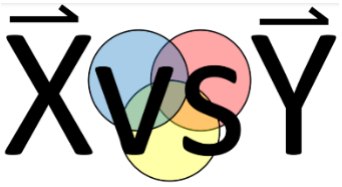


: pipeline to identify shared and distinct genomic signatures from multiple expression and protein-DNA interaction data sets over time

Isaac Nathoo

Advisor: Dr. Erica Larschan

Research Guided by Ashley Mae Conard

Senior Honors Thesis in Computational Biology

April 2021

Abstract:

Development is a complex process governed by coordinated gene regulation at the correct time and place. This complexity can be measured via multiple types of sequencing technologies, which come at continually lower costs and higher throughput. This fosters an exponential growth of sequencing data with increasingly multi-dimensional experimental designs, which is necessary to uncover the rules of the genomic landscape across different groups and sexes. There is an urgent need for methods to address the combinatorial complexity of comparing across multiple experiments, sexes, and/or time points. XvsY is the first streamlined and accessible command line tool to efficiently compare multi-dimensional RNA-seq experiments and integrate protein-DNA information to find evidence of true interaction. XvsY brings together differential expression (DE) analysis for the multi-dimensional inputs, followed by intersecting genomic sets. For each unique and shared set of genes across all dimensions, XvsY provides global expression and DE plots and information about mechanism through gene ontology and motif analysis. I illustrate the utility of XvsY to distinguish the temporal and sex-specific roles of two vital co-factors (MSL2 and CLAMP) on brain development in *Drosophila*.

Introduction:

Discovering the process by which transcription factors and their targets communicate at the DNA, RNA, and protein levels is critical to understanding signaling cascades of normal growth, and the conditions that promote the onset of disease. To provide new insight into the relationships between different regulatory factors in different groups, under different conditions, (such as knock out or knock down), and over time, researchers are now able to generate vast amounts of 'omics data at lowered costs with much higher throughput[28]. We urgently require methods that fully integrate these multi-dimensional data to model the cause-and-effect relationships between genes in regulatory cascades.

Transcription factors (TF) are proteins considered to be the master regulators of transcription, binding to DNA and driving gene activation and repression. TFs control the rate of transcription and bind to multiple sites in the genome, including at the gene promoters, enhancer regions and within the gene body. TFs also activate and repress genes at the right developmental time and place within an organism.

One way to study the functions of these TFs and their context specific roles are through RNA-seq experiments. A biologist can create a null mutant through techniques such as CRISPR or non-homologous end rejoining or use gene silencing methods, such as RNA interference (RNAi), to respectively knock out or knock down a transcription factor. Then they can perform RNA-seq to determine how the transcriptomic profile of the organism or cells have changed. Given the additional complexity of analyzing the context specific roles of transcription factors, it is crucial to have tools that enable a user to compare multi-dimensional RNA-seq experiments. This includes comparisons across multiple time points, tissues, sexes, and experiments, such as knock-down mutants and therapeutic drug testing.

Currently, no tools accurately address the combinatorial complexity of analyzing RNA-seq data from multiple experiments, different groups (i.e., sex, tissue), and over time. This analysis includes obtaining the differentially expressed genes, determining the fold change differences between the various experiments, finding the differences in the functions of the genes that were affected, determining the unique and shared mechanisms between such dimensions, and understanding the binding motifs in these genes to identify potential targets. Some tools available now that perform part of these functions are: IDEAMEX[14], Intervene[15], A-Lister[17], and TIMEOR[7] (Table 1).

IDEAMEX[14], standing for Integrative Different Expression Analysis for Multiple Experiments, allows a user with biological RNA-seq data to conduct 1) data analysis, 2) differential expression, and 3) results integration. Firstly, data analysis refers to the preliminary analysis for quality control to ensure that the experiments produced usable paired end or single end reads. Secondly, the differential expression step compares several Bioconductor packages, including DESeq2 and edgeR, for each of the experiments and generates reports for each method. Thirdly, results integration involves producing graphical outputs such as heatmaps, Venn diagrams, and text lists to compare each of the differential expression methods.

Intervene[15] allows users to overlap differentially expressed genes from disparate experiments given lists or genomic regions. A-Lister[17] accepts the inputs of the results of differential expression tools, such as DESeq2, and allows the user to filter by columns in the data, such as p-value and fold change. It also provides functionality of different set operations, such as AND, OR, and DIFF, to query these genomic intersections and compare experiments. The DIFF operator applied to two sets returns all the elements that are present in the first, but not the second. Lastly, TIMEOR[7] (Trajectory Inference and Mechanism Exploration with Omics data in R) uses time-series algorithms to characterize gene dynamics over time that integrates TF binding data to reconstruct gene regulatory networks (GRNs) from ordered RNA-seq data. A comparison of XvsY, IDEAMEX, Intervene and A-Lister is shown in Table 1.

	XvsY	IDEAMEX	Intervene	A-Lister
Differential Expression Analysis	✔	✔		
Method Comparisons	✔	✔		
Overlapping Gene Clusters	✔		✔	✔
Visualization of Fold Change Differences	✔			
Construction of X vs Autosome plots	✔			
GO Analysis	✔			
De Novo Motif Analysis	✔			
TF Enrichment	✔			

Table 1: A comparison of similar pipeline methods shows that XvsY provides new functionality for not only finding the DEGs but overlapping them and then characterizing the function and binding of these clusters.

However, each of these is missing an aspect for comparing complex, multi-dimensional RNA-seq experiments. IDEAMEX does not provide the users any information about global gene regulation across different chromosomes, gene functions and mechanisms, or about TF binding. Intervene and A-Lister are similar, and do not perform differential expression analysis or provide insights into how the genes work together. TIMEOR, although useful to analyze temporal RNA-

seq data, does not provide the utilities to analyze multiple RNA-seq experiments and does not address the challenge of combinatorial complexity of the data and gene intersection clusters.

Therefore, in my thesis, I present XvsY, an accessible command line tool enabling the user to easily and efficiently compare multi-dimensional RNA-seq experiments. This pipeline is suitable for the analysis of any experimental conditions that produce RNA-seq data, such as RNA interference or therapeutic drug testing. I will illustrate the utility of XvsY to distinguish the temporal and sex-specific roles of two vital co-factors (MSL2 and CLAMP), known for their roles in dosage compensation, on brain development in *Drosophila*.

Dosage compensation is a process that balances the expression of genes on the X chromosome and autosomes in males, which have one X and one Y chromosome as compared to the two X chromosomes in females. This process is essential for proper development and dysregulation can cause disorders such as Duchenne muscular dystrophy in humans[6], where in the brain there is neural shrinkage in regions of the cerebral cortex and brainstem[2]. Model organisms such as *Drosophila* provide a prime avenue to investigate the onset of such illnesses.

CLAMP (Chromatin-Linked Adapter for MSL Proteins) is a key transcription factor (TF) and chromatin remodeler for sex determination and development in *Drosophila*. CLAMP is essential for the recruitment of the Male Specific Lethal (MSL) Complex to the X chromosome and to initiate dosage compensation. The MSL complex is comprised of 5 proteins, MSL1, MSL2, MSL3, MOF (males-absent-on-the-first) and MLE (maleless) as well as two non-coding RNAs roX1 and roX2 (RNA on the X). The MSL complex assembles exclusively in male flies as part of dosage compensation and mediates the global acetylation of histone H4 lysine 16 (H4K16ac) on the single male X chromosome, which causes an upregulation of transcription[9].

