (Feb. 2012), pp. 487–494.

[164] *GitHub - zavolanlab/RCRUNCH: Workflow for automated (e)CLIP analysis. From raw fastq to peak calling and motif analysis*. en. https://github.com/zavolanlab/RCRUNCH. Accessed: 2022-7-5.

[165] J Koster and S Rahmann. *Snakemake–a scalable bioinformatics workflow engine*. 2012.

[166] Mark D Wilkinson et al. "The FAIR Guiding Principles for scientific data management and stewardship." en. In: *Sci Data* 3 (Mar. 2016), p. 160018.

[167] Sigrid D Auweter, Rudi Fasan, Luc Reymond, Jason G Underwood, Douglas L Black, Stefan Pitsch, and Frédéric H-T Allain. "Molecular basis of RNA recognition by the human alternative splicing factor Fox-1." en. In: *EMBO J.* 25.1 (Jan. 2006), pp. 163–173.

[168] Carrie A Davis et al. "The Encyclopedia of DNA elements (ENCODE): data portal update." en. In: *Nucleic Acids Res.* 46.D1 (Jan. 2018), pp. D794–D801.

[169] M Görlach, C G Burd, and G Dreyfuss. "The determinants of RNA-binding specificity of the heterogeneous nuclear ribonucleoprotein C proteins." en. In: *J. Biol. Chem.* 269.37 (Sept. 1994), pp. 23074–23078.

[170] Florian C Oberstrass et al. "Structure of PTB bound to RNA: specific binding and implications for splicing regulation." en. In: *Science* 309.5743 (Sept. 2005), pp. 2054–2057.

[171] E K White, T Moore-Jarrett, and H E Ruley. "PUM2, a novel murine puf protein, and its consensus RNA-binding site." en. In: *RNA* 7.12 (Dec. 2001), pp. 1855–1866.

[172] Xiaoqiang Wang, Juanita McLachlan, Phillip D Zamore, and Traci M Tanaka Hall. "Modular recognition of RNA by a human pumilio-homology domain." en. In: *Cell* 110.4 (Aug. 2002), pp. 501–512.

[173] Eric L Van Nostrand et al. "A large-scale binding and functional map of human RNA-binding proteins." en. In: *Nature* 583.7818 (July 2020), pp. 711–719.

[174] Rahul Siddharthan, Eric D Siggia, and Erik van Nimwegen. "PhyloGibbs: a Gibbs sampling motif finder that incorporates phylogeny." In: *PLoS Comput. Biol.* 1.7 (Dec. 2005), e67.

[175] Yoshinori Kariya, Kikuya Kato, Yoshihide Hayashizaki, Seiichi Himeno, Seiichiro Tarui, and Kenichi Matsubar. "Revision of consensus sequence of human Alu repeats—a review." In: *Gene* 53.1 (1987), pp. 1–10.

[176] Anna Knörlein, Chris Sarnowski, Tebbe de Vries, Moritz Stoltz, Michael Götze, Ruedi Aebersold, Frédéric Allain, Alexander Leitner, and Jonathan Hall. "Structural requirements for photo-induced RNA-protein cross-linking." en. In: *ChemRxiv* (June 2021).

[177] Qi Liu, Xue Zhong, Blair B Madison, Anil K Rustgi, and Yu Shyr. "Assessing Computational Steps for CLIP-Seq Data Analysis." en. In: *Biomed Res. Int.* 2015 (Oct. 2015), p. 196082.

[178] Chandani Warnasooriya, Callen F Feeney, Kholiswa M Laird, Dmitri N Ermolenko, and Clara L Kielkopf. "A splice site-sensing conformational switch in U2AF2 is modulated by U2AF1 and its recurrent myelodysplasia-associated mutation." en. In: *Nucleic Acids Res.* 48.10 (June 2020), pp. 5695–5709.

[179] Eric L Van Nostrand et al. "Principles of RNA processing from analysis of enhanced CLIP maps for 150 RNA binding proteins." en. In: *Genome Biol.* 21.1 (Apr. 2020), p. 90.

[180] Hana Antonicka and Eric A Shoubridge. "Mitochondrial RNA Granules Are Centers for Posttranscriptional RNA Processing and Ribosome Biogenesis." en. In: *Cell Rep.* 10.6 (Feb. 2015), pp. 920–932.

[181] Richard I Gregory, Kai-Ping Yan, Govindasamy Amuthan, Thimmaiah Chendrimada, Behzad Doratotaj, Neil Cooch, and Ramin Shiekhattar. "The Microprocessor complex mediates the genesis of microRNAs." en. In: *Nature* 432.7014 (Nov. 2004), pp. 235–240.

[182] *ENCORE Matrix*. en. https://www.encodeproject.org/encore-matrix/?type=Experiment&status=released&internal_tags=ENCORE. Accessed: 2022-7-5.

[183] Mohan T Bolisetty and Karen L Beemon. "Splicing of internal large exons is defined by novel cis-acting sequence elements." en. In: *Nucleic Acids Res.* 40.18 (Oct. 2012), pp. 9244–9254.

[184] Michael Uhl, Van Dinh Tran, and Rolf Backofen. "Improving CLIP-seq data analysis by incorporating transcript information." en. In: *BMC Genomics* 21.1 (Dec. 2020), p. 894.

