2018

Computational analyses of small silencing RNAs

BOSTON UNIVERSITY

GRADUATE SCHOOL OF ARTS AND SCIENCES

AND

COLLEGE OF ENGINEERING

Dissertation

COMPUTATIONAL ANALYSES OF SMALL SILENCING RNAS

by

YU FU

B.S., Ocean University of China, 2012

Submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

2018

Approved by

First Reader

Charles DeLisi, Ph.D.
Professor of Science and Engineering

Second Reader

Zhiping Weng, Ph.D.
Professor, Bioinformatics and Integrative Biology
University of Massachusetts Medical School

ACKNOWLEDGMENTS

This thesis would not have been possible without the incredible support of my family, for which I am forever grateful. I would like to thank my advisor Zhiping Weng for her unconditional support and scientific insights. I am greatly inspired by her passion for science. I would also like to thank Phillip Zamore, who not only taught me biology and writing, but also guided me to be an independent scientist. I would like to extend my gratitude to former and current Weng Lab members: Arjan van der Velde, Michael Purcaro, Shikui Tu, Wei Wang, Hao Chen, Xiao-Ou Zhang, Junko Tsuji, Jiali Zhuang, Eugenio Mattei, Jill Moore, Thom Vreven, Tyler Borrman for their valuable insights and technical support. I also have to thank my thesis committee members for their valuable feedback and mentorship over the past five years.

COMPUTATIONAL ANALYSES OF SMALL SILENCING RNAS

YU FU

Boston University Graduate School of Arts and Sciences

and College of Engineering, 2018

Major Professor: Zhiping Weng

Professor, Bioinformatics and Integrative Biology

UMass Medical School, Worcester MA

ABSTRACT

High-throughput sequencing is a powerful tool to study diverse aspects of biology and applies to genome, transcriptome, and small RNA profiling. Ever increasing sequencing throughput and more specialized sequencing assays demand more sophisticated bioinformatics approaches. In this thesis, I present 4 studies for which I developed computational methods to handle high-throughput sequencing data to gain insights into biology.

The first study describes the genome of High Five (Hi5) cells, originally derived from *Trichoplusia ni* eggs. The chromosome-level assembly (scaffold N50 = 14.2 Mb) contains 14,037 predicted protein-coding genes. Examination and curation of multiple gene families, pathways, and small RNA-producing loci reveal species- and order-specific features. The availability of the genome sequence, together with genome editing and single-cell cloning protocols, enables Hi5 cells as a new tool for studying small RNAs.

The second study focuses on just one type of piRNAs that are produced at the pachytene stage of mammalian spermatogenesis. Despite their abundance, pachytene piRNAs are poorly understood. I find that pachytene piRNAs cleave transcripts of protein-coding genes and further target transcripts from other pachytene piRNA loci. Subsequently, systematic investigation of piRNA targeting by integrating different types of sequencing data uncovers the piRNA targeting rule.

The third study describes computational procedures to map splicing branchpoints using high-throughput sequencing data. Screening >1.2 trillion RNA-seq reads determines >140,000 BPs for both human and mouse. Such branchpoints are compiled into BPDB (BranchPoint DataBase) to provide a comprehensive branchpoint catalog.

The final study combines novel experimental and computational procedures to handle PCR duplicates that are prevalent in high-throughput sequencing data. Incorporation of unique molecular identifiers (UMIs) to tag each read enables unambiguous identification of PCR duplicates. Both simulated and experimental datasets demonstrate that UMI incorporation increases the reproducibility of RNA-seq and small RNA-seq. Surveying 7 common variables in high-throughput sequencing reveals that the amount of starting material and sequencing depth, but not the number of PCR cycles, determine the PCR duplicate frequency. Finally, I show that removing PCR duplicates without UMIs leads to substantial bias into data analysis.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Ago ..... Argonaute

Aub ..... Aubergine

B. mori ..... *Bombyx mori*

BLAST ..... Basic Local Alignment Search Tool

BU ..... Boston University

BWT ..... Burrows-Wheeler transform

C. elegans: ..... *Caenorhabditis elegans*

CV ..... Coefficient of variation

D. melanogaster ..... *Drosophila melanogaster*

DNA ..... Deoxyribonucleic acid

dpc ..... Days post coitum

dpp ..... Days post partum

GEO ..... Gene Expression Omnibus

H3K9me3 ..... Trimethylation of lysine 9 on histone H3

H3K4me3 ..... Trimethylation of lysine 4 on histone H3

Hen1 ..... Hua Enhancer 1

miRNA ..... microRNA

mRNA ..... Messenger RNA

NCBI ..... National Center for Biotechnology Information

NIH ..... National Institutes of Health

piRNA ..... PIWI-interacting RNA

rRNA .....ribosomal RNA

siRNA .....Small interfering RNA

T. ni.....*Trichoplusia ni*

UCSC..... University of California, Santa Cruz

UMI ..... Unique molecular identifier

Chapter 1. General Introduction

1.1 Genome Assembly

Since the conception of the human genome project (International Human Genome Sequencing Consortium, 2001; Venter et al., 2001), genome sequences have become available for many species, such as mouse (Chinwalla et al., 2002) and fruit fly (Adams et al., 2000). Genome sequences are critical to analyses of high-throughput sequencing data, such as RNA-seq (Z. Wang, Gerstein, & Snyder, 2009), small RNA-seq (Lau, Lim, Weinstein, & Bartel, 2001), ChIP-seq (Park, 2009), and DNA-seq (Loman et al., 2012). Without them, none of the downstream analyses—such as transcript quantification, small RNA quantification, histone modification profiling, and SNP calling—would have been possible.

Previously the privilege of large genome consortia, DNA-seq has made its way into many labs. Advances in sequencing technologies and computation have made it much less costly to assemble large genomes. Genome assembly typically employs two approaches: overlap-layout-consensus (OLC), and de Bruijn Graph (DBG). Under the OLC paradigm, the assembly process begins by identifying read pairs that overlap well, and then stores the information into a graph where reads are represented by nodes, and overlapped read pairs by edges. OLC methods are useful for assembling long and accurate reads (e.g. sanger sequencing reads) to reconstruct the genome sequence. The DBG method starts by chopping reads into all possible substrings of length k (k-mers) and then stores such k-mers into a DBG, where each node is a k-mer and each edge represents two k-mers having an overlap of exactly length k-1. Subsequently, the genome

sequence can be constructed by traversing this graph. However, this process is highly dependent on sequencing error rates, as errors introduce wrong edges to the graph. It also depends on repeat structures in the genome: it cannot resolve repeat structures that are longer than *k*, since *k*-mers from the repeats with the same sequence collapse into the same set of nodes and edges. The advantage of DBG is that it avoids identifying overlaps among millions of reads and thus are more memory-efficient than OLC. DBG is more often used for short reads since it is typical to generate hundreds of millions of reads in one sequencing run.

1.2 Small silencing RNAs

1.2.1 miRNAs

microRNAs (miRNAs) are ~22 nucleotide (nt) RNAs that can be loaded into Ago proteins and identify mRNA targets via sequence complementarity. In animals, most miRNAs cause mRNA destabilization and translational repression (Ambros, 2004). miRNA genes are usually transcribed by RNA polymerase II (Pol II) to produce primary miRNA transcripts (pri-miRNA), which are capped at 5′ ends and polyadenylated at 3′ ends (He & Hannon, 2004). Subsequently, they are cleaved by Drosha into ~70 nt precursor miRNAs (pre-miRNAs) that typically have hairpin structures (Y. Lee et al., 2003). After pre-miRNAs are exported to the cytoplasm by Exportin-5, Dicer cleaves them into double-stranded RNAs (dsRNAs) with 2 nt overhang at 3′ ends (Hutvágner et al., 2001). Then one strand is loaded into RNA-induced silencing complex (RISC) and direct target repression.

Positions 2–7 are the most important regions (seeds) for miRNA targets (Lewis, Burge, & Bartel, 2005), so they are usually the most conserved region in miRNAs. In more recent literature, many other features—such as compensatory 3′ pairing (Friedman, Farh, Burge, & Bartel, 2008), centered pairing (Shin et al., 2010)—were found to contribute to the miRNA targeting specificity. miRNA target genes are often under the selective pressure to maintain the target sites, making conservation another feature to evaluate if a site can be targeted by a miRNA (Bartel, 2009).

1.2.2 siRNAs

Similar to miRNAs, siRNAs also derive from dsRNAs. dsRNAs are cleaved by Dcr-2 and loaded into Argonaute 2 (Ago2) to form an active RISC. Mature siRNAs are 2′-O-methylated at the 3′ ends by Hen1 in the RISC (Horwich et al., 2007, p. 1). siRNAs can be grouped into endogenous siRNAs (endo-siRNAs) and exogenous siRNAs (exo-siRNAs). In flies, endogenous siRNAs have at least 3 sources: transposon transcripts, cis-natural transcripts (cis-NATs), and hairpin RNAs (hpRNAs). Transposable elements (TEs), when active, can disrupt genes and other regulatory elements in the genome. TE-derived siRNAs can silence transposons (Ghildiyal et al., 2008). Another abundant source of siRNAs is the hairpin pathway (Katsutomo Okamura et al., 2008). Some genes produce hpRNAs that can be processed into ~21 nt siRNAs. Many of the siRNAs have confirmed targets (Katsutomo Okamura et al., 2008) and are under the selective pressure to coevolve with the targets in Drosophilids. A third source of siRNAs is cis-NATs. Such endogenous siRNAs come from mRNAs and are enriched in regions with overlapping mRNAs (e.g. two genes with convergent transcription with 3′ overlapping). These

endogenous siRNAs may be able to broadly regulate transcription (Katsutomo Okamura & Lai, 2008). siRNAs can also originate from exogenous RNAs, such as viruses, to protect the host (Tan & Yin, 2004). dsRNA intermediates produced by viruses can be fed into the siRNA biogenesis pathway to produce siRNAs, which guide Argonaute proteins to suppress viral transcription.

1.2.3 piRNAs

PIWI-interacting RNAs (piRNAs) are the most recently discovered type of small silencing RNAs. They are typically 23–31 nt, slightly longer than siRNAs and miRNAs. Even though piRNA sequences are not conserved, its presence in germline and its biogenesis pathway have been found conserved in a broad range of species, e.g., fly (Brennecke et al., 2007), mouse (A. A. Aravin, Hannon, & Brennecke, 2007), human (Ha et al., 2014), mosquito (Miesen, Girardi, & van Rij, 2015), zebrafish (Houwing et al., 2007), C. elegans (H.-C. Lee et al., 2012) and even hydra (Juliano et al., 2014). Critical to the piRNA pathway are Argonaute proteins, which can be divided into two clades: AGO and PIWI clades. Proteins from the former clades can load miRNAs and siRNAs, and proteins from the latter can load piRNAs. In fly germline, abundant piRNAs are derived from transposons and are loaded into PIWI clade proteins (Brennecke et al., 2007). Mutation of piRNA pathway components causes transposon derepression and genome instability. Thus, piRNAs suppress transposon activities by cleaving transposon transcripts the germline. This is critical, as transposon insertions in the germline likely pass deleterious mutations to the next generation. In fly, Rhino-Cutoff-Deadlock complexes are responsible for the transcription of piRNA clusters and suppress splicing

of these piRNA precursor transcripts (Mohn, Sienski, Handler, & Brennecke, 2014; Z. Zhang et al., 2014). However, orthologs of these genes only exist in a few Drosophilids that are closely related to Drosophila melanogaster, so it is likely that other species have different piRNA biogenesis pathways. Thus, it is important to examine a wider range of animals to better understand piRNA biogenesis. Three PIWI proteins—Piwi, Aubergine (Aub), and Argonaute 3 (Ago3)—can be found in the Drosophila germline and can load piRNAs. They are localized to different cellular components (Piwi almost exclusively localizes to the nucleus and other two to cytoplasm (Brennecke et al., 2007)). piRNAs in mammals have different behaviors. piRNAs in mouse testis can be categorized into prepachytene piRNAs, hybrid piRNAs and pachytene piRNAs, according to the stage of spermatogenesis (X. Z. Li et al., 2013). Most of these piRNA precursor transcripts are produced from non-coding genes. Pachytene piRNAs are particularly interesting to study: unlike fly piRNAs, most pachytene piRNAs can be uniquely mapped to the non-transposon portion of the genome. Thus, it is unclear if they cleavage non-transposon targets.

1.3 Transposons

Transposons are DNAs that are able to move in the genome. Such genomic "parasites" can insert into coding regions or regulatory elements, deleterious to the genome. In genomes of many species, they take up a considerable portion (e.g. approximately half of the human and mouse genomes, approximately one-third of the fly genome), reflecting their transposition history during evolution. The ability of transposons to change also make them a source of new genes. They are also of

therapeutic importance, as harnessing them may help alter genomic sequences to cure diseases caused by aberrant genomic sequences. For example, the piggyBac transposon, originally discovered in Trichoplusia ni, has been widely used for genetic manipulation (Bonin & Mann, 2004; Yusa, 2015). Transposons can be divided into two classes: retrotransposons and DNA transposon. Retrotransposons encode a reverse transcriptase and move via a "copy-and-paste" mechanism. Retrotransposons are transcribed from the genome to produce mRNA. With reverse transcriptase, the mRNA can be turned into DNAs and inserted into the genome. Retrotransposons can be further divided into transposons with long terminal repeats (LTR), long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). DNA transposons move via a "cut-and-paste". They encode transposases that can recognize certain genomic sites and cut out a transposon (or, sometimes, non-specifically, a piece of DNA) from the genome and insert it back to the genome. Because transposons are usually deleterious, organisms have evolved mechanisms to suppress their activities, transcriptionally or epigenetically. More than half of fly piRNAs derive from transposons and other repetitive elements to suppress transposons via cleaving transposon mRNAs (Brennecke et al., 2007). Epigenetically, transposons can be suppressed by DNA methylation (Alexei A. Aravin et al., 2008). In animals that do not possess the ability to methylate DNAs, histone modifications, such as H3K9me3, are used to shut down transposon transcription. Cells can also sense their suboptimal splicing and produce siRNAs against transposon transcripts (Dumesic et al., 2013).

Chapter. 2 The genome of the Hi5 germ cell line from *Trichoplusia ni*, an agricultural pest and novel model for small RNA biology

2.1 Introduction

Lepidoptera (moths and butterflies), one of the most species-rich orders of insects, comprises more than 170,000 known species, including many agricultural pests. One of the largest lepidopteran families, the Noctuidae diverged over 100 million years ago (mya) from the Bombycidae—best-known for the silkworm, *Bombyx mori* (Rainford, Hofreiter, Nicholson, & Mayhew, 2014). The Noctuidae family member cabbage looper (Trichoplusia ni) is a widely distributed generalist pest that feeds on cruciferous crops such as broccoli, cabbage, and cauliflower (Capinera, 2001). T. ni has evolved resistance to the chemical insecticide Dichlorodiphenyltrichloroethane (DDT; (McEwen & Hervey, 1956) and the biological insecticide Bacillus thuringiensis toxin (Janmaat & Myers, 2003), rendering pest control increasingly difficult. A molecular understanding of insecticide resistance requires a high-quality T. ni genome and transcriptome.

Hi5 cells derive from *T. ni* ovarian germ cells (Granados, Guoxun, Derksen, & McKenna, 1994). Hi5 cells are a mainstay of recombinant protein production using baculoviral vectors (Wickham, Davis, Granados, Shuler, & Wood, 1992) and hold promise for the commercial-scale production of recombinant adeno-associated virus for human gene therapy (Kotin, 2011). Hi5 cells produce abundant microRNAs (miRNAs) miRNAs, small interfering RNAs (siRNAs), and PIWI-interacting RNAs (S. Kawaoka et al., 2009) (piRNAs), making them one of just a few cell lines suitable for the study of all three types of animal small RNAs. The most diverse class of small RNAs, piRNAs

protect the genome of animal reproductive cells by silencing transposons (A. A. Aravin et al., 2007; Brennecke et al., 2007; Houwing et al., 2007; Shinpei Kawaoka et al., 2008; Vagin et al., 2006). The piRNA pathway has been extensively studied in the dipteran insect Drosophila melanogaster (fruit fly), but no piRNA-producing, cultured cell lines exist for dipteran germline cells. *T. ni* Hi5 cells grow rapidly without added hemolymph (Hink, 1970), are readily transfected, and—unlike *B. mori* BmN4 cells (Iwanaga et al., 2014), which also express germline piRNAs—remain homogeneously undifferentiated even after prolonged culture. In contrast to *B. mori*, no *T. ni* genome sequence is available, limiting the utility of Hi5 cells.

2.2 Methods

2.2.1 Genome assembly and annotation

Canu v1.3 (Koren et al., 2017) was used to assemble PacBio long reads into contigs, followed by two rounds of polishing using Quiver (https://github.com/PacificBiosciences/GenomicConsensus) and Pilon (Walker et al., 2014) to correct errors in the genome. The contigs were then assembled into chromosome-length scaffolds using Hi-C reads and LACHESIS (Burton et al., 2013). The mitochondrial genome was assembled separately using MITObim (six iterations, *D. melanogaster* mitochondrial genome as bait) (Hahn, Bachmann, & Chevreux, 2013).

The quality of the genome assembly was evaluated using BUSCO v3 (Simão, Waterhouse, Ioannidis, Kriventseva, & Zdobnov, 2015) with the arthropod profile and default parameters to identify universal single-copy orthologs. The genome quality was further evaluated using conserved gene sets: oxidative phosphorylation (OXPHOS) and

cytoplasmic ribosomal protein (CRP) genes. *B. mori* and *D. melanogaster* OXPHOS and CRP protein sequences were retrieved (Marygold et al., 2007; Porcelli, Barsanti, Pesole, & Caggese, 2007) and BLASTp was used to search for their *T. ni* homologs, which were further validated using InterPro (P. Jones et al., 2014; Mitchell et al., 2015). The *T. ni* genomes from male and female animals were also assembled separately using SOAPdenovo2 (kmer size 69) (Luo et al., 2012). These animal genome assemblies were compared to the *T. ni* genome assembled from Hi5 cells using QUAST (`–m 500`)(Gurevich, Saveliev, Vyahhi, & Tesler, 2013) and the nucmer and mummerplot (`––layout ––filter`) functions from MUMmer 3.23 (Kurtz et al., 2004, p. 201). The genomic variants were determined using HaplotypeCaller from GATK (McKenna et al., 2010, 2011; Van der Auwera et al., 2002) (`–ploidy 4 –genotyping_mode DISCOVERY`').

Annotation of the *T. ni* genome was performed in two steps: first masking repetitive sequences and then integrating multiple lines of evidence to predict gene models. RepeatModeler was used to produce repeat consensus sequences for the newly assembled genome. RepeatMasker (`–s –e ncbi`) was used to mask repetitive regions. 8S, 18S, 28S rRNA genes were predicted using RNAmmer (Lagesen et al., 2007), and 5.8S rRNA genes were predicted using Barrnap. Augustus v3.2.2 (Stanke, Tzvetkova, & Morgenstern, 2006) and SNAP (Korf, 2004) were used to computationally predicted gene models. Predicted gene models were compiled by running six iterations of MAKER (Campbell, Holt, Moore, & Yandell, 2014), aided with homology evidence of well annotated genes (UniProtKB/Swiss-Prot and Ensembl) and of transcripts from related

species, such as *B. mori* (Suetsugu et al., 2013) and *D. melanogaster* (Attrill et al., 2016). BLAST2GO (Conesa et al., 2005) was used to integrate results from BLAST, and InterPro (Mitchell et al., 2015) to assign GO terms to each gene. MITOS (Bernt et al., 2013) web server was used to predict mitochondrial genes. Genes of interest, such as small RNA pathway genes were manually curated in webApollo (E. Lee et al., 2013). Telomeres were searched by matching multiple variants of typical telomere sequences found in other species, such as $(TTAGG)_{200}$ (Robertson & Gordon, 2006) to the *T. ni* genome using BLASTn with the option '`-dust no`' and hits longer than 100 nt were kept. The genomic coordinates of these hits were extended by 10 kb to obtain the subtelomeric region.

2.2.2 Orthology

The predicted proteomes from 21 species (Appendix A) were compared to place genes into ortholog groups, using OrthoMCL (L. Li, Stoeckert, & Roos, 2003) with default parameters. MUSCLE v3.8.31 (Edgar, 2004) was used for strict 1:1:1 orthologs (n = 381) to produce sequence alignments. Conserved blocks (66,044 amino acids in total) of these alignments were extracted using Gblocks v0.91b (Castresana, 2000) with default parameters, and fed into PhyML 3.0 (Guindon et al., 2010) (maximum likelihood, bootstrap value set to 1000) to calculate a phylogenetic tree. The human and mouse predicted proteomes were used as an outgroup to root the tree. The tree was viewed using FigTree (http://tree.bio.ed.ac.uk/software/figtree/) and iTOL (Letunic & Bork, 2016).

2.2.3 An expectation-maximization algorithm to determine reads mapping to multiple loci

piRNA reads often map to repetitive regions in the genome, making it difficult to assign them to a genomic position. The same applies to other sequencing reads when they are not sufficiently long or originate from a repetitive region. To tackle this problem, I designed and implemented an expectation-maximization algorithm to determine the genomic sources of reads mapped to multiple loci (multimappers). Small RNA reads were first mapped to the genome as described (Han, Wang, Zamore, & Weng, 2014). Then piRNA abundance was calculated in each 5 kb genomic windows. For each window, uniquely mapped reads and multimappers are quantified by assigning reads using an expectation-maximization algorithm. Briefly, each window had the same initial weight. The weight was used to linearly apportion multimappers. During the expectation (E) step, uniquely mapped reads were unambiguously assigned to genomic windows; multimappers were apportioned to the genomic windows they mapped to, according to the weights of these windows. At the maximization (M) step, window weights were updated to reflect the number of reads (uniquely mapped reads plus multimappers) each window contained from the E step. The E and M steps were run iteratively until the Manhattan distance between two consecutive iterations was smaller than 0.1% of the total number of reads. Although this algorithm was used at 5 kb resolution, it can readily be generalized to determine the mapping positions at single nucleotide resolution.

