2020

A piRNA regulation landscape in C. elegans and a computational model to predict gene functions

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

COLLEGE OF ENGINEERING

Dissertation

A PIRNA REGULATION LANDSCAPE IN *C. ELEGANS* AND A COMPUTATIONAL MODEL TO PREDICT GENE FUNCTIONS

by

HAO CHEN

B.S., China Agricultural University, 2008

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2020

Approved by

First Reader

Charles DeLisi, PhD
Professor of Biomedical Engineering
Metcalf Professor, Science and Engineering

Second Reader

Zhiping Weng, PhD
Professor, Program in Bioinformatics and Integrative Biology,
University of Massachusetts Medical School

Third Reader

Simon Kasif, PhD
Professor of Biomedical Engineering

Fourth Reader

Zhenjun Hu, PhD
Research Associate Professor of Biomedical Engineering

知之为知之，不知为不知，是知也。

To know what it is that you know, and to know what it is that you do not know,----that is understanding.

论语

Analects of Confucius

Acknowledgments

I am fortunate to have a supportive environment for my *Ph.D.* training with mentors and friends. Among them, my most sincere thanks must go to my advisors, Prof. Charles Delisi and Prof. Zhiping Weng, and my other committee members, Prof. Simon Kasif, Prof. Kirill Korolev, and Prof. Zhenjun Hu. Their critics, open-minds, and patience are a joint, encouraging force for me to proceed further throughout the years.

Secondly, a good *Ph.D.* training life can not go without great mentors, helpful co-trainees, and friends. It is my great honor to have Dr. Xiaopeng Zhu, Prof. Shikui Tu, Prof. Qiong Wu, and Prof. Wen Zhang as my mentors and friends. As experienced researchers, they informally guided me through criticizing my ideas, cheering on my progress, and discussing a broad spectrum of scientific topics with drinks. I'd also like to thank all current and previous "zlab" members. While bearing my immature presentations and my competitions in getting computational resources, they are always remarkably friendly and helpful. In addition, I am truly grateful to have "zlab" members Dr. Micheal Purcaro and Dr. Arjan Vandervelde as system administrators. Their prompt and skilled responses are luxuries that should not be expected.

Also, I would like to pay my special regards to the administration staff, especially Dave King, Caroline C. Lyman, Rhonda O'Brein, Heidi Beberman and Barbara Bucciaglia, in both the bioinformatics program at Boston University and the bioinformatics and integrative biology program at the University of Massachusetts Medical School. Their dedicated work is essential for a smooth joint training.

I am also grateful for having my great parents, Mr. Jingbo Chen and Mrs. Feng Li. Their continuous care and love are the structural supports to my heart. At last but definitely not least, I really appreciate the countless help from my wife-to-be and

friend of life, Miss. Congcong Zhu. Without her company throughout the years, this journey would be much tougher with headwinds.

Though I wish to thank all the people whose assistance was necessary for the completion of this training, I am certain that I have missed many. So, I would like to end this acknowledgment with a ton of thanks to all of them.

Hao Chen

A PIRNA REGULATION LANDSCAPE IN *C. ELEGANS* AND A COMPUTATIONAL MODEL TO PREDICT GENE FUNCTIONS

HAO CHEN

Boston University,

Graduate School of Arts and Sciences,

College of Engineering, 2020

Major Professor: Charles Delisi, PhD
Professor of Biomedical Engineering
Metcalf Professor, Science and Engineering

ABSTRACT

Investigating mechanisms that regulate genes and the genes' functions are essential to understand a biological system. This dissertation is consists of two specific research projects under these aims, which are for understanding piRNA's regulation mechanism and predicting genes' function computationally.

The first project shows a piRNA regulation landscape in *C. elegans*. piRNAs (Piwi-interacting small RNAs) form a complex with Piwi Argonautes to maintain fertility and silence transposons in animal germlines. In *C. elegans*, previous studies have suggested that piRNAs tolerate mismatched pairing and in principle could target all transcripts. In this project, by computationally analyzing the chimeric reads directly captured by cross-linking piRNA and their targets in vivo, piRNAs are found to target all germline mRNAs with microRNA-like pairing rules. The number of targeting chimeric reads correlates better with binding energy than with piRNA abundance,

suggesting that piRNA concentration does not limit targeting. Further more, in mRNAs silenced by piRNAs, secondary small RNAs are found to be accumulating at the center and ends of piRNA binding sites. Whereas in germline-expressed mRNAs, reduced piRNA binding density and suppression of piRNA-associated secondary small RNAs targeting correlate with the CSR-1 Argonaute presence. These findings reveal physiologically important and nuanced regulation of piRNA targets and provide evidence for a comprehensive post-transcriptional regulatory step in germline gene expression.

The second project elaborates a computational model to predict gene function. Predicting genes involved in a biological function facilitates many kinds of research, such as prioritizing candidates in a screening project. Following the "Guilt By Association" principle, multiple datasets are considered as biological networks and integrated together under a multi-label learning framework for predicting gene functions. Specifically, the functional labels are propagated and smoothed using a label propagation method on the networks and then integrated using an "Error correction of code" multi-label learning framework, where a "codeword" defines all the labels annotated to a specific gene. The model is then trained by finding the optimal projections between the code matrix and the biological datasets using canonical correlation analysis. Its performance is benchmarked by comparing to a state-of-art algorithm and a large scale screen results for piRNA pathway genes in *D.melanogaster*.

Finally, piRNA targeting's roles in epigenetics and physiology and its cross-talk with CSR-1 pathway are discussed, together with a survey of additional biological datasets and a discussion of benchmarking methods for the gene function prediction.

Contents

List of Tables

List of Figures

List of Abbreviations

21U-RNAs	.....	piRNA 21nt long with a 5' uridine
22G-RNAs	.....	siRNA 22nt long with a 5' guanosine
AID	.....	Auxin-Inducible Degradation
BEAR	.....	Brand nEw Alphabet for RNA
BLASTP	.....	Basic Local Alignment Search Tool for Protein
AGO	.....	Argonautes
CCA	.....	Canonical Correlation Analysis
CLASH	.....	Crosslinking, Ligation, And Sequencing of Hybrids
CRISPR	.....	Clustered Regularly Interspaced Palindromic Repeats
ECOC	.....	Error correction Of Code
IP	.....	immunoprecipitation
Ka/Ks	.....	Synonymous/Nonsynonymous
KEGG	.....	Kyoto Encyclopedia of Genes and Genomes
modENCODE	.....	Model Organism ENCyclopedia Of DNA Elements
MIPS	.....	Munich Information Center for Protein Sequences
phyloP	.....	conservation or acceleration phylogenetic p-values
piRISC	.....	piRNA-induced Silencing Complex
piRNA	.....	PIWI interacting small RNA
PRINCE	.....	PRIoritizatioN and Complex Elucidation
RdRP	.....	RNA-dependent RNA Polimerase
RWR	.....	Random Walk with Restart
SOIS	.....	Second Order Iterative Stratification
SVM	.....	Support Vector Machine
WAGO	.....	Worm Argonautes
WS230	.....	Wormbase Sequence annotation version 230
WT	.....	Wild Type

Chapter 1

Introduction

Argonaute (AGO) proteins and their engaged small RNAs regulate genes at both transcriptional and post-transcriptional levels in many species. (Czech and Hannon, 2011; Ghildiyal and Zamore, 2009; Hutvagner and Simard, 2008; Meister, 2013; Siomi and Siomi, 2009; Thomson and Lin, 2009). Among them, PIWI proteins are members of the RNaseH-related Argonaute superfamily that engage small RNAs to function in animal gonad, as a piRNA-induced silencing complexes (piRISCs) (Czech and Hannon, 2016; Malone and Hannon, 2009; Weick and Miska, 2014). As the two projects in this dissertation are both aimed at exploring the piRNA pathway, I will start by introducing this pathway in two model organisms.

Although the genomic origin, sequence features vary a lot in *C. elegans* and *D. melanogaster*, some of their biological functions appear to be shared, such as transposon element silencing and fertility maintenance (Aravin et al., 2001; Batista et al., 2008; Siomi et al., 2011; Thomson and Lin, 2009). In *D. melanogaster*, piRNA originates from long non-coding RNAs precursors called piRNA clusters, with size up to 200kb in length and serve as the template for thousands of unique piRNAs (Huang et al., 2017). Once transcribed, the piRNA clusters are exported to cytoplasm through nuclear pores and processed into mature 23-29 nt piRNAs through 5' and 3' end nucleotide trimming and modifications. In the end, a so-called "ping-pong" cycle amplifies the amount of piRNAs for silencing.

However in *C. elegans* (See figure 1.1), piRNAs originate from many small non-

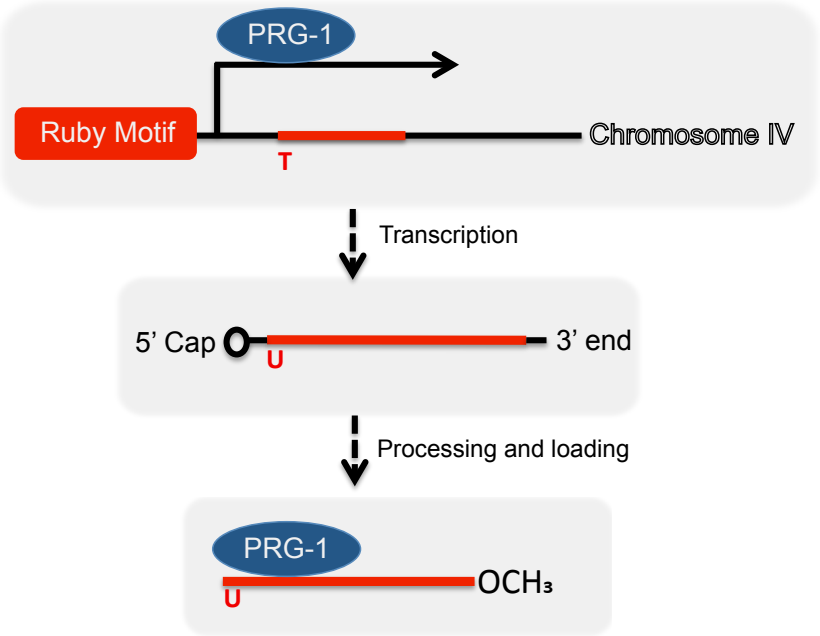


Figure 1·1: The piRNA's biogenesis in *C. elegans*.

coding gene precursors, each with a unique Ruby motif at the gene's up-stream (Batista et al., 2008; Weng et al., 2019). Once transcribed, these precursors go through a similar process that trims the piRNA into 21nt with a signature 5' uridine (21U-RNAs). Rather than using the "ping-pong" cycle to amplify silencing signal, the piRISC targets template RNAs and recruit an RNA-dependent RNA Polimerase (RdRP) for initiating the synthesis of secondary small RNAs. These secondary RNAs are typically 22nt long with a 5' guanosine (22G-RNAs) that engage an expended group of 12 worm Argonautes (WAGOs) to silence transposons and many endogenous genes.

1.1 piRNA regulation in C. elegans

In *C. elegans*, previous computational analysis suggests piRNA are targeting with mismatch tolerances, indicating thousands of endogenous mRNAs can be targeted

by them and silenced (Lee et al., 2012). On the other hand, the CSR-1 pathway is thought to be a "self" recognition pathway that protects endogenous mRNAs, which serves an anti-silencing mechanism to prevent or reduce the sensitivity of the piRISC silencing (Seth et al., 2013). CSR-1 also engages RdRP-derived smRNAs templates from nearly all germline expressed genes (Claycomb et al., 2009). However, it is unknown whether it directly competes with initial piRISC targeting or with the downstream WAGO recruitment, as shown in figure 1.2. Thus, direct genomewide

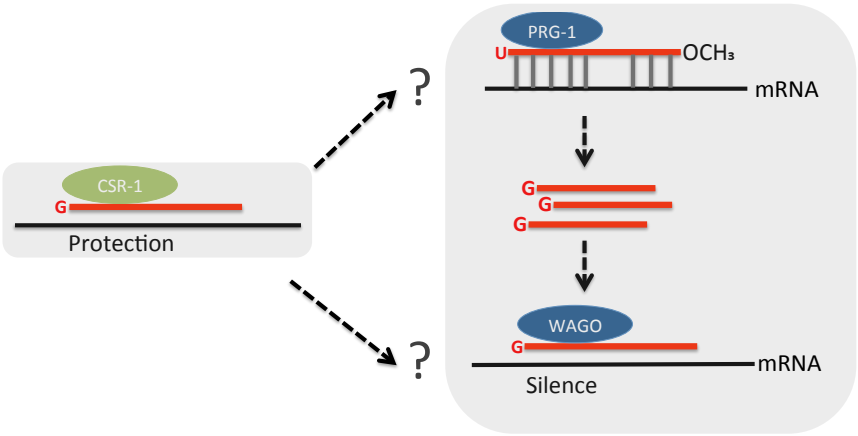


Figure 1·2: Diagram for CSR-1 protection pathway's potential cross-talks with piRNA regulation in *C. elegans*. The diagram on the right shows the piRNA biogenesis in *C. elegans*. And the arrows indicate CSR-1 protection pathway's potential cross-talk points.

identification of piRNA targets is essential for deciphering the mechanism for both sequence-directed immunity and germline gene regulations. In chapter 2, I will show a joint project that directly captured around 200,000 high-confidence piRNA-target interactions and deciphered the piRNA functioning mechanisms via computational analysis and experimental validations.

1.2 The efforts to predict piRNA pathway genes

Dozens of genes are involved in piRNA's biogenesis, exporting, and maturation processes. To further understand its function, it is essential to identify additional gene components in this pathway. Both computational and experimental efforts have been made in several model organisms. In *C. elegans*, Tabach *et al.* (Tabach et al., 2013) used a normalized phylogenetic profile to group genes involved in miRNA and siRNAs pathways and then experimentally validated around half of the candidates. In *D. melanogaster*, large scale RNAi screening projects have been conducted to identify novel piRNA pathway proteins (Czech et al., 2013). However, these screening projects are usually not complete, due to experimental limitations. In chapter 3, I will elaborate on a novel computational method that aims to predict additional co-functional genes and its application in predicting piRNA pathway genes in *D. melanogaster*. To provide additional view from this second angle, I will also introduce the background and motivation for the gene function prediction problem in computational biology in the following sections.

1.3 Computational models to integrate biological data

Understanding and modeling biological systems from a systematic view facilitates many biological research projects, such as experimental screening for genes in pathways and prioritizing disease-related genes testing for candidate biomarkers (Zitnik et al., 2019). The rapidly developing large-scale and high-throughput methods in recent years enable investigations from multiple aspects, such as cis-regulatory elements, topologically associated domains, gene expression, protein-protein interactions, and genetic interactions (Chatr-Aryamontri et al., 2017; ENCODE Project Consortium et al., 2012; Nora et al., 2012; Szklarczyk et al., 2019). In principle, an integration of these datasets compensates for the missing data and alleviate the noise in biological

datasets. Nevertheless, it also provides a more comprehensive approach for repurposing publicly available datasets. Several efforts have been made for summarizing these datasets, such as visAnt and stringDB (Hu et al., 2013; Jeanquartier et al., 2015). However, many methods are still designed for analyzing a single type of data, avoiding the different reliability and the inherent systematic bias from different experimental designs (Hwang et al., 2005).

On the other hand, various approaches were also proposed for integrating biological datasets. A recent review (Zitnik et al., 2019) classifies them as three types. As the first type, early integration methods typically concatenate and transform multiple inputs into one dataset, where feature selection methods or dimension reduction techniques are usually involved. On the other side, the late integration methods build separate models for each dataset and then combine their outputs for final predictions, where the aggregated ranks or p-values can be used. The method I proposed for gene function prediction is an intermediate integration method under this review. Using the network view of biological datasets and the "guilt by association" (GBA) principle, I used a multi-label approach for gene function prediction, which learns a joint representation of gene functions by optimizing correlations between the known gene functions and the input datasets. As a result, this learned latent representation can decode the biological datasets for predicting gene functions.

In the following sections, I will introduce this GBA principle, the biological datasets as networks, and a label propagation algorithm that propagates the functional labels on the networks.

1.4 "Guilt by association" (GBA) principle

The GBA principle assume that genes interacting or associated tend to share functions. Under this principle, we can propagate the functional annotation of a gene

to its not annotated neighbors on a gene-gene network. As one of the early work in the field, Schwikowski *et al.* show that a simple majority vote from physically interacting neighbors can correctly assign at least one function to 72% of the genes (Schwikowski et al., 2000). In addition to the physical interactions, other networks following GBA can also contribute to predicting gene functions. For instance, if two genes are involved in the same pathway, they may be co-evolving across multiple species (Pellegrini et al., 1999) and co-express at the same tissue type (Li et al., 2014). They might also be annotated with related GO terms (Resnik, 1999) and interacting with each other genetically (Costanzo et al., 2016). In principle, by quantifying these relationships and join them, we can build up a refined GBA view. And this view can help in propagating a functional annotation from known genes to others.

1.5 Biological data as "multi-view" networks

One conceptually meaningful way for integrating these GBA based relationships is to view them as networks, where each node is a gene or functioning element, and each edge is a quantified association between them. This transition is quite straightforward for some biological relationships. For instance, they can be discrete numbers from physical protein-protein interactions captured by experiments. They can also be continuous values from quantified genetic interaction of genes (Chatr-Aryamontri et al., 2017). In addition to these straightforward definitions, other biological data can also be viewed as networks. For instance, genes' phylogeny across species (Pellegrini et al., 1999) and their expression in different tissue types or developmental time points (Carlson et al., 2006) can be quantified as phylogenetic and expression profiles, respectively. These profiles can then quantify the relationships between genes via Pearson correlations. Also, using the frequencies of accumulated gene ontology (GO) annotations in a species, we can define a semantic similarity between genes (Resnik,

1999). Taken together, these biological datasets provide a "multi-view" of genes' functions. I will elaborate more on viewing phylogeny, expression, and GO annotation as networks below.

1.5.1 Genes' phylogeny as biological network

Phylogenetic profiling is a powerful way to identify evolutionary and functional associations between genes (Pellegrini et al., 1999). It takes advantage of the fact that nature tends to delete or maintain the whole functional gene set, rather than a random part of them. Thus, given a gene in a "center species", its phylogenetic profile can be defined as its present/absent pattern across multiple species. If two genes are sharing a similar pattern, they have a higher chance of co-evolving and co-function (Figure 1.3).

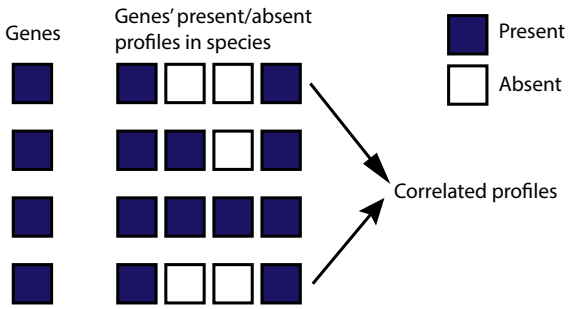


Figure 1·3: Diagram for associating similar phylogenetic profiles.

In Eukaryotes, Tabach *et al.* proposed a way to identify small silencing RNA pathway genes in *C. elegans* (Tabach et al., 2013). They defined phylogenetic profiles for genes in *C. elegans* as their sequence-level similarity to best-matching homologs in 95 other species. These similarities are quantified as the BLASTP bit scores, after normalizing them to sequence length and evolutionary distance (Zscore transformation for each species). In order to show the rationales behind my computational model,

and also as a preliminary result, I replicated this method and calculated the profiles for each *D. melanogaster* gene, using 52 evolutionarily distant species that are at least 500 million years away from each other (Kumar et al., 2017) (See appendix table 1). Then, as mentioned previously, the associations between genes are quantified using their profiles' Pearson correlation coefficients.

1.5.2 Genes' expression profile as biological network

Following the GBA principle, co-expression profiles is another approach for grouping functionally associated genes (Carlson et al., 2006). Similar to phylogenetic patterns, it assumes co-functioning genes tend to express cooperatively, thus correlated with each other with a higher Pearson correlation coefficient. Ideker *et al.* (Ideker et al., 2001) pioneered this idea by analyzing large-scale mRNA and protein level responses to 20 perturbations to genes involved in galactose utilization. In this study, they also showed that genes linked by physical interactions correlated more strongly than randomly chosen genes, indicating that these biological datasets are following GBA principle and they can compensate for each other. In 2002, Steffen *et al.* (Steffen et al., 2002) utilized co-expression profiles to reconstruct a MAP Kinase signaling network in yeast, together with its protein interaction map. More relevant to this work, Karaoz *et al.* (Karaoz et al., 2004) proposed a model based on a Markov random field and label propagation, which integrates expression and protein-protein interaction data for gene function prediction.

In the preliminary result in this chapter, expression profiles for *D. melanogaster* are built using normalized gene expression values from the modENCODE project, which cover mnay tissues types and developmental time points (See appendix tables 2 and 3) (Li et al., 2014).

1.5.3 Genes' similarity networks defined using GO annotations

The accumulated gene functional annotations can also define a similarity score between genes. Intuitively, GO terms enriched for cooccurrence in a species are more similar to each other. By summarizing the semantic similarities between GO terms, this similarity between any two genes can be estimated (Lord et al., 2003; Guzzi et al., 2012). Specifically, given the GO terms' tree-like structure and their empirical frequencies in a species, we can quantify the semantic similarity between two GO terms using their shared parental terms' frequencies.

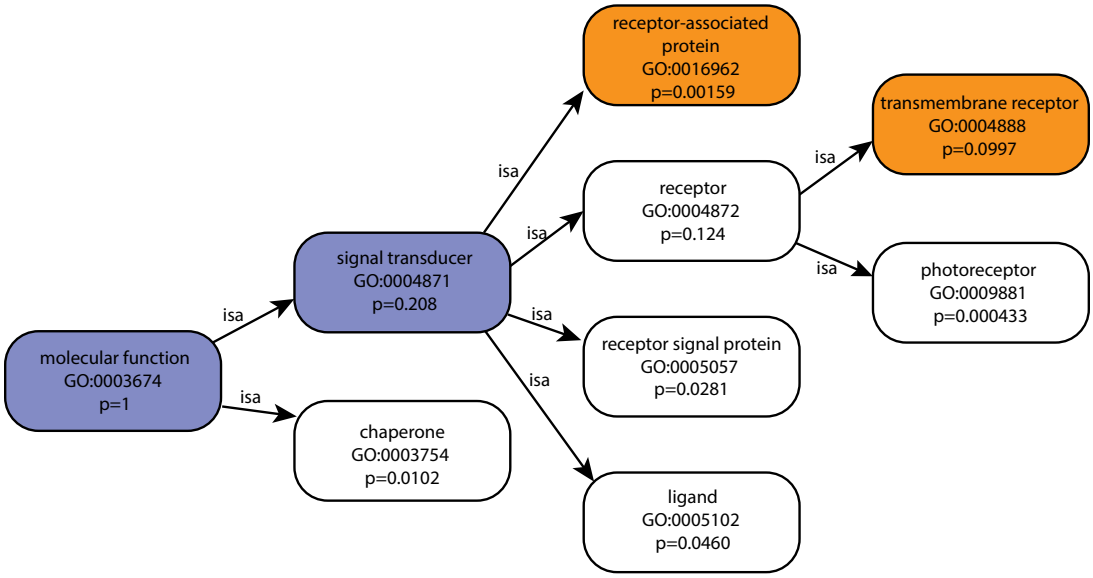


Figure 1·4: Diagram for GO terms' tree-like structure and empirical p-values. Orange indicates the two focused GO terms for similarity calculation. Blue indicates their parental terms. Figure reproduced using the toy example in Resnik *et al* (Resnik, 1999).

As shown by Resnik *et al.* (Resnik, 1999), two specific GO terms such as "transmembrane receptor (GO: 0004888)" and "photoreceptor (GO: 0009881)" are more similar to each other, compared to the most general term (root term) on the tree, which is "molecular function (GO: 0003674)". As the frequency of the GO term

monotonically increases from the leave terms to the root, the frequencies are normalized to a set of empirical p-values for each GO term, ranging from 0 to 1, where 1 is the p-value for the root GO term. Then, given the tree structure and p-values for each GO term, the similarity scores between any two GO terms are defined as the negative log of the smallest p-value of their shared parental terms.

$$similarity(c_1, c_2) = -log(min(p(S)))$$

Here, S denotes the set of shared parental terms' p-values (two blue-colored GO terms in figure 1.4). c_1 and c_2 denote the two focused terms for similarity calculation (two orange-colored GO terms in figure 1.4). In figure 1.4, the similarity of the two focused GO terms is 0.682, according to this definition.

With the GO terms' similarities defined, the genes' similarities are defined by integrating the GO term annotated to them. Various integration strategies have been proposed and compared (Guzzi et al., 2012). Among them, one simple way is to define the similarity between genes as the max of all GO term similarities. By considering each gene as a "node" and their similarities as edges, I generate a network view of genes based on their GO annotations. In the preliminary results, I calculated the GO term similarities for the three GO term trees, namely biological process, cellular component, and molecular function, and integrated them for gene's similarity scores using R package GOsemsim (Yu et al., 2010).