Previously, it had also be reported that CLAMP reduction leads to aberrant synaptic development in L3 larvae and in CLAMP-null mutants there are fewer boutons (synapses) formed at the neuromuscular junction (Figure 1A)[11,12]. However, within in the central nervous system, the Larschan lab found that the absence of CLAMP negatively affects brain development, causing shrinkage, in both sexes in *Drosophila* (Figure 1B and 1C)[27].This is interesting particularly because dosage compensation does not occur in females, thus suggesting that CLAMP has a different context specific role in females.

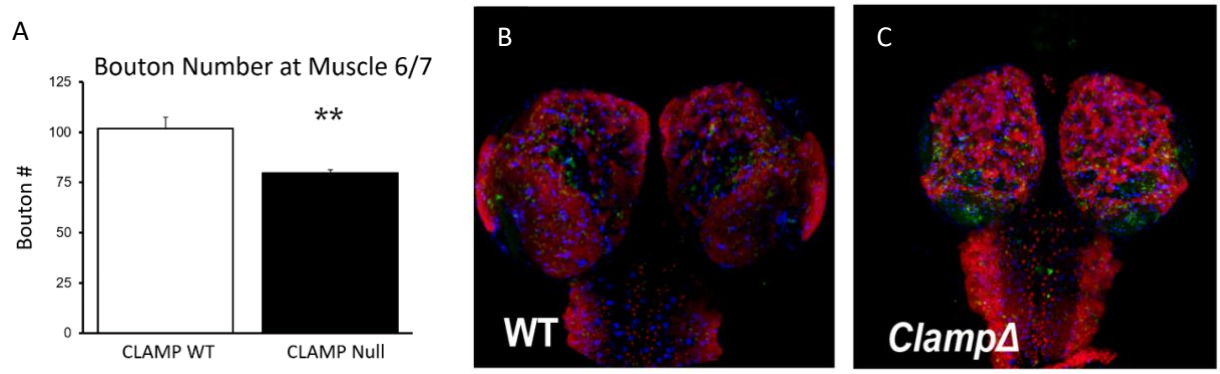


Figure 1: Effects of CLAMP removal through a null mutation on the peripheral and central nervous systems of Drosophila. CLAMP removal leads to fewer boutons at neuromuscular junctions (panel A) as well as atrophied and smaller brains compared to WT (panels B and C).

Overall, CLAMP's role is unknown in the brain with and without the MSL Complex, particularly its temporal and sex-specific effects on gene regulation. Furthermore, the role of CLAMP on the autosomes has also not been well characterized, even though it has been shown to bind ubiquitously across the genome[23]. My goal as well as the Larschan's lab goal is to understand CLAMP's role in the brain, and in combination with the MSL Complex. I show that XvsY is well-suited to efficiently analyze the multi-dimensional RNA-seq data produced to address our goal (Results Section).

Methods:

Here we present XvsY, the first method to comprehensively analyze complex multi-dimensional RNA-seq data to highlight shared and distinct genomic signatures. This pipeline can be broken down into four stages: 1) find differentially expressed genes and overlaps, 2) characterizing fold change differences between experimental groups, 3) characterizing function and mechanism, 4) characterizing motifs and binding for all intersections and unique groups (Figure 2). The inputs required to begin the analysis using XvsY are a read count matrix with the gene IDs as rows and experiment IDs as columns, and a metadata file specifying the experiment ID, condition, batch, and time point.

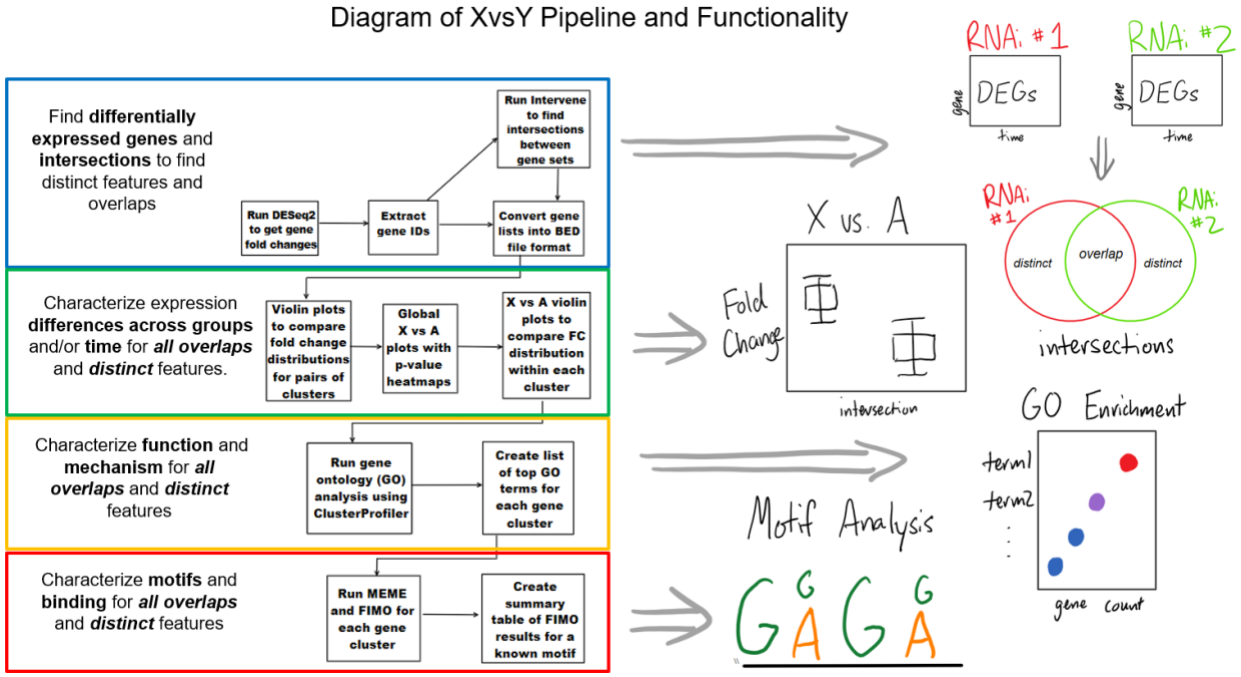


Figure 2: The XvsY pipeline enables users to compare multi-dimensional RNA-seq experiments easily and efficiently and perform various downstream analyses. The blue box corresponds to finding the DEGs and their intersections. The green box corresponds to determining expression differences. The yellow box corresponds to characterizing function and mechanism. The red box corresponds to characterizing motifs and binding.

In the first stage, DESeq2[18] in R is run for each of the RNA-seq experiments the user is interested in finding the differentially expressed genes for. The two inputs described above are

required to run DESeq2. Based on the type of experiments and the user's preference as a biologist, there are several important parameters the users can choose between. The first is the option between the Wald test and the Likelihood Ratio Test (LRT). The Wald test is better suited for categorical case versus control experiments, whereas the LRT is better suited for time course experiments [18]. Therefore, if a user is studying the role of TFs across development or how a system responds to perturbations over time and they want to evaluate the changes in differential gene expression between time points, it is recommended to use the LRT. In addition, there is also an option for the user to specify that they are inputting time course data which will use a reduced model to determine the effect of time. Next, the user can specify if they would like to control for a batch effect and this will add an extra term for batch to the model.

