Title: BindCompare: A Novel Integrated protein-nucleic Acid Binding Analysis Platform

Pranav Mahableshwarkar[1,2], Jasmine Shum[1,3], Mukulika Ray[1]*, and Erica Larschan[1,2]*

[1]Molecular Biology, Cell Biology, & Biochemistry Department, Brown University, Providence RI, USA
[2]Department of Computer Science, Brown University, Providence RI, USA
[3]Center for Computational Molecular Biology, Brown University, Providence RI, USA
*Co-corresponding

Abstract
Advanced genomic technologies have generated thousands of Protein-Nucleic acid binding datasets that have the potential to identify testable gene regulatory network (GRNs) models governed by combinatorial associations between factors. Transcription factors (TFs) and RNA binding proteins (RBPs) are nucleic-acid binding proteins regulating gene expression and are key drivers of GRN function. However, the combinatorial mechanisms by which the interactions between specific TFs and RBPs regulate gene expression remain largely unknown. To identify possible combinations of TFs and RBPs that may function together, developing a tool that compares and contrasts the interactions of multiple TFs and RBPs with nucleic acids to identify their common and unique targets is necessary. Therefore, we introduce BindCompare, a user-friendly tool that can be run locally to predict new combinatorial relationships between TFs and RBPs. BindCompare can analyze data from any organism with known annotated genome information and outputs files with detailed genomic locations and gene information for targets for downstream analysis. Overall, Bindcompare is a new tool that identifies TFs and RBPs that co-bind to the same DNA and/or RNA loci, generating testable hypotheses about their combinatorial regulation of target genes.

BindCompare is an open-source package that is available on the Python Packaging Index (PyPI, https://pypi.org/project/bindcompare/) with the source code available on GitHub (https://github.com/pranavmahabs/bindcompare). Complete documentation for the package can be found at both of these links.

1. Introduction

Spatio-temporal regulation of gene expression is a tightly regulated hierarchical cascade of interacting proteins binding to the genome and transcribing RNA molecules. Transcription factors (TFs) regulate RNA transcription from DNA, and RNA-binding proteins (RBPs) form the bulk of nucleic acid-binding proteins that control splicing and nuclear export. Multiple nucleic acid binding proteins function together to co-regulate specific sets of target genes, which are often called gene regulatory networks (GRNs)(Oksuz et al. 2023; Wang et al. 2022). High-throughput genomic technologies have generated hundreds of data sets that provide information regarding DNA and RNA targets of individual nucleic acid binding proteins, helping decipher the combinatorial function of TFs and RBPs in regulating specific GRNs (Salma et al. 2023; Martin et al. 2023; König et al. 2012). However, user-friendly analysis tools cannot generate high-throughput comparisons between multiple protein-nucleic acid binding datasets to determine common and unique targets.

We present a novel command line and interactive tool for analyzing coordinated protein binding to DNA and RNA called BindCompare that: i) compares multiple protein-nucleic acids binding datasets to identify interacting pairs of factors in an unbiased way; ii) identifies common targets between selected protein pairs identified by the user. Thus, BindCompare will enable the identification of inter-dependent and combinatorial RNA/DNA binding, generating new hypotheses about how single and dual RNA and DNA binding factors impact gene regulation. To ensure user accessibility, BindCompare has a user-friendly graphical user interface (GUI) that can be used to study any organism.