1.6 Propagate functional labels on a network

Following the "guilt by association" principle, the label propagation algorithms propagate genes' functional labels to their neighbors on a network. As an extension of the simple neighbor majority voting model for the yeast networks (Schwikowski et al., 2000), Hishigaki *et al.* (Hishigaki et al., 2001) further propagate the labels to genes in a radius. However, this approach does not consider local network topology restric-

tions. Vazquez *et al.*, and Karaoz *et al.* (Vazquez et al., 2003; Karaoz et al., 2004) take the topology into account by considering the functional prediction problem as a multi-way k-cut problem. However, as pointed out by Nabieva *et al.* (Nabieva et al., 2005), these methods don't reward local proximity. To overcome this limitation, they proposed a FunctionalFlow algorithm that used the idea of network flow, outperforming the methods above for gene function prediction in yeast. It controls the number of flow iterations with a parameter d, and typically used $d = 6$ for the yeast interaction network.

Similar to FunctionalFlow, PRIoritizatioN and Complex Elucidation (PRINCE) algorithm (Vanunu et al., 2010) is the network propagation method that I integrated into my computational model (Figure 1.5). It is a "Random Walk with Restart"

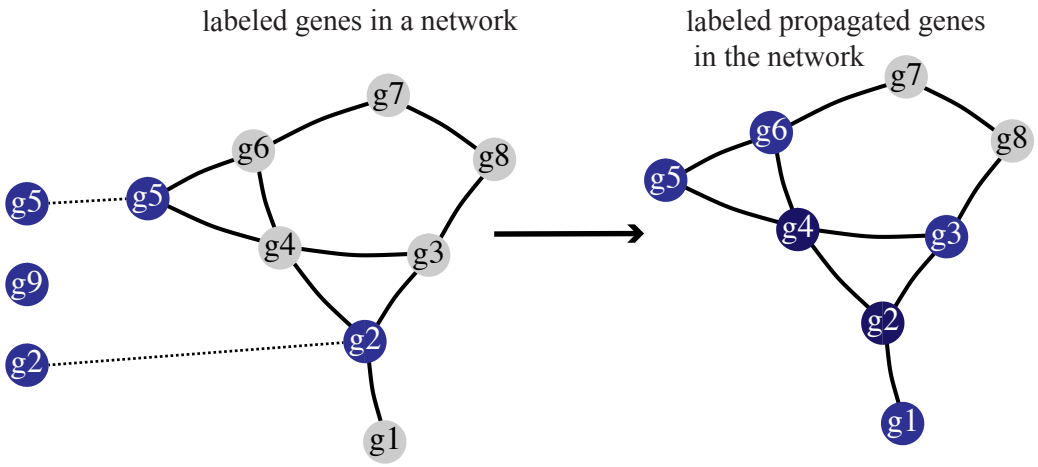


Figure 1·5: Diagram for propagating label on a gene-gene network. Blue indicates the labeled genes before and after the label propagation algorithm. In this example, gene g5 and g2 found in the network, and their associated label values are propagated to their neighbors. After reaching a stationary state, genes such as g4 are predicted to have the same label.

(RWR) approach that iteratively diffuses the label values into a network.

Given a set of co-function genes, the algorithm firstly finds them in a gene-gene

interaction network, and annotate these genes with a positive value, such as 1. Due to the network's topology restriction, they can only connect to a subset of other genes directly. PRINCE takes advantage of this restriction and iteratively propagates the label through weighted edges. At each iteration, propagation proceeds from genes with positive label values to all their neighbors. At the same time, channeled by the initially mapped genes, more label values are continuously infused into the network, which "restarts" the label propagations and smooths the distribution of label values. In PRINCE, a parameter α controls the proportion of restart label values that are infused at each iteration. In the end, the algorithm will terminate at a steady state when the amount of total changing label values between genes drops below a threshold. As a result, some previously unlabeled genes are annotated with positive label values and predicted.

1.7 Rationales for the proposed computational model

Although the mentioned methods following GBA principle are useful, computationally predict genes' function using the label propagation algorithms on multiple networks is challenging for several reasons. Firstly, the genes multi-function nature limit a model's prediction power, particularly for networks built following GBA principle. As multiple labels may co-exist for the same set of genes, it confounds the computational models that propagate one label at a time to additional genes. Nevertheless, missing annotations and class imbalance problems are also limiting the method performances. I will discuss them in detail in this section, with the preliminary results from my analysis on *D. melanogaster*'s networks.

1.7.1 The evolutionary origin of scale-free biological networks

The genes' multi-function nature might be rooted in the biological networks' distribution and evolutionary origin. Though being controversial (Broido and Clauset, 2019),

biological networks are generally considered to be scale-free, with a "power-law" distribution. This distribution can be denoted as the following.

$$P(k) \sim k^{-\gamma}$$

Here, k is the node degree and $P(k)$, the probability of randomly selecting a node with degree k, is proportional to $k^{-\gamma}$ (Barabási and Oltvai, 2004; Zhu et al., 2007). This is a long-tail distribution where genes with high degree are termed "hub" genes.

As discussed previously by Barabasi *et al.* (Barabási and Oltvai, 2004), the emergence of this scale-free property is probably correlated with gene duplication events, which produce identical proteins that interact with the same partners. While the original gene preserves its function, the network gains new gene functions as the newly duplicated genes evolve. As a result, genes highly connected in the network have a higher chance of gaining new interactions and functions.

1.7.2 Multi-function genes limit prediction power

Probably originated from the evolution of networks, the multi-function "hub" genes are hard cases in gene function predictions, as they typically correlated with a large number of partners with different functions. In the scenario of label propagation, it channels the label value to all its neighbors, resulting in a large number of false positives in prediction.

Here, as an preliminary result, a pathway to pathway relationship network is generated using iGraph (Csardi and Nepusz, 2006) and KEGG pathway (Kanehisa and Goto, 2000) annotation for *D. melanogaster*, where each node is one pathway and each edge denotes at least one multi-functioning genes is shared between the connected pathways.

This view indicates that many pathways are sharing genes with others, where

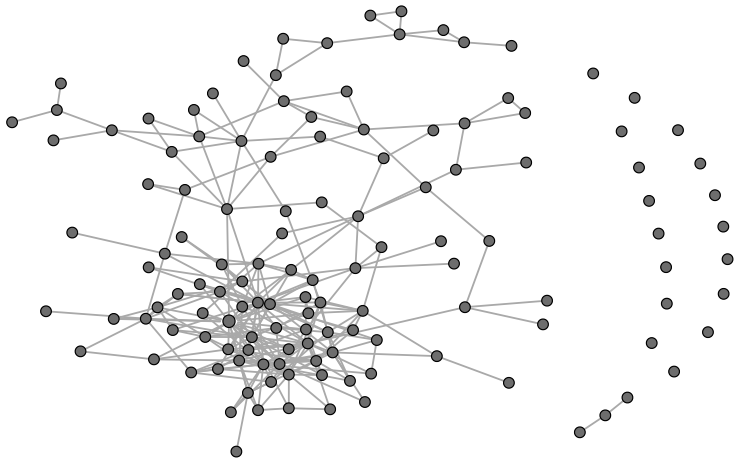


Figure 1·6: The view of KEGG pathway-pathway relationship in *D. melanogaster*. Each node is a KEGG pathay and each edge indicates that at least one gene is shared between two pathways.

some of them are densely connected. In other words, many genes are multi-funtional in *D. melanogaster*. To further support my hypothesis, I also calculated the average pathway-pathway similarities, using all possible gene pairs between two pathways. Specifically, To provide a "multi-view" of the pathways relationship from different views, I calculated genomewide associations or semantic similarities using phylogeny, co-expression, and GO terms frequencies in *D. melanogaster*.

In these case studies, similarities to a particular pathway from all pathways are calculated and used to color the nodes on the KEGG pathway view, including pathway "dme00020", the citrate cycle (TCA cycle, Krebs cycle), and pathway "dme04392", the hippo signaling pathway in *D. melanogaster* (Figure 1.6).

The TCA cycle is one of the "house-keeping" pathways that conserved across species. As shown in Figure 1.6A, it is highly associated with many other pathways in terms of co-evolution correlations. As a second example, the hippo signaling pathway is also correlating with other pathways, though to a lesser extent (Figure 1.6C). Interestingly, both pathways correlate with more others from the co-expression per-

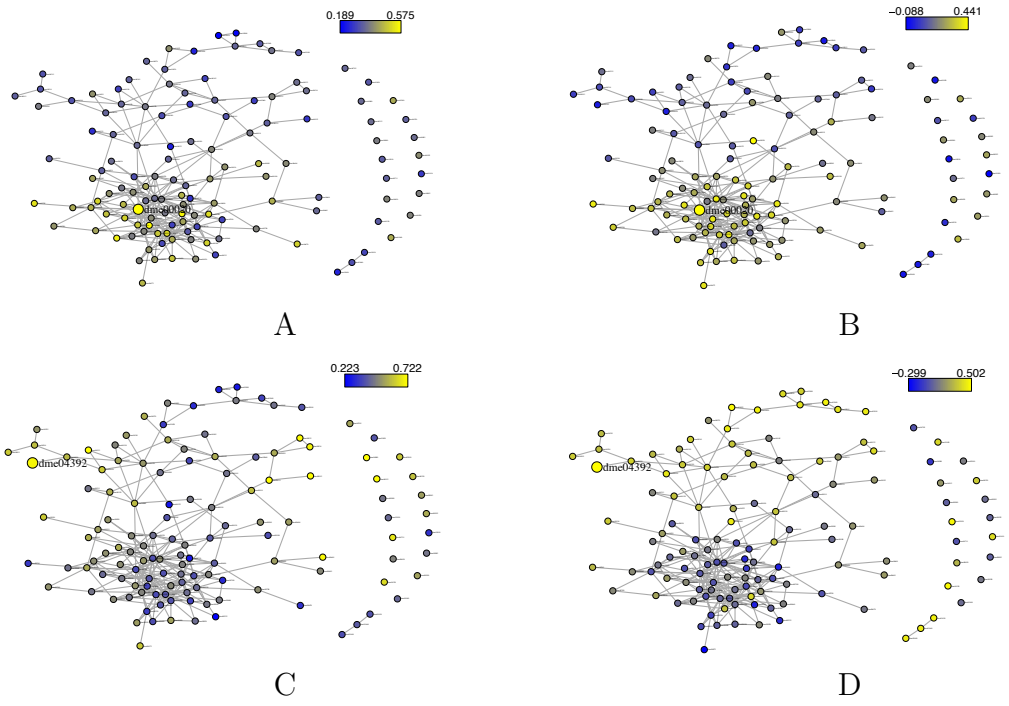


Figure 1·7: The view of KEGG pathway similarities to pathway "dme00020" and "dme04392" in *D. melanogaster*. Pathway nodes' are colored according to their associations to a focus pathway, which is indicated by a larger sized node. Pathways with lower associations are in blue, while those with higher associations are in yellow. The pathway relationships focusing on pathway "dme00020" and "dme04392" from phylogenetic profiles' perspective are shown in A and C; while those from co-expression profiles are shown in B and D.

spective (Figure 1.6B and 1.6D). Also, their network views from three different GO term trees are also diverse in their closely associated pathways (See appendix figure 1). This diversity suggests that this multi-view of networks indeed provides different views following GBA principle, which may compensate each other.

To summarize the case studies for this multi-function phenomenon, the pathway similarities in *D. melanogaster* are stratified into five groups based on their shortest path distances. As shown in Figure 1.8, the pathway similarities decrease as the pathway distances increase, indicating that they follow the GBA principle. The neighboring pathways with a distance of 1 and 2 do have a higher similarity to each other,

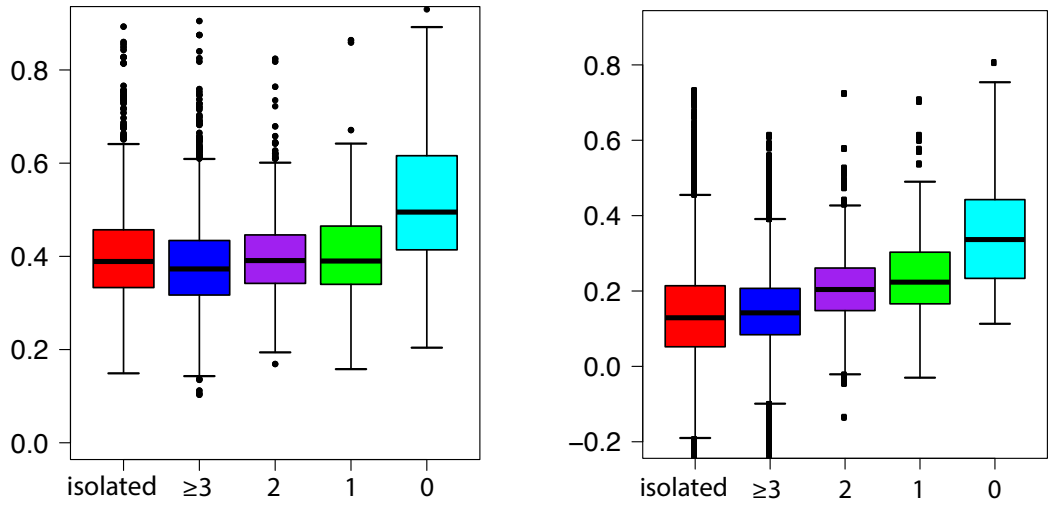


Figure 1·8: Similarities between KEGG pathways in *D. melanogaster* at different distances. A and B are for co-evolution and co-expression correlations, respectively. Cyan, green, purple, and blue boxplots are for pathway pair similarities with their shortest distances as 0, 1, 2, greater or equal to 3, respectively. Red boxplots are for self similarities of isolated pathways.

comparing to the distant pathways. This indicates that the GBA defined similarities are less effective in distinguishing the pathway with their neighboring pathways. In the proposed computational model, these relationships are taken as advantages and modeled as label dependencies.

1.7.3 Missing labels and highly imbalanced annotations

Since annotating gene function is still an on-going effort even for model organisms, missing labels are expected in training for gene function prediction algorithms (Liu and Thomas, 2019). For instance, while the yeast gene annotation can be considered as completed, human gene annotations are still largely missing.

These missing labels further confound the genome-wide gene functional predictions. And, this might also be the reason for the success of label propagation algo-

rithms in the past years (Cowen et al., 2017; Nelson et al., 2019; Vanunu et al., 2010; Nabieva et al., 2005; Karaoz et al., 2004; Vazquez et al., 2003), since label propagation does require true negative labels for training. As a result, missing negative labels have a moderate effect on them. However, missing positive labels can erroneously mark a top prediction as false positive, misleading the hyperparameter tuning process.

In addition to this missing label problem, the number of positive labels for a pathway is generally much less than the total number of genes in a species, resulting in a highly imbalanced dataset. In the case of piRNA pathway prediction in *D. melanogaster*, the number of known piRNA pathway genes is 33, much less than the nearly 12,000 genes in the species. In a worst-case scenario, an algorithm can reach over 99% accuracy by simply predicting all instances to be negative for this pathway.

1.8 Gene function prediction as a multi-label learning problem

Multi-label learning eases these problems by modeling multiple labels and their dependencies at the same time (Zhang and Zhou, 2014). To show the advantage of this multi-label learning framework, a toy example of six multi-label genes is prepared (Table 1.1). As shown in the table, this framework models a larger number of genes with positive labels, if compared to just 2 genes for each function label. The missing negative label problem is also reduced by the label dependency modeling process, as some pathways do not share genes, such as function 1 and 2. They are still not true negatives to each other, but by modeling the label distribution of the two functions, the partial negative dependencies can be taken into account. Also, to ease the impact of missing positive labels in this work, a robust Platt's scaling procedure is adopted in my computational model (see chapter 3 for detail).

Table 1.1: Multi-label learning schema for gene predictions. 1 indicates a gene is annotated with a function.

	function 1	function 2	function 3	function 4	function 5
gene 1	1	0	1	1	0
gene 2	0	1	0	0	1
gene 3	0	1	0	0	0
gene 4	1	0	0	1	0
gene 5	0	0	0	0	0
gene 6	0	0	1	0	1

1.9 "Error correction of code" for gene function prediction

Among multi-label learning methods, "Error correction of code" (ECOC) (Dietterich and Bakiri, 1994; Escalera et al., 2010) is quite straightforward to integrate single label propagation results, as a "codeword" can be an indicator vector of a gene' functions. For instance, gene 1 in Table 1.1 can be coded as "10110" for the five gene functions. Also, this framework corrects errors using repetitive codes, which is suitable for integrating the "multi-view" datasets in biology.

ECOC is originally from telecommunication community. In this field, it is important to minimize the errors in recovering original message while limiting the repetitive "codeword" length. Multiple coding designs were proposed for this purpose and multi-label learning, such as adding true negative labels, and randomly choosing additional label pairs as the repetitive codes (Escalera et al., 2010). Among them, one well-known schema is called "Turbo code"(Weiss and Freeman, 2001), which embeds a fixed pattern in the "codeword" during encoding and, upon receiving, decodes the original "codeword" back by maximizing the likelihood to this pattern through an iterative process.

For predicting gene functions using the "multi-view" datasets, the optimal design for transmitting efficiency is not as important as in telecommunications. Instead, it is important to fill a partially known "codeword" matrix with the help of biological

datasets. Here, the original codeword is partially known, and the biological datasets can be considered as the encoded repetitive codes through an unknown design. The goal is then to learn a coding design that fits both the known codewords and biological datasets. If it is learned, new function labels can be predicted by decoding the biological datasets from the same distribution. In chapter 3, I will propose a unified model that connecting label propagation algorithms and an ECOC based algorithm (Zhang and Schneider, 2011), which learns the coding design by maximizing the correlations between "codewords" and biological datasets.

Chapter 2

A piRNA regulation landscape in *C. elegans*

This chapter is based on a joint research project with Dr. Craig Mello's group in University of Massachusetts Medical School. Though this is a joint effort with shared thoughts, I primarily collaborated with Dr. Enzhi Shen, who mainly performed the experiments while I did the computational analysis (Shen et al., 2018).

As stated in chapter 1, the identification of piRNA targets is essential for deciphering the roles of piRNAs in both sequence-directed immunity and more broadly in the regulation of germline gene expression (Helwak et al., 2013; Van Nostrand et al., 2016; Vourekas and Mourelatos, 2014). Thus, it is of great interest to directly capture piRNAs and their targets *in vivo*. We optimized a crosslinking, ligation, and sequencing of hybrids (CLASH) protocol to identify piRNAs and its associated (candidate) target RNA binding sites in *C. elegans*. As a result, around 200,000 high-confidence piRNA–target site interactions are identified and the overwhelming majority of them were between piRNAs and mRNAs. The following bioinformatics analysis of the hybrids revealed that targets are enriched for energetically favorable Watson-Crick pairing with their associated piRNAs. Specifically, the seed sequence (i.e., positions 2 to 8) and supplemental nucleotides near the 3' end (positions 14 to 19) of the piRNA are important determinants of piRNA-target binding and silencing, suggesting that piRNA targeting resembles miRNA targeting.

We also find the piRNA target sites defined by CLASH show a non-random pattern

of secondary WAGO 22G-RNAs accumulation, which initiate at both ends and near the center (position 12) of the piRNA target site, consistent with local recruitment of RdRP. On the other side, analysis of CLASH hybrids obtained from CSR-1-depleted animals suggest that CSR-1 directly protects its targets from PRG-1 binding and WAGO-dependent silencing. These findings reveal that the entire germline mRNA transcriptome engages piRISC, and suggests how germline Argonaute pathways are coordinated to achieve comprehensive regulation and surveillance of germline gene expression.

2.1 PRG-1 CLASH experiment directly identifies piRNA-target chimeras

As mentioned previously, a modified CLASH approach was used to identify RNAs associated with the C. elegans PRG-1-piRISC complex. Briefly, CLASH involves the in vivo cross-linking of RNAs to a protein of interest followed by immunoprecipitation (IP), trimming of RNA ends, ligation to form hybrids between proximal RNAs within the crosslinked complex, cDNA preparation, library construction, and deep sequencing. In principle, this procedure should allow the recovery of hybrid-sequence reads formed when piRNAs are ligated to proximal cellular target RNAs within the cross-linked PRG-1 IP complex.

Given the strand specific short reads libraries from CLASH experiments, we firstly identified the chimeric reads with a full length piRNA and then trimmed the piRNA off the chimeric reads. Then, the remaining part of the short reads are mapped to the genome via BWT (Langmead et al., 2009) and classified into different groups using BEDtools (Quinlan and Hall, 2010) with Wormbase annotation (WS230) (Yook et al., 2012).

In two independent experiments, we found similar distributions of mapped se-

quence reads (Figure 2.1 and appendix figure 2), where more than 70% of the piRNA are targeting mRNA in both replicates. The unique piRNA sequences and the targeted mRNA transcripts are highly reproducible in terms of both sequence species and reads counts (Appendix figure 2).

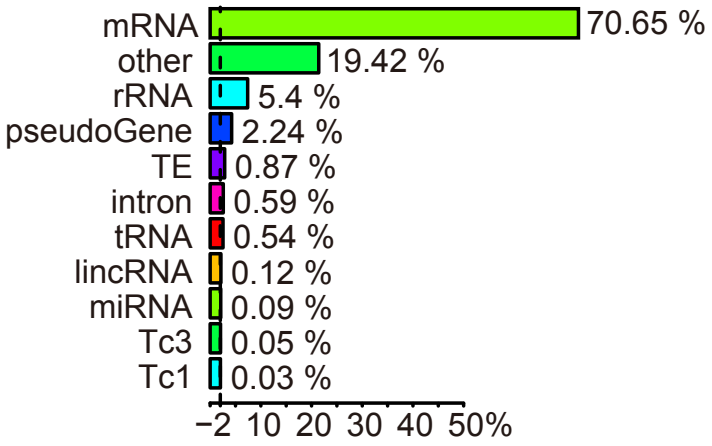


Figure 2·1: piRNA targets classification and abundance. The dashed line marks the 0% on the x-axis.

Together, these comprised a total of around 21 million reads, including a total of about 7 million reads corresponding to 17,192 different piRNAs. Most of these piRNA-containing reads lacked a hybrid sequence (1,083,172), or the hybrid sequences could not be mapped to the genome because they were too short, or for other reasons (3,946,162). We obtained 2,106,813 hybrid reads composed of a piRNA sequence and a genome-mapping sequence, of which around 1.5 million were composed of a single piRNA sequence fused to an mRNA. In addition to mRNA chimeras, we detected piRNAs fused to sequences corresponding to rRNA (137,322 reads), tRNA (11,231 reads), pseudogenes (48,208 reads), lincRNA (2,583 reads), miRNA (1,556 reads), introns (10,529 reads), and transposable elements (19,092 reads) (Figure 2.2).

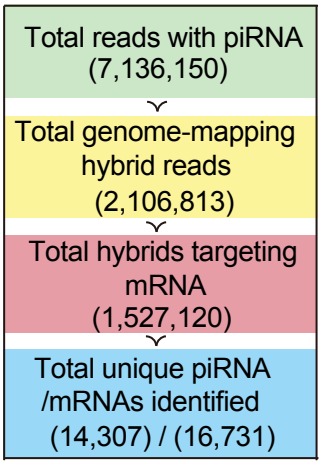


Figure 2·2: piRNA targets captured in two wild type experiments.

2.2 CLASH reveals piRNA target sites in germline mRNAs

Because mRNA chimeras were by far the most abundant type of hybrid read, we chose to focus on mRNA hybrids in the present study. Altogether, a total of 16,385 genes were represented among the piRNA hybrids (Figure 2.2). We found that "soma-specific" mRNAs were strongly under-represented in the CLASH data (Figure 2.3A) (Beanan and Strome, 1992; Li et al., 2014), consistent with the idea that CLASH captures interactions between piRNAs and mRNAs that occur in the germline, and not interactions that occur in lysates. On the other side, the frequency of recovering each piRNA by CLASH correlated with its level in the input sample as measured by small RNA sequencing (Figure 2.3B, $r = 0.58$, $P < 0.005$).

The nuclease treatment during the CLASH procedure was optimized to produce chimeras of approximately 40 nucleotides. Thus, each chimera potentially reflects a piRNA/target mRNA duplex ligated at, or near, one end of the duplex. We noted, however, that not all chimeras contained a full-length piRNA and that the recovered target regions varied in length, indicating some variability in nuclease trimming during

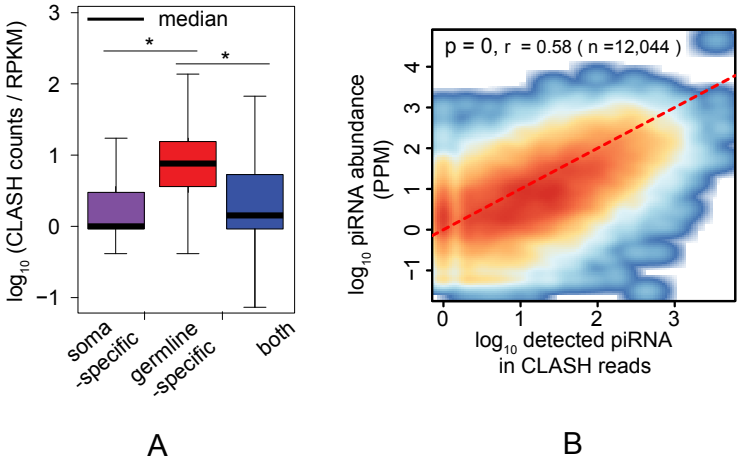


Figure 2·3: Density for piRNA and their targeting preference to soma and germline specific genes. (A) piRNA targetting density after normalizing the chimeric read counts to the targetted genes' expression level on "soma-specific," "germline-specific," and other genes, which are indicated in purple, red, and blue, respectively; (B) piRNA abundance in input small RNA sequencing correlates with CLASH captures piRNA abundance. * indicate t-test p-value smaller than $4.7 \cdot e^{-200}$.

the CLASH procedure. Therefore, prior to searching for base-pairing interactions, we implemented a customized computational protocol to extend the empirically defined target space by adding nucleotides to each end, using the initial partial piRNA-mRNA alignments (Figure 2.5B). In this way, we created an "ideal" piRNA/target RNA pairs.