Moreover, in this stage, the gene IDs are extracted from each list of differentially expressed genes output by DESeq2 and run through Intervene to find the intersections and distinct sets. Finally, a *GTF file* containing information about all the genomic locations of genes in the organism is used to generate BED files for the genes in each of these sets. This is important to get the gene name corresponding to the gene ID, chromosome location, and start and end sites.

In the second stage, the Seaborn and SciPy libraries in Python are used to generate three types of plots: 1) expression fold change violin plots for common differentially expressed genes, 2) X vs Autosome violin plots for genes that are upregulated and downregulated, and 3) global chromosome level notch plots with a heatmap significance table highlighting significant expression differences across chromosomes. These plots are useful to compare temporal and sex groups, or different RNAi experiments. The first type of plot is a comparison of the gene fold changes between every pair of differentially expressed gene sets and it also shows if there is a difference in the effect of different experimental conditions on the same set of genes. The second type of plot is a comparison of the fold change differences between the X chromosome versus the autosomes within each intersection and distinct set of differentially expressed genes. In the case of dosage compensation and sexual dimorphism studies, these types of plots are useful to determine if the genes on the X chromosome are significantly more differentially expressed than genes on the autosomes, which might indicate a sex specific effect. The third type of plot is global differentially expressed gene X-vs-Autosome plots for each experimental condition. The significance of the difference in expression between each pair of chromosomes is also computed and the resulting p-values are displayed in a heatmap with the darker values indicating stronger significance. Due to the issue of multiple comparisons, all p-values are adjusted using the Benjamini-Hochberg correction method [5].

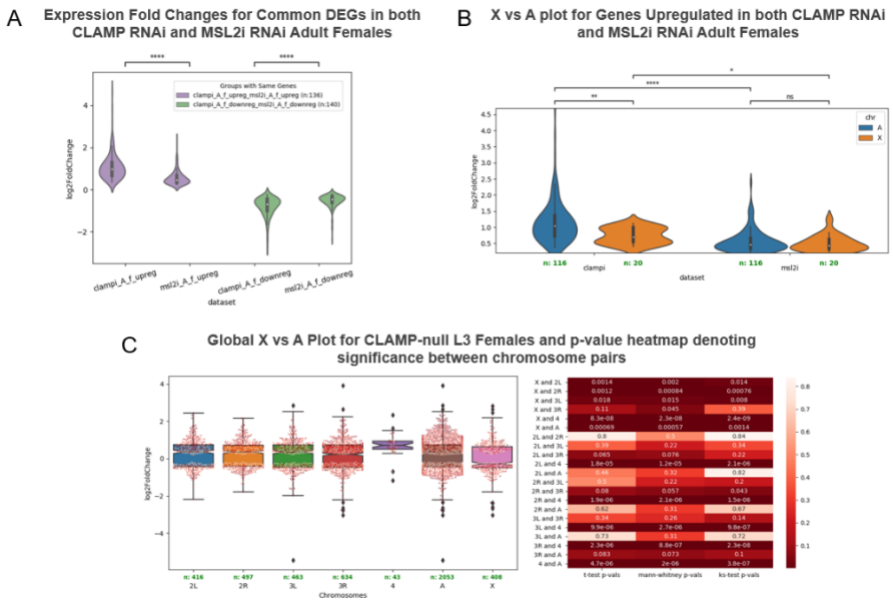


Figure 3: Examples of each type of plot generated by XvsY. In panel A, fold change comparison violin plots. In panel B, X vs A fold change violin plot for overlap. In panel C, global expression X vs A plot with p-value heatmap for t-test, Mann-Whitney test, and KS divergence test.

For these global plots (Figure 3C), the test statistics and p-values are computed using multiple tests, including the student's t-test[25], Mann-Whitney test[19], KS divergence test[20], and the Anderson-Darling test[1], to determine the significance of the differences in fold change between experiments and between genes on the X chromosome versus the autosomes. The t-test is a statistical test that is used to compare the means of two groups[25]. It is often used in hypothesis testing to determine whether two groups are different from one another. The Mann-Whitney test is a nonparametric test of the null hypothesis that for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X[19]. The Kolmogorov-Smirnov test is a nonparametric test of the equality of continuous, one-dimensional probability distributions that can be used to compare two samples and quantify the distance between the empirical cumulative distribution function of two samples[13,20]. The Anderson-Darling test is a statical test for testing whether several collections of observations or samples of data can be modeled as coming from a single population, where the probability distribution does not have to be specified[1,22]. Although the t-test and Mann-Whitney test are more commonly used in the analysis of biological data, they make assumptions about the distributions of the data and the sample sizes, that may not always be justified. Therefore, I include calculations of the significance as a p-value using the KS test and Anderson-Darling test, which are more robust and sensitive to the differences in the shapes of the probability distributions of multiple experiments[24].

In the third stage of the pipeline, ClusterProfiler[30] is used to characterize the biological processes (BP), molecular functions (MF), and cellular components (CC) of the genes in each intersection and distinct set by performing gene ontology (GO) term enrichment analysis. The results are output in several different graphical formats: a dot plot showing the gene ratio of different GO terms, a directed acyclic graph with the broadest GO terms near the root, an enrichment map, a bar graph, and a concept plot, which displays the linkages of genes and GO

terms as a network. These allow the user to visualize the mechanism of the genes in each set in multiple ways. To facilitate the analysis of time course data, the user can also select the parameter for time course data to run GO analysis for each timepoint separately. Once GO analysis has run for all gene overlaps and unique sets, summary tables are generated for BP, MF and CC containing the top GO term for each of these categories from each set.

In the fourth and final stage of the pipeline, the MEME[4] and FIMO tools from MEME Suite[3] are run to allow to user to better understand the motifs and transcription factor binding within each overlap and distinct set. Using MEME, XvsY will find the top three de novo motifs with widths between 8 and 50 base pairs using an objective function that scores motifs using an approximation to the E-value of the information content of the motif. Then if the user as a specific transcription factor or set of transcription factors that they are interested in and they want to determine if the genes in each cluster contain the binding motif for these TFs, the user can run FIMO. This will take in a position weight matrix (PWM) for the TFs of interest and will scan either the entire gene length or a 1 kilobase region around the transcription start site (TSS), based on the user's choice, for the sequence(s) encoded within the PWM. Summary tables will also be generated that describe the percentage of genes that contain the motif or each of the motifs in the PWM. This is particularly useful when the set of genes is large, making it time-consuming to see how many genes contain the motif for the TF of interest.

Overall, I built XvsY using the Snakemake[16] workflow management system, a command line tool to create reproducible and scalable data analyses. Within this pipeline, I intelligibly integrate: DESeq2[18] to determine which genes are differentially expressed in each experimental condition at each time point and Intervene[15] to identify different overlaps between these gene groups. Then for each distinct gene cluster and gene intersection cluster, XvsY provides results to highlight 1) shared mechanism, 2) enriched motifs to highlight potential binding TFs, and 3) altered expression patterns from the experiment. Specifically, XvsY uses 1) ClusterProfiler[30] to perform gene ontology (GO) analysis and examine the function of these genes and the biological processes they are involved in, and 2) the MEME Suite[3] to find the top three de novo motifs found in these genes and search for the presence of TF motifs. XvsY also constructs 3) global X-chromosome vs autosome boxplots to highlight differences in regulation across all chromosomes, and violin plots comparing the fold change differences between pairs of experiments. The final output is a comprehensive summary table of this information.