2.2.4 Sex determination and sex chromosomes

Since it was not known if T. ni females had ZW or only Z, whole genomes of T. ni males and females were separately sequenced. DNA-seq reads were mapped to the

genome assembly. Reads with poor mapping qualities (MAPQ <20) were removed to avoid the ambiguity. Then reads for each contig were quantified and further normalized by median coverage. The cutoff for coverage ratios (male:female ratios, M:F ratios) was empirically determined: M:F ratio >1.5 for Z-linked contigs and M:F ratio < 0.5 for W-linked contigs. Lepidopteran masc genes were obtained from Lepbase (Challis, Kumar, Dasmahapatra, Jiggins, & Blaxter, 2016). Z/AA ratio was calculated according to (Gu, Walters, & Knipple, 2017). Briefly, direct comparisons of gene expressions from different chromosomes are not statically reliable. To address this issue and provide a confidence interval, I used a bootstrap method. Certain number of genes were sampled from autosomes, Z-linked and W-linked contigs and the median ratios of expression levels were calculated. To produce confidence intervals for each ratio, this procedure was performed 10,000 times.

2.2.4 Gene families for detoxification and chemoreception

Multiple methods were used to curate genes related to detoxification and chemoreception. Seed alignments from Pfam (Finn et al., 2016) were obtained and hmmbuild was used to build HMM profiles of cytochrome P450 (P450), amino- and carboxy-termini of glutathione-*S*-transferase (GST), carboxylesterase (COE), ATP-binding cassette transporter (ABCs), olfactory receptor (OR), gustatory receptor (GR), ionotropic receptor (IR), and odorant binding (OBP) proteins (see (Y. Fu, Yang, et al., 2018)). These HMM profiles were then used to search for gene models in the *T. ni* genome and predicted proteome (hmmsearch, e-value cutoff: 1×10^{-5}). Reference sequences of P450, GST, COE, ABC, OR, GR, IR, OBP, and juvenile hormone pathway

genes were retrieved from the literature (Ai et al., 2011; Benton, Vannice, Gomez-Diaz, & Vosshall, 2009; Croset et al., 2010; Dermauw & Van Leeuwen, 2014; Gong, Zhang, Zhao, Xia, & Xiang, 2009; Goodman & Granger, 2005; Hekmat-Scafe, Scafe, McKinney, & Tanouye, 2002; Liu et al., 2011; van Schooten, Jiggins, Briscoe, & Papa, 2016; Wanner & Robertson, 2008; Xavier Bellés, David Martín, & Piulachs, 2005; Q. Yu et al., 2008; Q.-Y. Yu, Lu, Li, Xiang, & Zhang, 2009), and were aligned to the *T. ni* genome using tBLASTx (Altschul, Gish, Miller, Myers, & Lipman, 1990) and Exonerate (Slater & Birney, 2005) to search for potential homologs. Hits were manually inspected to ensure compatibility with RNA-seq data, predicted gene models, known protein domains using CDD (Marchler-Bauer et al., 2015) and homologs from other species. P450 genes were submitted to Dr. David Nelson (Nelson, 2009) for nomenclature and classification.

To determine the phylogeny of these gene families, the putative protein sequences from *T. ni* and *B. mori* genomes were aligned using MUSCLE (Edgar, 2004). The multiple sequence alignments were subsequently trimmed using TrimAl (Capella-Gutiérrez, Silla-Martínez, & Gabaldón, 2009) (with the option `-automated1`). Phylogenetic analysis was performed using PhyML 3.0 (Guindon et al., 2010) (with parameters: `-q --datatype aa --run_id 0 --no_memory_check -b -2`). Phylogenetic trees were visualized using FigTree (http://tree.bio.ed.ac.uk/software/figtree/).

To curate opsin genes, opsin mRNA and peptide sequences from other species (Feuda, Marlétaz, Bentley, & Holland, 2016; Futahashi et al., 2015) were used as homology evidence to search for counterparts in the *T. ni* genome. To discriminate opsin

genes from other G-protein-coupled receptors, it is required that the top hit in the NCBI non-redundant database and UniProt were opsins.

2.2.5 miRNA and siRNA analysis

mirDeep2 (Friedländer et al., 2008; Friedländer, Mackowiak, Li, Chen, & Rajewsky, 2012) was used to predict miRNA genes. Predicted miRNA hairpins were required to have homology (exact seed matches and BLASTn e-value < 1×10^{-5}) to known miRNAs or miRDeep2 scores ≥10. miRNAs were named according to exact seed matches and high sequence identities with known miRNA hairpins. To determine the conservation status of *T. ni* miRNAs, putative *T. ni* miRNAs were compared with annotated miRNAs from *A. aegypti*, *A. mellifera*, *B. mori*, *D. melanogaster*, *H. sapiens*, *M. musculus*, *M. sexta*, *P. xylostella*, and *T. castaneum*. Conserved miRNAs were required to have homologous miRNAs beyond Lepidoptera; Lepidoptera-specific miRNAs were required to be conserved in lepidopterans; *T. ni*-species miRNAs are those without homologs in other species.

To compare siRNA abundance in oxidized and unoxidized small RNA-seq libraries, siRNA read counts were normalized to piRNA cluster-mapping reads (piRNA cluster read counts had >0.98 Pearson correlation coefficients between oxidized and unoxidized libraries in all cases.) piRNA degradation products can be 20–22 nt long, so potential siRNA species that were prefixes of piRNAs (23–35 nt) were removed.

To detect viral transcripts in *T. ni*, viral protein sequences were retrieved from NCBI (http://www.ncbi.nlm.nih.gov/genome/viruses/) and used to perform tBLASTn to map to the *T. ni* genome and to the transcriptomes of Hi5 cells and five *T. ni* tissues. Hits

were filtered requiring percent identity ≥ 0.80, e-val ≤ 1×10^{-20}, and alignment length ≥100. To identify virus-mapping small RNAs, all small RNA-seq reads were mapped to the identified viral transcripts. Candidate genomic hairpins were defined according to (Katsutomo Okamura et al., 2008). And Candidate *cis*-NATs were defined according to (Ghildiyal et al., 2008).

2.2.6 piRNA analysis

To quantify piRNAs from each piRNA locus, the ppm and rpkm values were used (normalized to the total number of uniquely mapped reads). For analyses involving all mapped reads (uniquely mapped reads and multimappers), reads were apportioned by the number of times that they were mapped to the genome (i.e. if one read maps to two genomic positions, each position gets half a read). To make piRNA loci comparable across tissues, piRNA loci from ovary, testis, female and male thorax, and Hi5 cells were merged. For the comparison between female and male thoraces, the cluster on tig00001980 was removed as this contig is likely to be a mis-assembly. As for defining sex-linked contigs, M:F ratios were calculated and the same thresholds were used to determine whether a piRNA cluster was sex-linked. A piRNA locus was considered to be differentially expressed if the ratio between the two tissues was >2 or <0.5 and FDR <0.1 (after t-test). Splice sites were deemed to be supported by RNA-seq data when supported by at least one data set. AUGUSTUS (Stanke et al., 2006) was used with the model trained for *T. ni* genome-wide gene prediction, to predict gene models and their splice sites in *T. ni* piRNA clusters.

2.3 Results

2.3.1 Genome assembly and quality assessment

To assemble the *T. ni* genome, we used a strategy integrating long and short reads (Figure 2.1). We first utilized the PacBio long reads (46.4× coverage with reads longer than 5 kb) to obtain a high-quality contig set (1,976 contigs; contig N50 = 621.9 kb). The assembly was further polished by the same set of long reads and additional short reads. This assembly is already more contiguous than many published insect genome assemblies (e.g. contig N50 = 50.7 kb for the monarch butterfly (Zhan, Merlin, Boore, & Reppert, 2011), contig N50 = 10.0 kb for the diamondback moth (You et al., 2013)). However, since many transposons and piRNAs map to repetitive regions in the genome, a chromosome-level assembly is desired. Thus, Hi-C reads were used to further join the contigs into 1,031 scaffolds (N50 = 14.2 Mb). We found that more than 90% of the bases were assigned to one of the 28 major scaffolds. Meanwhile, karyotyping of Hi5 cells indicated 28 chromosomes, corresponding well to the results in our karyotyping experiments. Thus, we conclude that *T. ni* has 28 chromosomes and that we have a chromosome-level genome assembly.

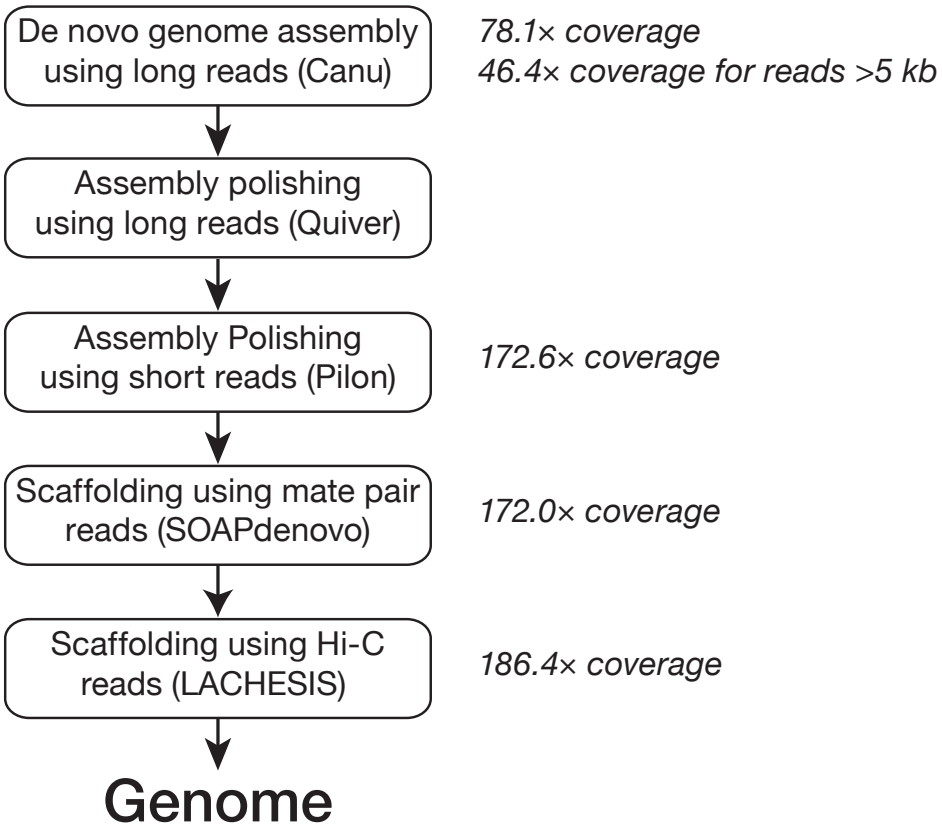

Figure 2.1. The genome assembly workflow. Each rounded rectangle indicates one step during the assembly, with the tools indicated in the parentheses. Genome coverage of sequencing data is indicated on the right.

Next, to evaluate the completeness of this genome assembly, we first used the Benchmark of Universal Single-Copy Orthologs (BUSCO v3; Arthropoda data set as the reference) (Simão et al., 2015) on multiple assemblies (*T.ni*, *B. mori*, *D. plexippus*, *P. xylostella*, *D. melanogaster*, *T. castaneum*, and *A. mellifera*). Our *T. ni* genome assembly captures 97.5% of the orthologs defined by BUSCO, better than the silkworm (95.5%) and monarch butterfly (97.0%) (Table 2.1). As a further test of the assembly quality, we searched for highly conserved genes encoding ribosomal proteins and genes belonging to

the nuclear oxidative phosphorylation pathway. Orthologs of all these genes can be identified, indicating that this genome assembly is highly complete.

Species	Complete	Fragmented	Missing
T. ni (cabbage looper)	97.5%	0.4%	2.2%
B. mori (silkworm)	95.5%	2.1%	2.5%
D. plexippus (monarch butterfly)	97.0%	1.9%	1.1%
P. xylostella (diamondback moth)	87.8%	2.7%	9.5%
D. melanogaster (fruit fly)	99.7%	0.2%	0.1%
T. castaneum (red flour beetle)	99.3%	0.5%	0.2%
A. mellifera (western honney bee)	97.8%	1.3%	0.9%

Table 2.1. BUSCOs found in genome assemblies of multiple species, including *T. ni*. The total number of BUSCO groups is 1,658.

2.3.2 Genome annotation

Knowing the genome sequence is the first step towards establishing Hi5 cells as a model for small RNA studies. Next, we de novo identified repeat consensus sequences for this genome using RepeatModeler and revealed 458 repeat families, including 44 DNA, 84 LINE, 14 LTR, 25 RC, and 26 SINE transposons. We then used RepeatMasker to identify and mask the genome. (This is a critical step before gene annotation, because without masking genomic repeats, automated gene annotation pipeline will produce inaccurate results, marking transposons as genes.) In total, 20.5% of the genome was masked as repetitive. Next, we annotated the *T. ni* genome using MAKER (Cantarel et al., 2008). To aid the identification of gene models, we used multiple sources of evidence (Figure 2.2): *T. ni* transcriptomes (assembled from RNA-seq data from multiple tissues using Trinity (Haas et al., 2013)), protein sequences from UniProtKB/Swiss-Prot, transcripts from related species (*Drosophila melanogaster*, and *Bombyx mori*). In total, 14,034 protein-

coding genes were annotated in the genome, similar to many other lepidopteran genomes (Challis et al., 2016).

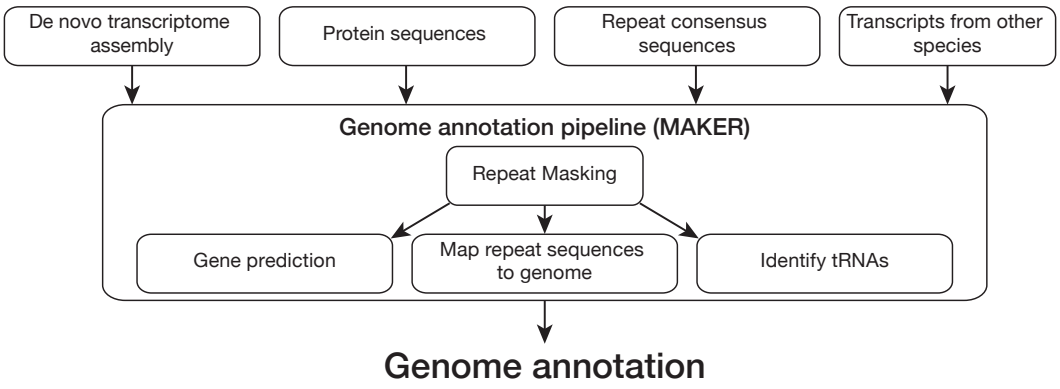

Figure 2.2. Genome annotation workflow.

Next, the homology of the predicted *T. ni* genes was determined by orthology analysis with 20 other species: *Acyrthosiphon pisum*, *Aedes aegypti*, *Anopheles gambiae*, *Apis mellifera*, *Atta cephalotes*, *Bombyx mori*, *Danaus plexippus*, *Drosophila melanogaster*, *Drosophila pseudobscura*, *Harpegnathos saltator*, *Homo sapiens*, *Linepithema humile*, *Mus musculus*, *Nasonia vitripennis*, *Pediculus humanus humanus*, *Plutella xylostella*, *Pogonomyrmex barbatus*, *Tetranychus urticae*, *Tribolium castaneum*, *Dendroctonus ponderosae* (see Appendix A for details). These species include common insect orders (Lepidoptera, Diptera, Coleoptera, and Hymenoptera). Proteomes of non-insect arthropods, and two mammals were also incorporated to serve as outgroups. OrthoMCL(L. Li et al., 2003) was used to assigned orthology groups. Using the numbers of genes and species in each orthology groups, orthology groups can be categorized. Genes in the 1:1:1 group are present in all species as just one copy (one absence or one

duplication in one species was allowed to alleviate the bias caused by unannotated genes). N:N:N orthologs are present in all species and have variable copy numbers (absence in one genome or two genomes was allowed). Lepidoptera-specific genes are those annotated in three or more lepidopteran genomes examined; Hymenoptera-specific genes are those in one or more wasp or bee genomes and one or more ant genomes. Coleoptera-specific genes are required to be present in both coleopteran genomes; Diptera-specific genes are those annotated in at least one fly genome and one mosquito genome. Insect-specific genes indicates other genes in insects, but their orthologs are not found in human and mouse. Mammal-specific genes are present in both mammalian genomes. 'Patchy' genes refer to those that are present in both arthropods and mammals but many species have lost them. SSD refers to those multi-copy genes that do not have orthologs in other species. SSS refers to those single-copy genes that do not have orthologs in other species. The orthology assignment is shown in Figure 2.3.

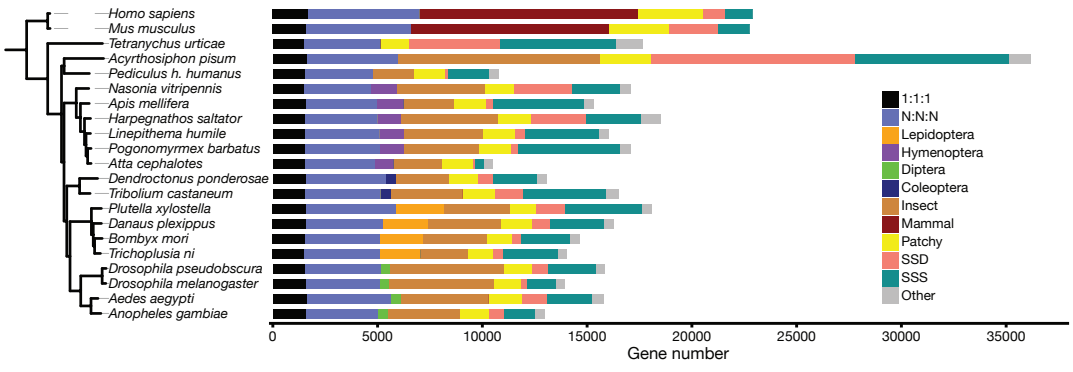

Figure 2.3. Orthology groups and phylogenetic tree of 21 species.

OrthoMCL defines 30,448 orthology groups using proteomes from these 21 species. There are 2,287 Lepidoptera-specific orthology groups, whereas Diptera,

Coleoptera, and Hymenoptera only has 404, 371, and 1,344, respectively, suggesting lepidopterans acquired new genes to adapt to their unique lifestyle. The *T. ni* genome further contains ~3,000 orphan genes (and, by definition, they do not have orthologs in other 20 species). About 450 of these orphan genes exist as two or more copies, likely due to recent gene duplication events. Some of these orphan genes might be false positives from gene predictors, as manual inspection revealed poor RNA-seq signals for them.

2.3.3 Genomic features

Basic genomic features, including transposons, centromeres, telomeres, and GC contents, are unique to each species and can hint at the quality of a genome assembly and other interesting genomic features. Approximately 20.5% of the *T. ni* genome assembly is repeats. A comparison with other assembled lepidopteran genomes shows that it fits well with the trend of repeat content vs genome size (Figure 2.4). One of the most notable transposons is a DNA transposon called piggyBac, for at least two reasons: it was originally discovered in a *T. ni* cell line due to its high level of activity (Fraser, Smith, & Summers, 1983); it has the potential to be a gene therapy vector due to its ability to transpose effectively in multiple species (Bonin & Mann, 2004; Lobo, Li, & Fraser, 1999; W. Wang et al., 2008; Yusa, 2015). In total, 262 copies of piggyBac transposons exist in the *T. ni* genome assembly, with a very low family divergence rate (0.17%; see Methods), substantially lower than other transposon families. Interestingly, 27% of these piggyBac copies exist only in the Hi5 cells, and these have even lower divergence rate than the

remaining piggyBac copies (0.22% vs 0.04%), indicating recent incorporation of piggyBac transposons into the *T. ni* genome and their highly effective transposition.

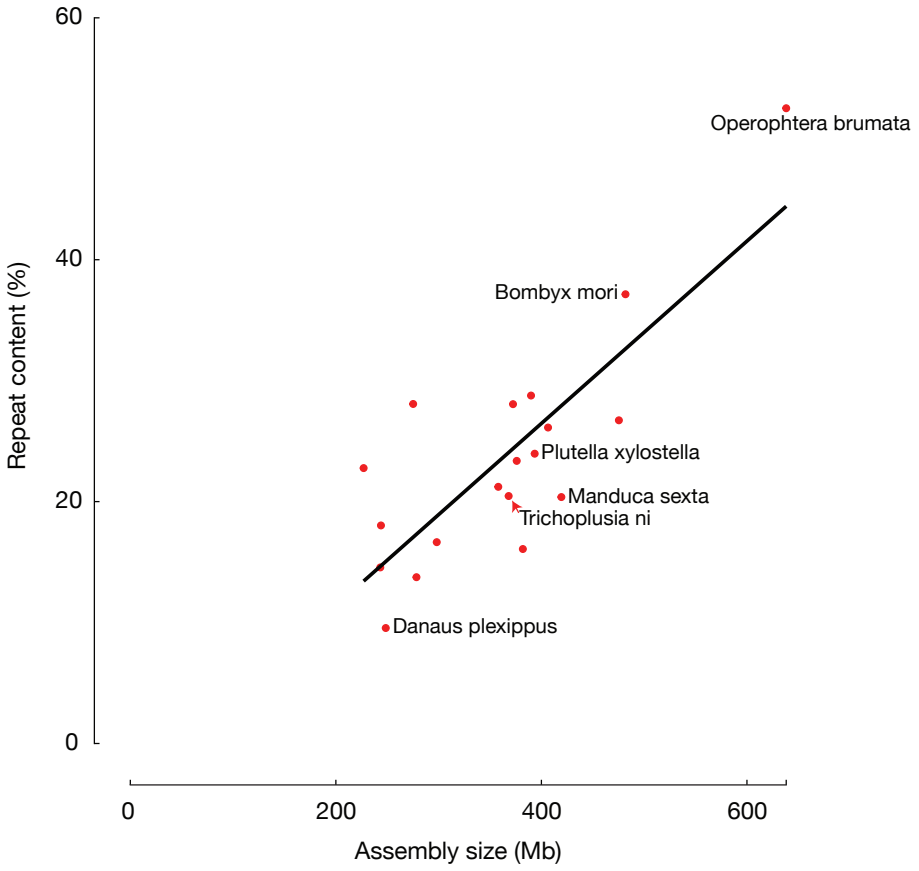

Figure 2.4. Repeat contents vs genome assembly size in lepidopteran genomes. Data for species other than *T. ni* were retrieved from Lepbase.