[185] Alexander Dobin, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. "STAR: ultrafast universal RNA-seq aligner." en. In: *Bioinformatics* 29.1 (Jan. 2013), pp. 15–21.

[186] Iakes Ezkurdia, Jose Manuel Rodriguez, Enrique Carrillo-de Santa Pau, Jesús Vázquez, Alfonso Valencia, and Michael L Tress. "Most highly expressed protein-coding genes have a single dominant isoform." en. In: *J. Proteome Res.* 14.4 (Apr. 2015), pp. 1880–1887.

[187] Alessia Galgano, Michael Forrer, Lukasz Jaskiewicz, Alexander Kanitz, Mihaela Zavolan, and André P Gerber. "Comparative analysis of mRNA targets for human PUF-family proteins suggests extensive interaction with the miRNA regulatory system." en. In: *PLoS One* 3.9 (Sept. 2008), e3164.

[188] Oriol Fornes et al. "JASPAR 2020: update of the open-access database of transcription factor binding profiles." en. In: *Nucleic Acids Res.* 48.D1 (Jan. 2020), pp. D87–D92.

[189] Daniel R Zerbino et al. "Ensembl 2018." en. In: *Nucleic Acids Res.* 46.D1 (Jan. 2018), pp. D754–D761.

[190] Marcel Martin. "Cutadapt removes adapter sequences from high-throughput sequencing reads." en. In: *EMBnet.journal* 17.1 (May 2011), pp. 10–12.

[191] The RNAcentral Consortium. "RNAcentral: a hub of information for non-coding RNA sequences." en. In: *Nucleic Acids Res.* 47.D1 (Jan. 2019), pp. D1250–D1251.

[192] *Picard*. http://broadinstitute.github.io/picard/. Accessed: 2022-2-1.

[193] Mikhail G Dozmorov et al. "Detrimental effects of duplicate reads and low complexity regions on RNA- and ChIP-seq data." en. In: *BMC Bioinformatics* 16.13 (Dec. 2015), pp. 1–11.

[194] Tom Smith, Andreas Heger, and Ian Sudbery. "UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy." en. In: *Genome Res.* 27.3 (Mar. 2017), pp. 491–499.

[195] Rob Patro, Geet Duggal, Michael I Love, Rafael A Irizarry, and Carl Kingsford. "Salmon provides fast and bias-aware quantification of transcript expression." en. In: *Nat. Methods* 14.4 (Apr. 2017), pp. 417–419.

[196] Rahul Siddharthan and Erik van Nimwegen. "Detecting regulatory sites using PhyloGibbs." en. In: *Methods Mol. Biol.* 395 (2007), pp. 381–402.

[197] Merkel. "Docker: lightweight linux containers for consistent development and deployment." In: *Linux J.* (2014).

[198] Brenton Graveley. *ENCSR550DVK*. Nov. 2014.

[199] Brenton Graveley. *ENCSR249ROI*. Mar. 2018.

[200] Brenton Graveley. *ENCSR993OLA*. Nov. 2014.

[201] Brenton Graveley. *ENCSR384KAN*. Aug. 2016.

[202] Brenton Graveley. *ENCSR981WKN*. Nov. 2014.

[203] Brenton Graveley. *ENCSR661ICQ*. Nov. 2014.

[204] Brenton Graveley. *ENCSR756CKJ*. Nov. 2014.

[205] Brenton Graveley. *ENCSR987FTF*. Nov. 2014.

[206] Mascha Jansen. *FAIR Principles*. en. https://www.go-fair.org/fair-principles/. Accessed: 2022-10-12. Nov. 2017.

[207] Ronny Lorenz, Stephan H Bernhart, Christian Höner Zu Siederdissen, Hakim Tafer, Christoph Flamm, Peter F Stadler, and Ivo L Hofacker. "ViennaRNA Package 2.0." en. In: *Algorithms Mol. Biol.* 6 (Nov. 2011), p. 26.

[208] Anna Knörlein, Chris P Sarnowski, Tebbe de Vries, Moritz Stoltz, Michael Götze, Ruedi Aebersold, Frédéric H-T Allain, Alexander Leitner, and Jonathan Hall. "Nucleotide-amino acid π-stacking interactions initiate photo cross-linking in RNA-protein complexes." en. In: *Nat. Commun.* 13.1 (May 2022), p. 2719.

[209] Piotr J Balwierz, Mikhail Pachkov, Phil Arnold, Andreas J Gruber, Mihaela Zavolan, and Erik van Nimwegen. "ISMARA: automated modeling of genomic signals as a democracy of regulatory motifs." en. In: *Genome Res.* 24.5 (May 2014), pp. 869–884.

[210] Viren Amin, Didem Ağaç, Spencer D Barnes, and Murat Can Çobanoğlu. "Accurate differential analysis of transcription factor activity from gene expression." en. In: *Bioinformatics* 35.23 (Dec. 2019), pp. 5018–5029.

[211] *Anaconda Documentation — Anaconda documentation*. en. https://docs.anaconda.com. Accessed: 2022-10-14.

[212] Madalina Giurgiu, Julian Reinhard, Barbara Brauner, Irmtraud Dunger-Kaltenbach, Gisela Fobo, Goar Frishman, Corinna Montrone, and Andreas Ruepp. *CORUM: the comprehensive resource of mammalian protein complexes—2019*. 2019.