The XvsY pipeline can also be broken down into a directed acyclic graph (DAG) of all the rules in each of these four main sections of the pipeline (Figure 4). This shows how each of the rules are connected to and depended on each other, and what rule produces an output that is required for another rule to run. The first rule *run_deseq2*, which is at the root of the DAG, only requires the read count data and metadata files to be present. However, the rules at the leaves of the DAG require all previous rules that lead up to it to also be run. Although the primary purpose of XvsY was to be able to start with running differential expression analysis and perform various types of downstream analysis, Snakemake[16] enables the user to start with any rule in the DAG, given that its inputs have already been generated by the pipeline or externally otherwise.

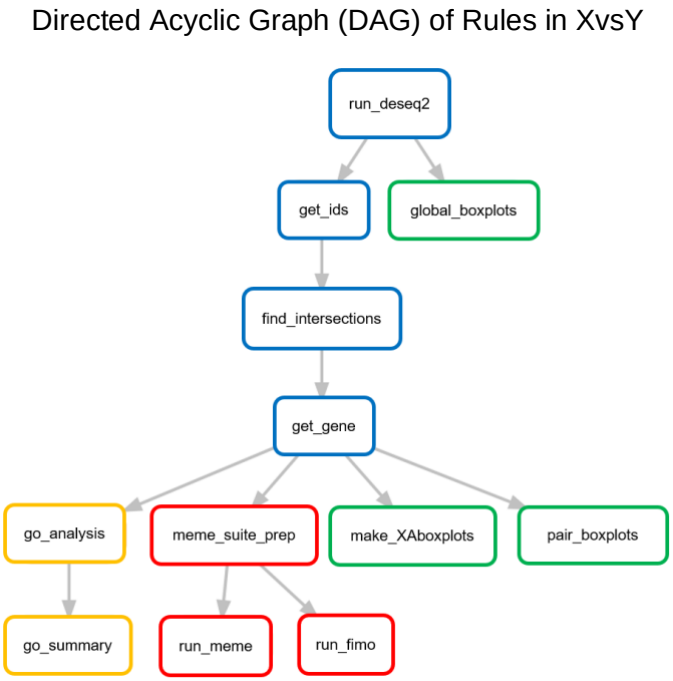


Figure 4: XvsY allows users to start at any rule (square) in the pipeline, given that its inputs are present, and run until any connecting rule. The pathways show the sets of rules to be executed and how the rules depend on each other.

To run through all of XvsY, the user needs to write one command from their command line terminal:

```
snakemake --R --until <RULE> --cores 1 --config outdir=<PATH/TO/OUTPUT/DIRECTORY>
```

To run XvsY, all the user needs to do is declare which rule they would like to run up to and what is the root directory that contains the inputs, which is where the outputs will also be generated. Importantly, XvsY also dynamically generates the folder structure as the different rules are run in an intelligible format for the user to probe at the end of the analysis.

Results:

This pipeline was tested using RNA-seq experiments after removing CLAMP (RNAi and knock-out) and MSL2 (RNAi) from the brain in different developmental stages (embryo, third instar larvae, and adult), and in both sexes in *Drosophila.* Data was available for 1) CLAMP RNAi, CLAMP-null, and MSL2 RNAi experiments in unsexed embryos, 2) CLAMP RNAi, CLAMP-null, and MSL2 RNAi experiments in larvae females, 3) CLAMP RNAi and MSL2 RNAi experiments in larvae males, 4) CLAMP RNAi and MSL2 RNAi experiments in adult females, and 5) CLAMP RNAi and MSL2 RNAi experiments in adult males (Figure 5).

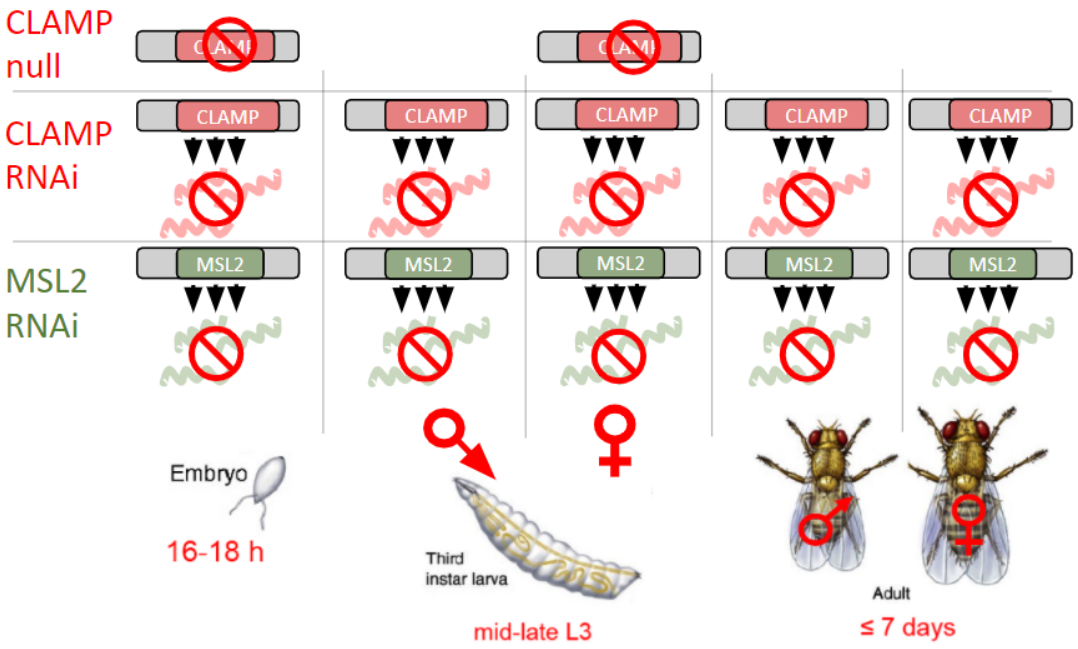


Figure 5: A schematic of the various RNA-seq experiments performed with different perturbations across time and sex.

For each of these five groups, I used XvsY to compare the different RNA-seq experiments conducted. I also performed comparisons across sexes in third instar larvae and adults, and across time from the embryo to adult stages. To illustrate the utility of XvsY with an example, I will show the results of the analysis for the experiments comparing experimental conditions at the embryo time point.