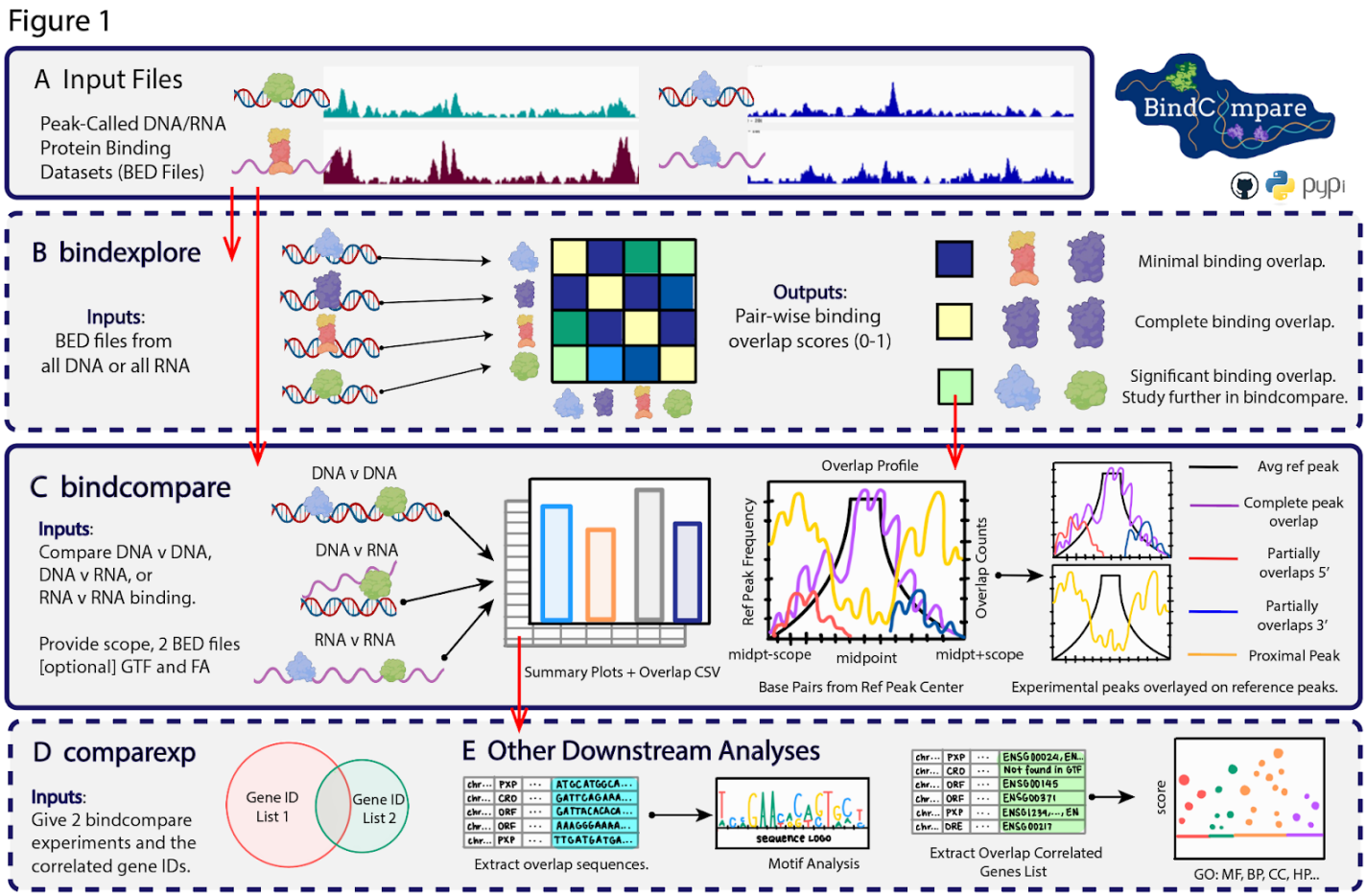

Figure 1: Schematic of the BindCompare workflow. A. Inputs to BindCompare are peak-called binding datasets from experiments such as ChIP-seq (DNA binding) and eCLIP (RNA binding) in BED format. **B**. If working with protein binding (BED files) to only DNA or RNA, the bindexplore function can identify pair-wise binding interactions between N data sets, helping identify candidate proteins to study with the **bindcompare** function. **C**. **bindcompare**, the core function, performs comprehensive binding analysis between two sets of BED files (DNA/DNA, RNA/RNA, or RNA/DNA), producing a diverse array of binding profiles and other outputs. **D**. **comparexp**, when paired with the gene GTF file, allows users to identify factors with overlapping binding sites. **E**. Gene lists and sequences correlated with binding overlaps can be used for Gene Ontology and motif analysis.

2. Method

The main inputs to BindCompare are peak-called Browser Extensible Data (BED) files (**Figure 1A**). The first module in BindCompare, **bindexplore**, identifies combinatorial binding patterns across the genome (**Figure 1B**), further studied with **bindcompare** (**Figure 1C**). **bindcompare** profiles all overlapping nucleic acid binding regions for any pair of two proteins and extracts related sequence and gene information. These annotated results allow downstream analysis through BindCompare's **comparexp** module and other tools (**Figure 1D, E**).

Protein-nucleic acid assays, including eCLIP, ChIP-seq, and CUT&RUN (Van Nostrand et al. 2016; Skene and Henikoff 2017) yield Peak-called BED files listing genomic binding locations. Platforms such as Model-based Analysis of ChIP-Seq (MACS2) or PureCLIP (Gaspar 2018; Krakau, Richard, and Marsico 2017) are also used for peak calling.