We next predicted the most energetically favorable piRNA-mRNA interactions from *in silico* folding of these "ideal" sequences and compared it with predicted binding energies in a control data set with randomly matched pairs (Figure 2.4), using RNAfold in Vienna package (Lorenz et al., 2011). By comparing binding energy distributions of *in silico* random interactions with CLASH defined interactions at different abundance, this analysis showed that stable base-pair interactions were strongly enriched in the recovered piRNA-mRNAs chimeras. In fact, when normalized for mRNA

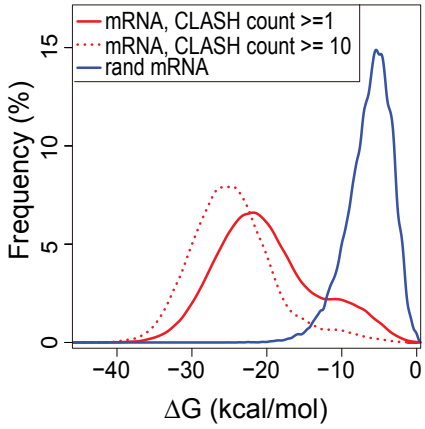


Figure 2·4: Distribution of the most favorable piRNA-mRNA interactions. Red solid line shows the binding energy distribution of CLASH defined interactions; Red dashed line shows the distribution of more stable interactions; The blue line shows the random control.

levels, we further found that hybrid read counts per target site correlated better with binding energy than with piRNA abundance (Figure 2.5A).

As a secondary validation, chimeras in which the piRNA 3' end was contiguous with mRNA sequence were found roughly 20-fold more frequent than chimeras ligated at piRNA 5' ends (Figure 2.5B). These findings are consistent with the idea that piRNA 3' ends are more available for ligation to their targets. Taken together, these findings support the idea that CLASH captures proximal mRNAs bound to piRISC via base-pairing interactions.

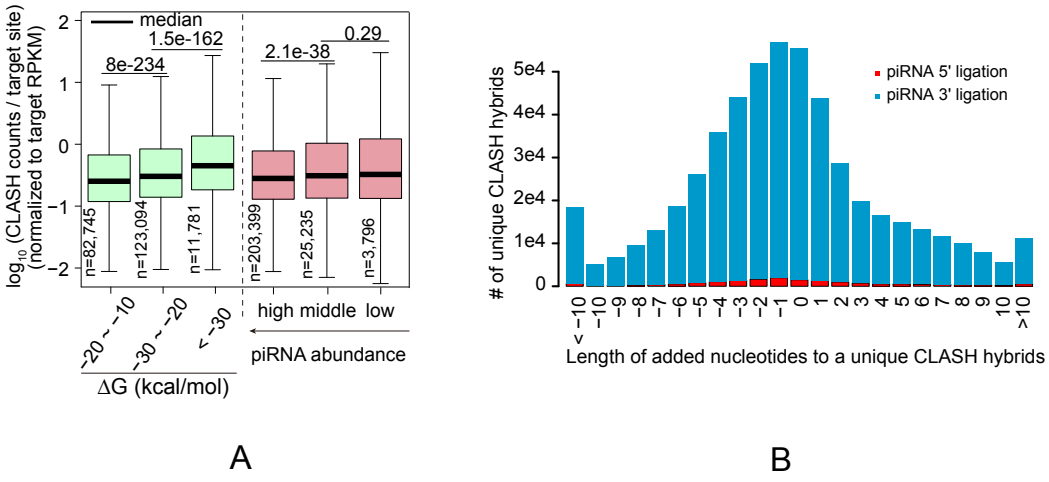


Figure 2·5: CLASH experiment validations by binding energy distribution and ligation event frequences. (A) Binding energy distribution of the chimeric reads at different read counts cut-off; (B) added length distribution of the piRNA 3' end to mRNA 5' end ligated target reads (in blue) vs others (red).

2.3 piRNA targets exhibit a pattern of discrete peaks in 22G-RNA levels

In *C. elegans*, piRISC recruits RdRP to its targets. Therefore, we wished to examine the pattern of RdRP-dependent 22G-RNA production near CLASH-defined piRNA target sites in both WT and prg-1 mutant worms, where prg-1 is the Argonaute protein that binds to piRNAs. To do this, we implemented a computational pipeline that normalized the 22G-RNA signal and aggregated their 5' end counts for a single base pair resolution. The aggregated 5' end counts are investigated within a 40-nt region centered on the piRNA targetting sites defined by CLASH. The 5' ends of 22G-RNAs are thought to be formed directly from RdRP initiating at C residues within the target mRNAs. We therefore normalized the 22G-RNA levels initiating at each position to the frequency of C residues within the CLASH-defined targets at each position.

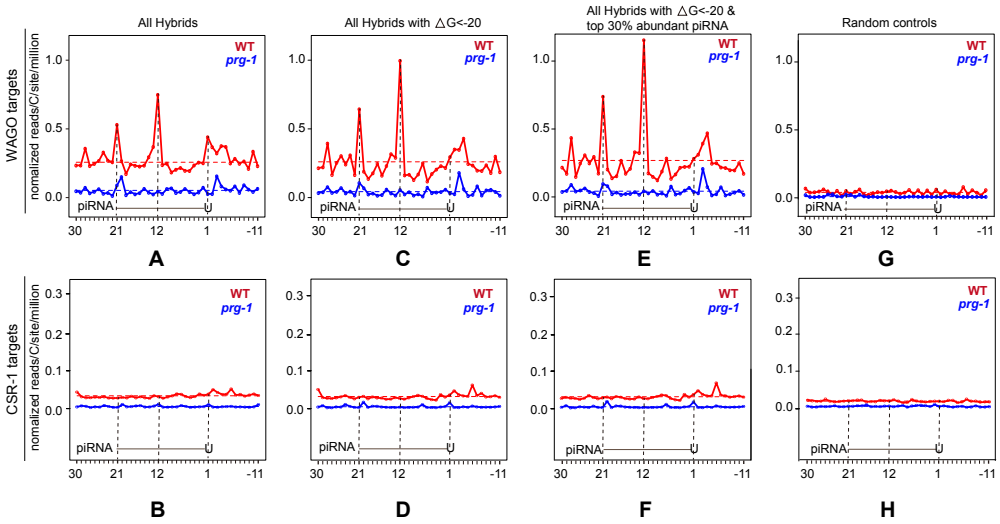


Figure 2·6: Normalized 22G-RNA signal over CLASH defined piRNA target sites. Solid lines show the normalized 5'G counts at and around the piRNA target sites while the dash lines show the average; Red and blue indicate wild type and prg-1 mutant background, respectively. subsets of target sites defined by binding energy and piRNA abundance cut-offs are shown in (C), (E), (D), and (F). (G) and (H) are generated using random target sites.

Because the CSR-1, and WAGO Argonaute pathway are thought to have opposing functions, resisting and supporting piRNA silencing (Seth et al., 2013; Wedeles et al., 2013), we separately considered predicted piRNA targets within previously defined WAGO and CSR-1 targeted mRNAs (Claycomb et al., 2009; Gu et al., 2009). As a control set, we considered a target region arbitrarily set more than 100 nts away (within each mRNA) from of the piRNA binding sites identified by CLASH. In WT animals, 22G-RNA levels were much higher for WAGO targets than for CSR-1 targets, as expected (Figures 2.6A and B). However, piRNA binding sites within both WAGO and CSR-1 targets showed a non-random distribution of 22G-RNA levels across the interval. By contrast, the control regions within the same target mRNAs, but offset from the hybrid sites, exhibited no such patterns (Figure 2.6G and H). WAGO targets exhibited a strong central peak, and clusters of peaks at either end of the piRNA target

sites. To describe these patterns, we refer to the mRNA sequences near the target site as follows: t1 through t30 includes the presumptive binding site (t1 to t21) plus 9 nucleotides 5' of the target site (t22 to t30). The mRNA region 3' of the target site consists of nucleotides t–1 through t–11. Strikingly, this analysis revealed a prominent peak in the center of the piRNA complementary region near t12, and smaller peaks centered at t1 and t21 (Figure 2.6A, C, and E). CSR-1 targets exhibited a cluster of much smaller peaks near the 3' end of the predicted target site, with the largest peak residing in sequences located near t–5 (Figure 2.6B, D, and F). The amplitudes of 22G-RNA levels on both the WAGO and CSR-1 targets correlated positively with the predicted free energy of piRNA binding and to a lesser extent with piRNA abundance (Figures 2.6A-F).

The amplitude and position of 22G-RNA peaks differed in prg-1 mutants. For WAGO targets, the central peak at t12 was completely depleted in prg-1 mutants, whereas the terminal peaks were reduced. In CSR-1 targets, the prominent peak located at t–5 disappeared, but new peaks at t1, t6, and t21 became evident (Figure 2.6). This analysis suggests that PRG-1 influences both the precise position, and the levels of 22G-RNAs on its targets, and that CSR-1 and WAGO targets differ strikingly in their accumulation of 22G-RNAs in response to piRNA targeting.

2.4 Patterns of piRNA targeting

Previous studies have revealed features of Argonaute/small RNA guided targeting, including the importance of "seed" pairing between the target and nucleotides 2 to 8 of the small RNA guide (Bartel, 2009). To explore patterns of piRNA-mediated targeting, we considered the *in silico* predicted piRNA-target folding within a high-confidence group of "ideal" pairs that were identified by at least 5 sequence reads. To identify preferred base-pairing patterns within this group of hybrids, we built up

a computational pipeline that convert the RNAfold output to a reduced BEAR RNA structural representation (Mattei et al., 2014) and then clustered them into clusters using Affinity Propagation (Frey and Dueck, 2007). This analysis revealed a clearly preferred interaction at the seed region and distinct base-pairing patterns at the 3' supplementary region (Figures 2.7A and B).

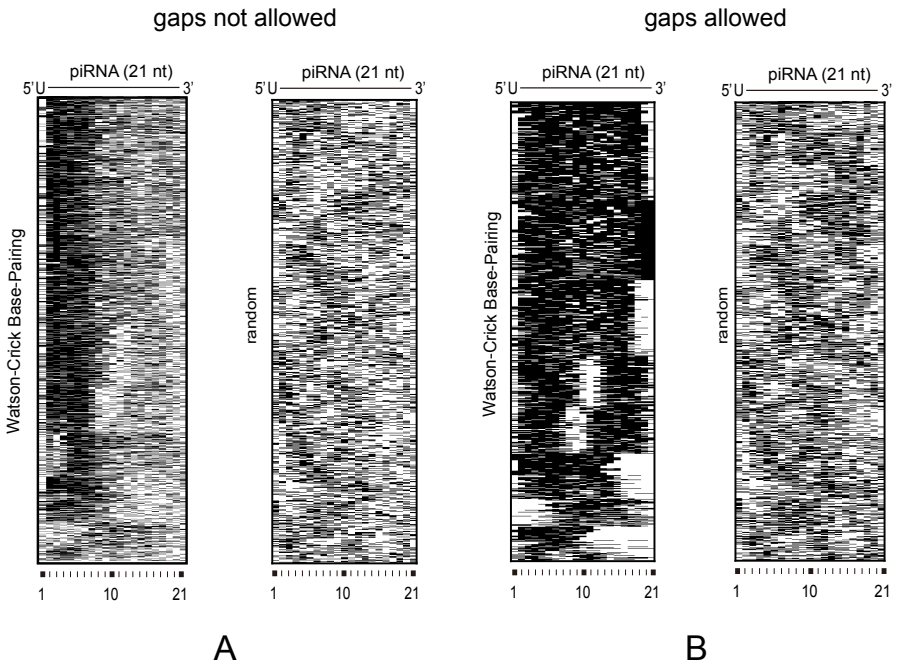


Figure 2·7: Heatmap for clustered matching (black) and mismatching (white) pattern for piRNA targeting. (A) allowing non-gapping "Watson-Crick" pairing only; (B) allowing alignment gaps; and corresponding random controls.

Notably, base-pairing frequencies declined from positions 9 to 13 of the piRNA and increased from positions 14 to 19 (Figure 2.8). As expected, these patterns were not enriched in a set of randomized piRNA target RNA pairs (Figure 2.7A-B, and Figure 2.8). These findings suggest that both seed pairing at positions 2 to 8 and supplementary pairing at positions 14 to 19 contribute to piRNA-target RNA binding (Shin et al., 2010; Wee et al., 2012).

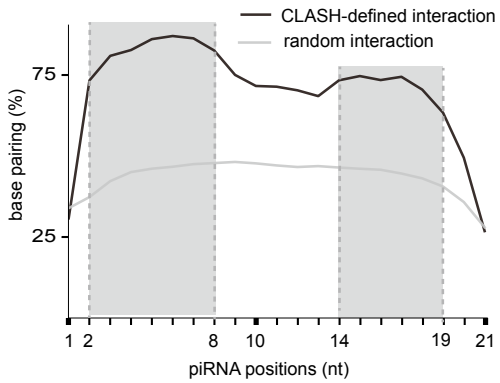


Figure 2·8: Aggregation plot for "Watson-Crick" pairing between piRNA and the targets. Solid line indicates the normalized pairing and the dashed line indicates the random control.

To further characterize piRNA–mRNA interactions, we analyzed A:U and G:C base-pair ratios at each position of the piRNA. We found no significant difference between the two base pair ratios within the seed region, but in other regions, we found a bias toward G:C pairing (Figure 2.9A). Notably, cytosine was strongly over-represented in the target strand immediately 3' of the seed complement opposite the 5' u, (defined as target strand position 1 cytosine, or "t1C") (Figure 2.9B).

This preference contrasts with t1A preferred by insect PIWI proteins (Wang et al., 2014) . To account for targeting sequences' potential composition bias, the "t1C" frequencies are also estimated after scrambling the target sequences via uShuffle (Jiang et al., 2008), while keeping the di-nucleotide and tri-nucleotide frequencies. In fact, an analysis of the C frequencies within the 40nt region centered by the CLASH target sites shows multiple cytosine over-represented positions (Figure 2.11D).

To further investigate the evolutionary conservation of the CLASH defined target sites, a 7way PhyloP score from multiple sequence alignments of 7 worm genomes are downloaded from UCSC genome browser (Kent et al., 2002). A following search covering 9400 genes with full PhyloP score coverage (Pollard et al., 2010) failed to reveal a preferential conservation for piRNA-mRNA target sites (Figure 2.10).

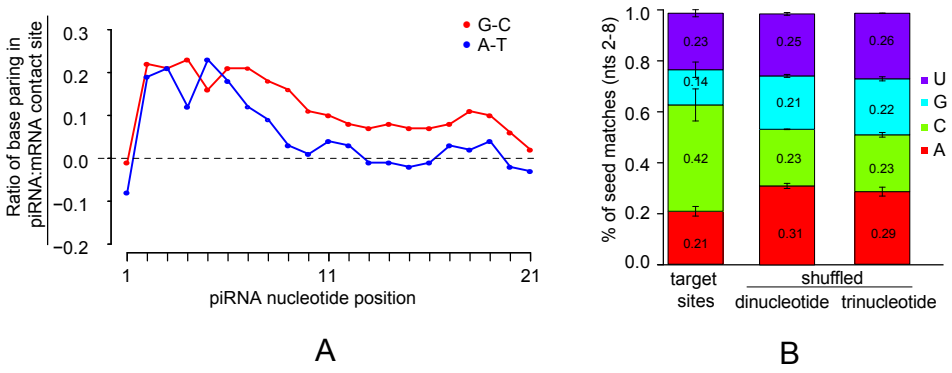


Figure 2·9: Nucleotides frequencies at the piRNA target sites. (A) aggregation plot for "G-C" pairing (red) and "A-T" pairing (blue) pattern for piRNA targeting pattern; (B) Nucleotide frequency at first target position, with scrambled seuqnces that keeping di-nucleotides and tri-nucleotides frenquencies as control.

Finally, we also separated the CLASH defined sites into WAGO targets and CSR-1 targets and compared the features of piRNA target interactions on them. The energetics of piRNA targeting, the patterns of seed and supplementary pairing, and the average C content along the target region were no different between these groups (Figure 2.11).

2.5 Seed and 3' supplementary pairing are required for target silencing

To determine the importance of base pairing along the length of the piRNA/target mRNA hybrid for piRNA silencing, we used CRISPR genome editing to systematically mutate positions 2 to 21 of an anti-gfp piRNA expressed from the 21ux-1 piRNA locus (See figure 3 in appendix) (Seth et al., 2018). We then assayed the ability of each 21ux-1(anti-gfp) mutant piRNA to silence a single-copy cdk-1::gfp transgene over a time course of up to 8 worm generations (See appendix figure 4). Strikingly, we found that individual mismatches in the seed region (i.e., m2 to m8) and 3' supplemental

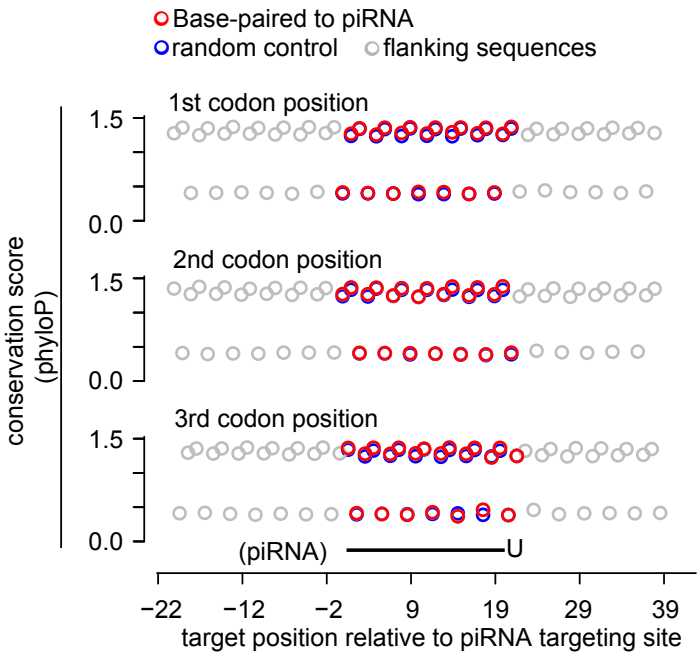


Figure 2·10: PhyloP score of the three codon position at the piRNA target sites. Red and gray indicate the average phyloP scores at and around the piRNA target sites; blue shows the scores at random sites as control.

region (i.e., m14 to m21) strongly reduced the ability of 21ux-1(anti-gfp) to silence cdk-1::gfp, but mismatches at the central region (m9 to m13) had a much more mild effect. By the F2 generation, when fully matched 21ux-1(anti-gfp) piRNA silences cdk-1::gfp by 70%, mismatches at positions 2 to 8 or 14 to 21 reduced silencing to less than 10% and 25% (respectively) of animals scored. By contrast, mismatches at positions 9 to 13 reduced silencing activity only slightly, to approximately 50% at the F2 generation. Mismatches at positions 2 or 3 prevented silencing of cdk-1::gfp, even after 8 generations, demonstrating that pairing at positions 2 and 3 is essential for piRNA-mediated silencing. Mutants with mismatches at any of the other 18 positions eventually silenced cdk-1::gfp over the 8 generation time course.

To further test the importance of pairing in these regions, we selectively mutated

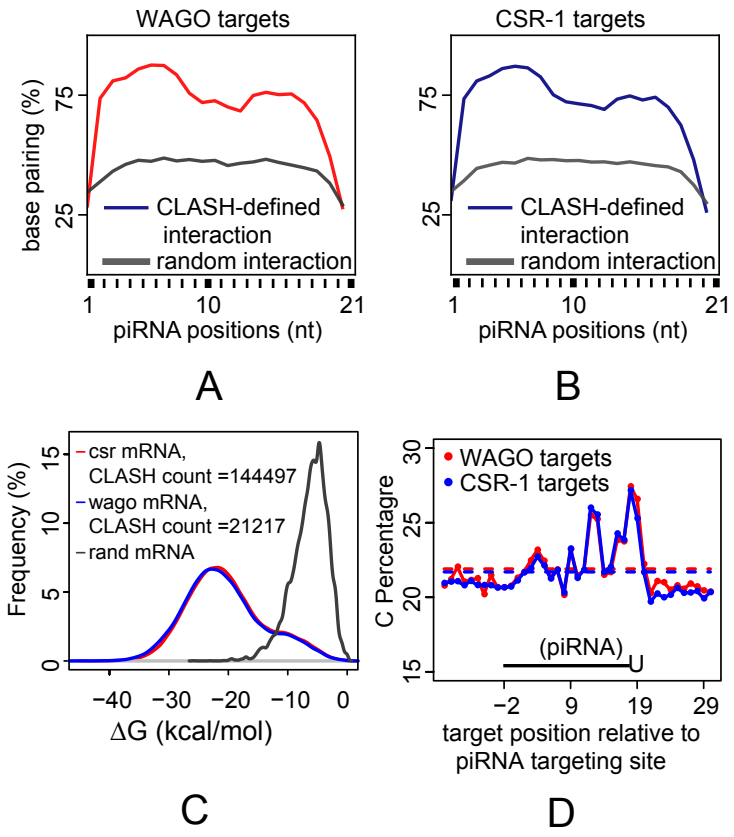


Figure 2·11: Comparison of WAGO and CSR-1 targets. (A) and (B) show the base-pairing rates at piRNA target sites and random sites in WAGO and CSR-1 target genes; (C) shows the deltaG distribution for all target sites on WAGO and CSR-1 target genes; (D) shows the distribution of C counts at and around piRNA target sites.

positions t3, t15, and t21 of the anti-gfp target site in cdk-1::gfp mRNA to compensate for anti-gfp piRNA mutations in guide-strand positions, g3, g15, and g21, each of which strongly diminished silencing. As expected, in the absence of 21ux-1(anti-gfp), these silent mutations did not affect the level of GFP expression. Strikingly, target mRNAs with "re-matching" mutations at t3, t15, and t21 were each rapidly silenced by piRNA strains bearing the corresponding guide mutations. Thus the failure of the g3, g15 and g21 point mutant piRNAs to silence wild-type cdk-1::gfp was caused specifically by the mismatches and not by changes in expression or piRISC loading

of the mutant piRNAs.

Lastly, we analyzed 22G-RNA induction for several 21ux-1(anti-gfp) point mutant strains. As expected, we found that 22G-RNA levels correlated with the degree of GFP silencing observed (Figure 2.12). Overall, these findings confirm the impor-

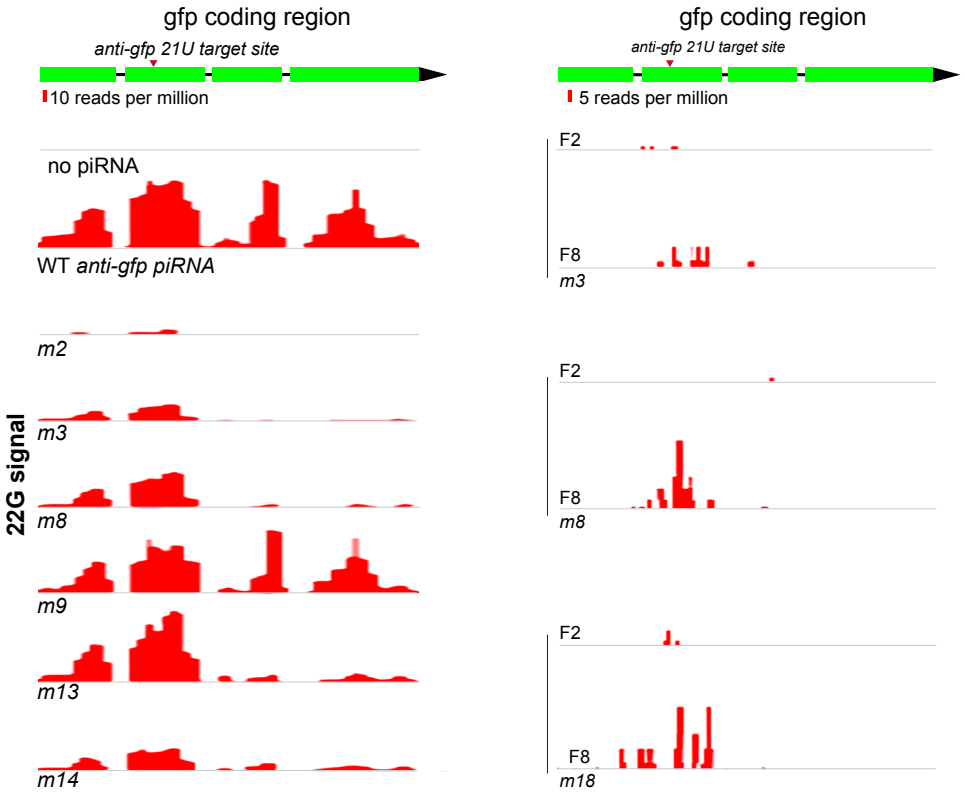


Figure 2·12: Schematic of 22G-RNAs targeting gfp in F2, F4 and F8 worms. Accumulated 22G-RNAs signal (red) targeting gfp gene (green) in F2, F4 and F8 cdk-1::gfp worms with the indicated single-nucleotide mismatches (m2 = position 2 mismatch, etc.). Positions from 5', central, and 3' regions of the piRNA were randomly chosen for analysis.

tance of base-pairing within the seed region (nucleotides 2 to 8) and within the 3' supplemental pairing region (nucleotides 14 to 21) for efficient piRNA targeting.