Using the read count matrices from the RNA-seq experiments for CLAMP RNAi, CLAMP-null (knock-out), and MSL2 RNAi in unsexed embryos, I was able to start at the beginning of the XvsY pipeline and run differential expression analysis. To accomplish this, DESeq2 was run using the Wald test with a beta prior for each experiment individually, then the sets of differentially expressed genes were broken down into those that are upregulated and downregulated. Volcano plots were generated to show the distribution of the fold change of the genes against the p-value of that difference. A minimum fold change cutoff was not used to ensure genes that are typically lowly expressed were not excluded from the downstream analysis, however, a p-value threshold of 0.05 was included. The volcano plot for CLAMP RNAi indicates that there were 2355 genes found to be differentially expressed, with some significantly downregulated genes being w, slo, and jus, which are involved in neurogenesis, and some significantly upregulated genes being CG12522 and Cp19 (Fig 6A). Next, the volcano plot for CLAMP-null mutants indicates that there were 3331 genes found to be differentially expressed, with some significantly downregulated genes being ldbr, Pep, and Blm, which are involved in cell cycle control and splicing, and some significantly upregulated genes being NT1 and lncRNA:CR44922[21], which are involved in nervous system development. Lastly, the volcano plot for MSL2 RNAi indicates that there were 101 genes found to be differentially expressed, with some significantly downregulated genes being CG42633 and GstO3, and some significantly upregulated genes being l(2)03659 (Fig 6B).

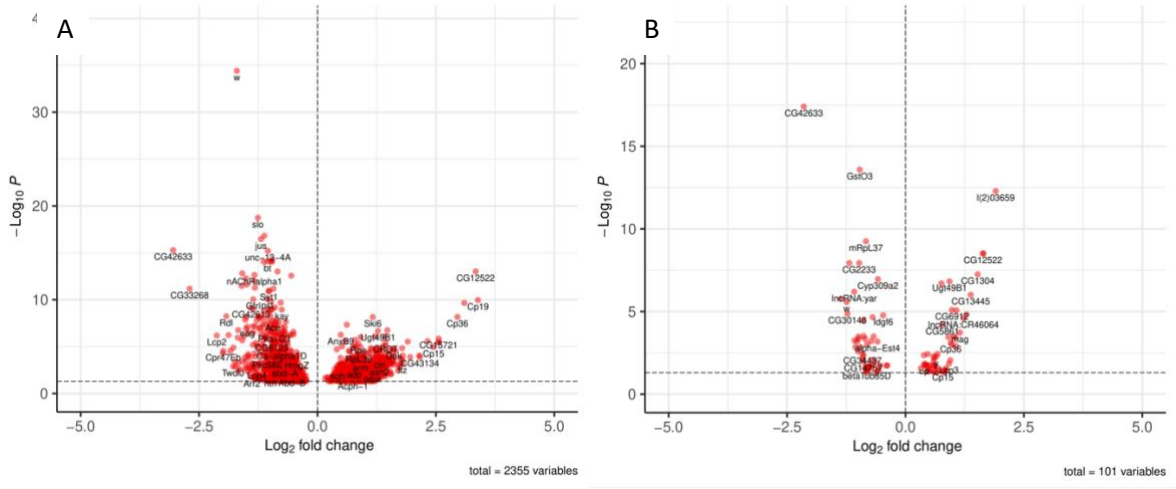


Figure 6: Volcano plots showing the log base-2 fold changes versus the negative log base-10 p-values for genes differentially expressed by CLAMP RNAi (panel A) and MSL2 RNAi (panel B).

Subsequently, Intervene was run by XvsY to overlap these differentially expressed gene sets to find intersections as well as distinct features, and an upset bar plot was produced to visualize these results (Fig 7). The four largest groups, which are non-overlapping gene sets, are CLAMP-null downregulated genes, CLAMP-null upregulated genes, CLAMP RNAi upregulated genes, and then CLAMP RNAi downregulated genes, in that order. Interestingly, and unexpectedly, the next two largest groups are the overlap between CLAMP-null downregulated and CLAMP RNAi upregulated genes, and between CLAMP-null upregulated and CLAMP RNAi downregulated genes. Much smaller are the overlaps between CLAMP RNAi upregulated and MSL2 RNAi upregulated genes, and between CLAMP RNAi downregulated and MSL2 RNAi downregulated genes. These intersections can provide biologists insights into how different TFs work together and are influenced by different types of conditions, such as sex.

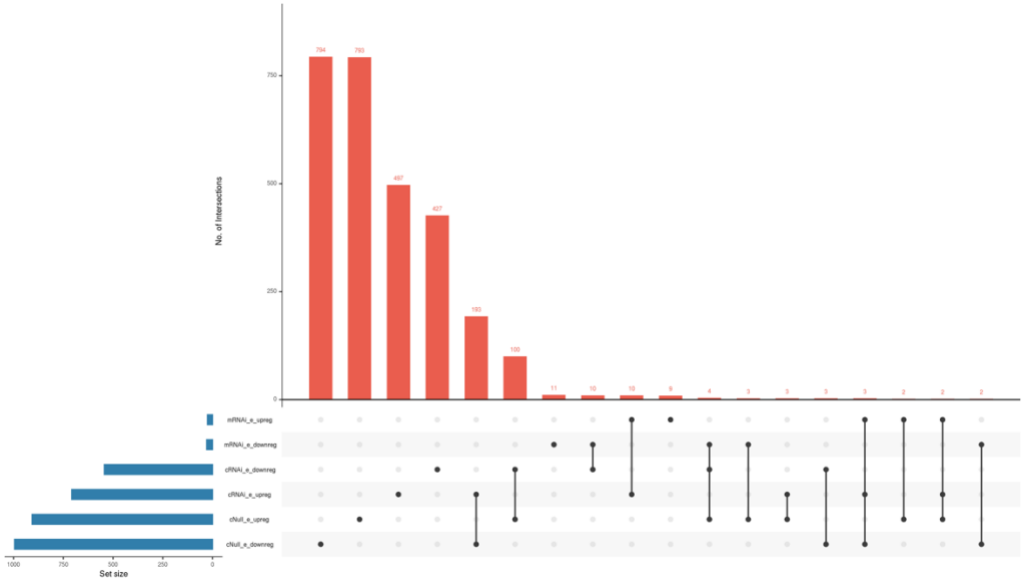


Figure 7: A upset bar plot showing the which gene sets overlap with each other and the number of genes found in each of these intersections or distinct features.

Here, the CLAMP-null mutation was present throughout the fly's entire body, whereas the CLAMP RNAi involved a brain-specific *elav-GFP* driver. Therefore, our results suggest that CLAMP has a unique role in the brain compared in other tissues of the body, namely that the same genes that might be affected in the brain are oppositely regulated in other tissues. Furthermore, the overlap of genes between CLAMP RNAi and MSL2 RNAi suggest that CLAMP and the MSL complex work together to regulate these genes because these genes are differentially expressed when the protein levels of either of these transcription factors is reduced. These genes are likely candidates to be involved in dosage compensation.

The next important figures generated are the plots illustrating the fold change differences between different overlaps or distinct features, and their corresponding experiments. To better understand the intersection of genes affected by both CLAMP RNAi and MSL2i, we can look at a plot illustrating these fold-change differences (Fig 8A). The genes that are found in these intersections are more strongly affected by CLAMP, as shown by the more negative (i.e., more downregulation) and the more positive (i.e., more upregulation) of different genes by CLAMP RNAi. The MSL complex leads to smaller differences in the fold change, which suggests that it relies on CLAMP to regulate these genes. These violin comparison plots also depict how there are common differentially expressed genes that show opposite regulation trends in CLAMP-null and CLAMP RNAi embryos (Fig 8B). The differences in the fold change are within each of these groups is significant and show that there are not just a large number of genes with close to a fold change of zero.