Next, I characterized telomeres and centromeres in the *T. ni* genome. Different species have different telomeric structures: human telomeres have (TTAGGG)n repeats (Morin, 1989); the fruit fly genome uses three telomere-specific retrotransposons (HeT-A, TART and Tahre) (Adams et al., 2000); the silkworm genome has (TTAGG)n repeats and retrotransposons (TRAS and SART) (Fujiwara, Osanai, Matsumoto, & Kojima, 2005). To determine the telomeric repeats in *T. ni*, I used Tandem Repeats Finder

(Benson, 1999) to search for simple repeats longer than 100 nt, and found 40 such repeats matching the pentanucleotide repeats (TTAGG)n, all of which are at or near the contig boundaries, indicating that the *T. ni* genome assembly captures sequences of many telomeres (See Figure 2.5 for an example). More than half of the sequences flanking (TTAGG)n repeats are transposons and approximately half of these transposons are homologous to TRAS and SART transposons in the silkworm genome (Figure 2.5), indicating that *T. ni* has a *B. mori*-like telomeres and subtelomeres (Fujiwara et al., 2005). Unlike telomeres, which have identifiable repeats or transposons, centromeres have different sequences even in closely related species, such as 14 yeast species (Varoquaux et al., 2015). Thus, instead of searching for the centromeric sequences, we searched for a gene related to centromeres: the centromeric histone H3 variant (CenH3). CenH3 was proposed to associate with monocentricity of chromosomes (a single centromere for the entire chromosome) and its loss in some species during evolution results in holocentricity (lack of coherent centromere) (Drinnenberg, deYoung, Henikoff, & Malik, 2014, p. 3). Searching for homologs of CenH3 reveals no CenH3 counterpart in *T. ni*. Thus, similar to other lepidopterans, *T. ni* chromosomes are holocentric.

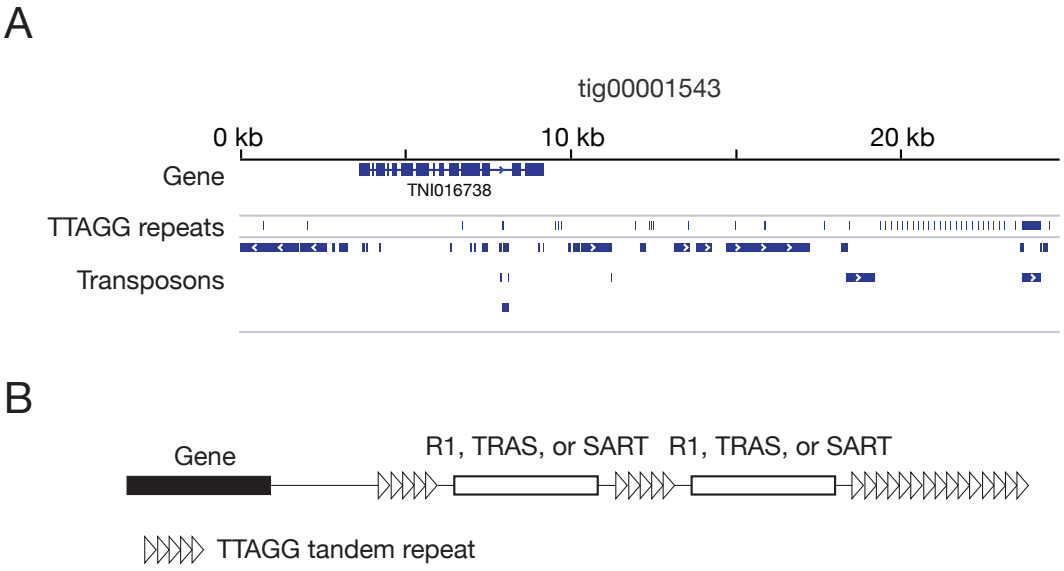

Figure 2.5. *T. ni* telomeres. (A) An example of *T. ni* telomeres on a contig (tig00001543). Three tracks show positions of the last gene on this contig, (TTAGG)n and identified transposons, respectively. (B) A schematic of *T. ni* telomere.

The GC content of the *T. ni* genome is then characterized. The *T. ni* genome has a GC content of 35.6%, close to that of *B. mori*. The observed/expected CpG ratios among these species can be categorized into three groups: honeybee shows high CpG ratios in the protein-coding genes and a bimodal distribution in the genome; fly shows low CpG ratios; other species, including *T. ni*, have similar CpG ratio distribution (Figure 2.6). This corresponds to the presence of DNMTs: the honeybee genome has two DNMTs (DNMT1 and DNMT3) whereas the fruit fly genome has neither. In contrast, other genomes inspected all have one DNMT (DNMT 1).

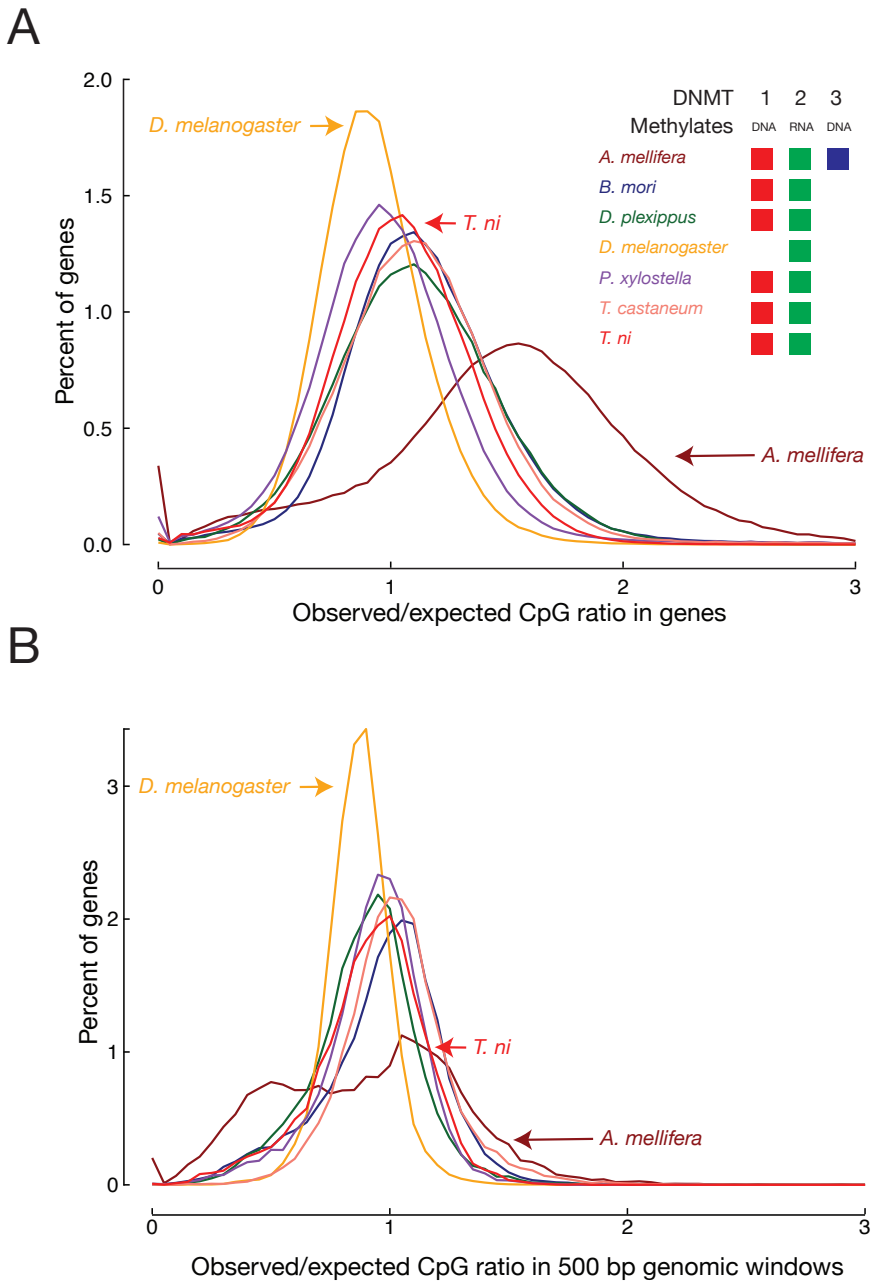

Figure 2.6. Observed/expected CpG ratios in genes and genomic windows in 7 species: *T. ni*, *T. castaneum*, *P. xylostella*, *D. melanogaster*, *D. plexippus*, *B. mori*, and *A. mellifera*.

2.3.4 Sex determination

Understanding the *T. ni* sex-determination pathway holds promise for engineering sterile animals for pest management. ZW and ZO chromosome systems determine sex in lepidopterans: males are ZZ and females are either ZW or ZO (Traut, Sahara, & Marec, 2008). To determine which system *T. ni* uses and to identify which contigs belong to the sex chromosomes, we sequenced genomic DNA from male and female pupae and calculated the male:female coverage ratio for each contig. We found that 175 presumably Z-linked contigs (20.0 Mb) had approximately twice the coverage in male compared to female DNA (median male:female ratio = 1.92; Figure 2.7A). Another 276 contigs (11.1 Mb) had low coverage in males (median male:female ratio = 0.111), suggesting they are W-linked. We conclude that sex is determined in *T. ni* by a ZW system in which males are homogametic (ZZ) and females are heterogametic (ZW).

For some lepidopteran species, dosage compensation has been reported to equalize Z-linked transcript abundance between ZW females and ZZ males in the soma, while other species show higher expression of Z-linked genes in males (Gu et al., 2017; Walters & Hardcastle, 2011). In the soma, *T. ni* compensates for Z chromosome dosage: transcripts from Z-linked genes are approximately equal in male and female thoraces (Z ≈ ZZ, Figure 2.7B). In theory, somatic dosage compensation could reflect increased transcription of the single female Z chromosome, reduced transcription of both male Z chromosomes, or silencing of one of the two male Z chromosomes.

To distinguish among these possibilities, we compared the abundance of Z-linked and autosomal transcripts (Z/AA in female and ZZ/AA in male). Z-linked transcripts in

the male thorax are expressed at lower levels than autosomal transcripts, but not as low as half (ZZ ≈ 70% AA). These data support a dosage compensation mechanism that decreases transcription from each Z chromosome in the T. ni male soma, but does not fully equalize Z-linked transcript levels between the sexes (Z ≈ ZZ ≈ 70% AA). In contrast, *T. ni* lacks germline dosage compensation: in the ovary, Z-linked transcript abundance is half that of autosomal transcripts (Z ≈ 50% AA), whereas in testis, Z-linked and autosomal transcripts have equal abundance (ZZ ≈ AA). We conclude that *T. ni*, like *B. mori* (Walters & Hardcastle, 2011), *Cydia pomonella* (Gu et al., 2017), and Heliconius butterflies (Walters & Hardcastle, 2011), compensates for Z chromosome dosage in the soma by reducing gene expression in males, but does not decrease Z-linked gene expression in germline tissues.

Little is known about lepidopteran W chromosomes. The W chromosome is not included in the genome assembly of Manduca sexta (Kanost et al., 2016) or *B. mori* (The International Silkworm Genome, 2008), and earlier efforts to assemble the silkworm W resulted in fragmented sequences containing transposons (Abe et al., 2005, 2008; Shinpei Kawaoka et al., 2011). The monarch genome scaffold continuity (N50 = 0.207 Mb versus N50 = 14.2 Mb for *T. ni*; (Zhan et al., 2011)) is insufficient to permit assembly of a W chromosome. Our genome assembly includes the 2.92 Mb T. ni W chromosome comprising 32 contigs (contig N50=101 kb). In *T. ni*, W-linked contigs have higher repeat content, lower gene density, and lower transcriptional activity than autosomal or Z-linked contigs (Figure 2.7B). Other lepidopteran W chromosomes are similarly

enriched in repeats and depleted of genes (Abe et al., 2005; Fuková, Nguyen, & Marec, 2005; Traut et al., 2008).

A search for *T. ni* genes that are homologous to insect sex determination pathway genes detected *doublesex* (*dsx*), *masculinizer* (*masc*), *vitellogenin*, *transformer 2*, *intersex*, *sex lethal*, *ovarian tumor*, *ovo*, and *sans fille*. *T. ni* males produce a four-exon isoform of dsx, while females generate a six-exon dsx isoform. The Lepidoptera-specific gene *masc* encodes a CCCH zinc finger protein. *masc* is associated with the expression of the sex-specific isoforms of *dsx* in lepidopterans, including silkworm (Katsuma, Sugano, Kiuchi, & Shimada, 2015). As in *B. mori*, T. ni *masc* lies next to the *scap* gene, supporting our annotation of T. ni masc. Lepidopteran *masc* genes are rapidly diverging and have low sequence identity with one another (30.1%). Figure 2.7C shows the multiple sequence alignment of the CCCH zinc finger domain of Masc proteins from several lepidopteran species.

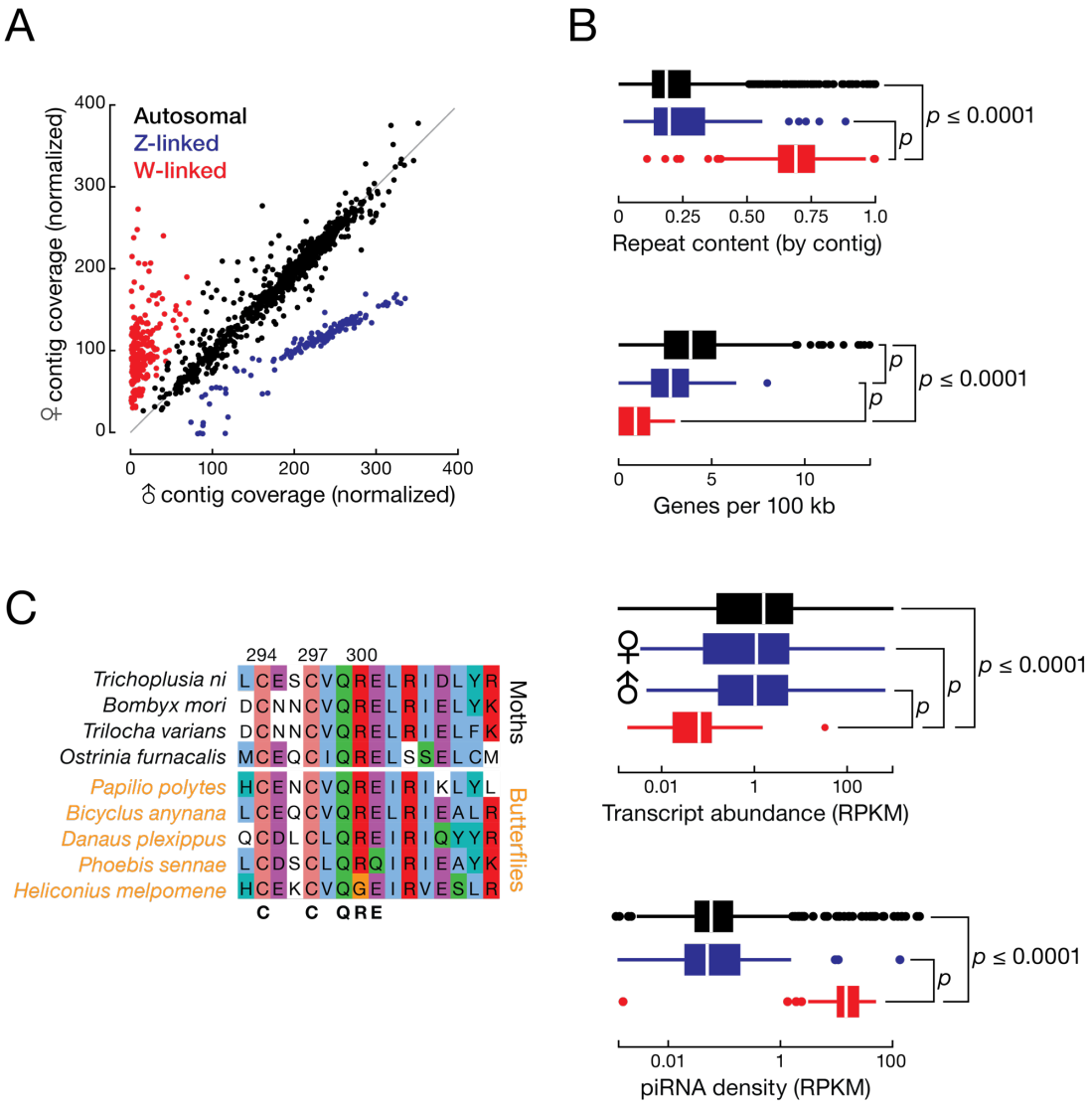

Figure 2.7. *T. ni* sex determination. (A) Normalized contig coverage in males and females. (B) Relative repeat content, gene density, transcript abundance (female and male thoraces), and piRNA density of autosomal, Z-linked, and W-linked contigs (ovary). (C) Multiple sequence alignment of the conserved region of the sex-determining gene *masc* among the lepidopteran species.

2.3.4 Multigene families

Genes are often grouped into multi-gene families according to their sequence homology. Genes belong to the same gene family often have similar biochemical functions, thus such grouping facilitates studies of their functions. Genes in the same biological pathway

are also often discussed together to better understand biological processes. To make *T. ni* genome a more useful resource, we computationally predicted and manually curated genes in notable families and pathways: opsin, cytochrome P450, glutathione S-transferase, carboxylesterase, ABC transporter, and chemoreception families (olfactory receptors, gustatory receptors, and ionotropic receptors), and genes in the juvenile hormone pathway.

Opsins are crucial for survival of insects because opsins enable the response to light cues (Terakita, 2005; Shichida & Matsuyama, 2009; Feuda et al., 2016). The *T. ni* genome contains opsins that mediate ultraviolet, blue and long-wavelength vision, suggesting that *T. ni*, albeit a nocturnal species, has color vision. In addition to the vision-related opsins, the *T. ni* genome also has orthologs of Rh7 opsin (Futahashi et al., 2015; Initiative, 2014) and pterosin (vertebrate-like opsin) (Velarde, Sauer, O. Walden, Fahrbach, & Robertson, 2005).

	Ultra-violet	Blue	Long-wavelength	Rh7	Pteropsin
Apis mellifera	1	1	2	0	1
Nasonia vitripennis	1	1	2	0	0
Tribolium castaneum	1	0	1	0	1
Aedes aegypti	1	1	6	1	1
Anopheles gambiae	1	1	7	1	2
Drosophila mojavensis	2	1	2	1	0
Drosophila melanogaster	2	1	3	1	0
Trichoplusia ni	1	1	2	1	1
Bombyx mori	1	1	2	1	1
Manduca sexta	1	1	1	1	1

Danaus plexippus	1	1	1	1	1
Heliconius melpomene	2	1	1	1	1
Pediculus humanus	1	0	1	1	0
Acyrthosiphon pisum	2	0	1	5	1

Table 2.2. Numbers of genes in 5 subfamilies of opsins in 14 species.

T. ni is a generalist herbivore and feeds on diverse plants; thus, it is constantly challenged by a variety of plant allelochemicals and synthetic insecticides. It is anticipated that *T. ni* maintains a battery of genes for detoxification. We surveyed four gene families known to play important roles in xenobiotic resistance: cytochrome P450s (P450s), glutathione-S-transferases (GSTs), carboxylesterases (COEs), and ATP-binding cassette (ABC) transporters (Figure 2.7) (Labbé, Caveney, & Donly, 2011; X. Li, Schuler, & Berenbaum, 2007).

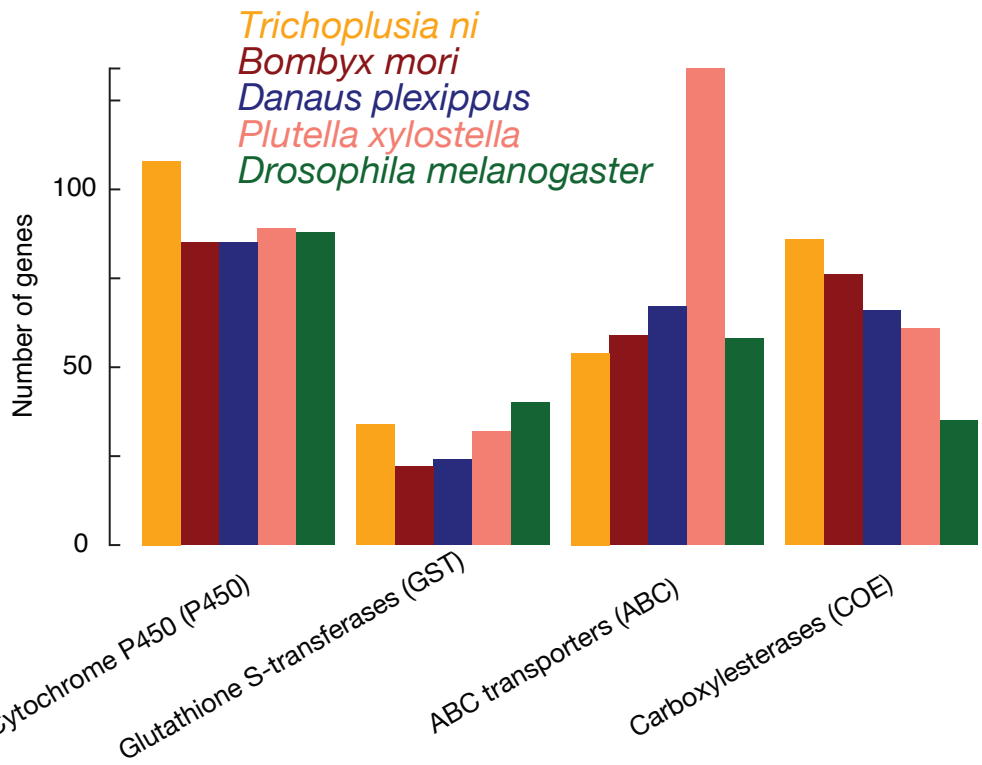

Figure 2.8. Counts of genes in 4 detoxification-related gene families in 5 species.