2.1 bindexplore - Multifactorial Binding Analysis

Users can input 'N = total number' of peak-called BED files for DNA or RNA into the **bindexplore** module, producing a heatmap of co-binding activity across all pairs of proteins (**Figure 1B, Supp. File 1**). Splitting the genome into user-defined bins (default size of 5000 bp) generates heatmaps. All binding sites are mapped into the bins. For each pair of protein binding data submitted, pairwise intersections of binding-correlated bins are determined. For each of the N^2 pairs, one protein is chosen as the reference, and the other is overlaid. Next, the number of times both proteins bind in the same bin is divided by the number of times the reference protein binds to provide a normalization.

The overlap between a pair of proteins targeting common nucleic acid target sequences is represented by a correlation value between 0-1 (blue-yellow) (**Figure 1B**, **Supp. File 2**, **Figure S2A, B**). This process reveals candidate co-regulators that can be further analyzed with **bindcompare** (Methods 2.2).

2.2 bindcompare - Pairwise Binding Analysis

With the **bindcompare** module, two sets of protein-nucleic data chosen by the user can be compared, including the following possible comparisons: i) protein-DNA and protein-DNA, ii) protein-RNA and protein-RNA, or iii) protein-DNA and protein-RNA **(Figure 1C**). In addition to these two inputs, the script requires a scope, the range at which the binding sites should be called as overlapping, and optional genome fasta and genes GTF files. The scope accounts for the co-regulatory interactions that sometimes occur at different genomic locations due to the resolution of the experimental technique and potential nucleic acid folding.

bindcompare uses the provided scope to look for overlapping binding sites upstream and downstream of each reference peak. These overlaps are then classified into four categories: 1) complete reference overlap (CRO); 2) Overlap Reference End or 3' overlaps (ORE); 3) Overlap Reference Front or 5' overlaps (ORF); 4) proximal peak (PXP) (**Supp. File 1**). This process is optimized using interval trees to store binding loci (Ang and Tan 1995; Halbert and Tretyakov 2020).

Then, the script outputs a visualization of the binding profiles of these proteins, plotting the average reference peak and the frequency of overlapping experimental peaks across the scoped region. In addition, the script provides the distribution of overlap types and visualization of binding profiles specific to each chromosome. The binding profile contains four colored lines over the average reference peak, each correlating to one of the four categories of overlap described above. The y-axis indicates the number of times that this specific overlap event occurred at a position over the scoped region. Additional plots produced by **bindcompare** place the co-regulatory activity in the context of overall binding patterns for both protein-binding datasets (**Figure S4, Table 1c-g**). Furthermore, **bindcompare** can directly compare DNA and RNA binding for specific TFs (**Figure 1C**), which is poorly understood even though most TFs bind to DNA and RNA (Oksuz et al. 2023).

2.3 comparexp and Downstream Analyses

The final module in BindCompare, **comparexp**, enables users to determine the impacts of genetic, RNAi, or small-molecule perturbations on locations of protein-nucleic acid binding (Figure 1D, Supp. File 1). comparexp identifies unique protein-nucleic acid interactions occurring in only one experimental condition and overlapping targets that occur in multiple conditions.

Furthermore, given a genome file or genes GTF file, **bindcompare** outputs all genes associated with the bound sequences that change after a perturbation. These outputs can be input into popular Gene Ontology tools, including GProfiler2 (Kolberg et al. 2020) and ShinyGO (Ge, Jung, and Yao 2020). Finally, sequence information can be studied using the MEMESuite (Bailey et al. 2015) (**Figure 1E**).

2.4 How to use BindCompare

BindCompare can be used either through i) the command line or ii) an interactive GUI. Users can install the tool using the Python Package Index (https://pypi.org/project/bindcompare) or GitHub (https://github.com/pranavmahabs/bindcompare). To increase the accessibility of the BindCompare pipeline to users who do not use the command line, the scripts can be run through a Python CustomTkinter GUI interface ((Schimansky 2024); **Supp. File 1**). The GUI platform and extensive documentation help to analyze the visual outputs.