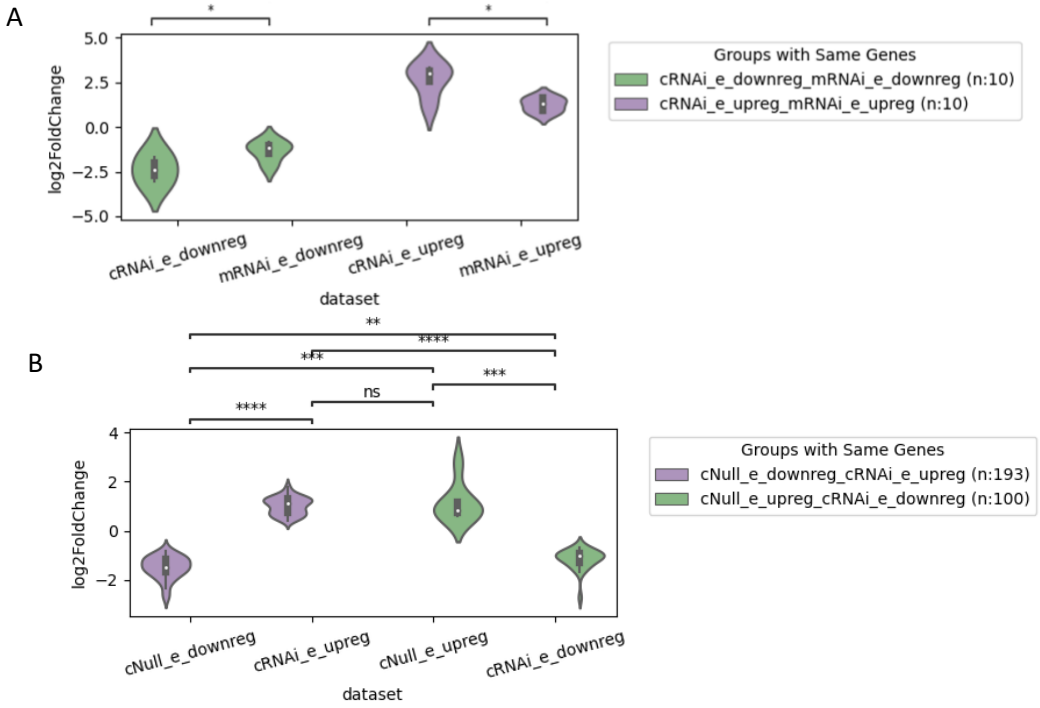


Figure 8: Violin plots comparing the fold change differences with p-values computed through the Mann-Whitney test for genes downregulated by both CLAMP RNAi and MSL2 RNAi versus genes upregulated by both CLAMP RNAi and MSL2 RNAi (panel A), and for genes downregulated by CLAMP RNAi and but regulated by CLAMP-null and vice versa (panel B).

However, further inspection of the fold change differences of the genes downregulated by CLAMP RNAi but upregulated by CLAMP-null mutation shows no difference between the magnitude of this difference between genes on the X chromosome versus the autosomes (Fig 9A). This indicates that for the common genes at this overlap, there are likely no sex-specific effects. Other groups, such as the genes downregulated by CLAMP RNAi only, show a significant difference between the fold change of the differentially expressed genes on the X chromosome versus the autosomes (Fig 9B). In this example, the genes on the autosomes are more negative fold changes, indicating that they are more downregulated, which might suggest that there are fewer dosage compensation genes effected by CLAMP alone.

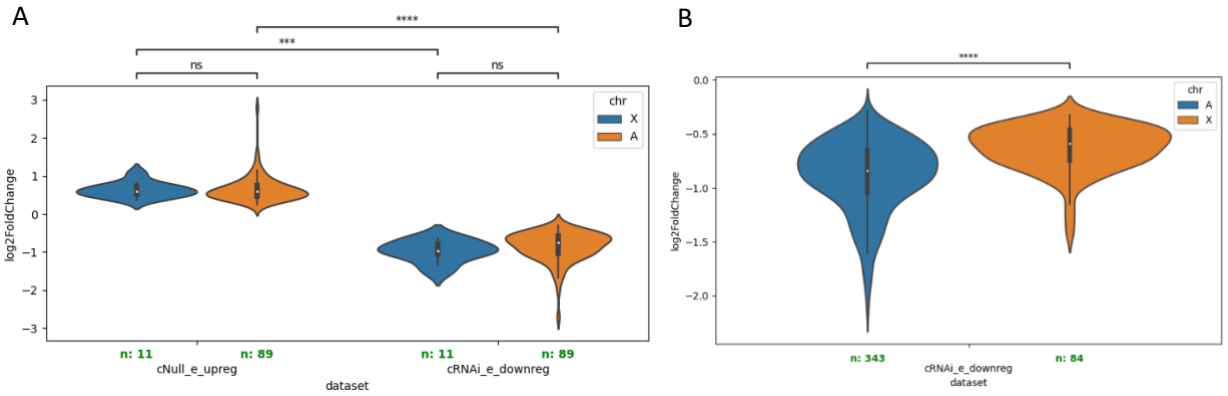


Figure 9: Violin comparing the fold change differences between genes on the X chromosome and the Autosomes with p-values computed through the Mann-Whitney test for genes downregulated by CLAMP RNAi and but regulated by CLAMP-null (panel A) and genes downregulated by CLAMP RNAi (panel B).

Based on the global expression fold change X vs Autosome plots for MSL2 RNAi data in embryos, there is no difference between the fold changes of the genes on each of the chromosomes. This might be due to the fact that the MSL2 protein plays an important epigenetic role across all the chromosomes and is responsible for global H4K16 acetylation. Another confounding factor may be that the embryos are unsexed, so any sex-specific effects or differences on the X chromosome, which would be more apparent in males, might be masked.

XvsY also facilitates continued downstream analysis by running GO analysis for each of these gene intersection clusters to help the user characterize the function and mechanism of the genes. GO term over-enrichment is performed for biological process, molecular function, and cellular component terms. Two clusters that were determined to be of interest are the genes downregulated by CLAMP RNAi alone, and the genes upregulated in CLAMP-null mutants but downregulated but CLAMP RNAi because of their involvement in neurogenesis. The genes in the former group are involved in cell-cell signaling and communication, which is particularly important in the brain and neurons for synaptic signaling, chemical synaptic transmission, and vesicle mediated transport (Fig 10A). The genes in the latter group are most significantly involved in central nervous system development, which is highlighted by the enrichment of GO terms such as neuron projection development, neuron projection morphogenesis, and cell morphogenesis involved in neuron differentiation (Fig 10B). These results indicate that in wild-

type *Drosophila*, CLAMP plays a significant role in cell-cell signaling and central nervous system development during the embryo timepoint.