Cytochrome P450 (CYP) is a large family of enzymes that can metabolize natural products and xenobiotics (Scott, 1999). They are also involved in many other physiological processes such as hormone and pheromone biosynthesis. Insect CYPs are grouped into four clades: CYP2, CYP3, CYP4, and mitochondrial P450s (Feyereisen, 2006). We found 6 CYP2, 60 CYP3, 32 CYP4,10 mitochondrial P450s in *T. ni* (Figure 2.9), forming a 108-gene CYP superfamily, more than 83 CYPs in *B. mori*, mainly due to expansions in the CYP6AE (21 vs 10), CYP6AN (5 vs 1), and CYP6B subfamilies (5 vs 2). Although, to the best of our knowledge, the functions of these subfamilies are not known, other members in the CYP6 family have been linked to insecticide resistance. For

example, CYP6G1 confers DDT resistance in Drosophila (Daborn et al., 2002); CYP6D1 and CYP6Z1 confers pyrethroid resistance in Musca domestica (Kasai & Scott, 2000) and Anopheles gambiae (Nikou, Ranson, & Hemingway, 2003), respectively; CYP6A1 confers organophosphate resistance in house flies (Andersen, Utermohlen, & Feyereisen, 1994). Thus, we speculate some of the additional *T. ni* P450s might aid its quick adaptation to hostile environments.

CYP2 and mitochondrial P450s predominantly show 1-to-1 relationships between *T. ni* and *B. mori* (Figure 2.9), including P450s for ecdysteroid biosynthesis and inactivation (CYP307A2, CYP306A1, CYP18A1, CYP302A1, CYP315A1, CYP314A1) (Iga & Kataoka, 2012), circadian rhythm (CYP49A1) (Sathyanarayanan et al., 2008), cuticle formation (CYP301A1) (Sztal et al., 2012), and juvenile hormone biosynthesis (CYP15C1) (Iga & Kataoka, 2012).

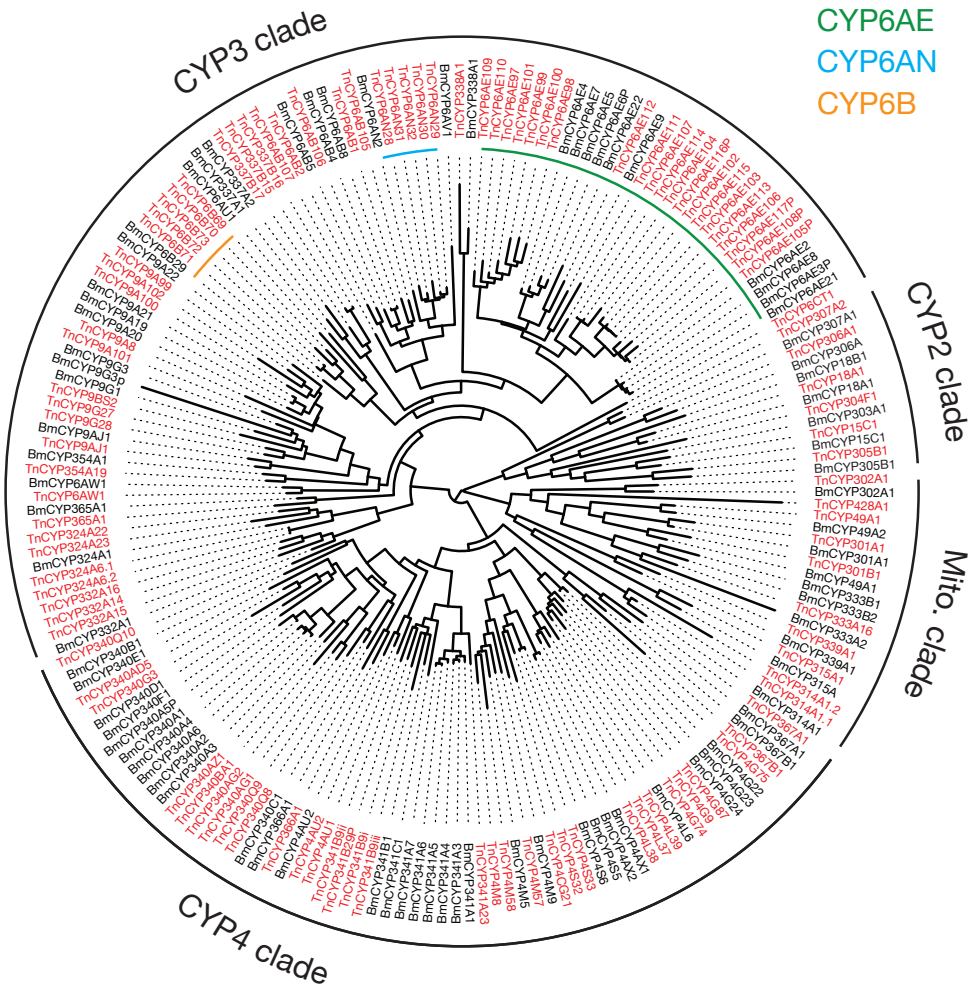

Figure 2.9. Phylogenetic tree of CYP genes in *T. ni* and *B. mori*. Black labels indicate *B. mori* genes and red labels indicate *T. ni* genes. Mito. clade: mitochondrial clade.

Glutathione S-transferases (GSTs) are a large family of enzymes that are important for detoxification. They catalyze the conjugation of glutathione to substrates, such as xenobiotics, to produce more soluble products which can be excreted (Strange, Spiteri, Ramachandran, & Fryer, 2001). Insects have six classes of GSTs: Delta, Epsilon, Omega, Sigma, Theta, and Zeta, with Delta and Epsilon being the two largest insect-specific classes (Enayati, Ranson, & Hemingway, 2005). Gene prediction followed by

manual review yielded 34 GSTs in *T. ni* (~90% are full length). In comparison, silkworm has 23 GSTs (Q. Yu et al., 2008). Phylogeny-based classification assigned 9, 14, 4, 2, 1, and 2 *T. ni* GSTs into the Delta, Epsilon, Omega, Sigma, Theta, and Zeta classes, respectively, among which Delta and Epsilon showed expansions in *T. ni* (Figure 2.10). Interestingly, many new GST genes formed clusters, suggesting that they are the result of recent gene duplication events, which may play important roles in facilitating the adaptation of *T. ni* to its ecological niches.

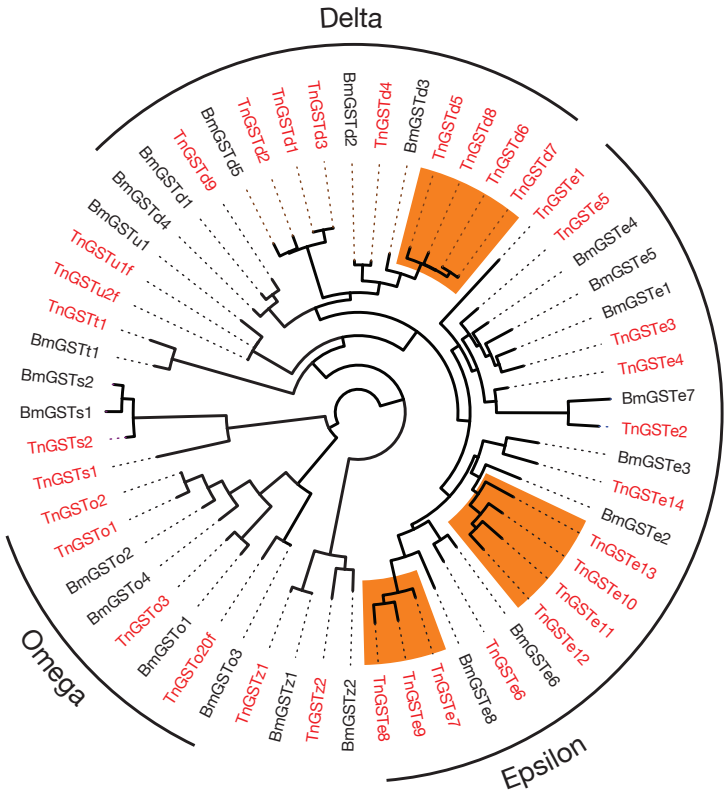

Figure 2.10. Phylogenetic tree of GST genes in *T. ni* and B. mori. Black labels indicate B. mori genes and red labels indicate *T. ni* genes. GST genes in Delta, Epsilon and Omega classes are marked.

Carboxylesterases (COEs) are a large protein family known to play important roles in metabolism of xenobiotics and pheromones (Ranson et al., 2002). Analysis of COEs may help understand how insects quickly become resistant to insecticides. We annotated 87 COE genes in the *T. ni* genome, more than the 76 putative COEs in the silkworm genome (Q.-Y. Yu et al., 2009). COEs are divided into 14 clades and 3 classes (intracellular catalytic, secreted catalytic, and neurodevelopmental classes) (A and B clades shown in Figure 2.10). *T. ni* COEs are distributed in 11 clades known to exist in Lepidoptera (Q.-Y. Yu et al., 2009).

In most clades, *T. ni* COEs have approximately one-to-one correspondence with their *B. mori* homologs. For example, Neuroligin and Neurotactin show perfect one-to-one orthologous relationships. Interestingly, even though the total numbers of α-esterases are roughly equal between *T. ni* and *B. mori*, subsets of the α-esterases form monophyletic groups in *T. ni*, suggesting these expanded independently after the two species diverged. The α-esterases in *T. ni* have an average amino acid identity of 33.3%. In contrast, both expansions in *T. ni* (Figure 2.11) are tightly clustered in the genome (11 COEs and 5 COEs in 280 Kb and 85 Kb, respectively) and have high sequence identity (62.4% and 65.3%), likely reflecting recent gene duplication events.

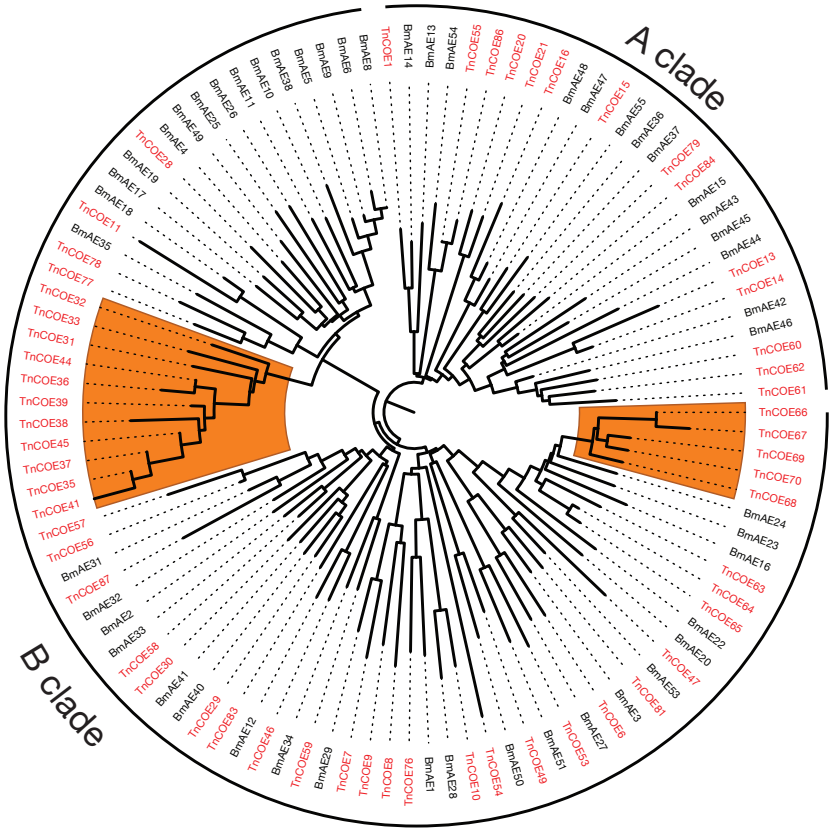

Figure 2.11. Phylogenetic tree of COE genes (A and B clades) in *T. ni* and B. mori. Black labels indicate B. mori genes and red labels indicate *T. ni* genes.

Insect ABC transporters play important roles in metabolism and insecticide resistance. ABC transporters in lepidopteran species can confer xenobiotic resistance, such as those in cotton bollworm Helicoverpa armigera, and tobacco hornworm Manduca sexta (Lanning et al., 1996; Murray, Quaglia, Arnason, & Morris, 1994). We identified

and manually annotated 54 ABC transporters in the *T. ni* genome. Previously, 5 ABC transporters (3 and 2 from the subfamilies ABCB and ABCC, respectively) were reported in *T. ni* (Labbé et al., 2011), all of which are consistent with our annotation. We annotated additional 49 members in the ABC transporter family, forming a total of 54 ABC transporters in the *T. ni* genome (Figure 2.12), which is within the range of ABC transporters in arthropods (Dermauw & Van Leeuwen, 2014). Notably, we annotated ABCC2, which is associated with Bt toxin resistance in *T. ni* (Baxter et al., 2011; X. Zhang, Tiewsiri, Kain, Huang, & Wang, 2012). We conclude that *T. ni* possess a unique repertoire of ABC transporters that provide insecticide resistance.

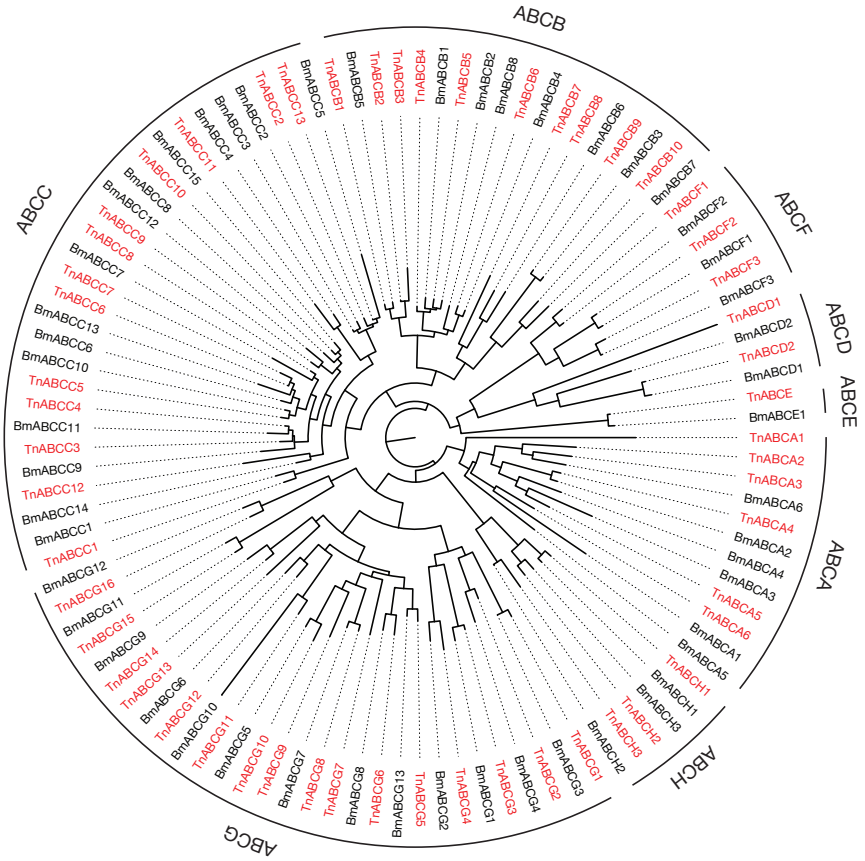

Figure 2.12. Phylogenetic tree of ABC genes in *T. ni* and B. mori. Black labels indicate B. mori genes and red labels indicate *T. ni* genes.

The success of insects is due in part to the variety of genes related to chemoreception (Leal, 2013). Agricultural pests, such as *T. ni*, use signal transduction cascades including the olfactory receptors (ORs), gustatory receptors (GRs), and ionotropic receptors (IRs) to utilize plant volatiles as chemical cues to recognize hosts.

We annotated 54 *T. ni* OR candidates, comparable to 66 ORs in *B. mori* (Figure 2.13), which are likely to be involved in activities, such as detecting pheromones for

sexual communication, detecting plant odors and long-range migration (Franklin, Ritland, & Myers, 2011). Interestingly, we found a paraphyletic group of 3 *T. ni* ORs (TnOR11, 15, and 18) that are close homologs to BmOR3, which recognizes sex hormones (Nakagawa, Sakurai, Nishioka, & Touhara, 2005). Thus, we hypothesize that this group may be involved in species-specific responses to sex hormones. We found another of *T. ni*-specific OR expansion of 3 genes (TnOR31, 33 and 40), the closest drosophila homologs of which can detect food odors (Laissue & Vosshall, 2008).

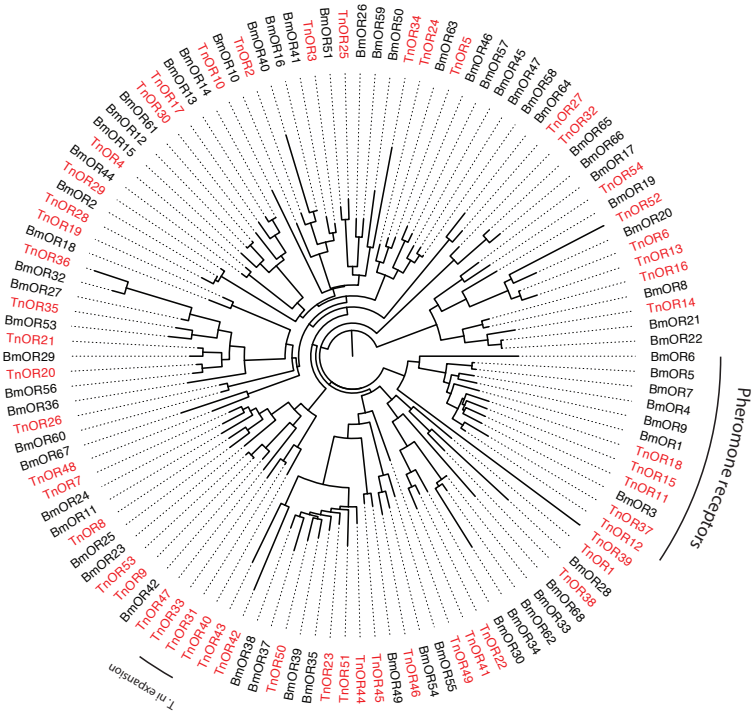

Figure 2.13. Phylogenetic tree of OR genes in *T. ni* and B. mori. Black labels indicate B. mori genes and red labels indicate *T. ni* genes.

Many insects rely on contact chemosensation to find host plants and to avoid toxic chemicals. Understanding gustatory receptors may help prevent agricultural pests, such as *T. ni*. We annotated 34 GR candidate genes (Figure 2.14), compared to 65 and 60 in *B. mori* and D. melanogaster (Montell, 2009; Wanner & Robertson, 2008). Despite the low number of GRs identified in *T. ni*, phylogenetic analysis indicated that 3 and 9 *T. ni* GRs cluster with their *B. mori* counterparts, respectively, forming putative carbon dioxide, sugar receptor branches (W. D. Jones, Cayirlioglu, Grunwald Kadow, & Vosshall, 2007; Slone, Daniels, & Amrein, 2007). Notably, both lepidopterans encode DmGR43a orthologs, which have been implicated in fructose perception in fruit fly (Sato, Tanaka, & Touhara, 2011). Additional 22 *T. ni* GRs form putative bitter receptors that have been proposed in other lepidopterans to be involved in detection of species-specific recognition of host plants (Wanner & Robertson, 2008; Zhan et al., 2011).

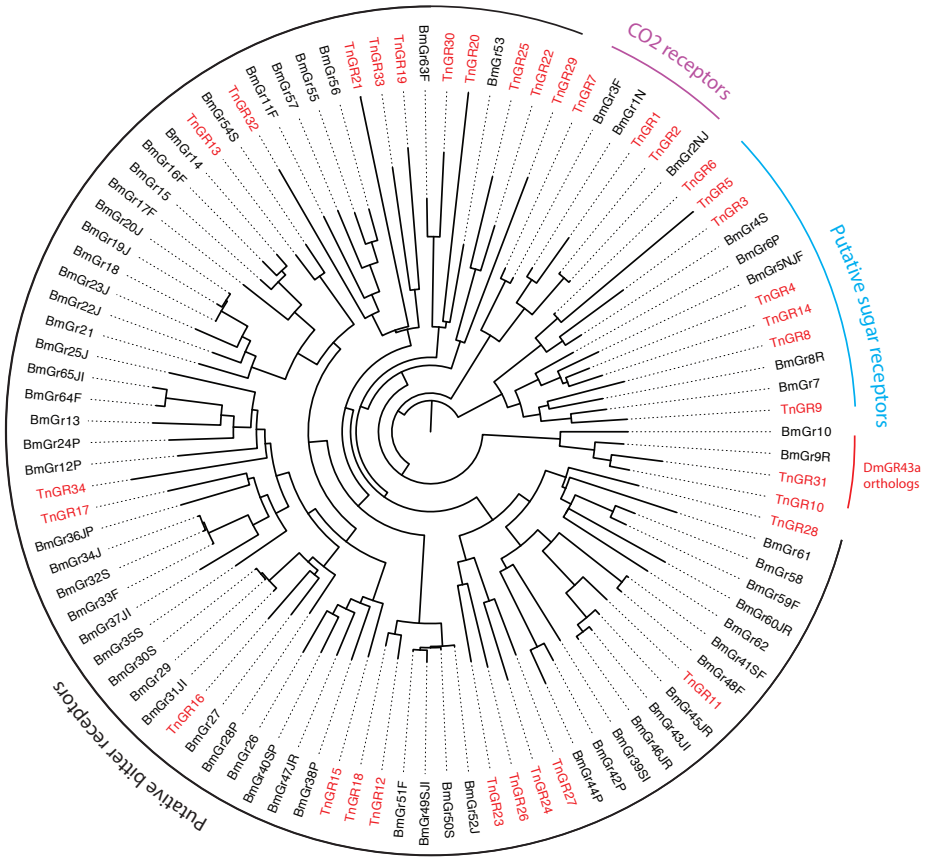

Figure 2.14. Phylogenetic tree of GR genes in *T. ni* (red) and B. mori (black).