3. Results

3.1 Comparing common targets of multiple proteins that bind to nucleic acids

We analyzed protein-DNA binding and protein-RNA binding datasets for 16 human RBPs from ENCODE project phase III (GSE77634) using BindCompare. ChIP datasets (protein-DNA) and eCLIP datasets (protein-RNA) were processed separately using the **bindexplore** module (**Figure 1B**), generating a correlation matrix (**Supp. File 2, Figure S2A, B; Supp. File 3, Table 1a,b**) for pairwise comparison between all sixteen RBPs. We found that RBP HnRNPA1 binding to RNA shows high correlation matrix values (**CMV**) (>0.50) with the following reference RBPs binding to RNA: Sfpq (0.57), Fus (0.63), Taf15 (0.64) and Hnrnpul1 (0.69). This is consistent with literature suggesting that HnRNPA1, Sfpq, Fus, and Taf15 have a common association with neurodegenerative diseases like ALS (Amyeloid Lateral Sclerosis) and regulate RNA splicing (Gordon et al. 2021; Xue et al. 2020). Interestingly, when we overlapped Fus, Sfpq, Hnrnpul1, and Taf15 binding on RNA with HnRNPA1-RNA as a reference, CMVs were <0.50. Thus, **bindexplore** generates hypotheses about the combinatorial function of factors by defining the fraction of common targets between specific protein pairs. The **bindexplore** correlation matrix for the DNA binding (ChIP) for the same set of RBPs (**Supp. File 2, Figure S2B; Supp. File 3, Table 1b**) shows that Taf15 has more common DNA targets with Ptbp1, Srsf1, and U2af1 despite having a low score for common RNA targets (**Supp. File 2, Figure S2A, B**). Thus, **bindexplore** can be used to build a network of combinations of TFs and RBPs targeting specific chromatin loci at the DNA and/or RNA levels to define a putative regulatory network at specific genomic locations containing target genes of interest.

Next, we used the **bindcompare** module (**Figure 1C**) to compare HnRNPA1 pairwise to RBPs with high correlation matrix scores defined by bindexplore. We identified common targets shared by HnRNPA1, Sfpq, Fus, and Taf15 (Supp. File 3, Table 1c). Understanding their combinatorial effect could help in better-managing diseases like ALS. **bindcompare** defines the following categories of overlapped peaks: 1) complete overlap between reference (Rbfox2-DNA) and overlapped (Ptbp1-DNA) protein-DNA binding regions (**CRO)**; 2) 5' front partial overlap with reference (**ORF**); 3) 3' end partial overlap with reference (**ORE)**; 4) proximal peaks ($\pm$ 1000bp) to reference (**Supp. File 2, Figure S3A-D**). Therefore, the user can also identify genome-wide shared targets for each pair of factors.

3.2 Relative binding of protein pairs to DNA and RNA: regulation of co-transcriptional RNA processing

RNA processing is co-transcriptionally integrated with RNA transcription, and many TFs and RBPs function in this process (Bentley 2014; Cordiner et al. 2023), but the mechanisms regulating it to the correct genomic locations still need to be better understood. We show that the **bindcompare** module (**Figure 1C**) can help define the co-transcriptional regulation of RNA processing using Taf15 and Fus, DNA/RNA binding proteins regulating transcription and RNA processing (Schwartz, Cech, and Parker 2015). Fus and Taf15 have more common DNA targets than RNA targets. Also, one-fourth of Fus and Taf15 DNA targets are associated with RNA binding (**Supp. File 2, Figure S4A-D**), suggesting that these proteins may regulate co-transcriptional RNA processing at a specific subset of loci (**Supp. File 3, Table 1d-g**).

bindcompare identifies the relative positions where RBPs/TFs bind to common nucleic acid targets. The frequency plot (overlap profiles) outputs of **bindcompare** (**Supp. File 2, Figure S4E-H)** show that Taf15 binds to RNA within 100 nucleotides of Fus binding at shared RNA targets. In contrast, at DNA targets, they bind side-by-side within a scope of 250 bp, with both Taf15 and Fus connecting RNA to DNA within 250 bp.

One can also do individual chromosome-wide binding analysis with BindCompare. We compared the binding of the male X-chromosome enriched *Drosophila* transcription factor CLAMP at genomic locations on DNA (# GSE220053) with its RNA targets (# GSE205987) in male cells (Ilik et al. 2013; Soruco et al. 2013; Ray et al. 2024). Using BindCompare for individual chromosome-wise overlap profiles, we show that CLAMP-DNA peaks overlapping with RNA peaks are more frequently distributed on X-chromosomes than autosomes (**Supp. File 2, Figure S5**), consistent with literature demonstrating that CLAMP is enriched on the male X-chromosome and interacts with RNA on the male-X.