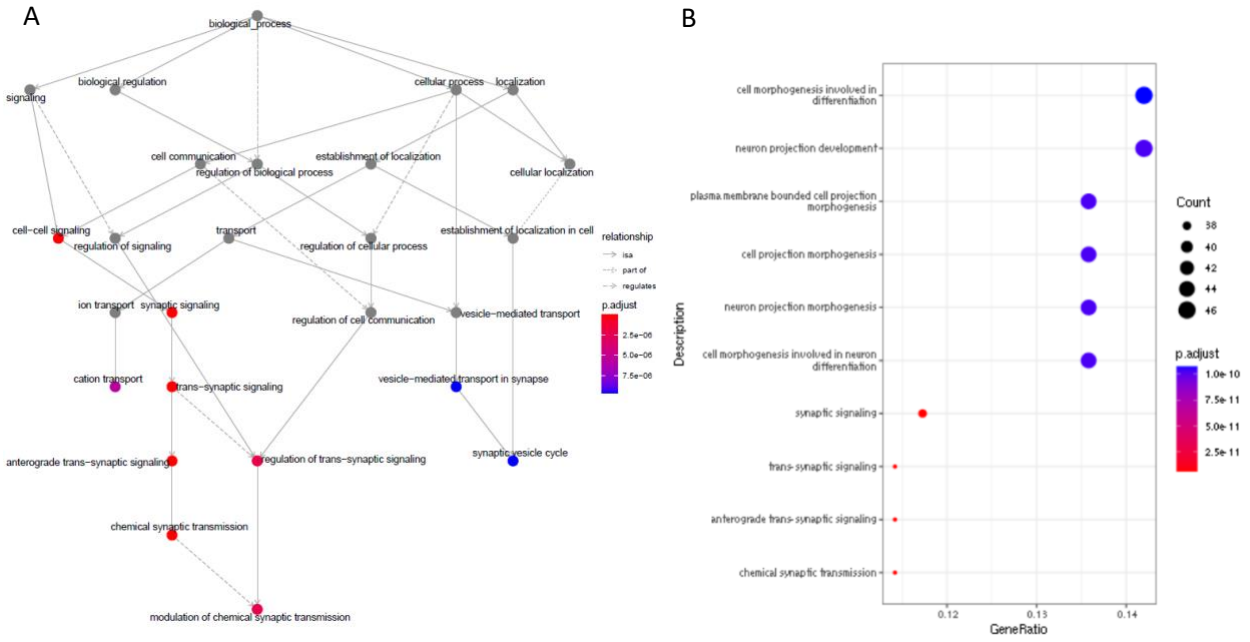


Figure 10: Results of gene ontology (GO) analysis for genes only downregulated by CLAMP RNAi shown as a type of DAG (panel A) and for genes upregulated in CLAMP-null mutants but downregulated but CLAMP RNAi shown as a dot plot (panel B).

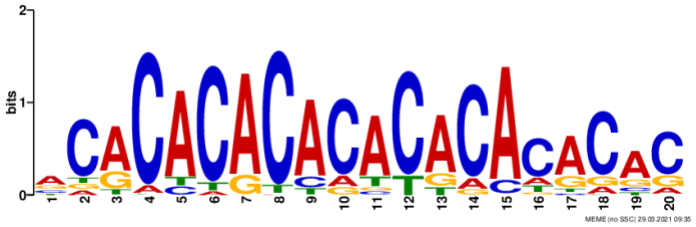
Comparisons across time suggest that in early development, CLAMP activates genes involved in neurogenesis, whereas later, during the third instar larvae and adult stages, CLAMP represses these genes based on GO analysis. Normal brain development requires the regulated growth of neurons, axons, and synapses. During these early stages, neurons usually develop superfluous connections, and this is resolved by pruning and the selective elimination of extensive processes[29]. Excessive synaptic growth can lead to various neurological and neurodevelopmental disorders. This trend is apparent in both females and males (slightly less so) at the later stages of life in *Drosophila*.

When looking at the overlap of genes between CLAMP RNAi and MSL2 RNAi in embryos, GO analysis reveals that genes involved in protein and amino acid metabolism, specifically the metabolism of methionine and glutathione, are downregulated in both conditions[8]. Thus, normally, CLAMP and the MSL complex upregulate these genes. This is notable because these are two important antioxidants and help to relieve oxidative stress in the brain. On the other hand, GO analysis also showed that genes involved in cholesterol homeostasis and sexual development are repressed by both CLAMP and MSL2 in embryos.

Lastly, the DNA sequences for each of the differentially expressed genes in each cluster was obtained and run through MEME to find the top 3 de novo motifs. For the genes downregulated by CLAMP RNAi in embryos, the motif "TTTTWTTWTWTWTTTWTWTTTWTWTWTWTTTWTWTTTTTTTTWTTTDTTT" was found with an E-value of 2.0e-178, the motif "ACACACACACACACACACAC" was found

with an E-value of 4.0e-44, and the motif “CKCCKBCSSCKYCNSCDSCBC” was found with an E-value of 1.5e-39 (Fig 11A). These motifs serve as potential candidates for other proteins to bind to and thus interact with CLAMP, our transcription factor of interest. For example, using TOMTOM[10] in the MEME Suite, a motif can be compared against a database of motifs and an alignment can be produced for each significant match. Inputting the first motif, the top match is jigr (jing interacting gene regulatory 1), which encodes a nuclear protein that is highly expressed in the central nervous system and is involved in transcriptional gene regulation[26]. Additionally, FIMO was run using the position weight matrix for CLAMP. In this same set of genes, only 51.86% of the genes have the GAGA binding motif for CLAMP (Fig 11B). This suggests that while some might be directly targeted by CLAMP, many differentially expressed genes are influenced downstream.

A



B

Motif	% Matches in Genes
RGAGAGMG	51.862464183381086
GTGTGTGTGTGTGTG	84.81375358166189
CKCTCTCTCKCTCTC	61.891117478510026

Figure 11: Results of motif and binding analysis show de novo motifs in genes downregulated by CLAMP RNAi alone as logos (panel A) and percentage of matches for motifs specified in a position weight matrix for CLAMP (panel B).

Discussion:

The XvsY pipeline allows users to complete an end-to-end analysis of multi-dimensional RNA-seq data quickly, starting from running differential expression to characterizing each gene cluster function and binding. As an efficient and accessible command-line tool, XvsY greatly simplifies the combinatorial complexity that arises from comparing multi-dimensional RNA-seq experiments. When biologists start to collect data for multiple different conditions, such as RNA interference or drugs, across multiple time points or in different sexes, the combinations of comparisons they can make grows rapidly.

In my case with data for RNA-seq experiments after removing CLAMP and MSL2 from the brain in three different developmental stages, and in both sexes in *Drosophila*, there are about 28 comparisons between all of the experiments that one can make. When looking just the embryo timepoint for the CLAMP RNAi, CLAMP-null, and MSL2 RNAi conditions, there are also 18 different gene set overlaps and unique features. If a biologist wanted to perform this same analysis on every comparison combination, there would be hundreds of gene sets to consider, which would be almost intractable to manually analyze. XvsY automatically runs this

complex analysis while providing the user flexibility for the parameters they would like to use and generating plots for each major rule in the pipeline.