We identified 29 ionotropic receptors in the *T. ni* genome, representing an intermediate size IR family, compared to 25 to 31 in other lepidopterans (31 in Heliconius melpomene, 25 in *B. mori*, and 27 in D. plexippus) (van Schooten et al., 2016). Phylogenetic comparison with IRs from other species revealed their homologous relationships (Figure 2.15). Interestingly, we observed an expansion of IR60 clade in *T. ni* (Figure 2.15, blue line), likely due to recent gene duplication events, as they are within

20 Kb genomic window. IR60 in D. melanogaster was shown to express in antenna and was designated “antennal IRs.”

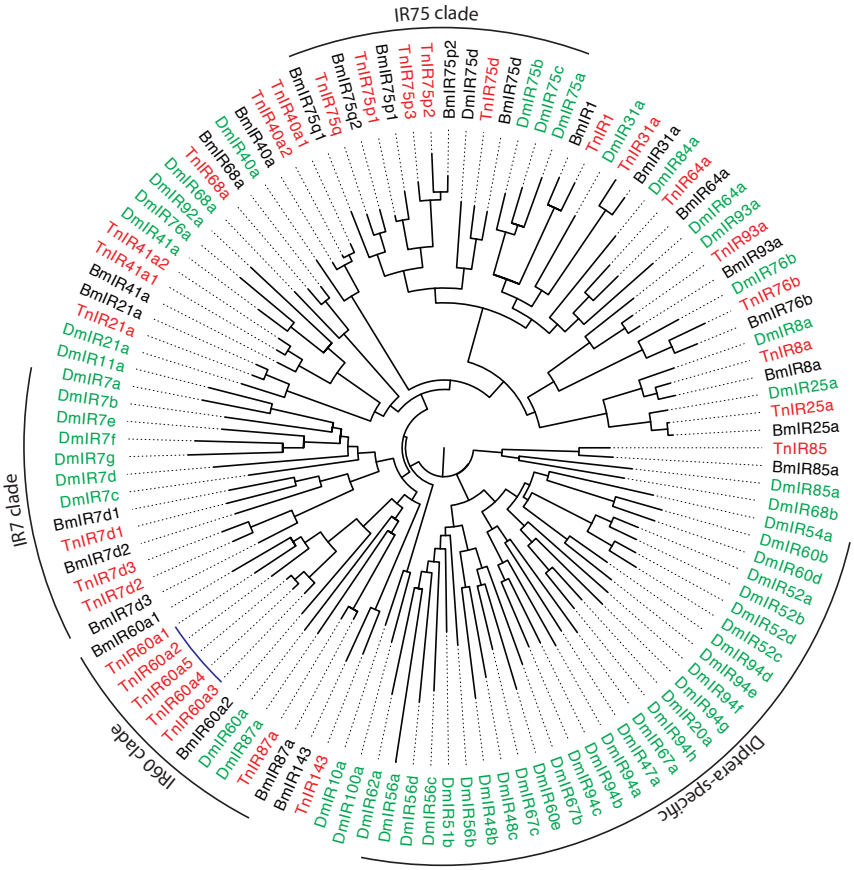

Figure 2.15. Phylogenetic tree of GR genes in *T. ni* (red), B. mori (black), and D. melanogaster (green).

Juvenile hormones regulate many physiological processes critical for insect survival, including metamorphosis, reproduction, and diapause. Some juvenile hormones are made exclusively by lepidopterans, indicating their finer control over some of the processes (Xavier Bellés et al., 2005) (Goodman & Granger, 2005). We searched for genes involved in the biosynthesis and degradation of Juvenile hormones and found that

T. ni possessed the entire repertoire of genes in this pathway (Figure 2.16). Notably, we found six copies of Isopentenyl pyrophosphate isomerases (IPPI 1-6) and five copies of farnesyl pyrophosphate synthases (FPPS 1-5), representing an expansion compared to one copy of IPPI and three copies of FPPS in silkworm (The International Silkworm Genome, 2008) Although the functions of the additional copies of the genes are unclear, we speculate that they may contribute to the production of Lepidoptera-specific and even *T. ni*-specific juvenile hormones.

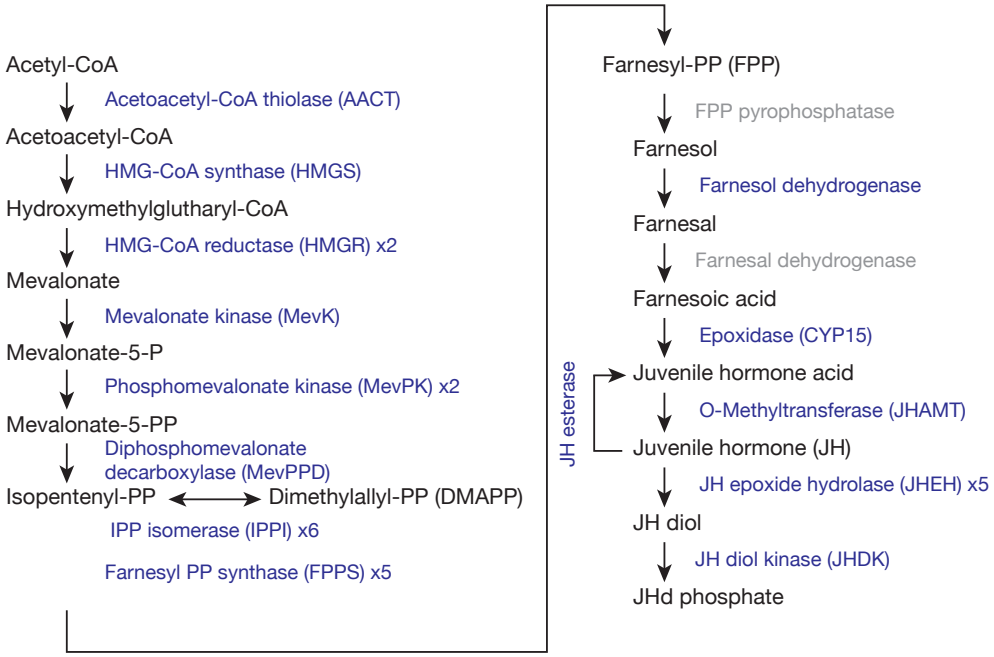

Figure 2.16. Genes in the juvenile hormone biogenesis and degradation pathways. Numbers after "x" indicates gene copy numbers. Gray gene names denote genes that have been proposed to reside in this pathway but their genomic loci are not known in any species.

2.3.5 miRNAs

Among the small silencing RNAs, microRNAs (miRNAs) are ~20–22 nt non-coding RNAs that regulate messenger RNAs (He & Hannon, 2004). In insects, miRNAs

may target genes in metamorphosis, reproduction, and other pathways of insect physiology and development (Lucas & Raikhel, 2013). Then, homology evidence and RNA-seq signals were used to determine miRNA pathway genes in the *T. ni* genome. Orthologs of all known miRNA pathway genes (including dcr-1, pasha, drosha, and ago2) can be found in this genome, suggesting that *T. ni* possess the ability to perform regulate genes via miRNAs. mirDeep2 was then used in conjunction with known miRNAs in other species and small RNA-seq data from *T. ni* ovary, testis, thorax and Hi5 samples to computationally predict miRNAs. In total, 295 miRNA genes (Figure 2.17, Supplementary file 3A and Supplementary file 4) were identified, including 77 conserved, 31 Lepidoptera-specific, and 187 *T. ni*-specific miRNAs (see (Y. Fu, Yang, et al., 2018)).

Then the miRNA expressions were compared in female and male thoraces (Figure 2.17). The majority of the expressed miRNAs (~82.2%) were not differentially expressed, indicating that normalization of miRNA counts is feasible. Forty-eight miRNAs are significantly differentially expressed between female and male thoraces (>2-fold change and FDR<0.1). Interestingly, miR-1—the highest expressed miRNA in both female and male thoraces—is differentially expressed: it was 2.2-fold more abundant in males. The function of miR-1 in *T. ni* is not known, but its homologs should provide some insights into its potential function in *T. ni*. miR-1 was previously characterized in fruit flies and has been shown to regulate muscle development (Sokol & Ambros, 2005). Thus, it is speculated that miR-1 is involved in sex-specific muscle development in female and male *T. ni*. An extremely conserved miRNA, let-7, was more abundant in males than female. *T. ni* let-7 has identical mature miRNA sequence with its homologs in many species,

including human, C. elegans. Since let-7 was previously characterized to play important roles in metamorphosis, the differentially expressed let-7 in *T. ni* might regulate sex-specific metamorphosis.

It is hypothesized that during miRNA evolution, newly formed miRNAs are first tested in limited tissues at low expression levels so that harmful miRNAs are selected against and useful ones are kept and increase their expression levels over time. These *T. ni* miRNAs provide an opportunity to test this idea. Indeed, more conserved miRNAs tend to have high expressions: the median expression is 320 ppm for conserved, 160 ppm for Lepidoptera-specific, and only 4.2 ppm for *T. ni*-specific miRNAs. It is worth noting that there are some newly involved miRNAs are highly expressed. For example, mir-novel1, mir-novel4 and mir-novel11 were highly expressed in both female and male thoraces.

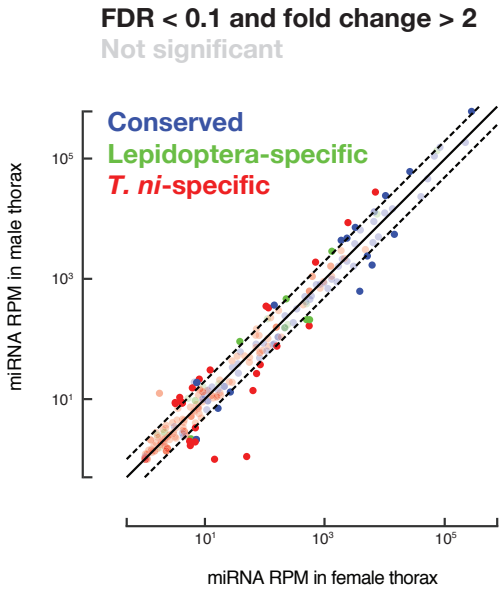

Figure 2.17. Expression of *T. ni* miRNAs in female and male thoraces. Colors indicate the level of conservation; solid dots indicate miRNAs that are significantly expressed.

Genes in miRNA/siRNA pathways	Gene ID in *T. ni*	*T. ni* gene name
CCR4-NOT transcription complex subunit 1 (not1, CNOT1)	TNI013924	CNOT1
CCR4-NOT transcription complex subunit 3 (not3, CNOT3)	TNI000261	CNOT3
CCR4-NOT transcription complex subunit 6-like (twin, CCR4, CNO6L)	TNI007086 *	CNO6L
CCR4-NOT transcription complex subunit 11 (not11, CNOT11)	TNI001169	CNOT11
hen1	TNI005148	hen1
ago1	TNI012430	ago1
ago2	TNI007888	ago2
microprocessor complex subunit DGCR8 (pasha)	TNI013094	DGCR8
Ribonuclease 3 (drosha, RNC)	TNI006564	RNC
exportin-5 (Ranbp21, exp5)	TNI002090	XPO5
GTP-binding nuclear protein Ran (ran)	TNI002740	RAN
endoribonuclease dcr-1	TNI002422	dcr-1
endoribonuclease dcr-2	TNI008774	dcr-2
RISC-loading complex subunit / Interferon-inducible double stranded RNA-dependent protein kinase activator A-A (loqs, PRKRA)	TNI009568	PRKRA
gawky (gw)	TNI003091	gawky

Table 2.3. *T. ni* genes in miRNA and siRNA pathways. Note that TNI007086 and TNI007087 were merged. The 3' UTR was curated to match RNA-seq signals.

2.3.6 siRNA characterization

siRNAs are another type of small silencing RNAs. They are typically 20–22 nt long and regulate gene expression, defend against viruses and suppress transposons (Agrawal et al., 2003; Chung, Okamura, Martin, & Lai, 2008; Chung et al., 2008; Czech et al., 2008; Ghildiyal et al., 2008; K. Okamura, Ladewig, Zhou, & Lai, 2013; Tam et al., 2008). siRNAs are processed from double-stranded RNAs into short double-stranded fragments with 2 nt overhang at 3′ ends. Unlike miRNAs, they require extensive sequence match between the guides and targets to facilitate target cleavage.

There are at least three sources of endogenous siRNAs (endo-siRNAs): transposon transcripts, *cis*-natural antisense transcripts (*cis*-NATs), and long hairpin

RNAs (hpRNAs) (Chung et al., 2008; Czech et al., 2008; Ghildiyal et al., 2008; Kawamura et al., 2008; K. Okamura et al., 2013; Katsutomo Okamura et al., 2008; Katsutomo Okamura & Lai, 2008; Watanabe et al., 2008). siRNAs from all three sources could be readily detected in *T. ni* tissues and Hi5 cells (Table 2.4), suggesting *T. ni* possess a functional siRNA pathway.

Species	Tissue/cell type	Endogenous siRNAs	Transposon siRNAs			Non-transposon siRNAs					
			Total	% total siRNAs	Hairpin siRNAs	Total	% total siRNAs	Hairpin siRNAs	% of non-transposon siRNA	cis-NAT siRNAs	% non-transposon siRNA
Trichoplusia ni	Hi5	912841	310761	34.0%	23692.7	602080	66.0%	25750	4.28%	147905	24.6%
	Ovary	313121	163940	52.4%	4085.26	149181	47.6%	1852.34	1.24%	17597.2	11.8%
	Testis	340933	145081	42.6%	14648.3	195852	57.4%	9037.56	4.61%	29663.4	15.1%
	Female thorax	563333	125809	22.3%	12724.2	437524	77.7%	3412.53	0.78%	138044	31.6%
	Male thorax	736512	152746	20.7%	12393.1	583766	79.3%	6186.64	1.06%	94731.5	16.2%
Fruit fly	Fly heads *	247377	129777	52.5%	18809	117600	47.5%	50104.5	42.6%	65683.6	55.9%

Table 2.4. Mapping statistics of *T. ni* siRNAs.

Next, exogenous siRNAs (exo-siRNAs) were characterized. Since one main function of exo-siRNAs is viral defense, a comprehensive search for viral transcripts was performed. No viral transcripts could be detected in *T. ni* tissues. However, highly abundant viral transcripts (FPKM of RNA1 and RNA2 of TNCL >5,000) could be detected in Hi5 cells, consistent with previous findings that Hi5 cells are latently infected with a positive-sense, bipartite virus (Tn5 Cell Line virus TNCL) (T.-C. Li, Scotti, Miyamura, & Takeda, 2007; Miller & Ball, 2012). Transcriptome assembly revealed that the detected viral transcripts had high sequence identity with previously characterized TNCL, further evidence of the existence of TNCL virus. Then all small RNAs that could not be mapped to the genome were mapped to the viral transcripts. Such virus-mapping small RNAs have the typical length distribution of siRNAs (median length = 21 nt, Figure 2.18A), suggesting Hi5 cells utilize siRNAs to defense against this virus. As a further test if they are bona fide siRNAs, we checked if these small RNAs bear 2 nt overhang at 3′ ends, by examining the distance from 5′ ends to 3′ ends on different strands. Such analysis revealed that such small RNAs tend to have 2 overhanging nucleotides at the 3′ ends (Figure 2.18B), hallmark of siRNAs. Next, another property of siRNAs is examined: siRNAs are typically produced in a processive manner. The distances from siRNA 3′ ends to 5′ ends are frequently zero ($p < 5.82 \times 10^{-5}$), and the length of a typical siRNA (20 nt), indicating that such small RNAs are made one after another. In summary, Hi5 cells use siRNAs for viral defense.

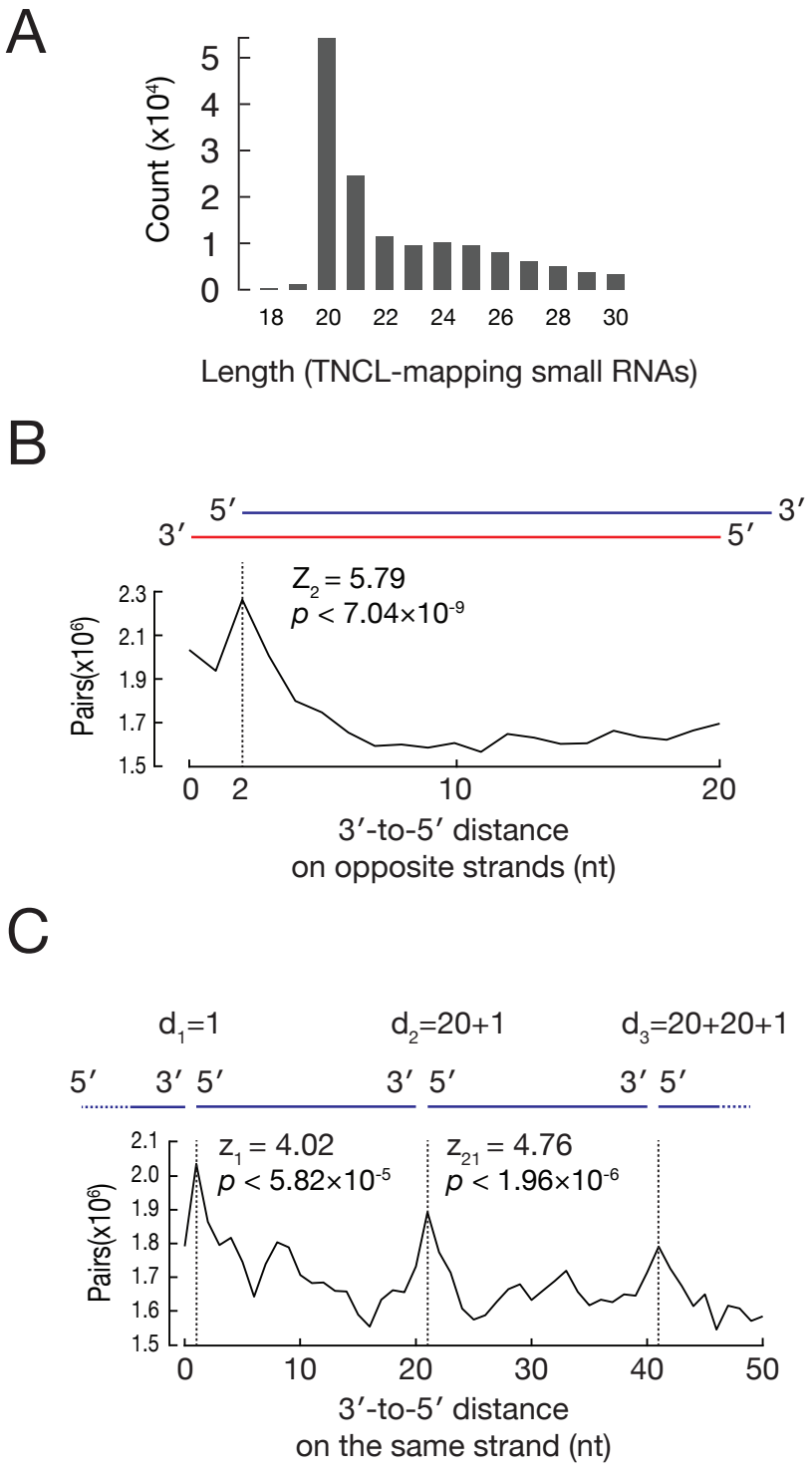
A
Count (x10^4)
5
4
3
2
1
0
18
20
22
24
26
28
30
Length (TNCL-mapping small RNAs)
B
5′
3′
3′
5′
Z_2 = 5.79
p < 7.04×10^-9
Pairs(x10^6)
2.3
2.1
1.9
1.7
1.5
0
2
10
20
3′-to-5′ distance
on opposite strands (nt)
C
d_1=1
d_2=20+1
d_3=20+20+1
5′
3′
5′
3′
5′
3′
5′
z_1 = 4.02
p < 5.82×10^-5
z_21 = 4.76
p < 1.96×10^-6
Pairs(x10^6)
2.1
2.0
1.9
1.8
1.7
1.6
1.5
0
10
20
30
40
50
3′-to-5′ distance
on the same strand (nt)

Figure 2.18. siRNA characterization. A. Length distribution of virus-mapping siRNAs. B. Distribution of 3′ to 5′ distances on opposite strands. C. Distribution of 3′ to 5′ distances on the same strand.

siRNAs and piRNAs, but not miRNAs, in fruit fly are 2′-O methylated, leading to the idea that siRNAs and piRNAs are 2′-O methylated in other insects. However, during the analysis of TNCL-mapping siRNAs, I noticed that siRNAs were almost depleted in oxidized small RNA-seq libraries (oxidization eliminated small RNAs without 2′-O methylation), suggesting that siRNAs are not 2′-O methylated in Hi5 cells. Length profiles of small RNAs sequenced from *T. ni* tissues and Hi5 cells indicated that 20-22 nt RNAs were abundant in unoxidized small RNA-seq libraries but depleted in oxidized small RNA-seq libraries. Thus, I conclude that siRNAs are not 2′-O methylated in *T. ni*.