2.6 Experimental validation of specific piRNA–mRNA interactions suppressing endogenous mRNA targets

To investigate how the base-pairing rules defined by our bioinformatics and transgene studies affect targeting of an endogenous mRNA, we edited the 21ux-1 target site, introducing single mismatches into the predicted 21ux-1/xol-1 target duplex (Fig-

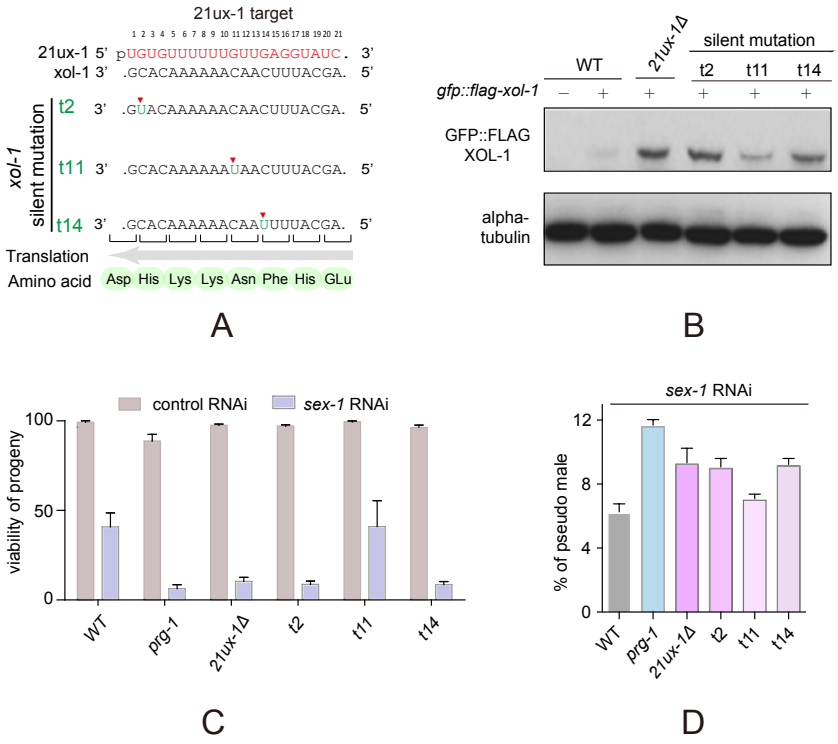


Figure 2·13: Validating 21ur-4863 and 21ux-1's suppression by mutating their target sites on xol-1. (A) indicate the mutations on xol-1; (B), (C), and (D) validate the xol-1 expression, viability of progeny and percentage of pseudo males, respectively.

ure 2.13A). XOL-1 is a key regulator of dosage compensation and sex determination in early zygotes, and xol-1 mRNA was recently shown to be regulated by the X chromosome-derived piRNA, 21ux-1(Tang et al., 2018). Consistent with our findings in the transgene studies, single-nucleotide mismatches within the seed and 3' supplemental pairing regions, but not within the central region, dramatically increased

expression of XOL-1 (Figure 2.13B). The 21ux-1 mutants with mismatches in the seed and 3' supplemental pairing regions were phenotypically similar to 21ux-1 null mutants and enhanced the dosage compensation and sex determination phenotypes (decreased brood size and masculinization of hermaphrodites) caused by silencing the X-signal element sex-1 (Figure 2.13C and D) (Carmi et al., 1998). Thus, mutating a single nucleotide in 21ux-1 dramatically increases both XOL-1 expression and activity.

Consistent with the observation that most germline mRNAs are targeted by multiple piRNAs, we identified a total of 166 CLASH hybrids containing xol-1 mRNA sequences, which are fused to 40 different piRNAs (Figure 2.14). However, given

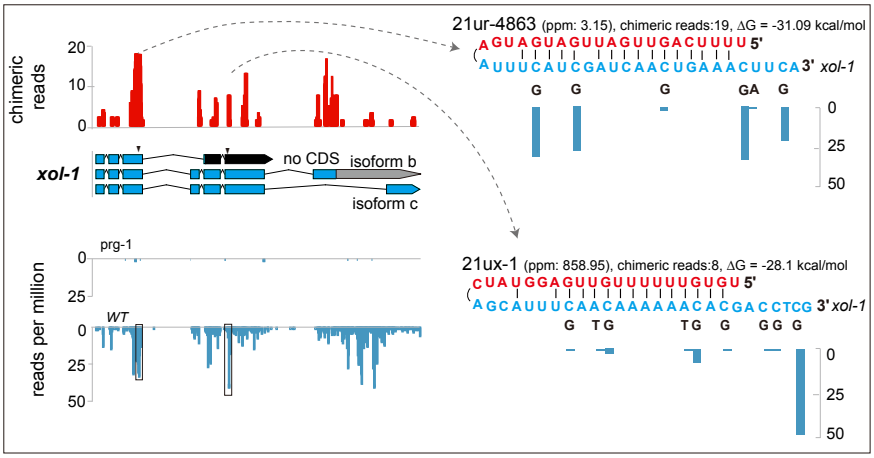


Figure 2·14: Distribution of chimeric xol-1 reads (red) identified by CLASH, and the distribution of xol-1 22G-RNAs (blue) in prg-1 mutant and WT worms. Locations of 21ur-4863 (upper) and 21ux-1 target sites in xol-1 gene indicated by inverted black triangles. Sequences and base pairing (right) of 21ur-4863:xol-1 (upper) and 21ux-1:xol-1 (lower) chimeras. piRNA expression level, number of chimeric reads, and binding energy (kcal/mol) indicated above each chimera. Distribution of 22G-RNAs at single-nucleotide resolution shown below each chimera.

the importance of 21ux-1 in regulating xol-1, and the fact that 21ux-1 is the most abundant piRNA, we were surprised to find that a different piRNA, 21ur-4863, was recovered in xol-1 chimeras at a frequency greater than twice that of 21ux-1 chimeras.

Specifically, we identified 8 reads with 21ux-1 fused to its xol-1 target site and 19 reads of 21ur-4863 fused to its xol-1 target site. This is consistent with the general observation that chimeric read abundance correlates better with binding energy, rather than the piRNA abundance (Figure 2.5A).

We therefore wished to ask if 21ur-4863 is also important for xol-1 regulation. Strikingly, deletion of 21ur-4863 resulted in the upregulation of both xol-1 mRNA and protein levels to a degree similar to that observed in 21ux-1 mutants (Figure

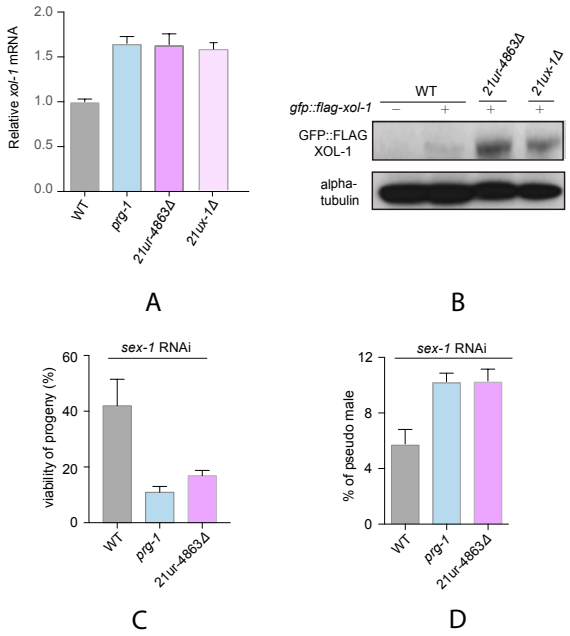


Figure 2·15: Validating 21ur-4863 and 21ux-1's suppression to xol-1 by mutating the two piRNAs. (A) Bar graph of xol-1 mRNA levels in WT, prg-1, 21ur-4863 deletion, and 21ux-1 deletion worms measured by RT-qPCR. actin mRNA served as the internal control. Data expressed as mean ± s.d. of three experiments; (B) Western blot (anti-FLAG) of GFP::FLAG::XOL-1 (top) levels in WT, 21ur-4863 deletion, and 21ux-1 deletion worms. Alpha-tubulin (bottom) was probed as a loading control; (C) and (D) are bar graphs of percent viable and pseudomale progeny of WT, prg-1, and 21ur-4863 deletion worms treated with sex-1(RNAi). n > 500 per experimental group. Data expressed as mean ± 2 s.e.m. of three experiments.

2.15A and B). Similar to the 21ux-1 mutant, the 21ur-4863 deletion mutant enhanced

defects in dosage compensation and sex determination caused by sex-1(RNAi): fewer progeny and dumpy (Dpy) body morphology (Figure 2.15C and D)(Carmi et al., 1998) . Thus, 21u-4863 and 21ux-1 are both required for xol-1 silencing, neither is sufficient, suggesting that piRNAs function cooperatively to silence xol-1.

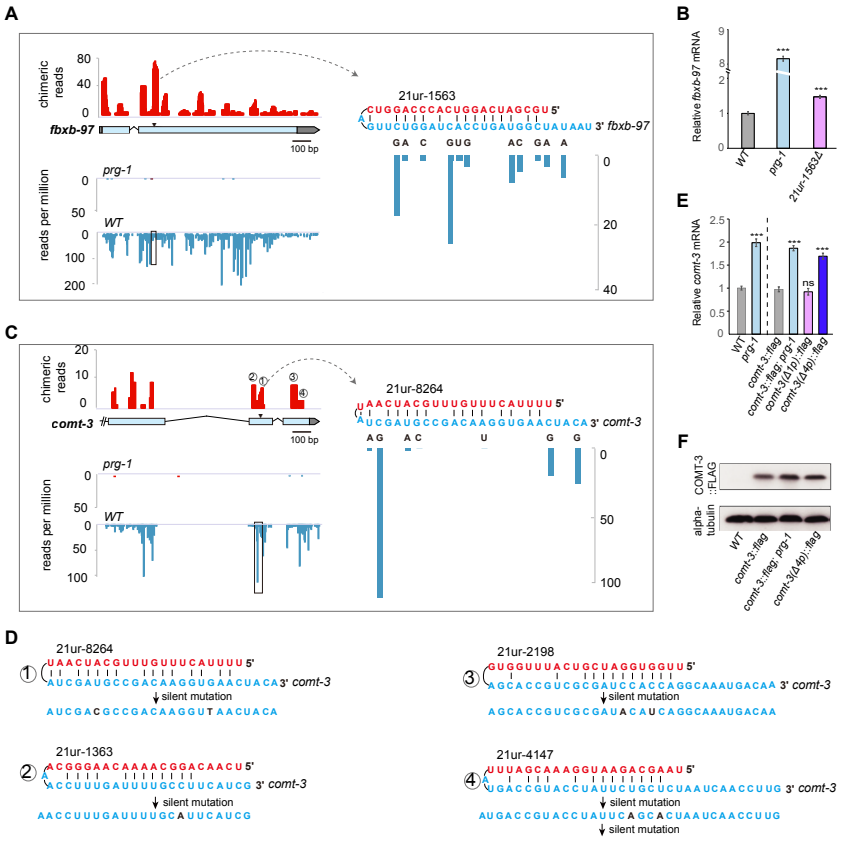


Figure 2·16: Analysis of piRNAs targeting fbxb-97 and comt-3. (A) and (C) are distributions of chimeric reads (red) identified by CLASH, and the distribution of 22G-RNAs (blue) in prg-1 mutant and WT worms, on gene fbxb-97 and comt-3, respectively. (D) indicate the mutations that disrupt 4 piRNA target sites; (B), (E), and (F) show the two genes' expression changes in the mutated worms.

We also examined the consequences of piRNA targeting on fbxb-97 and comt-3,

whose mRNAs are also regulated by PRG-1 (Bagijn et al., 2012; Batista et al., 2008; Gu et al., 2009; Lee et al., 2012). We identified 70 chimeric reads between 21ur-1563 and fbxb-97 (Figure 2.16A) and found fbxb-97 mRNA levels were upregulated 1.5-fold in a 21ur-1563 deletion mutant and ˜8-fold in the prg-1 mutant (Figure 2.16B). To analyze piRNA regulation of comt-3 (Figure 2.16C), we took the alternative approach of mutating target sequences. We introduced silent mutations into wobble-positions that maintain the comt-3 open reading frame but disrupt 4 piRNA target sites (Figure 2.16D). comt-3 mRNA levels were markedly increased in the prg-1 mutant and in the comt-3 quadruple-piRNA target site mutant, but were not elevated in a comt-3 single-piRNA target site mutant (Figure 2.16E). COMT-3::FLAG (introduced by CRISPR) was significantly elevated (by 1.5 fold) in the quadruple target site mutant (Figure 2.16F). Taken together, our findings suggest that individual piRNAs exhibit a range of regulatory effects and that multiple piRNAs cooperatively silence individual targets.

2.7 Competition between the CSR-1 and PRG-1 Argonaute pathways

Previous studies suggested that CSR-1 protects its germline mRNA targets from piRNA-mediated silencing (Seth et al., 2013; Shirayama et al., 2012; Wedeles et al., 2013). We sought to test whether CSR-1 protects its targets by preventing PRG-1 from binding. To do this, we used an auxin-inducible degradation (AID) system to conditionally deplete CSR-1 in young adult worms (Zhang et al., 2015), and then performed CLASH on CSR-1 depleted worms in two independent biological replicates. We compared the number of unique piRNA binding sites on CSR-1 targets from CSR-1 depleted and wild-type worms. Strikingly, we found that the number of unique piRNA binding sites significantly increased (about 2 fold) in the CSR-1 depleted worms compared to wild-type (Figure 2.17A,B). This increase did not result from

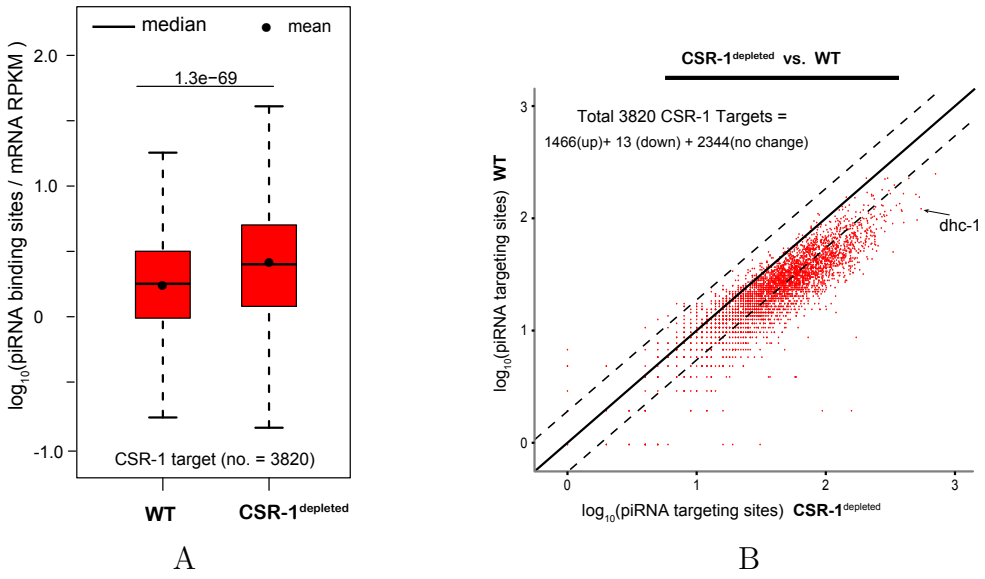


Figure 2·17: Competetion between CSR-1 and piRNA pathways. The numbers of both raw and normalized piRNA binding sites increase on CSR-1 targeted genes at a CSR-1 depleted background.

changes in target mRNA levels, which did not change dramatically during CSR-1 depletion (See figure 5 in appendix). Increased piRNA targeting is illustrated for dhc-1, whose mRNA levels did not appreciably change (˜1.5 fold), but whose piRNA targeting was elevated by >3.4-fold in CSR-1depleted worms (Figure 2.18). These results suggest that, when CSR-1 is depleted, mRNAs normally targeted by CSR-1 become bound by additional piRNAs.

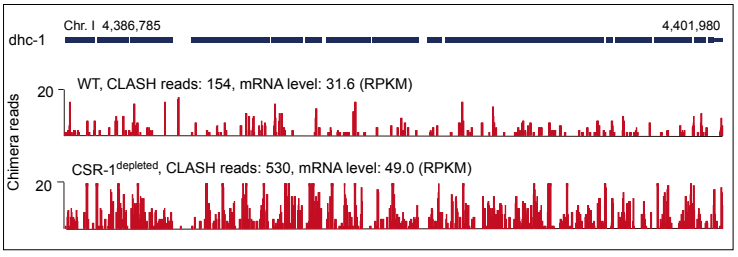
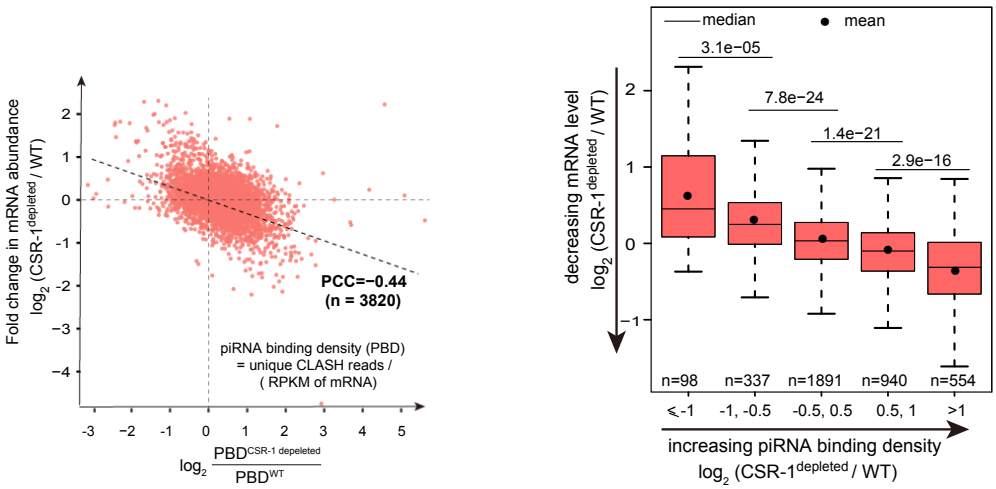


Figure 2·18: Change in piRNA target density in CSR-1 depleted worms versus wild type for gene dhc-1.

To determine whether increased piRNA binding correlates with decreased mRNA levels, we plotted the fold change in mRNA abundance (CSR-1depleted / WT) versus the fold change in piRNA-binding density (CSR-1depleted / WT) for 3,820 CSR-1 targets (Figure 2.19A) (Claycomb et al., 2009). We observed a negative correlation between increased piRNA-binding density and mRNA abundance in CSR-1depleted worms (r = - 0.44). To clearly visualize this relationship, we split the 3,820 CSR-1 targets into five bins of increasing piRNA binding density and plotted the change in mRNA abundance in CSR-1depleted versus wild type (Figures 2.19B). This analysis revealed that, as piRNA binding density increases, mRNA abundance decreases. These findings support the idea that CSR-1 functions, at least in part, upstream of PRG-1 to reduce piRNA targeting.



A B

Figure 2·19: Change in mRNA abundance between WT and CSR-1 depleted worms versus the change of piRNA binding. Scatterplot in (A) shows the change in mRNA abundance between WT and CSR-1 depleted worms versus the change of piRNA binding density for 3,820 CSR-1 targets. Boxplot in (B) shows the same trend with stratified piRNA binding density changes.

2.8 Non-mRNA piRNA interactions

Although mRNA target sites accounted for greater than 90% of CLASH hybrid reads, we also reproducibly identified CLASH reads mapping to a variety of non-coding RNA species (ncRNAs). For example, over 80,000 CLASH reads and hundreds of different

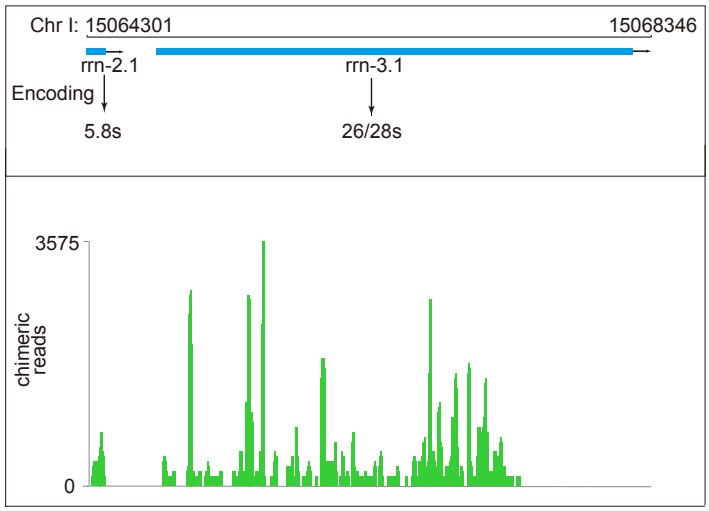


Figure 2·20: Distribution of chimeric rrn-2.1 and rrn-3.1 reads (green) identified by CLASH and genomic locus on the top (blue).

piRNAs were mapped to ncRNA hybrids, including sequences from a single region of the 26S rRNA (Figure 2.20). Interestingly, this rRNA region is also targeted by WAGO 22G-RNAs that were recently reported to downregulate rRNA levels in response to stress (Zhou et al., 2017). Our studies also identified many interactions between piRNAs and tRNA species. For example, tRNAGlu(CUC) and 21U-8377 formed highly reproducible chimeras that showed thermodynamically stable base-pairing (Figure 2.21). Altogether we identified piRNA-tRNA hybrids involving 474 different tRNAs and 1225 different piRNAs. The significance of these findings remains to be determined, but it is intriguing that in *Drosophila*, a mutation that leads to accumulation of misprocessed tRNA results in a collapse of Piwi-mediated transposon

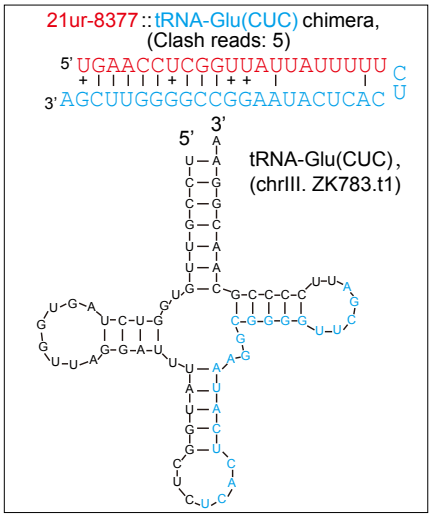


Figure 2·21: Putative interactions between piRNA and tRNA. 21ur-8377 reproducibly bind to the same region of tRNA-Glu(CUC), marked blue on the tRNA structure (chrIII. ZK783.t1).

silencing (Molla-Herman et al., 2015; Yamanaka and Siomi, 2015). We also identified hybrids between piRNAs and other ncRNAs including microRNAs and annotated long-noncoding, lncRNAs. The identification of these piRNA interactions provides a new lens through which to explore potential functions and regulation of germline ncRNA species.

Chapter 3

A computational model to predict gene functions

Given a set of co-functioning genes, can we predict their additional functional members? In this chapter, I introduce a computational model for answering this question. As stated in chapter 1, the rationales for this model are "guilt by association" principle and a multi-label learning framework called "Error correction of code" (ECOC), which is originally from the telecommunication community. By modeling multiple labels and their mutual dependencies, this method is more generalized for predicting labels for multi-function genes. Besides the error correction feature, this "ECOC" framework also in principle works better for gene function predictions that are highly imbalanced, as more genes with positive labels are involved in training for tuning two hyperparameters. To benchmark this method's performance, I did a comparison with a state-of-the-art algorithm "Mashup", using their benchmark dataset on yeast networks. Also, as a case study, it is used in predicting piRNA pathway genes in *D. melanogaster*, with a set of single-cell RNAseq pruned stringDB networks.

3.1 The multi-label learning model

As mentioned in chapter 1, this model uses the error-correcting codewords to represent the multi-function genes. With annotated codewords for each gene, it directly learns label dependencies from biological data, using canonical correlation analysis (CCA) that maximizes projections between these label codewords and the network

propagated label values. These codewords and the CCA projections form a coding structure that can be used to encode the gene function labels to biological datasets and decode the biological data back to labels. Thus, this model is capable of predicting function labels for genes from the same input biological distribution. Figure 3.1 provides an overview of this model.