The XvsY pipeline will generate 7 main folders in the personal directory specified by the user and saves all outputs based on this relative path. These folders are: 1) *data*, which contains the differentially expressed gene sets from DESeq2, 2) *other_deseq2_outputs*, which contains volcano plots for each experiment, Venn diagrams comparing the number of DEGs with the Wald test and LRT, and MA-plots, 3) *deg_sets*, which contains the upset bar plot, gene overlaps, and their corresponding BED files, 4) *comparisons*, which contains the pairwise boxplots, X vs Autosome plots, and global expression boxplots, 5) *go_analysis*, which contains all of the results from GO over-enrichment analysis and summary lists, 6) *motif_analysis*, which contains the outputs from MEME and FIMO, and 7) *summary_tables*, which contains comprehensive summary tables for each comparison. As a note on the summary tables, they contain 5 different columns: 1) overlap/unique set name, 2) percent of genes on X chromosome, 3) X vs A significance, 4) top GO terms for BP, and 5) percent of genes with FIMO motif. This will help the user, likely a biologist, to get an understanding of the network of genes affected by a given condition/perturbation.

From start to end, analysis takes about 11 minutes to run through every rule in XvsY, considering a total of 175 differentially expressed genes. For larger gene sets where there may be about 2000 DEGs, analysis takes about 18 minutes, excluding the run time for MEME. Within the automatic output inference through Snakemake, XvsY can be run very efficiently.

In the analysis of our embryo data, XvsY helped to reveal a role for CLAMP in the brain specifically for synaptic signaling and neuron projection development. Furthermore, CLAMP throughout the body of the fly is important for cell cycle regulation and RNA processing. It also showed that CLAMP and MSL2 both affect genes involved in metabolism that likely have important roles in brain function. The XvsY pipeline also provided insights into the motifs present in the gene clusters, suggesting that there may be other factors that interact with CLAMP, such as jigr1, and it highlighted that a nontrivial fraction of the genes differentially expressed by CLAMP RNAi and CLAMP-null do not have CLAMP binding sites, indicating that they are downstream targets. Lastly, XvsY helped to show that the removal of CLAMP has a larger effect on the difference in fold change that the removal of MSL2, and that CLAMP also effects many genes on the autosomes, thus its role extends beyond dosage compensation.

Future work on XvsY includes the incorporation of more multi-omics data. In particular, it would be useful to incorporate analysis of ChIP-seq or CUT&RUN data to supplement the motif and binding analysis, which would provide direct evidence of binding and interaction. This would provide users with evidence to support a more targeted and accurate gene regulatory network after analyzing their multi-dimensional RNA-seq data.

Citations:

1. Anderson, T. W., & Darling, D. A. (1954). A test of goodness of fit. *Journal of the American statistical association*, *49*(268), 765-769.
2. Anderson, J. L., Head, S. I., Rae, C., & Morley, J. W. (2002). Brain function in Duchenne muscular dystrophy. *Brain*, *125*(1), 4-13.
3. Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W., and Noble, W. S. (2009). MEME SUITE: tools for motif discovery and searching. *Nucleic acids research*, 37(Web Server issue), W202–W208.
4. Bailey, T. L., & Elkan, C. (1994). Fitting a mixture model by expectation maximization to discover motifs in bipolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, 28-36.
5. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1), 289-300.
6. Brockdorff, N., & Turner, B. M. (2015). Dosage compensation in mammals. *Cold Spring Harbor perspectives in biology*, *7*(3), a019406.
7. Conard, A. M., Goodman, N., Hu, Y., Perrimon, N., Singh, R., Lawrence, C., & Larschan, E. (2020). TIMEOR: a web-based tool to uncover temporal regulatory mechanisms from multi-omics data. *bioRxiv*.
8. Dringen, R. (2000). Metabolism and functions of glutathione in brain. *Progress in neurobiology*, *62*(6), 649-671.
9. Figueiredo, M. L., Kim, M., Philip, P., Allgardsson, A., Stenberg, P., & Larsson, J. (2014). Non-coding roX RNAs prevent the binding of the MSL-complex to heterochromatic regions. *PLoS Genet*, *10*(12), e1004865.
10. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L., & Noble, W. S. (2007). Quantifying similarity between motifs. *Genome biology*, *8*(2), 1-9.
11. Haussmann, I. U., White, K., & Soller, M. (2008). Erect wing regulates synaptic growth in Drosophila by integration of multiple signaling pathways. *Genome biology*, *9*(4), 1-17.
12. Haussmann, I. U., & Soller, M. (2010). Differential activity of EWG transcription factor isoforms identifies a subset of differentially regulated genes important for synaptic growth regulation. *Developmental biology*, *348*(2), 224-230.
13. Hodges, J. L. (1958). The significance probability of the Smirnov two-sample test. *Arkiv för Matematik*, *3*(5), 469-486.
14. Jiménez-Jacinto, V., Sanchez-Flores, A., & Vega-Alvarado, L. (2019). Integrative differential expression analysis for multiple experiments (IDEAMEX): a web server tool for integrated rna-seq data analysis. *Frontiers in genetics*, *10*, 279.
15. Khan, A., and Mathelier, A. (2017). Intervene: a tool for intersection and visualization of multiple gene or genomic region sets. *BMC Bioinformatics*, 18(1), 287.
16. Köster, J., & Rahmann, S. (2012). Snakemake—a scalable bioinformatics workflow engine. *Bioinformatics*, *28*(19), 2520-2522.
17. Listopad, S. A., & Norden-Krichmar, T. M. (2019). A-Lister: a tool for analysis of differentially expressed omics entities across multiple pairwise comparisons. *BMC bioinformatics*, *20*(1), 1-10.
18. Love, M. I., Huber, W., and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12), 550.

19. Mann, H. B., & Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, 50-60.
20. Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68-78.
21. McCorkindale, A. L., Wahle, P., Werner, S., Jungreis, I., Menzel, P., Shukla, C. J., ... & Zinzen, R. P. (2019). A gene expression atlas of embryonic neurogenesis in Drosophila reveals complex spatiotemporal regulation of lncRNAs. *Development*, *146*(6).
22. Scholz, F. W., & Stephens, M. A. (1987). K-sample Anderson–Darling tests. *Journal of the American Statistical Association*, *82*(399), 918-924.
23. Soruco, M. M., Chery, J., Bishop, E. P., Siggers, T., Tolstorukov, M. Y., Leydon, A. R., ... & Larschan, E. (2013). The CLAMP protein links the MSL complex to the X chromosome during Drosophila dosage compensation. *Genes & development*, *27*(14), 1551-1556.
24. Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American statistical Association*, *69*(347), 730-737.
25. Student. (1908). The probable error of a mean. *Biometrika*, 1-25.
26. Sun, X., Morozova, T., & Sonnenfeld, M. (2006). Glial and neuronal functions of the Drosophila homolog of the human SWI/SNF gene ATR-X (DATR-X) and the jing zinc-finger gene specify the lateral positioning of longitudinal glia and axons. *Genetics*, *173*(3), 1397-1415.
27. Tsiarli, M. A., Conard, A. M., Xu, L., Nguyen, E., & Larschan, E. N. (2020). The transcription factor CLAMP is required for neurogenesis in Drosophila melanogaster. *bioRxiv*.
28. Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP) Available at: www.genome.gov/sequencingcostsdata. Accessed March 31, 2021.
29. Williams, D. W., & Truman, J. W. (2005). Cellular mechanisms of dendrite pruning in Drosophila: insights from in vivo time-lapse of remodeling dendritic arborizing sensory neurons. *Development*, *132*(16), 3631-3642.
30. Yu, G., Wang, L. G., Han, Y., & He, Q. Y. (2012). clusterProfiler: an R package for comparing biological themes among gene clusters. *Omics: a journal of integrative biology*, 16(5), 284-287.