Then, another question comes up naturally: are siRNAs unmethylated in other lepidopteran species? To check this, I collected data from oxidized and unoxidized small RNA-seq libraries, and determined abundance ratios (ox/unox ratios) of siRNA species that exist in both versions of libraries. For fruit fly siRNAs, ox/unox should be close to 1 as such siRNAs are 2′-O methylated. If a species has unmethylated siRNAs, then ox/unox should be smaller than 1. Indeed, fruit fly siRNAs have ox/unox ratios close to 1, whereas siRNAs from other lepidopteran species have ox/unox ratios <0.23, much smaller than 1, indicating that siRNAs from these 3 lepidopterans are not methylated. Since *T. ni* and P. xylostella diverged more than 170 million years ago, this observation suggests that many other lepidopterans—and possibly all lepidopterans—lack the ability to 2′-O methylate siRNAs. This further raised the question of the purpose of siRNA methylation: if lepidopterans can survive without siRNA methylation, why do other species maintain

siRNA methylation? Profiling siRNAs from more species, especially basal species, can answer this question.

2.3.7 piRNAs

piRNAs, 23–32 nt long, exist in many animals to protect the germline genome by suppressing transposon activities (A. A. Aravin et al., 2007; Brennecke et al., 2007; Girard, Sachidanandam, Hannon, & Carmell, 2006; Lau et al., 2006; Vagin et al., 2006). In D. melanogaster, transposon-rich genomic loci (piRNA clusters) are transcribed to produce piRNA precursor transcripts, which are subsequently processed into piRNA and loaded into PIWI proteins (Piwi, Aubergine and Argonaute3). Piwi, when loaded with piRNAs, can direct installation of histone H3 lysine 9 tri-methylation (Brown et al., 2014; Le Thomas et al., 2014; Sienski, Dönertas, & Brennecke, 2012). In D. melanogaster cytoplasm, piRNAs guide Aub to find and cleave transposon mRNAs via sequence complementarity. The cleavage products can then be processed and loaded into Ago3 as sense piRNAs. Ago3, loaded with sense piRNAs can then cleave piRNA precursor transcripts from piRNA clusters, generating more piRNAs that can be loaded into Aub. This forms a feed-forward loop that efficiently amplify piRNAs and repress transposon activity (Brennecke et al., 2007).Ago3 cleavage can also initiate Piwi-bound piRNAs that effectively diversity piRNA pool, enhancing transposon suppression (Han, Wang, Li, Weng, & Zamore, 2015; Mohn, Handler, & Brennecke, 2015).

2.7.1 piRNA characterization

Most of the genes in the piRNA pathway are correctly predicted by computational methods. However, a few short genes and some UTR annotations were missing. Thus, D. melanogaster and *B. mori* piRNA pathway genes were used as references to detect piRNA pathway genes. Then all lines of evidence, such as RNA-seq coverage, BLAST results were loaded into WebApollo Figure 2.19. Then I manually ensured the compatibility of all evidence by modifying gene models. Such gene curation revealed that the *T. ni* genome contains a full repertoire of piRNA pathway genes (Appendix B). Many genes were expressed in both germline and somatic tissues, though the germline expression tend to be higher than that in the soma (median ratios: ovary/thorax = 14, testis/thorax = 3, and Hi5/thorax = 5, Figure 2.20). The expression of piRNA pathway components in both the germline and Hi5 cells suggests that Hi5 cells have the intact piRNA pathway.

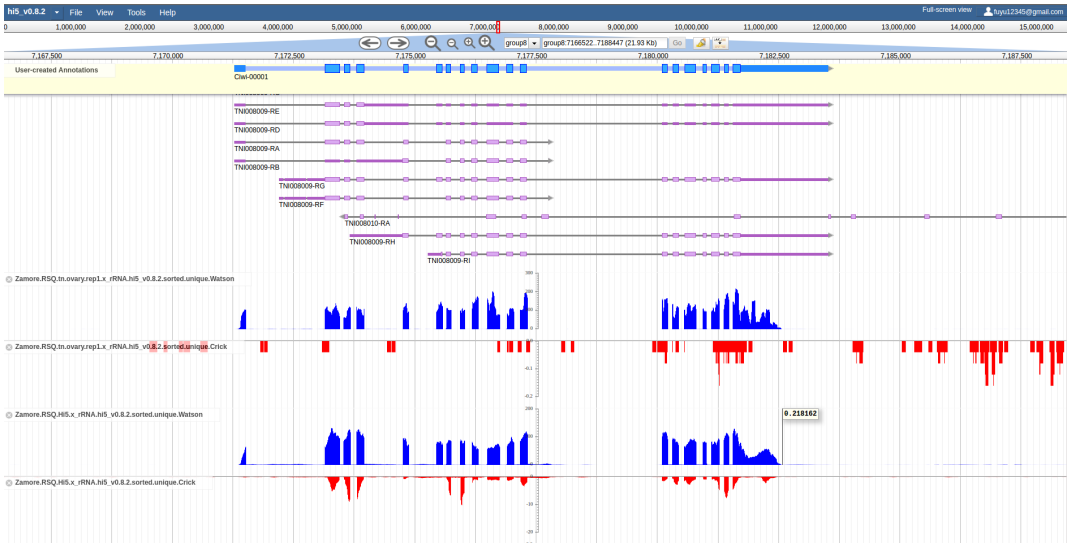

Figure 2.19. Screenshot of Apollo showing the ciwi gene.

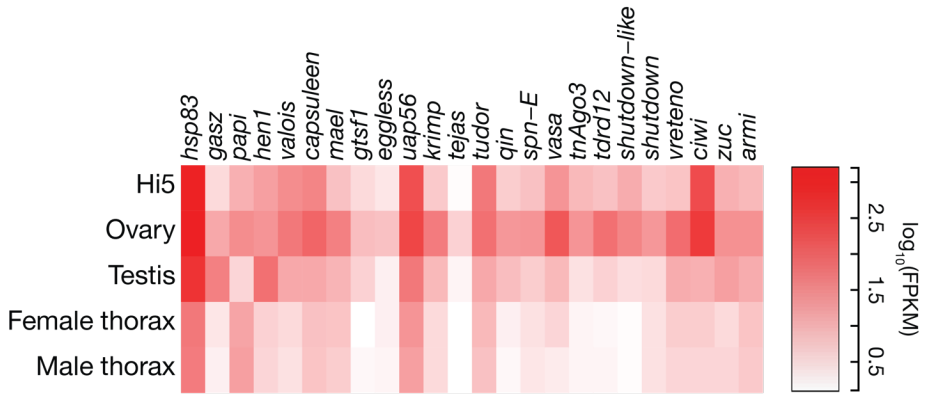

Figure 2.20. Expression of piRNA pathway genes in 4 *T. ni* tissues and Hi5 cells.

In terms of orthology, most genes in the *T. ni* piRNA pathway have one-to-one correspondence with D. melanogaster orthologs. However, *T. ni* genome encodes only two PIWI proteins, TnPiwi and TnAgo3 instead of 3 in D. melanogaster. (I originally named TnPiwi as Ciwi [cabbage looper piwi] to follow the naming convention of Siwi [silkworm Piwi] in *B. mori*, but later, due to the potential scalability issues pointed out by reviewers, TnPiwi replaced Ciwi.) Without further experiments, it is unknown if TnPiwi functions more like Aub or Piwi in D. melanogaster. Another noticeable difference is that D. melanogaster genome encodes Rhino, Cutoff and Deadlock to mark piRNA clusters but *T. ni* genome encodes none. Furthermore, the trio of genes is known to be poorly conserved, indicating that how fruit fly marks piRNA clusters is highly unlikely to be a universal mechanism and that *T. ni* is possibly a better representation of how insects mark piRNA clusters.

2.3.8 Characterization of piRNA clusters

In D. melanogaster, piRNAs are produced in the germline but not in the soma. *T. ni*, however, produces piRNAs from discrete genomic loci in both the germline and the soma. piRNAs are short and thus are often map to multiple genomic loci, making it difficult to resolve these multimappers. To solve this, I designed an expectation-maximization algorithm that resolves such multimappers and applied this method to datasets from different tissues and Hi5 cells. In total, piRNA-producing loci comprise 10.7 Mb in ovary, 3.1 Mb in testis, 3.0 Mb in Hi5 cells and 2.4 Mb in thorax (Figure 2.21). For each tissue or cell type, these clusters can explain >70% of uniquely mapped piRNAs and all piRNAs when using expectation-maximization mapping. Interestingly, 1.5 Mb of piRNA clusters are active in both the germline and the soma, suggesting that these are required for *T. ni* development.

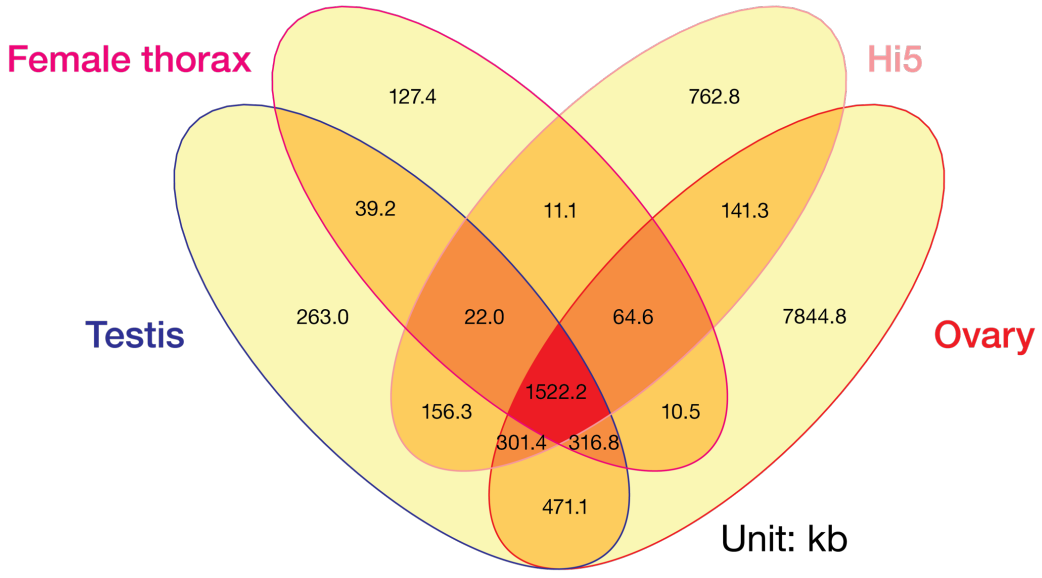

Figure 2.21. Expression of piRNA pathway genes in 4 *T. ni* tissues and Hi5 cells.

T. ni piRNA cluster have substantially different sizes. In *T. ni* ovary, more than half of the bases in piRNA clusters are in 67 piRNA clusters (median length = 53 kb). The largest five piRNA clusters are longer than 200 kb and the smallest one is 38 kb. Some *T. ni* piRNA clusters produce abundant piRNAs. For example, the cluster on chromosome 13 produce the most piRNAs among all piRNA clusters and can explain 7.8% uniquely mapped piRNAs (~50,000 piRNA species) (Figure 2.22). piRNAs originate from limited genomic loci. The top 20 piRNA clusters in *T. ni* ovary can explain >50% uniquely mapped piRNAs.

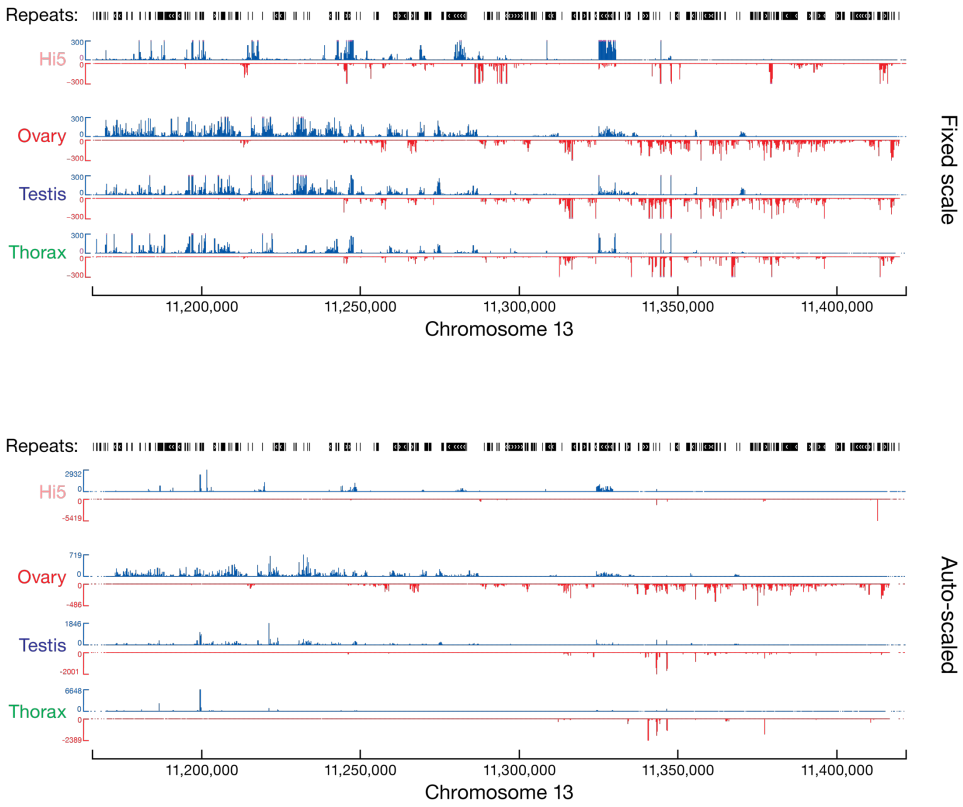

Figure 2.22. Small RNA signals along the most productive piRNA cluster on chromosome 13.

Much of the knowledge on piRNAs were obtained from studying fly piRNAs. In the fly ovary germline, most piRNA clusters generated piRNAs from both strands. Such piRNAs can fuel the "Ping-Pong" amplification cycle and robustly promote piRNA production (Brennecke et al., 2007). Some piRNA clusters, such as flamenco piRNA cluster (Brennecke et al., 2007; Malone et al., 2009), produce piRNAs from one strand only, without Ping-Pong amplification. Such uni-strand piRNA clusters are the only sources of piRNAs in the somatic follicle cells in the fly ovary.

The *T. ni* genome contains both types of piRNA clusters. In ovary, about 20% of piRNA clusters are dual-strand. And they collectively produce 35.9% of uniquely mapped piRNAs (and 22.8% of all piRNAs). piRNAs from dual-strand clusters are mostly antisense to transposons (71.6%). The remaining 286 piRNA clusters are uni-strand and can explain 54.8% of uniquely mapped piRNAs and 36.7% of all piRNAs. Similar to piRNAs mapped to dual-starnd clusters, piRNAs from uni-strand clusters are also mostly antisense to transposons (74.8%), which reflects that piRNAs suppress transposon transcripts. The antisense bias of piRNAs from uni-strand piRNA clusters is likely to originate from positive selection for antisense insertions. Collectively for uni-strand clusters, 57.1% of transposons insertions are opposite to the direction of piRNA precursor transcriptions. Dual-strand clusters, on the other hand, lack such bias: 49.5% of transposon insertions in dual-strand clusters are in the antisense direction.

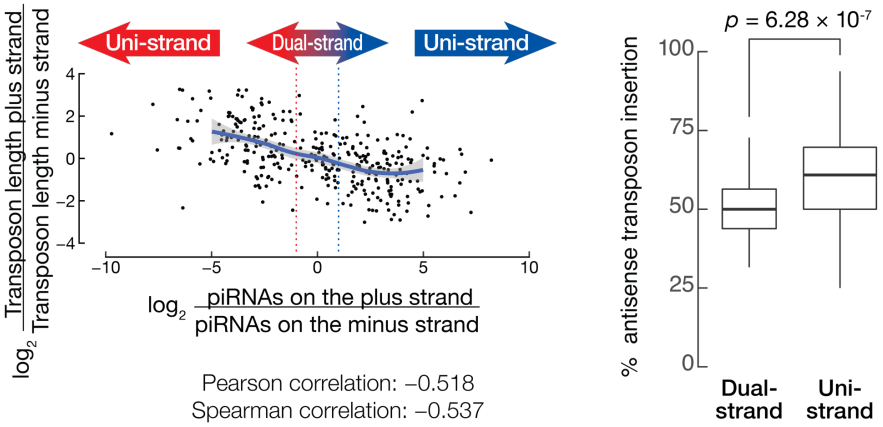

Figure 2.23. Transposon insertion bias in dual- and uni-strand piRNA clusters

2.3.9 The entire W chromosome as a major source of piRNAs

The W chromosome was not well understood due to the difficulty to assemble it. The availability of the *T. ni* W chromosome allows the first opportunity to globally characterize it. The largest piRNA cluster is a 462 kb region on the W chromosome, consistent with the observation that W chromosome produces a substantial portion of piRNAs. This is likely to be a underestimation of this piRNA clusters due to the mappability problem. (70.8% of the bases in the flanking regions are not uniquely mappable.) As a matter of fact, 85.1% of the bases between W-linked piRNA clusters are not uniquely mappable. Thus, these gaps between piRNA clusters are likely due to the limitation of mappability, and we propose that those piRNA clusters are likely just one giant piRNA cluster.

To further test if the W chromosome is a major source of piRNAs, we calculated piRNA abundance (normalized to contig length) using piRNA reads that could be uniquely mapped to the all contigs. W-linked contigs produced much more piRNAs than Z-linked and autosomal contigs, consistent with our notion that W is a major source of

piRNAs. In *T. ni*, 27.2% of uniquely mapped piRNAs in the ovary derived from W-linked contigs, even though these contigs compose only 2.8% of the genome (Figure 2.24). The W chromosome is likely to produce much more piRNAs, due to the unassemblable part of the W chromosome.

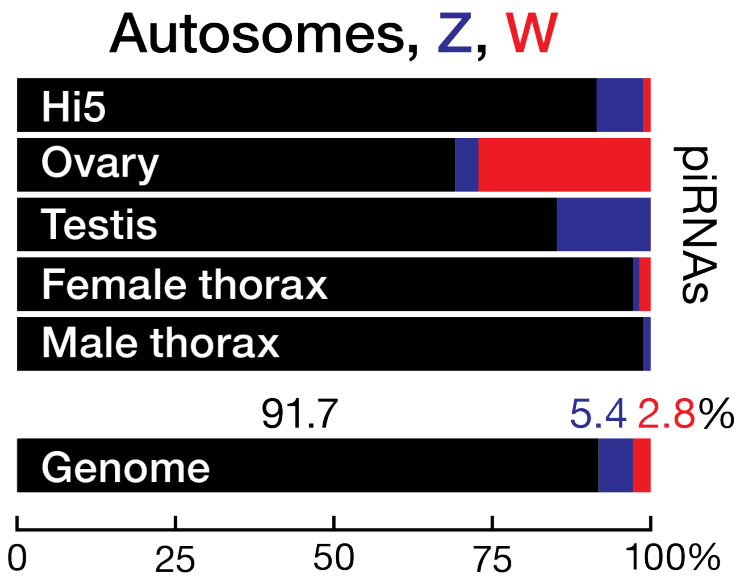

Figure 2.24. piRNA that could be uniquely mapped to the genome (first 5 bars) and the proportions of the genome that are autosomal (black), Z-linked (blue) and W-linked (red).

The reviewers of this paper raised the question if every base of the W-linked regions produces piRNAs, since rigorously speaking, it could be that some W-linked regions do not produce any piRNAs. To determine this, I search for W-linked regions that are not covered by any piRNAs and found that 11.0% of the W-linked bases do not produce any piRNAs. The reviewers also asked if the predicted miRNAs and coding genes on the W produce piRNAs. Since it is possible that some of these predictions were wrong, I manually curated all annotations on the W chromosome. All 9 predicted miRNAs produce small RNAs showing the Ping-Pong signature, suggesting that these are

likely misannotated as miRNAs and are instead piRNA-producing loci. All 74 predicted protein-coding genes on the W chromosome were further categorized into 4 groups: orphan genes (no homologs found), transposons (good homology to transposons), uncharacterized/hypothetical proteins, and potential protein-coding genes with homology. Those with transposon homology tend to produce more piRNAs (median = 44.9 ppm), with uncharacterized/hypothetical proteins and potential protein-coding genes produce fewer piRNAs (Figure 2.25). Those orphan genes produce the fewest piRNAs, with some putative genes produce no piRNAs at all. We thus conclude that though some W-linked loci are devoid of piRNAs, nearly the entire W chromosome is devoted to piRNA production.

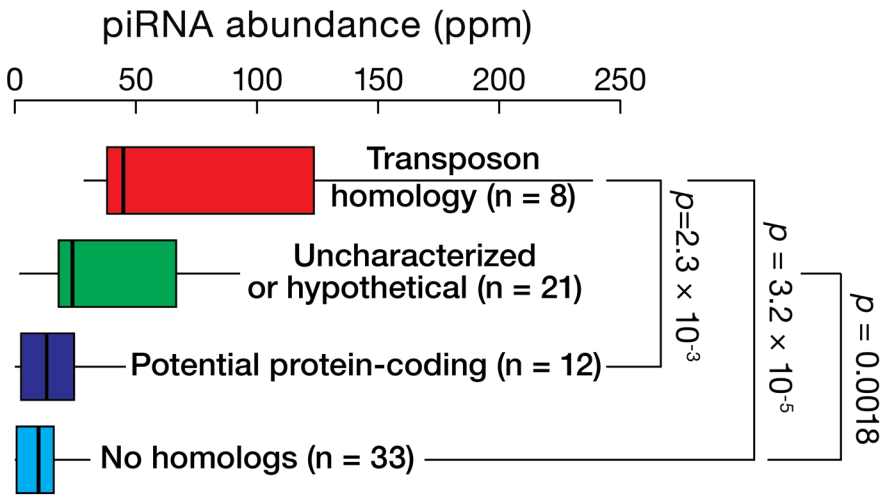

Figure 2.25. piRNA abundance for W-linked genes (categorized according to their homology to existing annotations.