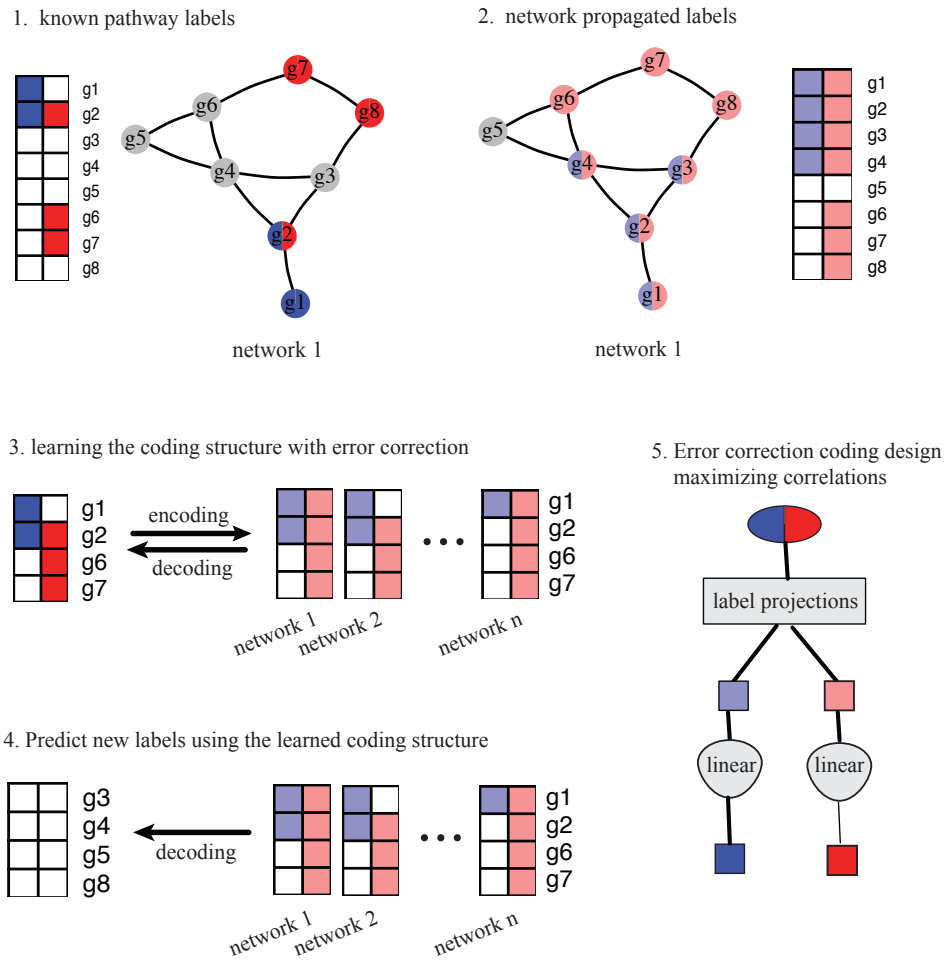


Figure 3·1: The multi-label learning framework.

3.1.1 Multi-label propagation

Firstly, provided with multi-functional genes and edge weight normalized networks, this model implemented PRINCE (Vanunu et al., 2010) algorithm to propagate the known genes' labels to their topologically close genes in a network. Given q known labels for p genes, the label space is expressed as a $p \cdot q$ matrix Y. In figure 3.2, this

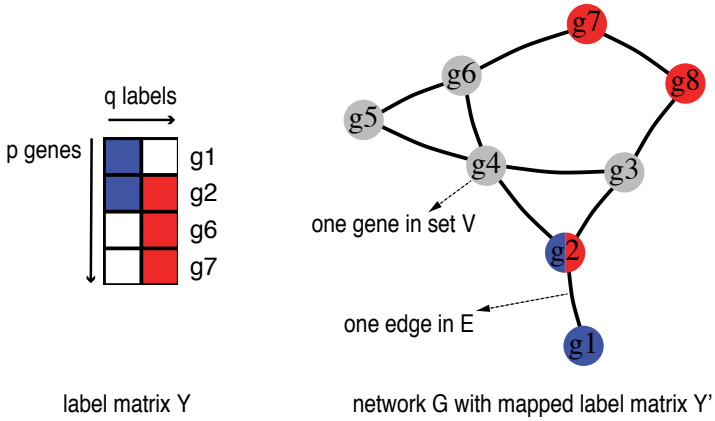


Figure 3·2: Multi-label genes mapped to a network.

known label set is denoted in red and blue. And the label matrix Y can be written as the following.

$$Y = \{y_{1,},\ y_2,\ \ldots\ ,y_j,\ \ldots\ ,\ y_q\}$$

Where each y_j is a p element column vector filled with 1s and 0s, denoting whether a gene is labeled or not. For each of the q label vectors, we can map them to a gene-gene interaction network and get a newly mapped label vector y'_j, which are the column vectors in the mapped label matrix Y'.

$$y'_{ij} = y_{ij}{=}1,\ \text{ if gene in row } i \text{ is found in network } G$$

This network G is consists of genes as nodes and weighted edges representing the genes' associations.

$$G = \{V, E, W'\}$$

Here, V denotes the genes in the network, E denotes the edges connecting the genes. and W' is the normalized weights of the edges. This normalization is done by solving the following equation (Vanunu et al., 2010), where D is the diagonal matrix and W is the raw edge weights.

$$W' = D^{-1/2} W D^{-1/2}$$

The next step is to propagate the mapped label matrix Y' on the edge weight normalized networks. Initially, each labeled gene in the network is assigned with prior value 1. Then, these label values are propagated from labeled genes (such as u) to their neighbor genes (such as v). These iterations are tuned by a parameter a. As stated in chapter 1, this parameter balances the influences between the infused label values from the initial gene and the other label values propagated on the network. Specifically in this model's RWR process, $(1 - a) \cdot y'$ additional label values are injected into the network at each iteration.

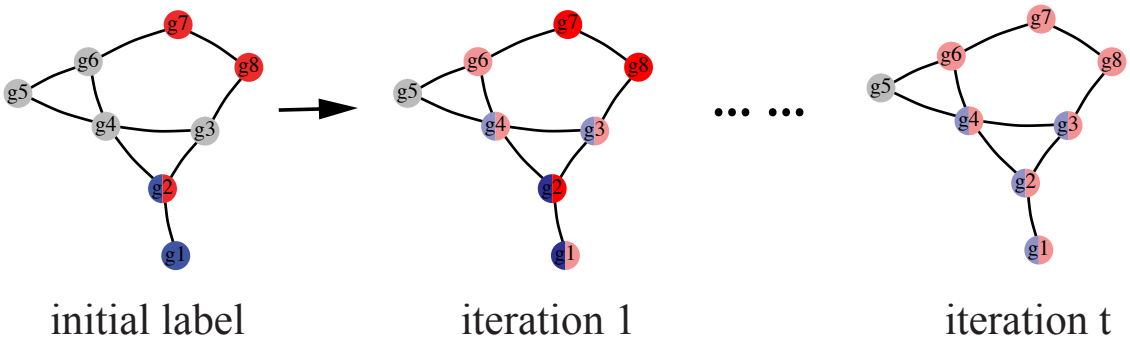


Figure 3·3: Multi-label propagation on a network.

The following equation summarized this iteration process for updating a gene v's label value from its neighbors.

$$F(v) = a\left[\sum_{u \in Neighbor(v)} F(u)\, w'(v,u)\right] + (1-a)\, y'$$

Here, y' is the mapped labels on the network and $w'(v,u)$ denotes the normalized edge weight. The algorithm continues this iterative process until reaching a maximum number of iterations or the termination criteria, which checks whether the label values are distributed at a stationary state on the network. To simplify, the above equation can also be rewritten in a matrix form.

$$F^t = aW'F^{t-1} + (1-a)\, y'$$

Here, W' is the weight matrix in G and t denotes the number of iterations. After termination, the final propagated values of a gene function label can be denoted as a data vector x for each gene in the network. In figure 3.3, this corresponds to the light red and blue label values at iteration t.

$$x = F^t$$

3.1.2 Integrating the propagated label values

Then, aiming to model a set of labels on multiple networks, each of the column label vectors in Y are mapped and propagated on n networks, generating a data vector x. These vectors are summarized as data matrix X.

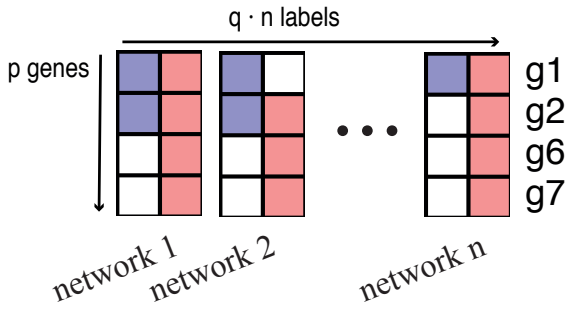


Figure 3·4: Data matrix summarized from multiple label propagations.

Probably as a result of different numbers of iterations and different numbers of the initial known labels, the propagated data vectors are in different scales. Thus, to facilitate the training process, they are normalized back to a $\{0, m\}$ scale, where maximum value is defined as $m = \sum y_i/p$, which is the proportion of annotated genes of label i on a network. Here, p is the number of genes in that network. Also, if a labeled gene is missing in a network, it is filled with this estimated maximum value m in the data matrix X. And finally, following the error-correcting design, propagation values from different networks are stacked together as the final data matrix X.

$$X = (x_1, x_2, ..., x_{q.n})$$

3.1.3 Codeword design and learning process

Following Zhang *et al.* (Zhang and Schneider, 2011), this model learns a coding structure by maximizing projections between the codewords Y and the data matrix X. As shown below, a matrix Z representing this structure is designed as the label matrix Y with additional label dependency terms (Figure 3.5).

$$\begin{aligned} Z = &\ (Y \qquad\qquad\qquad\quad , V_1^T Y, V_1^T Y, ..., V_k^T Y) \\ = &\ (y_1\ , y_2, \ ...\ , \ y_q \quad , V_1^T Y, V_1^T Y, ..., V_k^T Y) \end{aligned}$$

Here the $V_k^T Y$ denote the label combination which is most predictable using feature matrix X, so that the projection vector V mainly capture the conditional label dependency given X. This vector is estimated using Canonical Correlation Analysis (CCA), which finds the two projection matrices V and U that maximize the correlation between label matrix Y and data matrix X.

To derive this VY term, the CCA object function can be rewritten as the followings, which minimizes the sum of squared errors with constraints.

$$argmin \|XU - YV\|^2$$

$$st.\ U^T X^T XU = 1$$

$$st.\ V^T Y^T YV = 1$$

We can see that VY can be expressed as the variable we'd like to predict using projection matrix U and data matrix X.

$$VY = UX$$

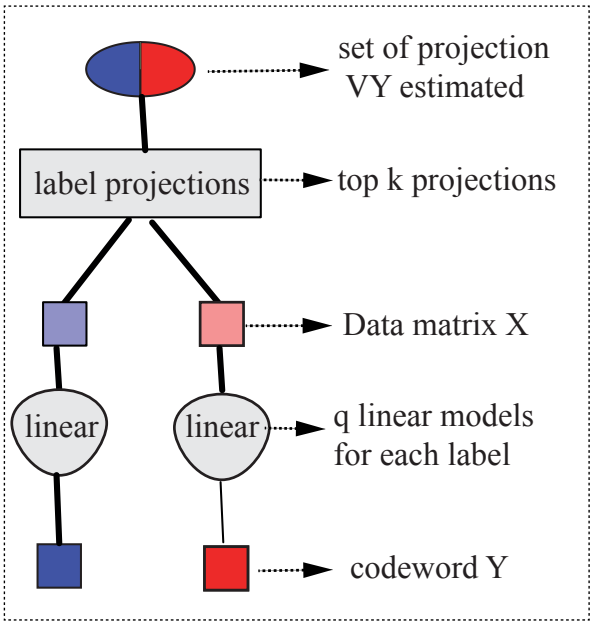


Figure 3·5: Code Matrix designed as codeword and additional dependency terms.

As shown in figure 3.5, each of the q labels can be encoded as data matrix X using a linear transfermation, which is learned as a linear model.

$$p_j \leftarrow learn\_classifier\left(x_i, z_j\right), j \in 1, ..., q$$

Each of the k additional VY variable can be encoded as well, using a Gaussian regression model.

$$m_j \leftarrow learn\_regression\left(X, z_j\right), j \in q+1, ..., q+k$$

3.1.4 Decoding process for gene function prediction

The final step is to decode a data matrix X to functional label matrix Y, using the learned q linear models and k label dependency terms. If this learned structure decodes a data matrix of functionally unknown genes, it will generate a multi-label matrix as gene function predictions. Specifically, this model calculates the probability of assigning each of the labels to a gene, using Bernoulli distributions for the probability estimated from the linear models and Gaussian potentials for the estimations from the label dependency terms.

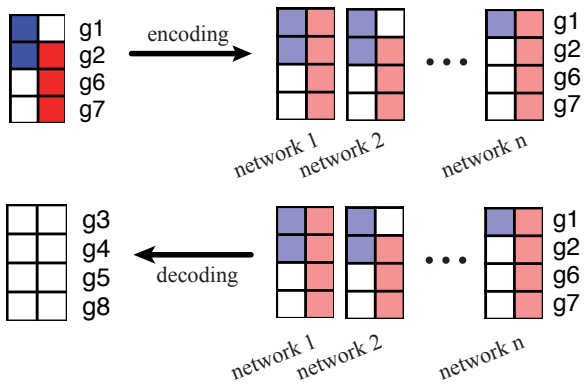


Figure 3·6: Predicting gene labels from data matrix.

This joint probability $P(y)$ of labeling genes can be summarized using the following equation.

$$logP(y) = -logZ + \sum_{k=1}^{d} \log \psi_k(y) + \lambda \sum_{j=1}^{q} \log \phi_j(y_i)$$

Here, Z is the partition function and λ balances the two types of potentials. As mentioned, the second term $\phi_j(y_i)$ is a Bernoulli distribution for a label y_i .

$$\phi(y_i) = p_j(x)^{y_j} (1 - p_j(x))^{1-y_j}, j = 1, 2, ..., q$$

And the third term $\psi_k(y)$ is a Gaussian model

$$\psi_k(y) \sim exp^{-\left(V_k^T y - m_k(x)\right)^2 / 2\sigma_k^2}, \; , k = 1, 2, ..., k$$

As each Gaussian potential $\psi_k(y)$ usually involves all q labels, I followed Zhang *et al.* (Zhang and Schneider, 2011) to use a mean-field approximation of $P(y)$ as the following.

$$Q(y) = \prod_{j=1}^{q} Q_j(y_j)$$

It is fully factorized and each $Q_j(y_j)$ in $Q(y)$ is a Bernoulli distribution on the label y_j. The best approximation $Q(y)$ can be obtained by minimizing the KL divergence between $Q(y)$ and $P(y)$.

$$KL(Q(y)|P(y))$$

Thus the fixed-point equation for updating each $Q_j(y_j)$ in $Q(y)$ can be written as the following.

$$Q_j(y_j) \leftarrow 1/Z_j + exp\{\lambda log\phi_j(y_i) + \sum_{n=1}^{k} E_{y\sim Q}[log\psi_n(y)|y_i]\}$$

3.2 Model specialized for genes function predictions

In addition to the multiple-label learning process introduced above, several algorithms are also integrated into this model for this gene function prediction problem, which is unique for several reasons.

Firstly, many gene function labels are rare in a species, making gene stratification a non-trivial task. For example, a simple random stratification of genes can generate subsets with no positive instances for some functional labels. It leaves some of the measures undefined for these labels, such as the area under a precision-recall curve. Also, to further facilitate potential biological experiments, it is essential to calibrate the calculated probabilities and to define an optimal threshold for the predicted confidence scores, in addition to a ranked gene list. They are described in detail in the following sections.

3.2.1 Second order iterative stratification

A second-order iterative stratification (SOIS) procedure (Szymański and Kajdanowicz, 2017; Sechidis et al., 2011) is implemented in this model for stratifying the training and testing datasets. In addition to the measure definition problems mentioned, additional cautions are also necessary for multi-label learning. Specifically, the data subsetting procedure should also preserve the label relationships, especially for these imbalanced gene labels. This SOIS procedure iteratively distributes the gene instances using criteria that estimate the most demanding multi-label genes for each subset at each iteration and evenly distribute them. This stratification procedure also distributes genes with fewer labels by minimizing label dependency differences between the subsets and the whole dataset.

As mentioned, two hyperparameters exist in this model, which are the number of label dependency terms k and λ that balances the estimated probabilities from the linear model and the label dependency terms. They are also estimated via a nested SOIS cross-validation procedure in this model.

3.2.2 Robust probability calibration using Platt's scaling

Transferring the output scores of a classifier to reliable probabilities is also not a trivial task, as machine learning methods tend to produce skewed probability distributions (Platt, 1999). In the case of multi-label learning, the differences in probability distributions can bias the metrics for benchmarking the algorithm performances and hyperparameter tuning.

To account for this problem, I implemented a modified version of Platt's scaling for this model, according to the pseudocode from Hsuan-Tien Lin *et al* (Lin et al., 2007). Platt's scaling is a simple probability calibration method that uses logistic regression to calibrate the probabilities, using known positive and negative labels in

training (Platt, 1999). Also, as stated in chapter 1, missing positive labels in the training dataset may confuse the calibrating process. To ease this potential problem, a further modification that removes a fraction of top scores before Platt's scaling is used, which alleviates the effect of high probabilities assigned to false-negative genes in calibration (Rüping, 2006).

3.2.3 Finding the optimal threshold using F-score

In addition to the probability calibration, my model also finds an optimal threshold for separating positive and negative predictions, which is the probability that maximizes F-score along a precision and recall curve. It firstly records the cut-offs that maximizes F-score in the training dataset (Fan and Lin, 2007). Then, it estimates the probability distribution in testing data and rescales the optimal cut-off accordingly (Zou et al., 2016). As the β value for F-score is adjustable, the optimal threshold can be tuned for either a better precision or recall.

3.3 Highly customizable implementation

To facilitate users building up their own functional prediction pipelines, this method is implemented in GO with a command-line interface and hosted on Github[1] with an open-source MIT license, which is easy to distribute and install on all major operating systems. As it supports simple plain text tab-delimited matrix as inputs, it is also highly customizable for the choice of input networks, label definition, with a built-in automatically hyperparameter tuning procedure.

[1] `https://github.com/chenhao392/ecoc`

3.4 Benchmarking on the yeast neworks and labels

The method performance is benchmarked by comparing to a state-of-art algorithm Mashup (Cho et al., 2016), which also integrate multiple biological networks for gene function prediction. Rather than using a multi-label learning framework, it "mashes up" the networks into one joint representation and predict one label at a time using a radius SVM kernel.

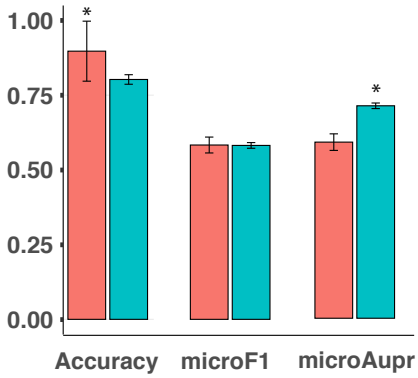


Figure 3·7: Model performance compared with "Mashup". Barplot for comparson between this gene function prediction method (red) and "Mashup" (cyan). The performace is compared using the fraction of correct top prediction for each gene (accuracy), the harmonic mean of precision and recall for all labels (microF1) and microAupr for all labels. * indicates t-test p-value < 0.05 from 10 random hold-outs.

To compare with this mehod, the exact same benchmark dataset for level 1 MIPS annotations in yeast is generated, using Mashup's Matlab implementation downloaded from their website. As shown in Figure 3.7, my multi-label learning method performs equally well with Mashup. Specifically, my model outperformed Mashup in prediction accuracy but performed worse in microAupr. Here, the accuracy, microF1 and microAupr metrics are defined in the same way as they are used in "Mashup". Specifically, accuracy measures the percentage of correctly predicted top label for each gene; microF1 calculated the overall F1 score after keeping top 3 predictions for each gene;

and microAupr calculate the area under the precision and recall Curve after stacking all predicted probabilities together.

3.5 Benchmark by predicting piRNA pathway genes in *D. melanogaster*

As stated in chapter 1, piRNA silences transposon activities in fly's gonad. Given the publicly available single cell RNA-seq dataset in recent years, it is possible to dissect tissue-specific networks out of the general networks for specific pathway gene prediction problems. As an case study, I pruned the publicly available stringDB networks (Szklarczyk et al., 2019) into testis specific ones using a recently published scRNA-seq dataset (Witt et al., 2019). Specifically, I only kept genes expressed in the early spermatogonia stage in the stringDB networks, as most known piRNA pathway genes peak their expression at this stage (Figure 3.8). As an result, a set of much smaller networks are generated specifically for piRNA pathway gene prediction (Table 3.1).

network	# genes	# stage genes	# interactions	# stage genes
co-expression	12,173	5,991	1,330,835	718,685
database	5,128	2,904	128,517	65,196
experimental	11,280	5,543	749,503	310,226
fusion	3,973	1,735	7,643	2,216
neighborhood	3,512	1,742	205,801	86,836
cooccurence	2,344	1,024	30,603	4,560
Total	11,828	6,121	1,298,440	812,196

Table 3.1: Stage-specific networks for predicting piRNA pathway genes.

To model the piRNA pathway with other associated labels, a g:profiler (Raudvere et al., 2019) based GO enrichment analysis was performed using these genes. Then, to avoid numerical difficulties in training, GO terms containing the exact same set genes were grouped and only the term with a smallest in-group p-value is selected. Also,

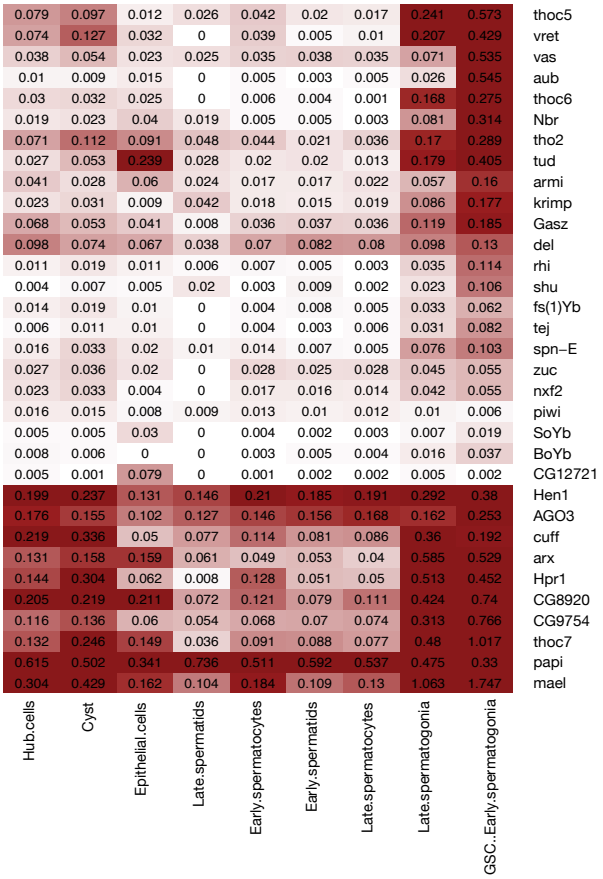


Figure 3·8: Heatmap for piRNA pathway gene expression in different developmental time at the single cell resolution. Darker red color indicates higher gene expression values. each row is a gene and each column is a tissue type or developmental stage.

further p-value (e^{-6}) and target size (25) cut-offs were used to remove unnecessary redundancies. As an result, 29 GO terms were selected as the associated labels (See appendix table 4). The model for piRNA pathway gene prediction is then trained on these pruned stringDB networks, using a multi-label definition consist of 30 labels.

After making predictions (See appendix table 5 and 6), they are benchmarked by a large-scale experimental screening effort for the piRNA pathway components, where genes' expression is knocked down (Czech et al., 2013). As the piRNA pathway is known to silence transposon activities, the possibilities for genes functioning in

piRNA pathway are indicated by four transposons' activities in this screen, namely HeTA, THARE, blood, and burdock. As shown in figure 3.9, the top 50 predictions have a significantly enriched transposon up-regulation, compared to the transposons' activities for all screened genes as background distribution.