3.3 Comparing common protein-nucleic acid binding targets between different experimental conditions

Many studies focus on the dependence of the binding of a single RBP or TF to its targets on another RBP/TF. Using **comparexp (Figure 1D),** we compared common DNA targets for the interacting partner's *Drosophila* TF CLAMP and the *Drosophila* RBP RNA helicase MLE (Maleless) (Tikhonova et al. 2024) in the presence and absence of TF CLAMP (**Supp. File 2, Figure S6**). **comparexp** determined that at 537 CLAMP binding loci, MLE targets DNA only in the absence of CLAMP, i.e., CLAMP inhibits MLE binding to DNA at these ectopic binding loci. Thus, **comparexp** determined how the binding to chromatin of one RBP, MLE, is promoted or inhibited at different loci by its binding partner, the CLAMP TF.

Overall, we show that BindCompare is a new tool that provides essential information for identifying and studying the combinatorial binding of TFs and RBPs. Moreover, BindCompare has the potential to reveal novel mechanisms regulating co-transcriptional RNA processing because it is the only available tool that simultaneously measures DNA and RNA binding to chromatin.

Author Contributions

PM and MR conceived the project idea. PM implemented the BindCompare platform under the guidance of MR and EL. JS and PM ran analyses on the BindCompare platform and developed figures. PM and MR wrote the manuscript, and EL provided revisions.

Conflict of Interest

None declared.

Funding

This project was funded by NIH (National Institutes of Health) R35 GM126994 to EL.

Data Availability

BindCompare is an open-source package that is available on the Python Packaging Index (PyPI, https://pypi.org/project/bindcompare/) with the source code available on GitHub (https://github.com/pranavmahabs/bindcompare). Complete documentation for the package can be found at both of these links.

References

Ang, Chuan-Heng, and Kok-Phuang Tan. 1995. "The Interval B-Tree." *Information Processing Letters* 53 (2): 85–89.

Bailey, Timothy L., James Johnson, Charles E. Grant, and William S. Noble. 2015. "The MEME Suite." *Nucleic Acids Research* 43 (W1): W39–49.

Bentley, David L. 2014. "Coupling mRNA Processing with Transcription in Time and Space." *Nature Reviews. Genetics* 15 (3): 163–75.

Cordiner, Ross A., Yuhui Dou, Rune Thomsen, Andrii Bugai, Sander Granneman, and Torben Heick Jensen. 2023. "Temporal-iCLIP Captures Co-Transcriptional RNA-Protein Interactions." *Nature Communications* 14 (1): 696.

Gaspar, John M. 2018. "Improved Peak-Calling with MACS2." *bioRxiv*. https://doi.org/10.1101/496521.

Ge, Steven Xijin, Dongmin Jung, and Runan Yao. 2020. "ShinyGO: A Graphical Gene-Set Enrichment Tool for Animals and Plants." *Bioinformatics* 36 (8): 2628–29.

Gordon, Patricia M., Fursham Hamid, Eugene V. Makeyev, and Corinne Houart. 2021. "A Conserved Role for the ALS-Linked Splicing Factor SFPQ in Repression of Pathogenic Cryptic Last Exons." *Nature Communications* 12 (1): 1918.

Halbert, Chaim Leib, and Konstantin Tretyakov. 2020. "Intervaltree." Python Package Index. August 3, 2020. https://pypi.org/project/intervaltree/.

Ilik, Ibrahim Avsar, Jeffrey J. Quinn, Plamen Georgiev, Filipe Tavares-Cadete, Daniel Maticzka, Sarah Toscano, Yue Wan, et al. 2013. "Tandem Stem-Loops in roX RNAs Act Together to Mediate X Chromosome Dosage Compensation in Drosophila." *Molecular Cell* 51 (2): 156–73.

Kolberg, Liis, Uku Raudvere, Ivan Kuzmin, Jaak Vilo, and Hedi Peterson. 2020. "gprofiler2 -- an R Package for Gene List Functional Enrichment Analysis and Namespace Conversion Toolset g:Profiler." *F1000Research* 9 (July). https://doi.org/10.12688/f1000research.24956.2.