2.3.10 piRNA cluster expression

In the *T. ni* germline, different piRNA cluster produce wildly different piRNAs, but the top 5 piRNA clusters consistently produce the most piRNAs, indicating that these

constitutive piRNA clusters are the master loci for piRNA production and transposon silencing. Many of the other piRNA clusters show tissue-specific expression. One observation is that the W chromosome produces plurality of piRNAs in ovary, but not in Hi5 cells (Figure 2.26A), which could be due to a) Hi5 cells reduced the ability to produce W-linked piRNAs, or b) Hi5 cells was derived from one type of germline cells, and such cells are under-represented in the ovary. A comparison between female and male thorax reveals that some Z-linked clusters produce more piRNAs in male (Figure 2.25B).

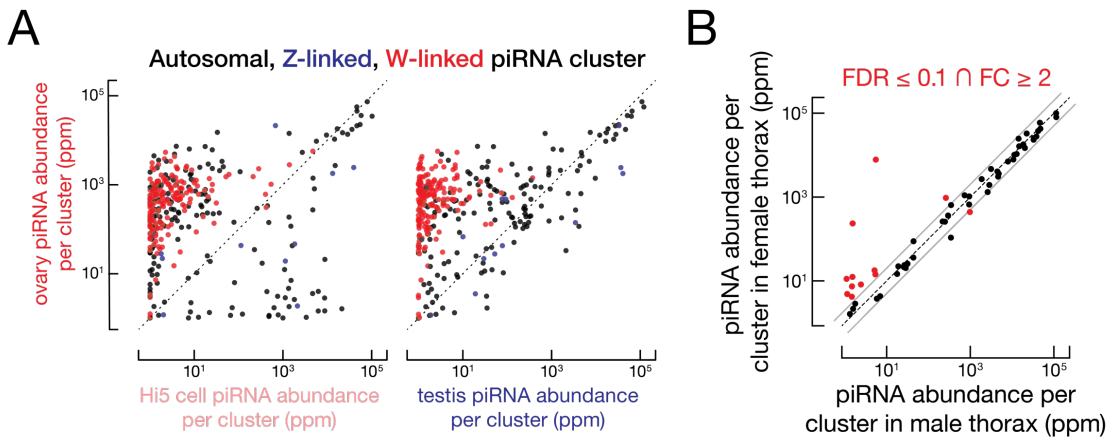

Figure 2.26. Comparisons of piRNA abundance (A) among ovary, testis and Hi5, and (B) between female and male thorax.

Forty clusters produce piRNAs in Hi5 cells, but not in ovary, which raises the question if these clusters were newly gained by Hi5 cells during immortalization. To test this, I looked for new transposon insertions in these Hi5-specific clusters by integrating WGS data from male and female individuals. Of these 40 clusters, 12 contain 74 Hi5-specific transposon insertions, suggesting that the insertions of these transposons transformed the inserted regions into piRNA clusters. To test if such transposons

evolutionarily young, I calculated sequence divergence rates and found that, compared to the transposons shared by ovary and Hi5, these 74 Hi5-specific transposons have lower sequencing divergence rate (Figure 2.27). The conclusion is that the Hi5-specific piRNA producing loci were likely caused by transposon insertions after the derivation of Hi5 cells, suggesting that *T. ni* and other animals can readily evolve new piRNA clusters to protect their genomes against transposon insertions.

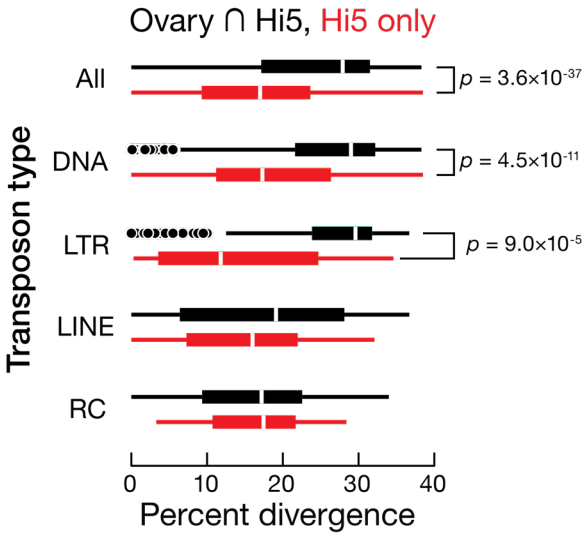

Figure 2.27. Sequence divergence rate for Hi5-specific transposons and transposons shared between ovary and Hi5

Next, somatic piRNA clusters in *T. ni* were examined. In fruit fly, somatic piRNAs are much less abundant than other types of small RNAs, suggesting that fly does not utilize the piRNA machinery in somatic tissues. Surprisingly, *T. ni* somatic tissues produce abundant piRNAs, suggesting that somatic piRNAs may play important roles for transposon suppression and gene regulation. In female and male thorax, piRNA clusters compose only ~0.57% of the genome and can explain the majority of uniquely mapped piRNAs (86.8% and 89.5 for females and males, respectively). These somatic piRNA

clusters are mostly the same with germline piRNA clusters (>90% of bases in somatic clusters in germline clusters), supporting the notion that some piRNA clusters are always active during development. A comparison of piRNA clusters between female and male thorax reveals that, in addition to most piRNA clusters with comparable expression levels, 12 clusters are differentially expressed. Nine of these 12 clusters are W-linked produce significantly more piRNAs in female than in male thorax (Figure 2.26B).

2.3.11 The lack of splicing of piRNA precursor transcripts

In fruit fly, splicing of piRNA precursor transcribed from dual-strand piRNA clusters is suppressed by Rhino, Cutoff and Deadlock (Mohn et al., 2014; Z. Zhang et al., 2014). Uni-strand piRNA clusters, on the hand, behave like canonical PolII transcribed genes (Brennecke et al., 2007). A search for the three genes reveals no hit, which seemingly would predict the presence of piRNA precursor splicing. To answer this question, I identified splicing events using RNA-seq data, by looking for reads that map across exon-exon junctions, requiring that read counts ≥10 to ensure enough coverage and splicing entropy ≥2 to exclude PCR duplicates (Graveley et al., 2011). Even though there are >100 piRNA clusters, only 27 splice sites could be detected from all tissues (Figure 2.28). Of these 27, 19 reside in uni-strand piRNA clusters. We conclude that transcripts from *T. ni* dual-strand piRNA clusters are rarely splices and that transcripts from uni-strand clusters undergo infrequent splicing.

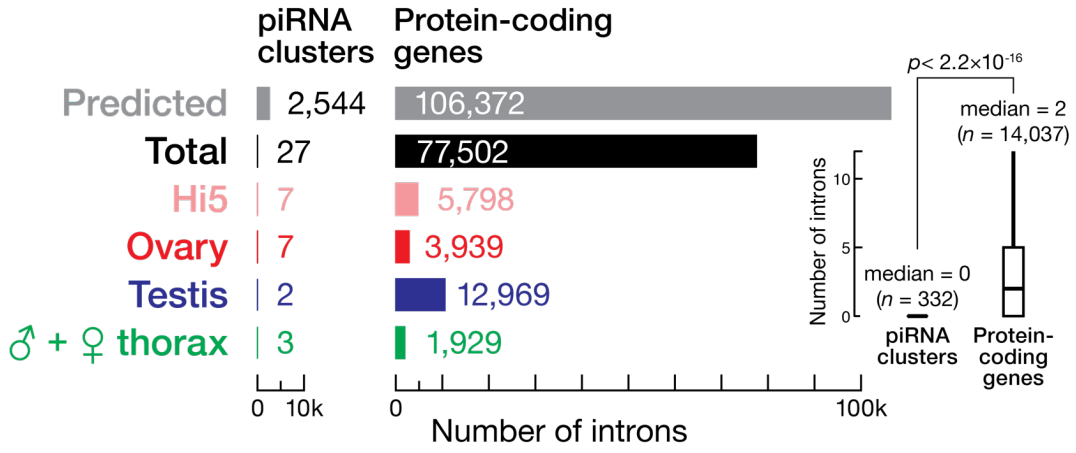

Figure 2.28. Splice sites in piRNA clusters and protein-coding genes. The first bar is derived from gene prediction. The remaining bars show the splice sites supported by RNA-seq. The boxplot shows the number of introns supported by RNA-seq.

The lack of splicing could be due to an active suppression mechanism, or lack of splice sites. To distinguish these two possibilities, I computationally predicted gene models (requiring peptide length >200 amino acids) in piRNA clusters using the same parameters trained for genome-wide gene prediction. This round of gene prediction was done without masking the genome as the majority of the bases in clusters fall into repetitive regions. This method predicted a total of 1,332 gene models containing 2,544 introns with good splicing signals. Notably, ~90% of these gene models had high sequence homology with transposons, indicating that many transposons in piRNA clusters have intact splice sites. Splicing efficiency was then measured by the ratio of spliced to unspliced reads for each of the splice sites supported by RNA-seq. Compared to the control set of introns (i.e. introns from protein-coding genes), splicing efficiency in piRNA clusters was lower (9.67-fold lower in ovary, 2.41-fold lower in testis, 3.23-fold lower in thorax, and 17.0-fold lower in Hi5 cells) (Figure 2.29), indicating that piRNA precursor transcripts in *T. ni* are inefficiently spliced. To test if the splicing efficiency is

different in dual- and uni-strand piRNA cluster transcripts, I compared the splice sites supported by RNA-seq and found that dual-strand cluster transcripts had lower splicing efficiency compared to uni-strand cluster transcripts (Figure 2.29). In conclusion, piRNA clusters transcripts are rarely and inefficiently spliced and dual-strand cluster transcripts have lower splicing efficiency.

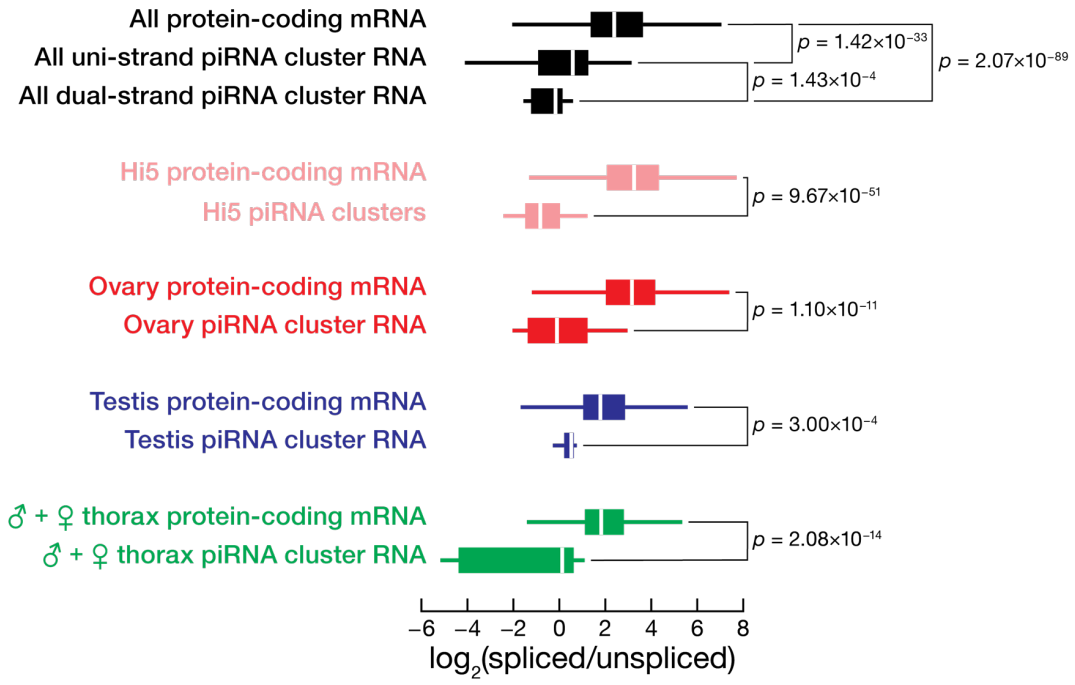

Figure 2.29. Splicing efficiencies in *T. ni* tissues and Hi5 cells.

2.4 Discussion

T. ni is a common and destructive pest that feeds on many plants, such as cabbage and broccoli. Using Hi5 cells, we sequenced and assembled the *T. ni* genome. Computational prediction followed by manual curation reveals the expansions of detoxification-related gene families. Further characterization of these genes may provide insights into pesticide development. In addition, *T. ni* is a pest found worldwide, so the

availability of its genome enables the research on its genetic diversity and population structure. Furthermore, much research was done in Diptera such as *D. melanogaster* and having the *T. ni* genome sequence would facilitate studies that could be applicable to more species.

Two vital steps for this genome assembly are contig construction using PacBio long reads and scaffolding using Hi-C reads. The PacBio reads make it possible to obtain highly contiguous contigs, which serve as a foundation for scaffolding. Hi-C reads can often connect two loci that are far apart (e.g. 10 kb, 100 kb, or even >1 Mb), and such information is critical for the scaffolding process and can produce a chromosome-level assembly. The genome assembly strategy should be readily applicable for other species and enable quick and cost-efficient genome assemblies of other species.

Previously assembled lepidopteran genomes usually do not contain W chromosomes, or have very fragmented W-linked sequenced, due to its repetitiveness. For example, the silkworm genome project only included males (ZZ) (Biology analysis group, 2004; The International Silkworm Genome, 2008). The monarch butterfly genome project included both males and females, but the genome assembly is fragmented (Zhan et al., 2011), hindering the characterization of the W chromosome. In contrast, our *T. ni* genome assembly not only captured many W-linked sequences, but also assembled them into highly contiguous chromosome-length scaffolds, which, to our knowledge, is the first chromosome-level assembled of a lepidopteran W chromosome. The availability of the W chromosome, together with the Z chromosome and autosomes, provides a unique opportunity to look into sex determination and dosage compensation.

Hi5 cells produce siRNAs from the RNA genome of an alphanodavirus and such siRNAs are produced in a one-after-another (i.e. processive) manner, consistent with previous characterization of fruit fly siRNAs. However, unlike siRNAs in fruit fly, *T. ni* siRNAs are not 2′-O-methylated at 3′ ends. We currently do not understand the implications of the lack of methylation. The commonalities and difference between *T. ni* and widely studies D. melanogaster should enable molecular dissection of the deeply conserved and rapidly evolving components of small RNA pathways.

One motivation of the genome assembly project is to establish a cell culture model and provide a counterpoint for studying small RNAs. The genome assembly and the gene-editing procedures (see (Y. Fu, Yang, et al., 2018)) can enable the use of Hi5 cells to study small RNA biogenesis. A systematic search of piRNA pathway genes reveals all known piRNA pathway genes (except those Drosophilid-specific ones). The piRNA clusters in Hi5 cells, ovary, testis, thorax should facilitate the next steps (e.g. genome-wide screen of potential piRNA pathway genes). The fact that the same set of 5 most productive piRNA clusters is present in ovary, testis, and Hi5 cells also suggests that Hi5 cells can recapitulate the piRNA pathway. Additionally, Hi5 cells evolved to gain extra piRNA clusters that are not present in *T. ni*, suggesting that one could potentially create new piRNA clusters in Hi5 cells and study such clusters from an evolutionary prospective.

Despite the lack of the trio (rhino, cutoff and deadlock) responsible for splicing suppression of piRNA cluster transcripts, *T. ni* piRNA cluster transcripts are rarely and inefficiently spliced, suggesting that *T. ni* has other mechanisms to perform this task.

Since this trio is not found outside Drosophilids, Hi5 cells are likely a better and more general model for studying piRNAs. Notably, almost the entire W chromosome is devoted to piRNA production. As more genomes are assembled, we shall be able to tell if this is a general feature of lepidopteran W chromosomes and even other animals.

Procedures for genome editing and single-cell cloning are also established to facilitate further studies, making the Hi5 cell line a powerful tool to study small RNAs. (Procedures can be found in (Y. Fu, Yang, et al., 2018).) In principle, the genome-editing procedures can be readily applied to the cabbage looper embryos or eggs to generate genetically modified *T. ni* strains, or to implement pest management to contain this agricultural pest.

Chapter 3. Characterization of pachytene piRNAs during mouse spermatogenesis

3.1 Introduction

piRNAs are 23–35 nt small RNAs that are abundant during germline development. Drosophila piRNAs often map to transposons and can protect the germline genome by suppressing transposons. Mouse piRNAs are abundant in testis and mutating important piRNA pathway genes often causes male sterility. Mouse piRNAs can be divided into two waves: prepachytene and pachytene piRNAs (X. Z. Li et al., 2013). Dedicated loci in the mouse genome give rise to pachytene piRNA precursors, which are subsequently processed into mature piRNAs. Most of these piRNA-producing loci are depleted of transposons and piRNAs often map to non-transposon regions in the genome, suggesting that they may have functions other than transposon suppression. Previously, studies of such piRNAs have come to different and sometimes contradictory conclusions (Goh et al., 2015; Gou et al., 2014; Vourekas, Alexiou, Vrettos, Maragkakis, & Mourelatos, 2016). Some conclude that these piRNAs find their targets in a sequence-specific manner, requiring certain level of complementarity while others conclude that piRNA sequences are not important. To better understand the function of pachytene piRNAs, the 5 most productive piRNA clusters were knocked down using CRISPR. Removing piRNAs may reveal a phenotype that provides a clue as to their function.

3.2 Methods

3.2.1 Experiment design

To determine targets of pachytene piRNAs, three types of high-throughput sequencing data were extensively used: degradome-seq, small RNA-seq and RNA-seq. Presumably, piRNAs—like other types of small silencing RNAs, such as miRNAs and siRNAs—can cleave the target mRNAs and leave cleavage products with 5′-monophosphate. Such degraded RNAs can be enriched and sequenced using degradome-seq, which provides crucial clues to identify potential targets. Small RNA-seq provides the identity of pachytene piRNAs, important clues for figuring out the guides. RNA-seq profiles stable expression levels of RNA and can be used to detect differentially expressed genes. These three types of data, obtained for mutant and wildtype mouse testis, were then integrated to predict targets of piRNAs.

3.2.2 Definition of seed and non-seed regions of piRNAs

miRNA targeting rules are well studied (Agarwal, Bell, Nam, & Bartel, 2015; Bartel, 2004; Friedman et al., 2008; Garcia et al., 2011; Grimson et al., 2007, 2008, 2008; Lewis et al., 2003). Currently, multiple metrics were used to score miRNA:target relationships, with the most important one being seed matching. Seed matching requirements for miRNAs vary but positions 2–7 of miRNAs almost always require full complementarity. Since miRNAs and piRNAs are both bound Ago-clade proteins with

similar structures (Matsumoto et al., n.d.), it is reasonable to assume that piRNAs also require positions 2–7 to be fully complementary to targets.

Previous biochemistry experiments surveying the pairing requirements of piRNAs provide some clues as to the non-seed region (Reuter et al., 2011). piRNA positions 2–21 are critical for targeting, as a mismatch in these positions causes piRNAs to abolish the cleavage activity, whereas mutations in positions beyond 21 have little effect on piRNA targeting. Thus, in this study, positions 8–21 are defined as the non-seed region.

3.2.3 Determination of piRNA targets

First, all potential targets were extracted from degradome-seq data. Degradome-seq reads were mapped to the genome using parameters previously described (Han et al., 2014). Only 5′ ends of degradome-seq reads were aggregated for each genomic position and were considered as potential cleavage sites (required >1 RPM). Such potential cleavage sites were extended by 50 nt upstream and downstream to serve as the sequences that piRNAs may map to. Next, all potential guide piRNAs were determined by obtaining abundant piRNA species from small RNA-seq (RPM >1).

Guide:target pairs were then determined by requiring perfect seed matches and reporting the number of matches in the non-seed region. Importantly, the offset of 10 nt was not required (i.e. Ping-Pong signature is not required) to provide the background for calculating Ping-Pong Z-scores. The targets were further stratified by the features of interest, e.g. the number of GU wobbles, folding energy, and the number of perfect matches. To validate these guide:target pairs using independent datasets, transcript abundance was quantified using RNA-seq data. Specifically, genes that contain good

piRNA target sites were grouped to compare with genes with poor or no piRNA target sites.

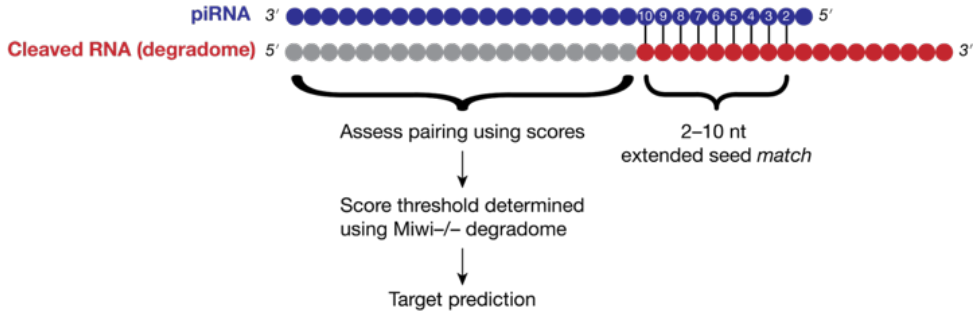

Figure 3.1. Schematics of piRNA target discovery

3.3 Results

3.3.1 piRNA loci and piRNAs are depleted of repeats

Fly piRNA loci are mostly transposons and other repeats. To check if it holds for mouse piRNA loci, repeat levels of piRNA loci were checked. Compared to the genome background, piRNA loci are depleted of repeats (Figure 3.2), suggesting that the main function of these piRNAs is not transposons suppression. However, it is still possible that piRNAs produced from the repetitive regions in these loci are enriched. To determine if piRNAs frequently map to repeats, the proportions of piRNAs mappable to repeats were calculated for each piRNA loci, which demonstrate that piRNAs often map to non-repetitive regions (Figure 3.2). Both lines of evidence suggest that pachytene piRNAs have functions other than transposon suppression.