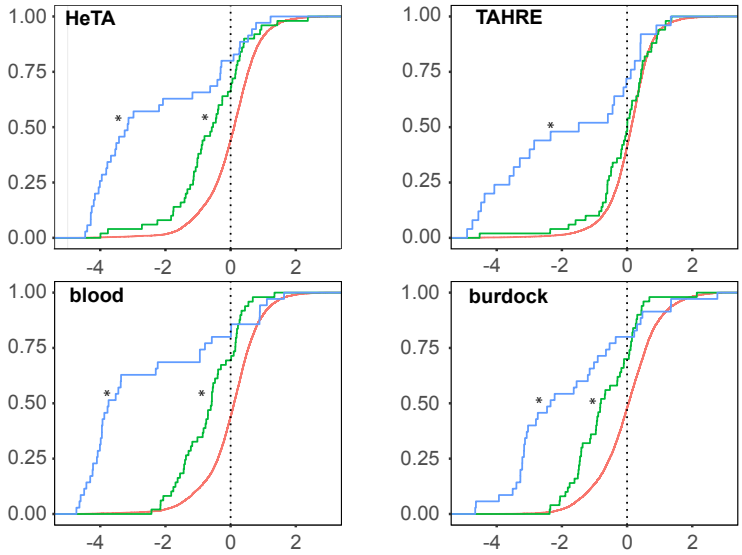


Figure 3·9: Transposon activities after knocking down known piRNA pathway genes and predicted genes. Empirical cumulative distribution for transposon activities after knocking down known piRNA pathway genes (blue), top 50 predictions (green), and all genes as background (red) for 4 transposons. * indicates t-test p-value $<$ 0.001 when compared with background genes.

piRNA pathway genes are known to be fast-evolving (Sarkies et al., 2015). However, it also contains evolutionarily conserved genes such as UAP56 and THO complex (Zhang et al., 2018) in *D. melanogaster*. Aiming for a secondary validation, I calculated the genome-wide Ka/Ks ratio of between *D. melanogaster* and the other 11 Drosophila species available in FlyBase (Thurmond et al., 2019), using FlyBase defined ortholog groups and a genomewide Ka/Ks ratio calculation pipeline (Wang et al., 2010; Zhang et al., 2013).

Interestingly, the same top 50 predicted genes have significantly lower Ka/Ks

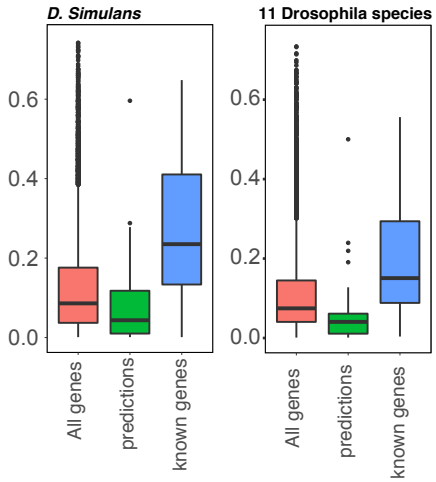


Figure 3·10: Ka/Ks ratio for known piRNA pathway genes and predicted genes. Boxplots for known piRNA pathway genes (blue), top 50 predictions (green), and all genes as background (red) between (A) *D. melanogaster* and *D. simulans*; and the average ratio between (B) *D. melanogaster* and 11 *Drosophila* species.

ratios between *D. melanogaster* and the other 11 species, including its most close relative *D. simulans*, compared to all genes' ratios as the background (Figure 3.10). Taken together, the two benckmarks indicate that this multi-label framework captures more conserved multi-function components in piRNA pathway.

Chapter 4

Discussion and future direction

As reported in previous chapters, piRISC binds to its target with a matching rule similar to miRNAs, though its supplementary pairing region (position 15-18) shift around 2 nucleotides towards the 3' end. With this binding preference, they densely covered many genes' CDS regions, such as xol-1, fbxb-97, comt-3 and dhc-1. Mutating one of the many piRNAs targeting these genes can de-silence them, indicating these piRNAs are silencing genes cooperatively. Also, cytosines are found to be enriched in the first target position and throughout the target sites in both WAGO and CSR-1 target mRNAs. This preference for C may help to trigger the secondary 22G-RNAs amplification, using the targeted RNAs as templates.

Previous studies have shown that PIWI, WAGO and CSR-1 pathways propagate the epigenetic memories of gene expression states across generations. Our multi-generation validation experiments further indicate the importance of the 2nd and 3rd target position, as they are still important for silencing the GFP target genes at F9 generation. Also, it is interesting to find the enrichment of 22G-RNAs at both ends of piRNA target sites persists in a prg-1 mutant background, though to a less extent, indicating an independent silencing mechanism does not require piRNA.

Our analysis also shows that CSR-1 pathway directly competes with piRNA binding, consistent with the finding that CSR-1 protects its targets. However, it is also known that CSR-1 can regulate its target by slicing. Surprisingly, our further experiment for WAGO IP at the csr-1 catalytic mutant background shows enrichment of

WAGO binding at presumably CSR-1 target genes' 3' UTR. In the future, it is of great interest to investigate the WAGO and CSR-1 targeting preference and mechanism, following this lead at the csr-1 catalytic mutant background.

In chapters 1 and 3, I introduced a computational model for gene function prediction. To be consistent with previous publications, I chose to use the same benchmark data and measures as in "Mashup". Alternatively, it is also possible to validate these predictions through organized consortiums' efforts or through orthogonal datasets, such as GWAS cohorts for diseases. Also, following the GBA principle, additional biological datasets, such as genes' genomic collinearity and their expression pattern at a single-cell resolution, can be integrated into the model. It is of great interest to see how this computational model performs in these consortiums' settings and to what extent that these additional biological datasets can further help the prediction. I will also discuss these in detail in this chapter.

4.1 Rules governing piRNA Targeting

As shown in chapter 2, our analysis of the CLASH chimeric reads suggests that animal PIWI and AGO-clade Argonautes have broadly similar patterns of base-pairing. As previously described for miRNA RISC, we find that pairing in the seed region is important for piRISC to function, and to a lesser extent in 3′ supplemental region (Shin et al., 2010). The most notable difference we observe is a shift in 3′ supplementary pairing. It is from positions 13 to 16 in miRISC, while it is from positions 15 to 18 in PRG-1 piRISC (Grimson et al., 2007), perhaps consistent with structural differences between miRISC and piRISC (Matsumoto et al., 2016).

In addition to base-pairing interactions, both AGO and Piwi Argonautes make direct contact with their target RNAs, including specific amino acid contacts with the t1 nucleotide. Human AGO2 and insect Piwi proteins (i.e., Siwi and Aubergine)

exhibit a strong preference for adenosine at t1 (t1A), which differs from our finding that PRG-1 prefers t1C. This preference for C may help ensure that PRG-1 target sites often have optimal positioning of a C residue that can serve as a start site for RdRP-dependent amplification of 22G-RNAs. A comparison of the region in PRG-1 that corresponds to the t1 binding pocket in other Argonautes suggests a possible structural basis for this discrimination for t1C. Whereas the polar hydrophobic amino acid Thr640 in Siwi and Aubergine is thought to bind t1A (Matsumoto et al., 2016), the corresponding position in PRG-1 is a non-polar hydrophobic leucine.

Using a sensitive epigenetic silencing assay, we were able to directly validate the importance of pairing at each position of the seed and 3' supplemental pairing regions. Silencing was most sensitive to the loss of pairing at positions 2 and 3, suggesting that targeting is initiated by the first half of the seed region. Remarkably, with the exception of positions 9 to 13, which had very little effect on silencing, single nucleotide substitutions at any other location from positions 2 to 8 or 14 to 21 dramatically reduced silencing over the first several generations. Mutants with mismatches in the 3' supplementary pairing region eventually silenced the target in later generations, but mutants with mismatches in the seed region, especially at g2 and g3, never exhibited full silencing of the target. Thus, seed and 3' supplementary pairing are of key importance to piRNA targeting. Even single-nucleotide changes dramatically reduced targeting and extended the number of generations required for penetrant silencing.

4.2 The physiology of piRNA targeting

In most animals, PIWI mutants are completely sterile, likely due at least in part, to loss of transposon regulation. In worms, most transposons appear to be silenced by epigenetic mechanisms—i.e., WAGO 22G-RNAs and heterochromatin pathways—that maintain transgenerational silencing downstream of PRG-1 (Ashe et al., 2012;

Bagijn et al., 2012; Gu et al., 2009; Lee et al., 2012; Shirayama et al., 2012). This additional layer of epigenetic silencing may explain why prg-1 mutants exhibit relatively minor transposon activation and fertility defects during early generations, but exhibit declining fertility over multiple generations (i.e., a mortal germline phenotype) (Simon et al., 2014).

PRG-1 is nevertheless constantly required to maintain silencing at some loci. Transgenes exposed to both positive (i.e., CSR-1-dependent) and negative (i.e., piRNA-dependent) signals can achieve a balanced state of regulation, where PRG-1 targeting becomes essential to maintain silencing (Seth et al., in press). At least a few hundred endogenous mRNAs are significantly up-regulated in prg-1 mutants, with a concomitant loss of robust 22G-RNAs levels. One such gene, xol-1, is silenced in the hermaphrodite germline by an X-chromosome expressed piRNA, 21ux-1(Seth et al., 2018; Tang et al., 2018). Silencing of xol-1 ensures that hermaphrodite offspring respond robustly to signals that initiate dosage compensation and sex determination in the early embryo. Although 21ux-1 is by far the most abundant piRNA species, a piRNA with average abundance (21ur-4863) binds xol-1 more efficiently based on the frequency of CLASH hybrid identification. 21ur-4863 is predicted to bind xol-1 with higher binding energy than predicted for 21ux-1, highlighting the importance of binding energy rather than abundance in driving piRNA targeting. Surprisingly, both 21ur-4863 and 21ux-1 are required to maintain xol-1 silencing, suggesting that they—and perhaps other—piRNAs cooperatively silence xol-1. Consistent with this idea, the pattern of 22G-RNA induction along xol-1 extends beyond the regions proximal to these two piRNA target sites, suggesting that additional piRNAs likely contribute to the cooperative regulation of xol-1 mRNA. Indeed, our CLASH experiments identified 40 piRNAs that target different sites in xol-1 mRNA. Similarly, because we tagged the endogenous prg-1 gene with GFP and FLAG to permit tandem-affinity purification,

we were able to identify 92 different piRNAs that target sites distributed along the length of gfp. Cooperative targeting by these piRNAs could explain why 22G-RNA accumulation occurs broadly along silenced gfp transgenes (Shirayama et al., 2012; Seth et al., 2018). Remarkably, even though multiple piRNAs regulate xol-1, changing a single nucleotide within the seed or 3' supplementary pairing regions of 21ux-1 can disrupt silencing of xol-1 and thus affect the regulation of dosage compensation and sex determination.

In summary, our findings show that piRNAs target the entire germline transcriptome. Together with findings from previous and parallel studies our findings also suggest that piRNAs are remarkably versatile in their control of gene expression. piRNAs can act decisively in one generation to initiate epigenetic silencing that persists for multiple generations without need for further piRNA targeting. piRNAs can act cooperatively to silence germline mRNAs (e.g., xol-1) that would otherwise reactivate in each generation. And finally, piRNAs can act gradually, over multiple generations, to progressively silence a germline mRNA. Understanding how piRNAs achieve these nuanced modes and tempos of regulation may shed light on whole new vistas of post-transcriptional and epigenetic regulation in animal germlines.

4.3 Molecular cross-talk between germline Argonaute pathways

As shown chapter 2, we took the unbiased approach of directly cross-linking piRNAs to target RNAs *in vivo*. The resulting transcriptome-wide snap-shot of piRNA/target-RNA interactions reveals that all germline mRNAs undergo piRNA surveillance. Our findings are consistent with a model for germline gene regulation where mRNAs undergo comprehensive post-transcriptional scanning by Argonaute systems. More than 10,000 distinct piRISCs access hundreds of thousands of target sequences on germline

mRNAs. Our finding that binding energy was better correlated with hybrid formation than was piRNA abundance, suggests that, for most piRNAs, piRISC concentration is not limiting. Thus surveillance by piRISC is both transcriptome-wide and remarkably efficient. Perhaps as yet unknown features of the enigmatic P-granules, where piRISC resides and presumably functions, create an environment that facilitates this seemingly daunting task of comprehensive mRNA surveillance (See figure 7 in appendix; see also (Seth et al., 2018)).

Previous genetic studies have revealed interactions between the Piwi pathway and two Argonaute pathways that propagate epigenetic memories of gene expression states: the WAGO pathway, which targets silenced genes, and the CSR-1 pathway, which targets expressed genes. Targeting by WAGO and CSR-1 Argonautes is readily apparent since both engage 22G-RNAs templated directly from the target RNA by RdRP. Therefore, the comprehensive identification of PRG-1/piRNA target sites affords an opportunity to explore how piRNA targeting correlates with 22G-RNA levels across annotated WAGO and CSR-1 targeted mRNAs. A striking and unanticipated pattern of 22G-RNA levels emerged from this analysis. On WAGO-targeted mRNAs, piRNA target sites were correlated with three predominant 22G-RNA peaks, one in the center at t12, and one on each side of the targeted site. Interestingly, the central peak at t12 was completely dependent on PRG-1, while the flanking peaks were much less dependent on PRG-1. The flanking peaks that persist in prg-1 mutants may reflect piRNA-initiated 22G-RNAs that function in WAGO-mediated trans-generational silencing. Consistent with this idea, analyses of data from published WAGO IP experiments indicate that 22G-RNAs at these somewhat prg-1-independent flanking sites associate with Argonautes required for propagating piRNA-induced epigenetic silencing (WAGO-1 and WAGO-9) (See figure 7 in appendix). Interestingly, the strongly prg-1-dependent 22G-RNAs generated at t12 associate with WAGO-1 only.

Thus, it will be interesting to learn why WAGO-1 but not WAGO-9 binds these t12-associated species and whether their biogenesis depends on PRG-1-dependent mRNA slicing which is predicted to occur between t10 and t11.

Our findings also shed light on the relationship between PRG-1 and CSR-1 targeting. Depletion of CSR-1 resulted in an increase in both unique and total piRNA hybrid reads on mRNAs targeted by CSR-1. These findings are consistent with genetic findings that CSR-1 protects its targets from PRG-1-induced silencing (Seth et al., 2013; Wedeles et al., 2013). Moreover, piRNA target regions on CSR-1 target mRNAs exhibit a pattern of 22G-RNA accumulation that is strikingly different from that observed on WAGO-targeted mRNAs. Instead of a central peak and twin flanking peaks, as in WAGO targets, a small but reproducible 22G-RNA peak, positioned just 5 nucleotides 3' of the piRNA target site (See figure 7 in appendix), was evident in CSR-1 target mRNAs. The 22G-RNA distribution around piRNA target sites in CSR-1 mRNAs remained unchanged after the short period of CSR-1 depletion. This finding suggests that the effect of CSR-1 depletion on patterns of WAGO 22G-RNA accumulation, if any, is less rapid and perhaps less direct than its effect on piRNA targeting of these mRNAs. Unfortunately, depletion of CSR-1 leads to adult sterility, precluding a longer-term multi-generational analysis of 22G-RNA patterns. Taken together, our findings suggest that CSR-1 protects its targets from piRNA silencing in two ways; first by reducing the frequency of PRG-1 piRISC binding, and second, perhaps more indirectly, by preventing 22G-RNA accumulation at t12 and flanking regions correlated with WAGO-1 and WAGO-9 targeting.

4.4 On-going investigation on WAGO and CSR-1 pathway interactions

Both CSR-1 and WAGOs are Argonautes proteins that engage 22G-RNAs to silence their RNA targets by slicing their targets. As shown in chapter 2, we found the CSR-1 pathway directly competes with piRNA binding while the WAGO silencing pathway can be triggered by the piRNA targeting. We also found a CSR-1 22G-RNA signal at the upstream of the piRNA binding sites, indicating a potential correlation, though to a less extent. Surprisingly, a further experiment at the CSR-1 catalytic mutant background shows an enriched WAGO 22G-RNAs at genes' 3' end UTR on the CSR-1 target genes, while the 22G-RNA abundances in WAGO targets are not changing. This phenomenon indicates the WAGO pathway can compensate CSR-1's post-transcriptional regulation roles, particularly at their regulatory targets' 3' end UTRs. It is of great interest to investigate how WAGO Argonautes target these regions, instead of the whole gene body, as they are known to target the mRNAs of more than 1000 endogenous WAGO target genes (Gu et al., 2009).

4.5 Benchmarks for gene function prediction methods

A fair benchmark is essential to compare algorithms. However, it is not easily achievable for gene function prediction methods, as the fundamental validation is originated from biological experiments. As an alternative approach, we can stratify the whole datasets into training and testing subsets, where the algorithm's performance can be measured via cross-validation and leave-one-out settings. However, these settings are sometimes not consistent with each other, resulting in debates between research groups and confusions to users (Murali et al., 2006). Cautions are also necessary for developing benchmark datasets, as the annotations from different sources may overlap significantly. For instance, stringDB networks are already optimized to KEGG anno-

tations (von Mering et al., 2005). If a benchmark dataset is prepared using KEGG annotation with stringDB networks, algorithms are likely to overfit, rather than being correctly benchmarked.

In order to avoid these problems and to be consistent with previous methods, I reproduced the same stratified subsets used by Mashup, which are consist of GO term annotations for genes in stringDB. I also used the same measures that are used by Mashup and GeneMania (Cho et al., 2016; Mostafavi et al., 2008), which are accuracy for top prediction to each gene, microF1 for top three predictions for each gene, and microAupr for all predictions. While accuracy measures the top prediction, microF1 and microAupr measure the correctness in predicting multiple labels for the same gene.

Alternative approaches are also used in recent years for benchmarking. In the most recent Dialogue for Reverse Engineering Assessments and Methods (DREAM) challenge for disease module identification (Choobdar et al., 2019), the predicted modules are tested using a collection of 180 genome-wide association studies (GWAS). This approach avoids any potential cross-talk between annotation databases since GWAS datasets were not used for gene function annotations. Also, as a community approach, the Critical Assessment of protein Function Annotation (CAFA) consortium holds the submitted predictions for several months. It then uses the accumulated new gene annotations during this period for testing. As random walk based algorithms have appeared in the DREAM challenges for module detection and CAFA consortium also provides a fair method benchmark, it is of great interest to participate in these challenges in the future for benchmarking this multi-label learning model.

4.6 Additional datasets following "guilt by association" principle

In chapter 1 and 3, I introduced a multi-label learning framework that uses multiple biological networks for gene function prediction. Specifically, following the GBA principle, phylogenetic profiles, co-expression profiles, protein-protein interaction, genetic interactions, and semantic similarity from GO annotations are introduced. Though not mentioned directly, fusion genes and neighborhood genes are also summarized in stringDB (Szklarczyk et al., 2019), which are used as part of the benchmark dataset in Mashup (Cho et al., 2016).

If focusing on more specific species, additional biological datasets are also available, such as defined operons and syntenic regions. For instance, in microbial genomes, Zheng *et al* predicted operon structures using a graph representation for pathways (Zheng et al., 2002). Though to much less extend on higher-order organisms, operon and syntenic regions can associate co-functioning genes. For instance, 15% genes are believed to function as operons in *C. elegans* (Blumenthal and Gleason, 2003; Guiliano and Blaxter, 2006). In 2012, Proost *et al* found that conserved collinearity regions between close Eukarotes species are functional coherence (Proost et al., 2012).

In recent years, other datasets are also emerging at the genomic scale, such as the topologically associated domain, and cis-regulatory elements (ENCODE Project Consortium et al., 2012; Nora et al., 2012). These regulatory elements can be integrated into GBA based networks as nodes, with the co-regulation association as edges (Zhu et al., 2007). For example, Tian *et al.* suggested a MOCHI algorithm (Tian et al., 2020) to identify heterogeneous interactome modules, representing a set of gene loci that contain in co-regulated genes.

In addition to these "horizontal" approaches that integrate biological datasets, "vertical" strategies that prunes the GBA datasets into the context-specific subsets

are also emerging (Zitnik et al., 2019). For example, Kotlyar *et al.* pruned the tissue-specific databases for model organisms (Kotlyar et al., 2016). As a case study in chapter 3, additional piRNA pathway gene are predicted using a set of testis-specific networks in *D. melanogaster*. As the number of single-cell datasets is rapidly increasing in recent years (Svensson et al., 2019), this context-specific approach is promising for many customized pathway prediction problems.

Appendix A

Appendices

Species	
Ajellomyces Dermatitidis	*Giardia Intestinalis*
Amphimedon Queenslandica	*Globodera Pallida*
Aplysia Californica	*Haemonchus Contortus*
Arabidopsis Thaliana	*Hydra Vulgaris*
Arthroderma Otae	*Ixodes Scapularis*
Ascaris Suum	*Leishmania Major*
Babesia Bovis	*Lodderomyces Elongisporus*
Bison Bison	*Malassezia Globosa*
Bombus Impatiens	*Microplitis Demolitor*
Bombus Terrestris	*Neosartorya Fischeri*
Bos Mutus	*Nippostrongylus Brasiliensis*
Brugia Malayi	*Panagrellus Redivivus*
Bursaphelenchus Xyliphilus	*Penicillium Chrysogenum*
Camelina Sativa	*Phaeodactylum Tricornutum*
Candida Glabrata	*Plasmodium Falciparum*
Chlamydomonas Reinhardtii	*Pristionchus Pacificus*
Ciona Intestinalis	*Pyrus X*
Ciona Savignyi	*Romanomermis Culicivorax*
Clavispora Lusitaniae	*Saccoglossus Kowalevskii*
Coprinopsis Cinerea	*Scheffersomyces Stipitis*
Crassostrea Gigas	*Schizosaccharomyces Japonicus*
Cryptosporidium Parvum	*Schizosaccharomyces Pombe*
Dictyostelium Discoideum	*Strongylocentrotus Purpuratus*
Entamoeba Histolytica	*Tarenaya Hassleriana*
Erythranthe Guttata	*Theileria Annulata*
Fopius Arisanus	*Trichinella Spiralis*

Table A.1: Species for building phylogenetic profiles of genes in *D. melanogaster*. 52 Eukaryotes species that are differentiated at least 500 million years ago with *D. melanogaster* and also 500 million years away from each other are used for building the phylogenetic profiles of *D. melanogaster*.

Time course libraries		
Embryo0-2h	Embryo20-22h	Prepupae+24h
Embryo2-4h	Embryo22-24h	Prepupae+2d
Embryo4-6h	L1	Prepupae+3d
Embryo6-8h	L2	Prepupae+4d
Embryo8-10h	L3+12h	Male+1d
Embryo10-12h	L3PS1-2	Male+5d
Embryo12-14h	L3PS3-6	Male+30d
Embryo14-16h	L3PS7-9	Female+1d
Embryo16-18h	Prepupae	Female+5d
Embryo18-20h	Prepupae+12h	Female+30d

Table A.2: modENCODE time course libraries for building gene expression profiles in *D. melanogaster*.

Tissue cell libraries		
(Embryo) 1182-4H	TestesMatedMale+4d	(L3 wing disc) ML-DmD16-c3
(Embryo) GM2	FatL3	(L3 prothoracic leg disc) CME_L1
(Embryo) Kc167	FatPrepupae	(L3 eye-antennal disc) ML-DmD11
(Embryo) S1	FatPrepupae+2d	(L3 haltere disc) ML-DmD17-c3
(Embryo) S3	SalivaryGlandsL3	(L3 mixed imaginal discs) ML-DmD4-c1
(Embryo) Sg4	SalivaryGlandsPrepupae	CarcassL3
HeadsVirginFemale+1d	ImaginalDiscsL3	CarcassMixedMaleFemale+1d
HeadsVirginFemale+4d	CNSL3	CarcassMixedMaleFemale+4d
HeadsVirginFemale+20d	CNSPrepupae+2d	CarcassMixedMaleFemale+20d
HeadsMatedFemale+1d	(L3 wing disc) CME_W2	(Tumorous blood cells) MBN2
HeadsMatedFemale+4d	(L3 wing disc) ML-DmD8	AccessoryGlandsMatedMale+4d
HeadsMatedFemale+20d	(L3 wing disc) ML-DmD9	DigestiveSystemL3
HeadsMatedMale+1d	(L3 wing disc) ML-DmD21	DigestiveSystemMixedMaleFemale+1d
HeadsMatedMale+4d	(L3 wing disc) ML-DmD32	DigestiveSystemMixedMaleFemale+4d
HeadsMatedMale+20d	(L3 CNS) ML-DmBG1-c1	DigestiveSystemMixedMaleFemale+20d
OvariesMatedFemale+4d	(L3 CNS) ML-DmBG2-c2	

Table A.3: modENCODE tissue cell libraries for building gene expression profiles in *D. melanogaster*.