König, Julian, Kathi Zarnack, Nicholas M. Luscombe, and Jernej Ule. 2012. "Protein–RNA Interactions: New Genomic Technologies and Perspectives." *Nature Reviews. Genetics* 13 (2): 77–83.

Krakau, Sabrina, Hugues Richard, and Annalisa Marsico. 2017. "PureCLIP: Capturing Target-Specific protein–RNA Interaction Footprints from Single-Nucleotide CLIP-Seq Data." *Genome Biology* 18 (1): 240.

Martin, Vincentius, Farica Zhuang, Yuning Zhang, Kyle Pinheiro, and Raluca Gordân. 2023. "High-Throughput Data and Modeling Reveal Insights into the Mechanisms of Cooperative DNA-Binding by Transcription Factor Proteins." *Nucleic Acids Research* 51 (21): 11600–612.

Oksuz, Ozgur, Jonathan E. Henninger, Robert Warneford-Thomson, Ming M. Zheng, Hailey Erb, Adrienne Vancura, Kalon J. Overholt, Susana Wilson Hawken, Salman F. Banani, and Richard Lauman. 2023. "Transcription Factors

Interact with RNA to Regulate Genes." *Molecular Cell* 83 (14): 2449–63. e13.
Ray, Mukulika, Julia Zaborowsky, Pranav Mahableshwarkar, Smriti Vaidyanathan, Jasmine Shum, Renjith Viswanathan, Annie Huang, et al. 2024. "Dual DNA/RNA-Binding Factor Regulates Dynamics of hnRNP Splicing Condensates." *bioRxiv : The Preprint Server for Biology*, January. https://doi.org/10.1101/2024.01.11.575216.
Salma, Mohammad, Charlotte Andrieu-Soler, Virginie Deleuze, and Eric Soler. 2023. "High-Throughput Methods for the Analysis of Transcription Factors and Chromatin Modifications: Low Input, Single Cell and Spatial Genomic Technologies." *Blood Cells, Molecules & Diseases* 101 (July): 102745.
Schimansky, Tom. 2024. "CustomTkinter: A Modern and Customizable Python UI-Library Based on Tkinter." Python Package Index. January 9, 2024. https://pypi.org/project/customtkinter/.
Schwartz, Jacob C., Thomas R. Cech, and Roy R. Parker. 2015. "Biochemical Properties and Biological Functions of FET Proteins." *Annual Review of Biochemistry* 84: 355–79.
Skene, Peter J., and Steven Henikoff. 2017. "An Efficient Targeted Nuclease Strategy for High-Resolution Mapping of DNA Binding Sites." *eLife* 6 (January). https://doi.org/10.7554/eLife.21856.
Soruco, Marcela M. L., Jessica Chery, Eric P. Bishop, Trevor Siggers, Michael Y. Tolstorukov, Alexander R. Leydon, Arthur U. Sugden, et al. 2013. "The CLAMP Protein Links the MSL Complex to the X Chromosome during Drosophila Dosage Compensation." *Genes & Development* 27 (14): 1551–56.
Tikhonova, Evgeniya, Anastasia Revel-Muroz, Pavel Georgiev, and Oksana Maksimenko. 2024. "Interaction of MLE with CLAMP Zinc Finger Is Involved in Proper MSL Proteins Binding to Chromosomes in Drosophila." *Open Biology* 14 (3): 230270.
Van Nostrand, Eric L., Gabriel A. Pratt, Alexander A. Shishkin, Chelsea Gelboin-Burkhart, Mark Y. Fang, Balaji Sundararaman, Steven M. Blue, et al. 2016. "Robust Transcriptome-Wide Discovery of RNA-Binding Protein Binding Sites with Enhanced CLIP (eCLIP)." *Nature Methods* 13 (6): 508–14.
Wang, Chuhui, Xueqing Zong, Fanjie Wu, Ricky Wai Tak Leung, Yaohua Hu, and Jing Qin. 2022. "DNA- and RNA-Binding Proteins Linked Transcriptional Control and Alternative Splicing Together in a Two-Layer Regulatory Network System of Chronic Myeloid Leukemia." *Frontiers in Molecular Biosciences* 9 (August): 920492.
Xue, Yuan Chao, Chen Seng Ng, Pinhao Xiang, Huitao Liu, Kevin Zhang, Yasir Mohamud, and Honglin Luo. 2020. "Dysregulation of RNA-Binding Proteins in Amyotrophic Lateral Sclerosis." *Frontiers in Molecular Neuroscience* 13 (May): 78.