GO term	P-value	Target size	Description
GO:0003724	4.55e-7	42	RNA helicase activity
GO:0031047	8.89e-23	86	gene silencing by RNA
GO:0035194	2.43e-14	64	posttranscriptional gene silencing by RNA
GO:0007281	1.34e-13	754	germ cell development
GO:0022412	3.45e-13	912	cellular process involved in reproduction
GO:0048477	1.62e-12	627	oogenesis
GO:0030717	2.26e-11	31	oocyte karyosome formation
GO:0010608	4.57e-11	225	posttranscriptional regulation of gene expression
GO:0010629	1.63e-10	690	negative regulation of gene expression
GO:0048519	4.92e-10	1637	negative regulation of biological process
GO:0051321	5.25e-10	272	meiotic cell cycle
GO:0006403	5.28e-10	208	RNA localization
GO:0010467	7.46e-10	2526	gene expression
GO:0010468	1.207e-9	1704	regulation of gene expression
GO:0010605	2.283e-9	936	negative regulation of metabolic process
GO:0090304	1.15e-8	2318	nucleic acid metabolic process
GO:0016070	1.19e-8	2099	RNA metabolic process
GO:0071427	2.55e-8	41	mRNA-containing ribonucleoprotein complex
GO:0009994	3.32e-8	163	oocyte differentiation
GO:0060255	6.03e-8	2254	regulation of macromolecule metabolic process
GO:0046843	1.48e-7	52	dorsal appendage formation
GO:0034641	1.63e-7	3114	cellular nitrogen compound metabolic process
GO:0048599	2.55e-7	141	oocyte development
GO:0016246	7.30e-7	35	RNA interference
GO:0007315	8.37e-7	66	pole plasm assembly
GO:0043186	5.23e-30	38	P granule
GO:1990904	1.59e-11	655	ribonucleoprotein complex
GO:0043232	6.08e-9	1755	intracellular non-membrane-bounded organelle
GO:0044424	7.74e-7	6846	intracellular part

Table A.4: GO terms selected as piRNA associated labels in *D. melanogaster.*

Gene	probability	HeTA	THARE	blood	burdock
CG6418	1.000	-1.127	-1.581	-0.748	-1.409
CG10333	0.943	-1.413	-0.443	-1.019	-1.492
Dbp45A	0.927	0.020	-0.606	0.333	-0.218
CG6227	0.913	-1.020	0.426	-0.145	-0.959
Rm62	0.901	-0.864	-0.651	-2.154	-1.892
Art1	0.871	0.725	0.744	1.345	0.311
Cpsf73	0.859	-0.390	-0.190	-0.609	-1.364
RpII140	0.852	-1.024	0.902	-0.535	-0.668
l(2)37Cb	0.845	-1.378	-0.554	-1.371	-0.683
Prpk	0.833	-0.453	-0.487	-0.398	-0.294
RpII18	0.824	-1.135	-0.606	-1.177	-0.885
Rpb5	0.815	-1.151	1.184	-0.573	-1.510
Rbp1	0.805	-0.627	-0.857	-0.297	-0.525
Chd3	0.799	-0.435	0.343	0.149	0.150
IntS11	0.793	-3.995	-4.524	-0.697	-0.822
Gem3	0.785	0.348	0.961	-1.441	-0.828
mle	0.778	0.398	-0.029	0.441	0.184
Tbp	0.773	-0.086	0.007	-0.580	0.438
CG10907	0.764	0.903	0.397	0.016	0.502
Art3	0.754	0.004	0.539	-0.479	-0.915
CG7878	0.750	-0.888	-1.797	0.296	-1.478
CG5800	0.743	-0.372	0.005	-1.400	-0.800
RpII15	0.738	-1.327	0.758	-1.220	-2.072
CG9344	0.732	-1.072	0.152	-1.757	-0.950
Rs1	0.727	-1.827	-0.599	-1.823	-1.456

Table A.5: Top 25 piRNA pathway gene prediction in *D. melanogaster*.

Gene	probability	HeTA	THARE	blood	burdock
CG9630	0.722	-0.255	1.170	-0.856	-0.177
CG14683	0.714	0.290	0.366	0.273	0.122
SF2	0.710	1.434	0.921	0.225	0.684
CG17266	0.705	0.055	-0.493	0.183	0.056
cyp33	0.700	-0.579	0.476	-0.549	-0.303
Cypl	0.697	0.212	0.433	0.160	-0.095
RpII33	0.693	-1.752	-1.273	-1.692	-1.762
hrg	0.691	-0.994	-0.747	-1.553	-2.056
CG32344	0.687	-1.241	0.010	-1.292	-1.404
Aos1	0.684	-2.722	-2.359	-1.427	-0.928
x16	0.682	-0.257	0.467	-0.695	-1.613
SC35	0.680	0.157	0.628	0.195	0.109
tra2	0.678	0.945	0.360	0.680	0.267
CG3645	0.674	0.274	0.142	0.143	0.288
CG4338	0.672	0.141	-0.059	0.546	0.160
Rpt5	0.670	-1.763	0.281	-2.066	-2.355
Rpt3	0.668	-0.538	0.097	-0.573	-1.819
Dis3	0.666	-2.247	-0.152	-2.148	-2.376
CG3225	0.664	0.412	1.362	-0.770	2.139
Dbp73D	0.661	-0.911	0.079	-0.831	-1.137
me31B	0.658	0.180	-0.178	0.095	0.447
Rpt6	0.652	-0.803	-0.785	-1.626	-1.444
Pabp2	0.647	-3.765	-0.099	-2.430	-1.113
CG7747	0.645	2.369	-0.647	3.440	0.071
pont	0.642	-1.535	-0.310	-0.385	-0.102

Table A.6: piRNA pathway gene prediction ranked 26-50 in *D. melanogaster.*

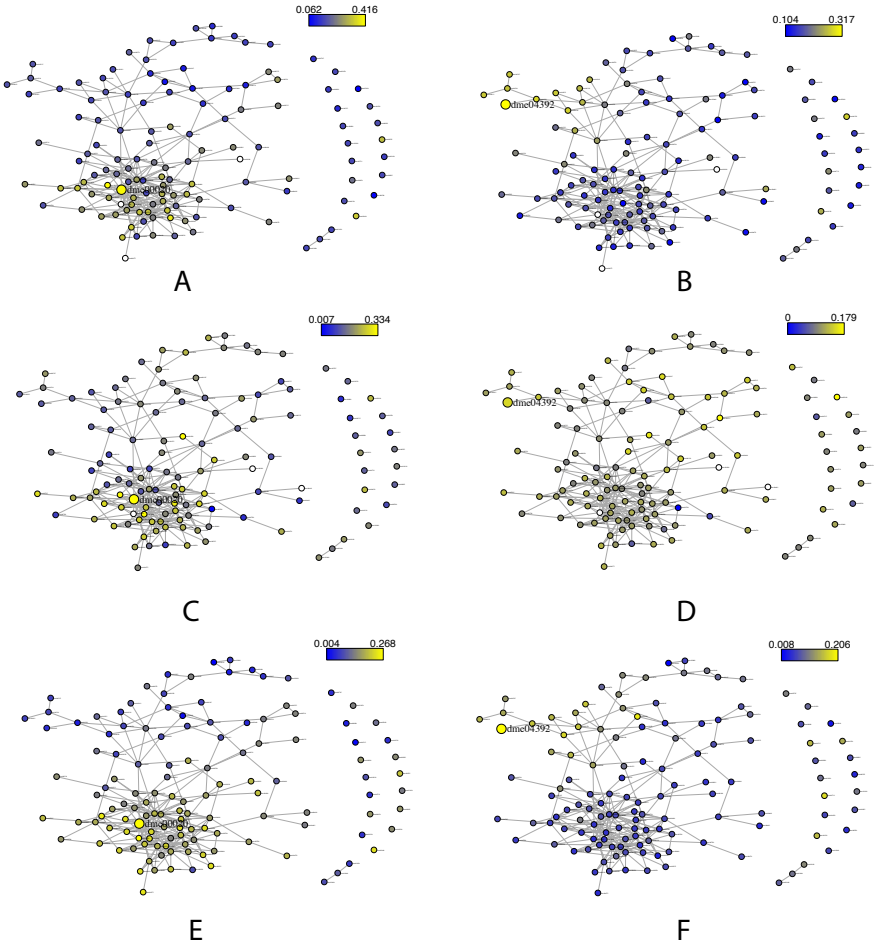


Figure A·1: The view of KEGG pathways in *D. melanogaster* from GO terms' semantic similarities. the pathway nodes' are colored according to their average gene associations to genes in a focus pathway, which is indicated by a larger node. Pathways with lower associations are in blue, while with higher associations are in yellow. The pathway relationships focusing on pathway "dme00020" and "dme04392" from biological process are shown in (A) and (B); those from cellular components are shown in (C) and (D); and those from molecular functions are shown in (E) and (F).

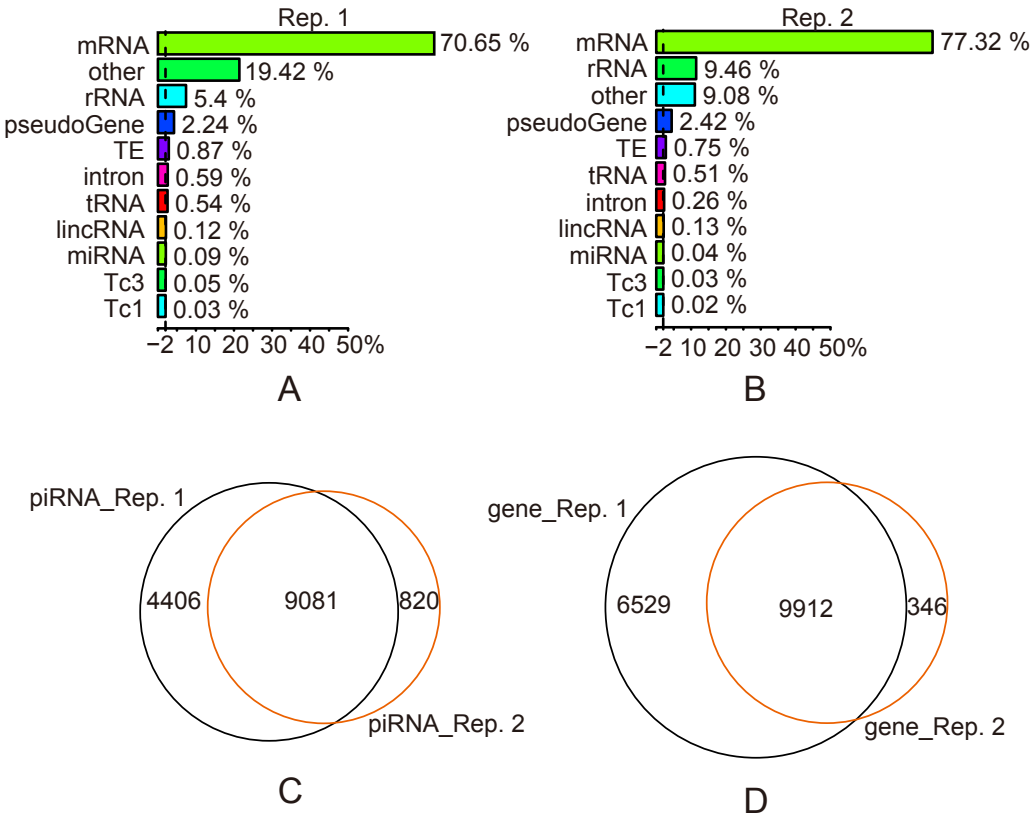


Figure A·2: piRNA targets classification, abundance, and overlapping in two independent CLASH experiments. (A) distribution of piRNA targets in percentages; and (B) reproducbility for piRNA species and their mRNA targets.

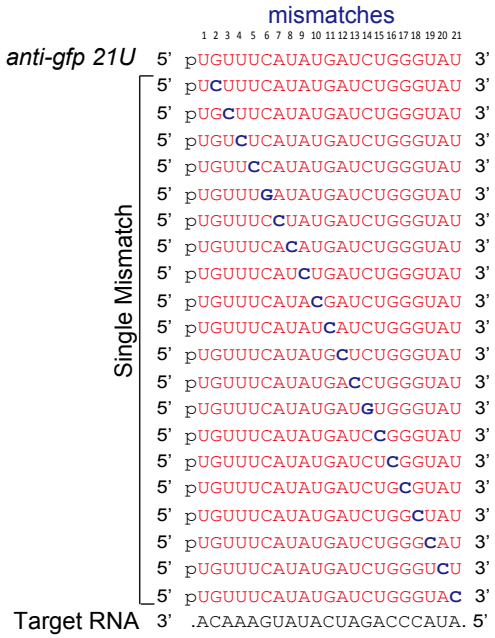


Figure A·3: Seed and 3' supplementary pairing are required for silencing. anti-gfp piRNA (red) and single-nucleotide mismatches (blue) from positions 2 to 21 on the piRNA target site in cdk-1::gfp (black).

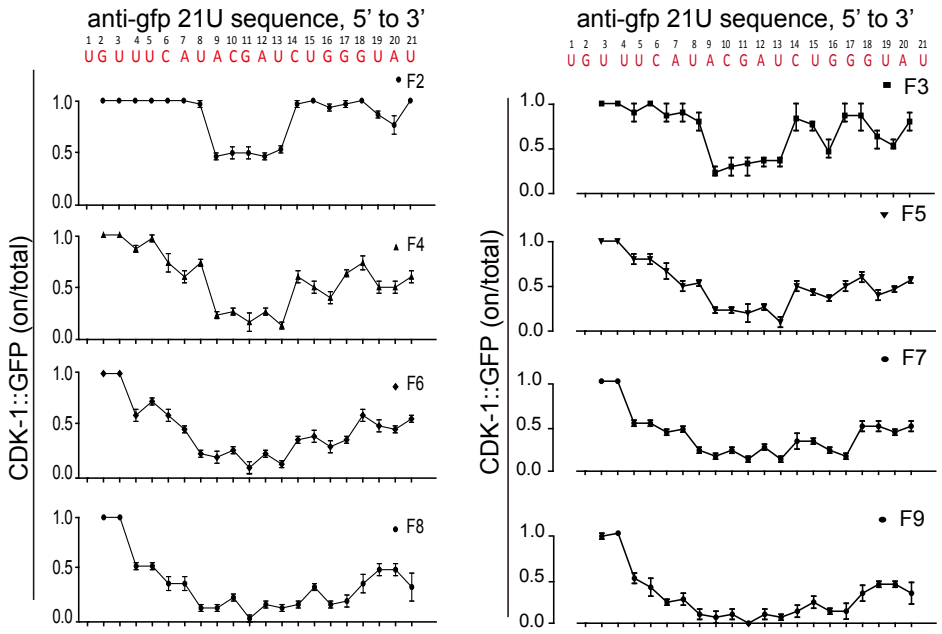


Figure A·4: CRISPR experiments validate the piRNA target rule in C. elegans across multiple generations. Eight plots show the effect of piRNA de-silencing for F2 to F9 offsprings. Each plot shows the ratio of CDK-1::GFP turned-on worms with single-base mutations on the piRNA, from t2-t21.

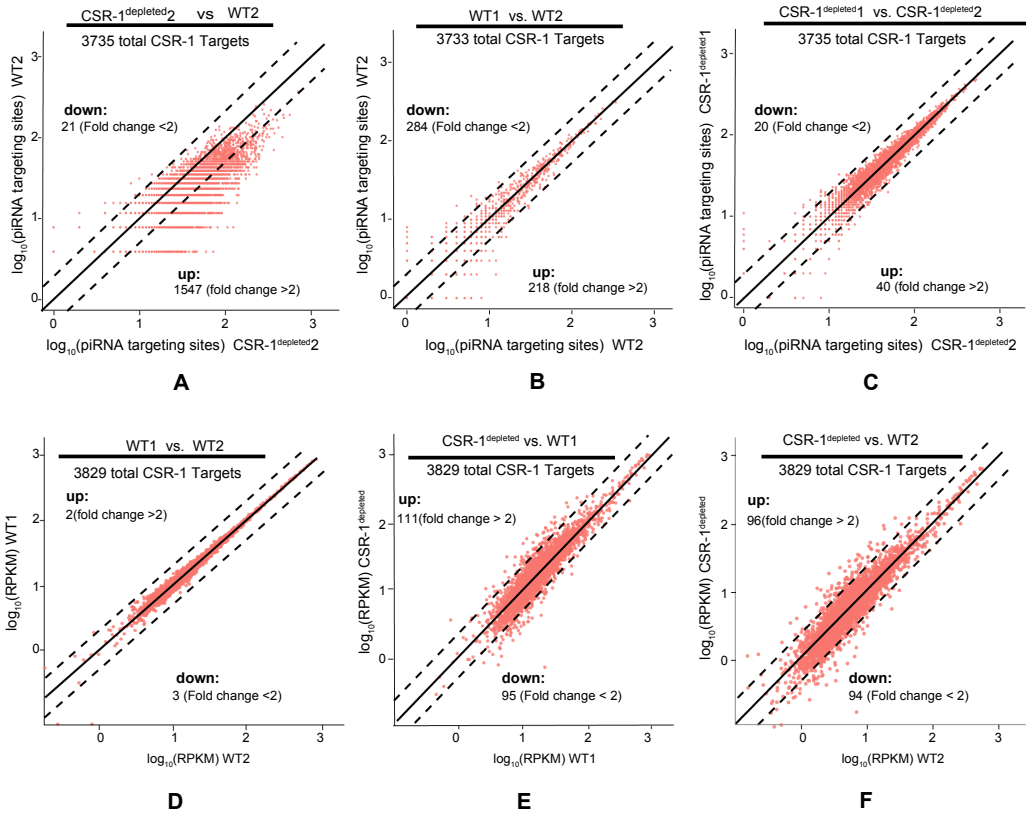


Figure A·5: piRNA targeting density on CSR-1 targets and corresponding gene expression levels in WT and CSR-1 depletion backgrounds. (A), (B), and (C) show the number of CLASH defined target sites on CSR-1 target genes in the two backgrounds; (D), (E), and (F) show the gene expression levels of these genes in both backgrounds.

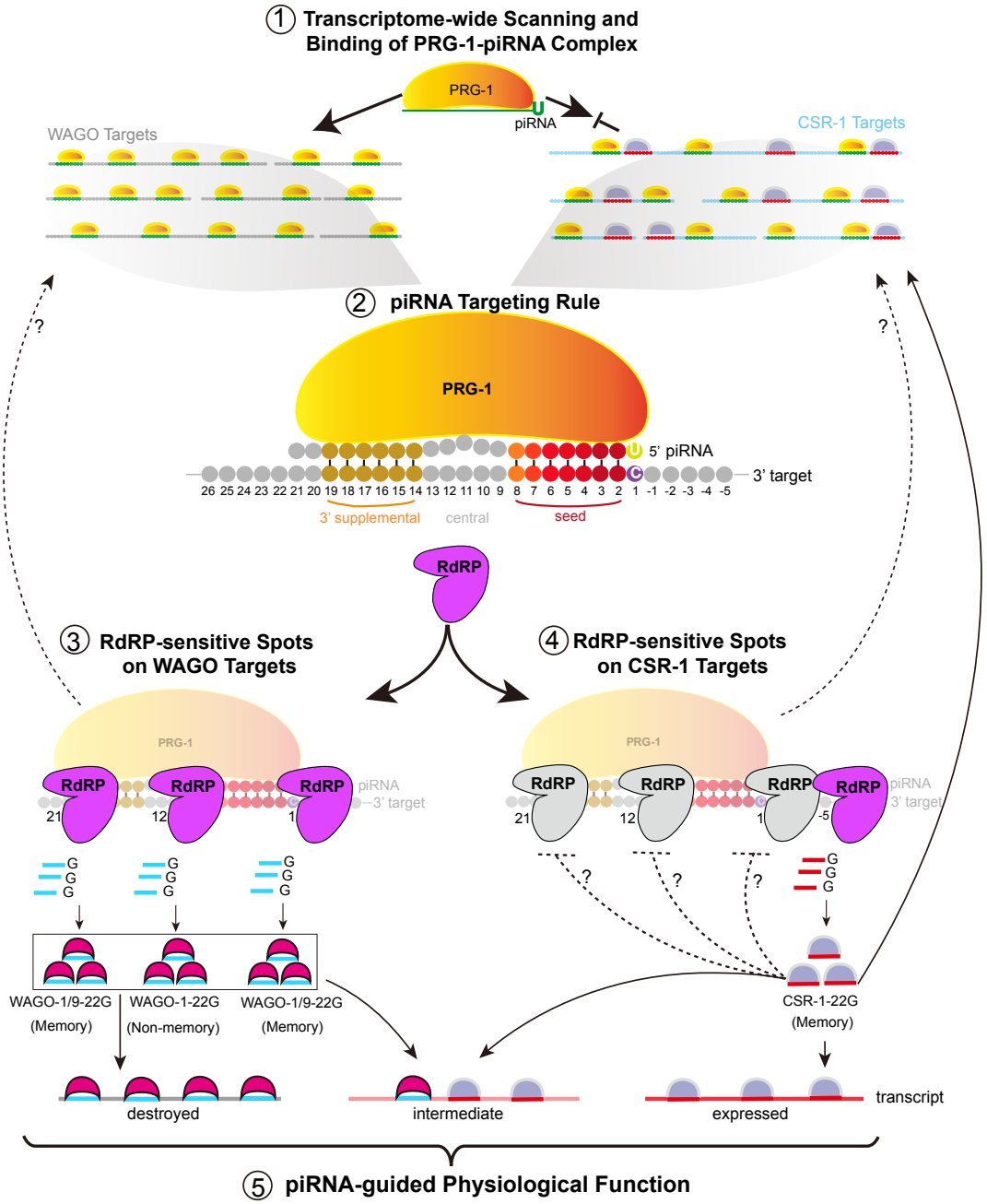


Figure A·6: Model for a regulatory landscape of piRNAs in the *C. elegans* germline.

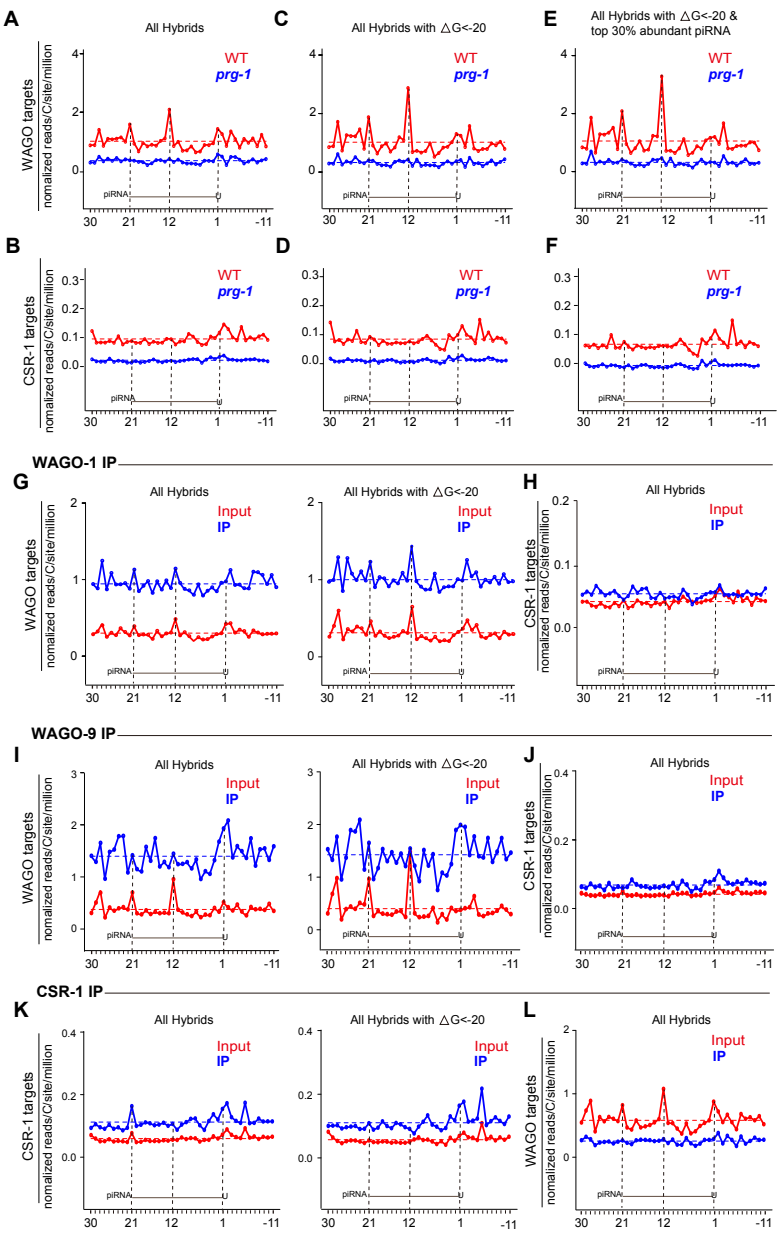


Figure A·7: 22G-RNAs distribution at CLASH defined piRNA target sites in small RNAseq, WAGO and CSR-1 IP libraries. Solid lines show the normalized 5'G counts at and around the piRNA target sites while the dash lines show the average; Red and blue indicate wild type and prg-1 mutant background, respectively. subsets of target sites defined by binding energy and piRNA abundance cut-offs are shown in (E), (F), (G), and (I). small RNAseq Replicates are shown in (A) - (F); WAGO-1 and WAGO-9 IP libraries are shown in (G) - (J); CSR-1 IP library is shown in (K) and (L).

References

Aravin, A. A., Naumova, N. M., Tulin, A. V., Vagin, V. V., Rozovsky, Y. M., and Gvozdev, V. A. (2001). Double-stranded RNA-mediated silencing of genomic tandem repeats and transposable elements in the D. melanogaster germline. *Current biology : CB*.

Ashe, A., Sapetschnig, A., Weick, E.-M., Mitchell, J., Bagijn, M. P., Cording, A. C., Doebley, A.-L., Goldstein, L. D., Lehrbach, N. J., Le Pen, J., Pintacuda, G., Sakaguchi, A., Sarkies, P., Ahmed, S., and Miska, E. A. (2012). piRNAs can trigger a multigenerational epigenetic memory in the germline of C. elegans. *Cell*.

Bagijn, M. P., Goldstein, L. D., Sapetschnig, A., Weick, E.-M., Bouasker, S., Lehrbach, N. J., Simard, M. J., and Miska, E. A. (2012). Function, targets, and evolution of Caenorhabditis elegans piRNAs. *Science (New York, N.Y.)*.

Barabási, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature Reviews Genetics*.

Bartel, D. P. (2009). MicroRNAs: target recognition and regulatory functions. *Cell*.

Batista, P. J., Ruby, J. G., Claycomb, J. M., Chiang, R., Fahlgren, N., Kasschau, K. D., Chaves, D. A., Gu, W., Vasale, J. J., Duan, S., Conte, D., Luo, S., Schroth, G. P., Carrington, J. C., Bartel, D. P., and Mello, C. C. (2008). PRG-1 and 21U-RNAs interact to form the piRNA complex required for fertility in C. elegans. *Molecular Cell*.

Beanan, M. J. and Strome, S. (1992). Characterization of a germ-line proliferation mutation in C. elegans. *Development (Cambridge, England)*.

Blumenthal, T. and Gleason, K. S. (2003). Caenorhabditis elegans operons: form and function. *Nature Reviews Genetics*.

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature communications*.

Carlson, M. R. J., Zhang, B., Fang, Z., Mischel, P. S., Horvath, S., and Nelson, S. F. (2006). Gene connectivity, function, and sequence conservation: predictions from modular yeast co-expression networks. *BMC Genomics*.

Carmi, I., Kopczynski, J. B., and Meyer, B. J. (1998). The nuclear hormone receptor SEX-1 is an X-chromosome signal that determines nematode sex. *Nature*.

Chatr-Aryamontri, A., Oughtred, R., Boucher, L., Rust, J., Chang, C., Kolas, N. K., O'Donnell, L., Oster, S., Theesfeld, C., Sellam, A., Stark, C., Breitkreutz, B.-J., Dolinski, K., and Tyers, M. (2017). The BioGRID interaction database: 2017 update. *Nucleic acids research*.

Cho, H., Berger, B., and Peng, J. (2016). Compact Integration of Multi-Network Topology for Functional Analysis of Genes. *Cell systems*.

Choobdar, S., Ahsen, M. E., Crawford, J., Tomasoni, M., Fang, T., Lamparter, D., Lin, J., Hescott, B., Hu, X., Mercer, J., Natoli, T., Narayan, R., DREAM Module Identification Challenge Consortium, Subramanian, A., Zhang, J. D., Stolovitzky, G., Kutalik, Z., Lage, K., Slonim, D. K., Saez-Rodriguez, J., Cowen, L. J., Bergmann, S., and Marbach, D. (2019). Assessment of network module identification across complex diseases. *Nature Methods*.

Claycomb, J. M., Batista, P. J., Pang, K. M., Gu, W., Vasale, J. J., van Wolfswinkel, J. C., Chaves, D. A., Shirayama, M., Mitani, S., Ketting, R. F., Conte, D., and Mello, C. C. (2009). The Argonaute CSR-1 and its 22G-RNA cofactors are required for holocentric chromosome segregation. *Cell*.

Costanzo, M., VanderSluis, B., Koch, E. N., Baryshnikova, A., Pons, C., Tan, G., Wang, W., Usaj, M., Hanchard, J., Lee, S. D., Pelechano, V., Styles, E. B., Billmann, M., van Leeuwen, J., van Dyk, N., Lin, Z.-Y., Kuzmin, E., Nelson, J., Piotrowski, J. S., Srikumar, T., Bahr, S., Chen, Y., Deshpande, R., Kurat, C. F., Li, S. C., Li, Z., Usaj, M. M., Okada, H., Pascoe, N., San Luis, B.-J., Sharifpoor, S., Shuteriqi, E., Simpkins, S. W., Snider, J., Suresh, H. G., Tan, Y., Zhu, H., Malod-Dognin, N., Janjic, V., Przulj, N., Troyanskaya, O. G., Stagljar, I., Xia, T., Ohya, Y., Gingras, A.-C., Raught, B., Boutros, M., Steinmetz, L. M., Moore, C. L., Rosebrock, A. P., Caudy, A. A., Myers, C. L., Andrews, B., and Boone, C. (2016). A global genetic interaction network maps a wiring diagram of cellular function. *Science (New York, N.Y.)*.

Cowen, L., Ideker, T., Raphael, B. J., and Sharan, R. (2017). Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*.

Csardi, G. and Nepusz, T. (2006). The igraph software package for complex network research. *InterJournal, Complex Systems*.

Czech, B. and Hannon, G. J. (2011). Small RNA sorting: matchmaking for Argonautes. *Nature Publishing Group*.

Czech, B. and Hannon, G. J. (2016). One Loop to Rule Them All: The Ping-Pong Cycle and piRNA-Guided Silencing. *Trends in biochemical sciences*.

Czech, B., Preall, J. B., McGinn, J., and Hannon, G. J. (2013). A transcriptome-wide RNAi screen in the Drosophila ovary reveals factors of the germline piRNA pathway. *Molecular Cell*.

Dietterich, T. G. and Bakiri, G. (1994). Solving Multiclass Learning Problems via Error-Correcting Output Codes. *Journal of Artificial Intelligence Research*.

ENCODE Project Consortium, Dunham, I., Kundaje, A., Collins, P. J., Davis, C. A., Frietze, S., Landt, S. G., Lee, B.-K., Pauli, F., Sabo, P., Safi, A., Sanyal, A., Trinklein, N. D., Dong, X., Greven, M., Hoffman, M. M., Iyer, S., Kelllis, M., Kheradpour, P., Lassman, T., Merkel, A., Parker, S. C. J., Schlesinger, F., Thurman, R. E., Wilder, S. P., Wu, W., Yip, K. Y., Bernstein, B. E., Pazin, M. J., Lowdon, R. F., Dillon, L. A. L., Adams, L. B., Kelly, C. J., Zhang, J., Wexler, J. R., Green, E. D., Good, P. J., Feingold, E. A., Birney, E., Elinitski, L., Farnham, P. J., Hubbard, T. J., Myers, R. M., Starnatoyannopoulos, J. A., Tennebaum, S. A., White, K. P., Wrobel, J., Gunawardena, H. P., Kuiper, H. C., Maier, C. W., Xie, L., Giddings, M. C., Epstein, C. B., Shoresh, N., Ernst, J., Mikkelsen, T. S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M. J., Durham, T., Ku, M., Truong, T., Ward, L. D., Altshuler, R. C., Eaton, M. L., Kellis, M., Djebali, S., Dobin, A., Lassmann, T., Mortazavi, A., Lagarde, J., Lin, W., Xue, C., Marinov, G. K., Khatun, J., Zaleski, C., Rozowsky, J., Röder, M., Abdelhamid, R. F., Alioto, T., Antoshechkin, I., Baer, M. T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chen, X., Curado, J., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M. J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., Kingswood, C., Park, E., Preall, J. B., Presaud, K., Ribeca, P., Risk, B. A., Robyr, D., Ruan, X., Sammeth, M., Sandu, K. S., Schaeffer, L., See, L.-H., Skancke, J., Suzuki, A. M., Takahashi, H., Tilgner, H., Trout, D., Wang, H., Yu, Y., Hayashizaki, Y., Harrow, J., Reymond, A., Antonarakis, S. E., Hannon, G. J., Ruan, Y., Wold, B., Carninci, P., Gingeras, T. R., Rosenbloom, K. R., Sloan, C. A., Learned, K., Malladi, V. S., Wong, M. C., Barber, G. P., Cline, M. S., Dreszer, T. R., Heitner, S. G., Karolchik, D., Kent, W. J., Kirkup, V. M., Meyer, L. R., Long, J. C., Maddren, M., Raney, B. J., Furey, T. S., Song, L., Grasfeder, L. L., Giresi, P. G., Lee, B.-K., Battenhouse, A., Sheffield, N. C., Simon, J. M., Showers, K. A., London, D., Bhinge, A. A., Shestak, C., Schaner, M. R., Kim, S. K., Zhang, Z. Z., Mieczkowski, P. A., Mieczkowska, J. O., Liu, Z., McDaniell, R. M., Ni, Y., Rashid, N. U., Kim, M. J., Adar, S., Zhang, Z., Wang, T., Winter, D., Iyer, V. R., Lieb, J. D., Crawford, G. E., Li, G., Sandhu, K. S., Zheng, M., Wang, P., Luo, O. J., Shahab, A., Williams, B. A., Gertz, J., Reddy, T. E., Vielmetter, J., Partridge, E. C., Varley, K. E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K. M., Anaya, M., Cross, M. K., King, B., Muratet, M. A., Newberry, K. M., McCue, K., Nesmith, A. S., Fisher-Aylor, K. I., Pusey, B., DeSalvo, G., Parker, S. L., Balasubramanian, S., Davis, N. S., Mead-

ows, S. K., Eggleston, T., Gunter, C., Newberry, J. S., Levy, S. E., Absher, D. M., Wong, W. H., Blow, M. J., Visel, A., Pennachio, L. A., Elnitski, L., Margulies, E. H., Petrykowska, H. M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Chrast, J., and Dav... (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*.

Escalera, S., Pujol, O., and Radeva, P. (2010). On the Decoding Process in Ternary Error-Correcting Output Codes. *IEEE transactions on pattern analysis and machine intelligence*.

Fan, R.-E. and Lin, C.-J. (2007). A study on threshold selection for multi-label classification. *Department of Computer Science, National Taiwan University*.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *Science (New York, N.Y.)*.

Ghildiyal, M. and Zamore, P. D. (2009). Small silencing RNAs: an expanding universe. *Nature Publishing Group*.

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., and Bartel, D. P. (2007). MicroRNA targeting specificity in mammals: determinants beyond seed pairing. *Molecular Cell*.

Gu, W., Shirayama, M., Conte, D., Vasale, J., Batista, P. J., Claycomb, J. M., Moresco, J. J., Youngman, E. M., Keys, J., Stoltz, M. J., Chen, C.-C. G., Chaves, D. A., Duan, S., Kasschau, K. D., Fahlgren, N., Yates, J. R., Mitani, S., Carrington, J. C., and Mello, C. C. (2009). Distinct argonaute-mediated 22G-RNA pathways direct genome surveillance in the C. elegans germline. *Molecular Cell*.

Guiliano, D. B. and Blaxter, M. L. (2006). PLOS Genetics: Operon Conservation and the Evolution of trans-Splicing in the Phylum Nematoda. *PLoS Genet*.

Guzzi, P. H., Mina, M., Guerra, C., and Cannataro, M. (2012). Semantic similarity analysis of protein data: assessment with biological features and issues. *Briefings in bioinformatics*.

Helwak, A., Kudla, G., Dudnakova, T., and Tollervey, D. (2013). Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*.

Hishigaki, H., Nakai, K., Ono, T., Tanigami, A., and Takagi, T. (2001). Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*.

Hu, Z., Chang, Y.-C., Wang, Y., Huang, C.-L., Liu, Y., Tian, F., Granger, B., and DeLisi, C. (2013). VisANT 4.0: Integrative network platform to connect genes, drugs, diseases and therapies. *Nucleic acids research*.

Huang, X., Fejes-Toth, K., and Aravin, A. A. (2017). piRNA Biogenesis in Drosophila melanogaster. *Trends in Genetics*.

Hutvagner, G. and Simard, M. J. (2008). Argonaute proteins: key players in RNA silencing. *Nature Reviews: Molecular Cell Biology*.

Hwang, D., Rust, A. G., Ramsey, S., Smith, J. J., Leslie, D. M., Weston, A. D., de Atauri, P., Aitchison, J. D., Hood, L., Siegel, A. F., and Bolouri, H. (2005). A data integration methodology for systems biology. *Proceedings of the National Academy of Sciences of the United States of America*.

Ideker, T., Thorsson, V., Ranish, J. A., Christmas, R., Buhler, J., Eng, J. K., Bumgarner, R., Goodlett, D. R., Aebersold, R., and Hood, L. (2001). Integrated Genomic and Proteomic Analyses of a Systematically Perturbed Metabolic Network. *Science (New York, N.Y.)*.

Jeanquartier, F., Jean-Quartier, C., and Holzinger, A. (2015). Integrated web visualizations for protein-protein interaction databases. *BMC Bioinformatics*.

Jiang, M., Anderson, J., Gillespie, J., and Mayne, M. (2008). uShuffle: a useful tool for shuffling biological sequences while preserving the k-let counts. *BMC Bioinformatics*.

Kanehisa, M. and Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. *Nucleic acids research*.

Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., and Kasif, S. (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. *Proceedings of the National Academy of Sciences of the United States of America*.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at UCSC. *Genome research*.

Kotlyar, M., Pastrello, C., Sheahan, N., and Jurisica, I. (2016). Integrated interactions database: tissue-specific view of the human and model organism interactomes. *Nucleic acids research*.

Kumar, S., Stecher, G., Suleski, M., and Hedges, S. B. (2017). TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular biology and evolution*.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S. L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome biology*.

Lee, H.-C., Gu, W., Shirayama, M., Youngman, E., Conte, D., and Mello, C. C. (2012). C. elegans piRNAs mediate the genome-wide surveillance of germline transcripts. *Cell*.

Li, J. J., Huang, H., Bickel, P. J., and Brenner, S. E. (2014). Comparison of D. melanogaster and C. elegans developmental stages, tissues, and cells by modENCODE RNA-seq data. *Genome research*.

Lin, H.-T., Lin, C.-J., and Weng, R. C. (2007). A note on Platt's probabilistic outputs for support vector machines. *Machine Learning*.

Liu, M. and Thomas, P. D. (2019). GO functional similarity clustering depends on similarity measure, clustering method, and annotation completeness. *BMC Bioinformatics*.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*.

Lorenz, R., Bernhart, S. H., Höner Zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for molecular biology : AMB*.

Malone, C. D. and Hannon, G. J. (2009). Small RNAs as guardians of the genome. *Cell*.

Matsumoto, N., Nishimasu, H., Sakakibara, K., Nishida, K. M., Hirano, T., Ishitani, R., Siomi, H., Siomi, M. C., and Nureki, O. (2016). Crystal Structure of Silkworm PIWI-Clade Argonaute Siwi Bound to piRNA. *Cell*.

Mattei, E., Ausiello, G., Ferrè, F., and Helmer-Citterich, M. (2014). A novel approach to represent and compare RNA secondary structures. *Nucleic acids research*.

Meister, G. (2013). Argonaute proteins: functional insights and emerging roles. *Nature Publishing Group*.

Molla-Herman, A., Vallés, A. M., Ganem-Elbaz, C., Antoniewski, C., and Huynh, J.-R. (2015). tRNA processing defects induce replication stress and Chk2-dependent disruption of piRNA transcription. *The EMBO journal*.

Mostafavi, S., Ray, D., Warde-Farley, D., Grouios, C., and Morris, Q. (2008). GeneMANIA: a real-time multiple association network integration algorithm for predicting gene function. *Genome biology*.

Murali, T. M., Wu, C.-J., and Kasif, S. (2006). The art of gene function prediction. *Nature biotechnology*.

Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*.

Nelson, W., Zitnik, M., Wang, B., Leskovec, J., Goldenberg, A., and Sharan, R. (2019). To Embed or Not: Network Embedding as a Paradigm in Computational Biology. *Frontiers in genetics*.

Nora, E. P., Lajoie, B. R., Schulz, E. G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N. L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., and Heard, E. (2012). Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*.

Pellegrini, M., Marcotte, E. M., Thompson, M. J., Eisenberg, D., and Yeates, T. O. (1999). Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*.

Platt, J. C. (1999). Probabilistic Outputs for Support Vector Machines and Comparisons to Regularized Likelihood Methods. In *ADVANCES IN LARGE MARGIN CLASSIFIERS*.

Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. *Genome research*.

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., and Vandepoele, K. (2012). i-ADHoRe 3.0–fast and sensitive detection of genomic homology in extremely large data sets. *Nucleic acids research*.

Quinlan, A. R. and Hall, I. M. (2010). BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*.

Resnik, P. (1999). Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *J Artif Intell Res(JAIR)*.

Rüping, S. (2006). Robust Probabilistic Calibration. In *Machine Learning: ECML 2006*.

Sarkies, P., Selkirk, M. E., Jones, J. T., Blok, V., Boothby, T., Goldstein, B., Hanelt, B., Ardila-Garcia, A., Fast, N. M., Schiffer, P. M., Kraus, C., Taylor, M. J., Koutsovoulos, G., Blaxter, M. L., and Miska, E. A. (2015). Ancient and novel

small RNA pathways compensate for the loss of piRNAs in multiple independent nematode lineages. *PLoS biology*.

Schwikowski, B., Uetz, P., and Fields, S. (2000). A network of protein-protein interactions in yeast. *Nature biotechnology*.

Sechidis, K., Tsoumakas, G., and Vlahavas, I. (2011). On the stratification of multi-label data. *Machine Learning and Knowledge Discovery in Databases*.

Seth, M., Shirayama, M., Gu, W., Ishidate, T., Conte, D., and Mello, C. C. (2013). The C. elegans CSR-1 argonaute pathway counteracts epigenetic silencing to promote germline gene expression. *Developmental Cell*.

Seth, M., Shirayama, M., Tang, W., Shen, E.-Z., Tu, S., Lee, H.-C., Weng, Z., and Mello, C. C. (2018). The Coding Regions of Germline mRNAs Confer Sensitivity to Argonaute Regulation in C. elegans. *Cell reports*.

Shen, E.-Z., chen, h., Ozturk, A. R., Tu, S., Shirayama, M., Tang, W., Ding, Y.-H., Dai, S.-Y., Weng, Z., and Mello, C. C. (2018). Identification of piRNA Binding Sites Reveals the Argonaute Regulatory Landscape of the C. elegans Germline. *Cell*.

Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A., and Bartel, D. P. (2010). Expanding the microRNA targeting code: functional sites with centered pairing. *Molecular Cell*.

Shirayama, M., Seth, M., Lee, H.-C., Gu, W., Ishidate, T., Conte, D., and Mello, C. C. (2012). piRNAs initiate an epigenetic memory of nonself RNA in the C. elegans germline. *Cell*.

Simon, M., Sarkies, P., Ikegami, K., Doebley, A.-L., Goldstein, L. D., Mitchell, J., Sakaguchi, A., Miska, E. A., and Ahmed, S. (2014). Reduced insulin/IGF-1 signaling restores germ cell immortality to Caenorhabditis elegans Piwi mutants. *Cell reports*.

Siomi, H. and Siomi, M. C. (2009). On the road to reading the RNA-interference code. *Nature*.

Siomi, M. C., Sato, K., Pezic, D., and Aravin, A. A. (2011). PIWI-interacting small RNAs: the vanguard of genome defence. *Nature Reviews: Molecular Cell Biology*.

Steffen, M., Petti, A., Aach, J., D'haeseleer, P., and Church, G. (2002). Automated modelling of signal transduction networks. *BMC Bioinformatics*.

Svensson, V., da Veiga Beltrame, E., and Pachter, L. (2019). A curated database reveals trends in single-cell transcriptomics. *bioRxiv*.

Szklarczyk, D., Gable, A. L., Lyon, D., Junge, A., Wyder, S., Huerta-Cepas, J., Simonovic, M., Doncheva, N. T., Morris, J. H., Bork, P., Jensen, L. J., and Mering, C. v. (2019). STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic acids research*.

Szymański, P. and Kajdanowicz, T. (2017). A Network Perspective on Stratification of Multi-Label Data. In *Proceedings of the First International Workshop on Learning with Imbalanced Domains: Theory and Applications*.

Tabach, Y., Billi, A. C., Hayes, G. D., Newman, M. A., Zuk, O., Gabel, H., Kamath, R., Yacoby, K., Chapman, B., Garcia, S. M., Borowsky, M., Kim, J. K., and Ruvkun, G. (2013). Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature*.

Tang, W., Seth, M., Tu, S., Shen, E.-Z., Li, Q., Shirayama, M., Weng, Z., and Mello, C. C. (2018). A Sex Chromosome piRNA Promotes Robust Dosage Compensation and Sex Determination in C. elegans. *Developmental Cell*.

Thomson, T. and Lin, H. (2009). The biogenesis and function of PIWI proteins and piRNAs: progress and prospect. *Annual review of cell and developmental biology*.

Thurmond, J., Goodman, J. L., Strelets, V. B., Attrill, H., Gramates, L. S., Marygold, S. J., Matthews, B. B., Millburn, G., Antonazzo, G., Trovisco, V., Kaufman, T. C., Calvi, B. R., and FlyBase Consortium (2019). FlyBase 2.0: the next generation. *Nucleic acids research*.

Tian, D., Zhang, R., Zhang, Y., Zhu, X., and Ma, J. (2020). MOCHI enables discovery of heterogeneous interactome modules in 3D nucleome. *Genome research*.

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

Vanunu, O., Magger, O., Ruppin, E., Shlomi, T., and Sharan, R. (2010). Associating genes and protein complexes with disease via network propagation. *PLoS computational biology*.

Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction from protein-protein interaction networks. *Nature biotechnology*.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*.

Vourekas, A. and Mourelatos, Z. (2014). HITS-CLIP (CLIP-Seq) for mouse Piwi proteins. *Methods in Molecular Biology*.

Wang, D., Zhang, Y., Zhang, Z., Zhu, J., and Yu, J. (2010). KaKs Calculator 2.0: a toolkit incorporating gamma-series methods and sliding window strategies. *Genomics, proteomics & bioinformatics*.

Wang, W., Yoshikawa, M., Han, B. W., Izumi, N., Tomari, Y., Weng, Z., and Zamore, P. D. (2014). The initial uridine of primary piRNAs does not create the tenth adenine that Is the hallmark of secondary piRNAs. *Molecular Cell*.

Wedeles, C. J., Wu, M. Z., and Claycomb, J. M. (2013). Protection of germline gene expression by the C. elegans Argonaute CSR-1. *Developmental Cell*.

Wee, L. M., Flores-Jasso, C. F., Salomon, W. E., and Zamore, P. D. (2012). Argonaute divides its RNA guide into domains with distinct functions and RNA-binding properties. *Cell*.

Weick, E.-M. and Miska, E. A. (2014). piRNAs: from biogenesis to function. *Development (Cambridge, England)*.

Weiss, Y. and Freeman, W. T. (2001). Correctness of belief propagation in Gaussian graphical models of arbitrary topology. *Neural computation*.

Weng, C., Kosalka, J., Berkyurek, A. C., Stempor, P., Feng, X., Mao, H., Zeng, C., Li, W.-J., Yan, Y.-H., Dong, M.-Q., Morero, N. R., Zuliani, C., Barabas, O., Ahringer, J., Guang, S., and Miska, E. A. (2019). The USTC co-opts an ancient machinery to drive piRNA transcription in C. elegans. *Genes & Development*.

Witt, E., Benjamin, S., Svetec, N., and Zhao, L. (2019). Testis single-cell RNA-seq reveals the dynamics of de novo gene transcription and germline mutational bias in Drosophila. *eLife*.

Yamanaka, S. and Siomi, H. (2015). Misprocessed tRNA response targets piRNA clusters. *The EMBO journal*.

Yook, K., Harris, T. W., Bieri, T., Cabunoc, A., Chan, J., Chen, W. J., Davis, P., de la Cruz, N., Duong, A., Fang, R., Ganesan, U., Grove, C., Howe, K., Kadam, S., Kishore, R., Lee, R., Li, Y., Muller, H.-M., Nakamura, C., Nash, B., Ozersky, P., Paulini, M., Raciti, D., Rangarajan, A., Schindelman, G., Shi, X., Schwarz, E. M., Ann Tuli, M., Van Auken, K., Wang, D., Wang, X., Williams, G., Hodgkin, J.,

Berriman, M., Durbin, R., Kersey, P., Spieth, J., Stein, L., and Sternberg, P. W. (2012). WormBase 2012: more genomes, more data, new website. *Nucleic acids research*.

Yu, G., Li, F., Qin, Y., Bo, X., Wu, Y., and Wang, S. (2010). GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics*.

Zhang, C., Wang, J., Long, M., and Fan, C. (2013). gKaKs: the pipeline for genome-level Ka/Ks calculation. *Bioinformatics*.

Zhang, G., Tu, S., Yu, T., Zhang, X.-O., Parhad, S. S., Weng, Z., and Theurkauf, W. E. (2018). Co-dependent Assembly of Drosophila piRNA Precursor Complexes and piRNA Cluster Heterochromatin. *Cell reports*.

Zhang, L., Ward, J. D., Cheng, Z., and Dernburg, A. F. (2015). The auxin-inducible degradation (AID) system enables versatile conditional protein depletion in C. elegans. *Development (Cambridge, England)*.

Zhang, M.-L. and Zhou, Z.-H. (2014). A Review on Multi-Label Learning Algorithms. *IEEE Transactions on Knowledge and Data Engineering*.

Zhang, Y. and Schneider, J. (2011). Multi-Label Output Codes using Canonical Correlation Analysis. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*.

Zheng, Y., Szustakowski, J. D., Fortnow, L., Roberts, R. J., and Kasif, S. (2002). Computational identification of operons in microbial genomes. *Genome research*.

Zhu, X., Gerstein, M., and Snyder, M. (2007). Getting connected: analysis and principles of biological networks. *Genes & Development*.

Zitnik, M., Nguyen, F., Wang, B., Leskovec, J., Goldenberg, A., and Hoffman, M. M. (2019). Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*.

Zou, Q., Xie, S., Lin, Z., Wu, M., and Ju, Y. (2016). Finding the Best Classification Threshold in Imbalanced Classification. *Big Data Research*.

Curriculum Vitae