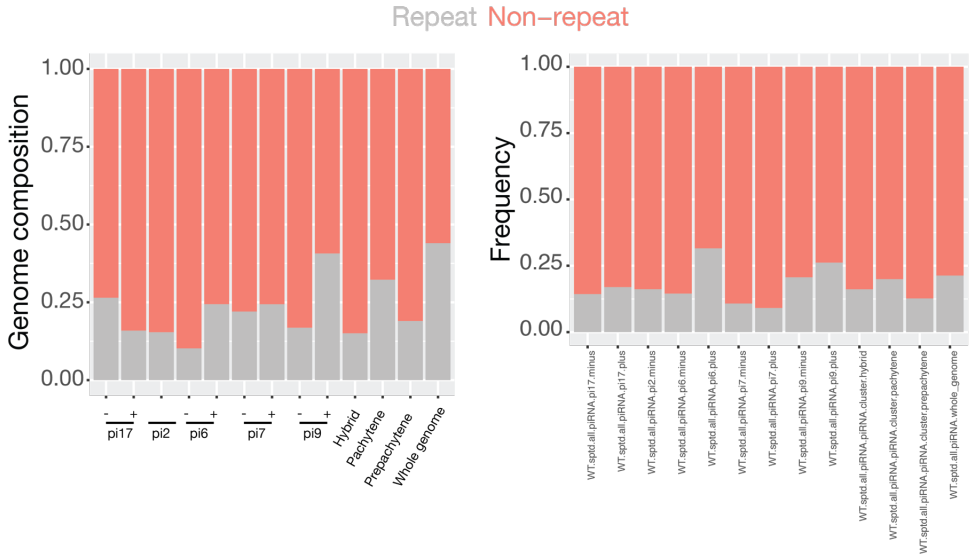

Figure 3.2. Repeat levels for piRNA-producing loci (left) and proportions of piRNAs mapping to repeats (right).

3.3.2 trans-Ping Pong analysis of pachytene piRNAs

To determine if cleavage products can be identified, all predicted targets with ≥10 matches in the non-seed region (i.e. perfect matches of the seed [position 2–7] and ≥10 matches in non-seed [position 8–21]) were obtained to calculate the Ping Pong signal. The idea is simple: if the predicted targets are truly piRNA targets, then one should observe more targets when the 5′ ends of degradome-seq reads were used; on the contrary, when the 5′ ends of degradome-seq reads were shifted, then one should observe a depletion of targets. Figure 3.3 indicates strong signals when 5′ degradome-seq reads were used ($Z = 32.4$, $p < 6 \times 10^{-255}$). In conclusion, degradome-seq reads are enriched for piRNA targets and provide important clues as to the exact cleavage position of piRNAs.

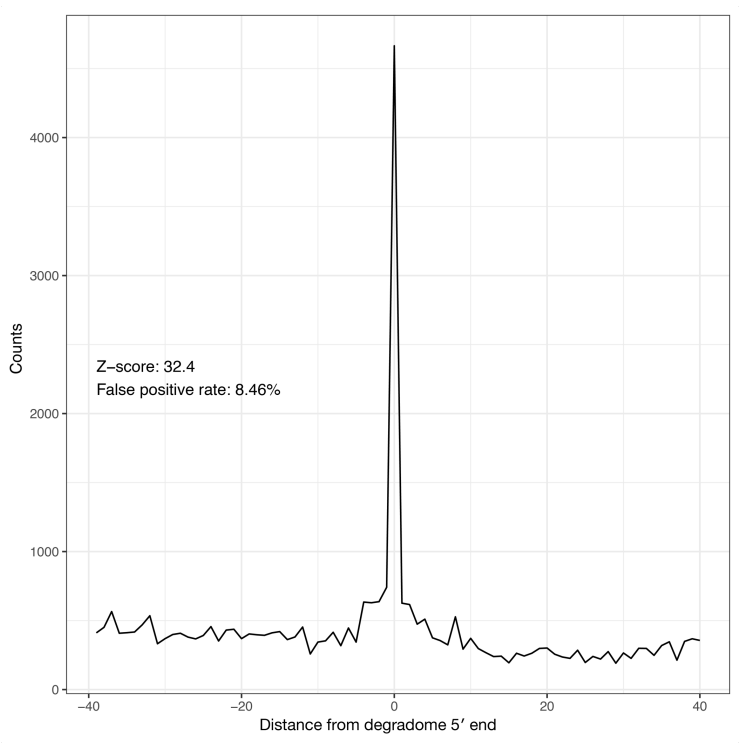

Figure 3.3. *trans*-Ping Pong signals for targets with good matched in the non-seed region. Enrichment at x = 0 indicates that piRNA targets are enriched at the 5′ ends of degradome-seq reads.

The trans-Ping Pong analysis also provides an opportunity to determine the false positive rates: $\frac{y_0-\overline{y}}{y_0}$. Grouping targets by the number of non-seed matches reveals false positive rates for different stringencies: 56.1% for 9 or more non-seed matches, 25.1% for 10 or more non-seed matches, 8.46% for 11 or more non-seed matches. This reflects the trade-off between stringencies and number of targets: more strict cutoffs are more likely to reveal highly confident targets, but reveal fewer piRNA targets. To further test the existence of trans-Ping Pong, degradome-seq and small RNA-seq data from rat testis were used to perform similar analysis. This analysis revealed the similarly significant *trans*-Ping Pong signals ($Z = 28.5$, $p < 6 \times 10^{179}$), indicating that piRNAs can cleave targets via *trans*-Ping Pong in both mouse and rat.

3.3.3 More non-seed matches lead to better cleavage

To evaluate the piRNA cleavage, changes of cleave products were examined. A bona fide piRNA cleavage product should decrease or disappear in a piRNA gene mutant. To check this, each potential cleavage targets was evaluated by ratios of degradome-seq signals in mutant over in wildtype. Predicted piRNA target sites tend to have lower degradome signals in Miwi mutant compared to wildtype (Figure 3.4). Interestingly, when targets are grouped by their complementarity with guide piRNAs (i.e. number of non-seed matches), targets with better complementarity (e.g. 13 or 14 non-seed matches) show more "shifts" than targets with less complementarity (e.g. 10 or 11 non-seed matches). This further confirms the previous biochemistry experiment showing that piRNAs require extensive complementarity in the positions 2–21 (Reuter et al., 2011).

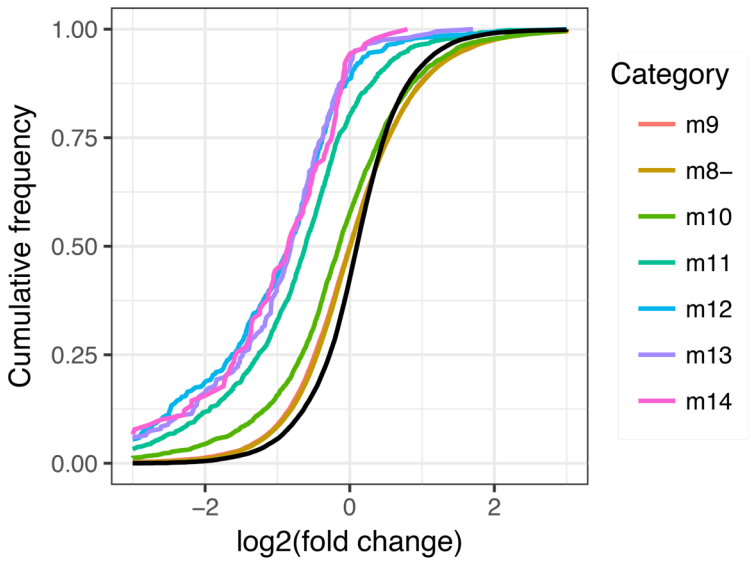

Figure 3.4. piRNA target sites have lower degradome-seq signals in the Miwi mutant. X-axis indicates the log2(fold change) of reads from predicted target sites (mutant / heterozygous).

GU wobbles are known to enhance miRNA targeting. To check if the same applies to piRNA targeting, we grouped targets according to the number of GU wobbles (0, 1, 2) while keeping constant the number of matches (10) (Figure 3.5). Indeed, target sites with more GU wobbles have better degradome-seq response, indicating that GU wobbles—compared to mismatches—enhance piRNA targeting, similar to miRNA targeting. To further check if GU wobbles are equally effective with matches, we grouped targets according to the number of GU wobbles (0, 1, 2) while keep constant the total number of matches and GU wobbles (11) (Figure 3.5). When matches are replaced by GU wobbles, the cleavage efficiency go downs. In summary, we conclude that GU wobbles are better than mismatches and worse than matches.

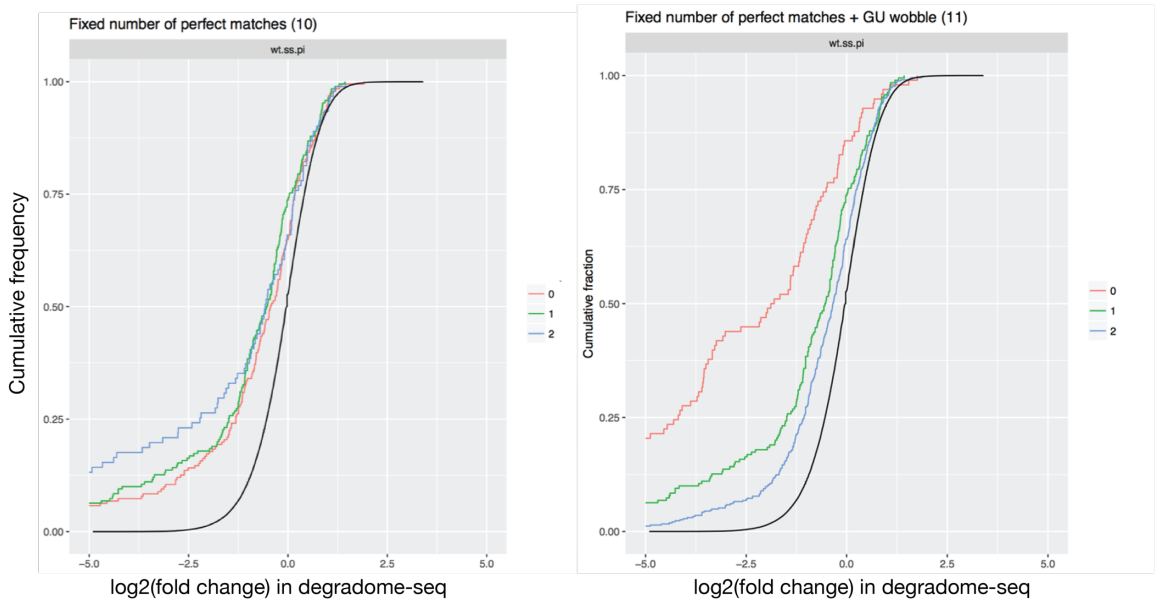

Figure 3.5. GU wobbles are better than mismatches for piRNA targets. X-axis indicates the log2(fold change) of reads from predicted target sites (mutant / heterozygous).

3.4 Discussion

By integrating degradome-seq and small RNA-seq data, we found that pachytene piRNAs can cleave their targets when there exists extensive complementarity. Better complementarity leads to better cleavage, suggesting that extensive complementary promotes target cleavage. Pairing at position 1 is not required, but if the first nucleotide is U, it enhances cleavage. We also found that GU wobbles are better than mismatches, though they are not as good as perfect matches. Although we do not have the experimental evidence, but we speculate that in the case of less extensive complementarity, piRNAs may still bind the targets but not cleave the targets. In summary, pachytene piRNAs can regulate genes during mouse spermatogenesis.

Chapter 4. Genome-wide identification and characterization of branch points in human and mouse

4.1 Introduction

The majority of genes in higher vertebrates contain introns. When genes are transcribed to produce transcripts, introns are removed in a process called splicing and exons are joined to form the mature mRNAs to direct protein synthesis. Introns and exons are well annotated by both automated processes and manual curation, thanks to gene prediction algorithms and massive numbers of RNA-seq datasets. During the early steps of splicing, spliceosome ligates each of 5′ introns to a branchpoint via *trans*-esterification to form a circular structure called a lariat. The spliceosome can subsequently recognize the downstream 3′ splice site and excise the intron lariat via another *trans*-esterification process. Branchpoints are important signals for splicing and mutations at branchpoint can often cause disease (Khan et al., 2004; M. Li, Kuivenhoven, Ayyobi, & Pritchard, 1998; Padgett, 2012), so mapping branchpoints is a critical step to better understand genes. However, in contrast to exons and introns, branchpoints are poorly annotated. This is partially because of the difficulty to computationally predict the branchpoints, which have high sequence degeneracy.

Previously, efforts have been made to annotated branchpoints by exploiting the rare reads that traverse 5PRME splice site/branchpoint junction (Taggart et al., 2017; Taggart, DeSimone, Shih, Filloux, & Fairbrother, 2012), but these did not make use of all RNA-seq data available, limiting the completeness of branchpoint annotation. Experimental methods, such as CaptureSeq (Mercer et al., 2015), to enrich such rare

reads are highly efficient in terms of producing branchpoint-supporting reads but these were done in limited tissues/cell types, and cannot map branchpoints of genes with little or no expression in the surveyed tissue/cell type.

To comprehensively annotated branchpoints, I screened >1.2 trillion RNA-seq reads from ENCODE and NCBI SRA and determined the genomic positions of ~150k branchpoints for both human and mouse, forming the largest catalog of branchpoints to date. To facilitate queries and visualization of these branchpoint, I built a database and a website that can quickly return informative results.

4.2 Methods

4.2.1 The branchpoint discovery pipeline

The first step is to filter out mappable reads, since the vast majority of RNA-seq reads do not derive from the 5′ splice site/branchpoint junctions. STAR was used with the following parameters: `--runMode alignReads --runThreadN $CPU --outFilterScoreMin 0 --outFilterScoreMinOverLread 0.89 --outFilterMatchNmin 0 --outFilterMatchNminOverLread 0.89 --outFilterMultimapScoreRange 1 --outFilterMultimapNmax -1 --outFilterMismatchNmax 10 --outFilterMismatchNoverLmax 0.05 --alignIntronMax 0 --alignIntronMin 21 --outFilterIntronMotifs None --genomeLoad NoSharedMemory --outSAMunmapped None --outReadsUnmapped Fastx --outSJfilterReads Unique --seedSearchStartLmax 20 --seedSearchStartLmaxOverLread 1.0 --chimSegmentMin 20`. The

unmapped reads were kept for next steps. Since many submitters do not include strand specificity of their RNA-seq datasets, strand specificity was determined by first mapping 100,000 reads of each fastq file to the genome and then calculating the ratio of reads mapping to the sense vs the antisense strands. If the ratio is ≥2, the RNA-seq reads are considered to have derived from the transcripts; if the ratio is ≤0.5, the RNA-seq reads are considered to have derived from antisense strands of RNAs; if the ratio is between 0.5 and 2, then the RNA-seq dataset is not strand-specific.

In the second step, unmapped reads were screened to obtain those that traverse the 5′ splice site/branchpoint. Unmapped reads were mapped to the 5′ introns using bowtie2 with parameters: `--local --score-min L,45,0 -D 20 -R 2 -N 0 -L 20 -i L,1,0`, which ensures the sensitivity by trying all possible seeds for sequence alignment. Next, portions of reads mappable to 5′ introns are clipped and the remaining portions are mapped to the genome. Then the alignments were further filtered by requiring that both portions map to the same intron in Gencode annotation. The overview of this pipeline is visualized in Figure 4.1.

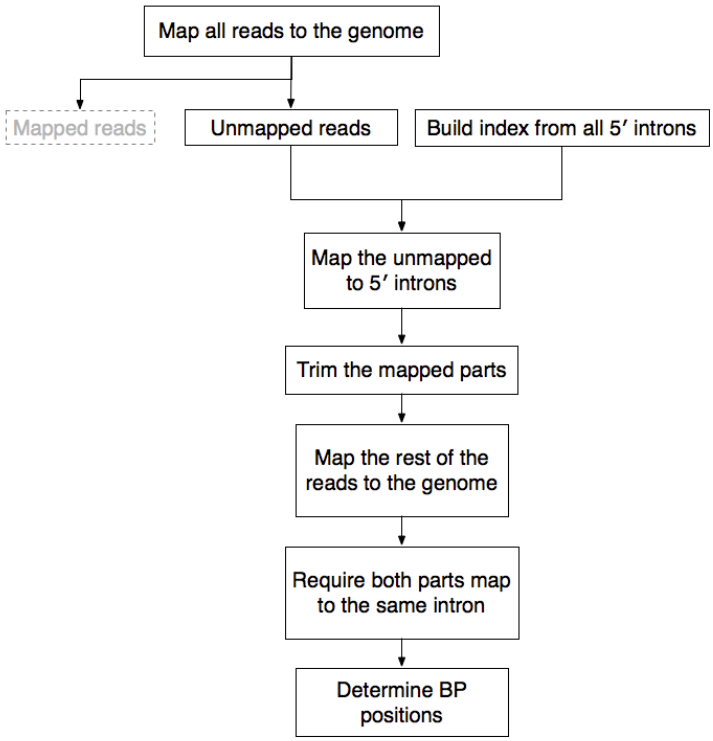

Figure 4.1. Workflow of the branchpoint discovery pipeline.

Since a read seemingly support a branchpoint may come from a regular RNA-seq reads with a few mutations, extra efforts were made to exclude such bogus reads. According to the branchpoint and corresponding 5′ splice sites from the previous step, the potential lariat sequences were constructed and then mapped to the genome. If a potential lariat can be mapped to the genome within certain edit distances, it is removed.

4.2.2 Alternative splicing analysis

Alternative splicing events were extracted using an R package "SplicingGraphs". Exon skipping, alternative acceptor, alternative donor and intron retention events were extracted using '0,1-2^', '1-,2-', '1^,2^', '0,1^2-'. To quantify the strength of splice sites, their seqlogos and k-mer were compared.

4.2.3 Database schema

SQLite was used to store all branchpoint-related information. A total of 7 tables were used: bp (basic branchpoint information, such as coordinates and flanking sequences), intron (including intron coordinate, intron ID, parent transcript ID and parent gene ID), exon (including exon coordinate, exon ID, parent transcript ID and parent gene ID), gene (gene coordinate, gene names, species, etc.), transcript (transcript coordinate, transcript ID, and gene ID), species (describing species names and IDs), and bp_src (describing the source datasets of branchpoints). Specifically, the bp table was built first as three smaller tables, each of which holds the branchpoint ID. These three tables were later joined to produce the final bp table. For the SQLite script to create the database, see Appendix C.

4.2.4 Website

The website was built using the Python package Flask, a micro web framework. The front-end framework is Bootstrap with customized styles. User queries were parsed by Python scripts and results were return in the JSON format. The tables were implemented using Bootstrap Table. Data visualization was performed using D3.js on the client end.

4.3 Results

4.3.1 Branchpoint annotation and characterization

By screening for reads traversing 5′ splice site/branchpoint junctions, 153,303 human and 148,282 mouse branchpoints. Examination of these branchpoint revealed the

TnA motif and polypyrimidine tract (PPT) (Figure 4.2), consistent with previous characterization of branchpoints (Gao, Masuda, Matsuura, & Ohno, 2008).

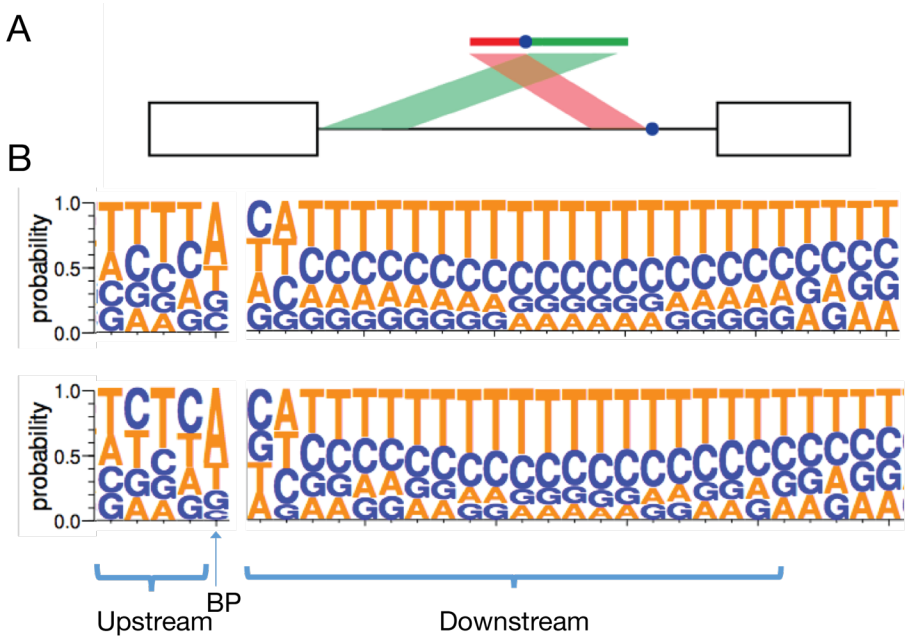

Figure 4.2. Overview of branchpoints. A. Schematic of a lariat-supporting read mapped to the genome. The blue dot indicates the position of the branchpoint. B. SeqLogo showing the motifs at and around branchpoints.

To determine the frequency of lariat-supporting reads, the ENCODE RNA-seq datasets were used, since these data have better metadata (e.g. tissue, strand-specificity, and cellular compartment) that facilitate the analysis. To quantify the frequency, we calculated the number of unique branch points per billion reads (BPB), for each library. Human RNA-seq datasets generated a median of 607 BPB whereas mouse RNA-seq datasets generated a median of 365 BPB, indicating that lariat-supporting reads are rare and that a massive number of reads are required to obtain a comprehensive branchpoint annotation.

To quantify which set of libraries are the most informative in terms of mapping branchpoints, ENCODE RNA-seq data were grouped by biosample type and were

quantified using the BPB measure. In total, there are 1,474 categorized as “immortalized cell line”, 40 categorized as “in vitro differentiated cells”, 12 categorized as “induced pluripotent stem cell line”, 413 categorized as “primary cell”, 57 categorized as “stem cell”, and 138 categorized as tissue. Stem cell RNA-seq data were the most informative group whereas the immortalized cell RNA-seq data were the least informative group (Figure 4.3A). Stem cells are more likely to actively produce nascent transcripts, and thus have more lariats accumulate in cells. The ENCODE RNA-seq data can be further grouped by cellular compartment. RNA extracted from nucleus are more likely to support lariats and branchpoints, compared to RNA from cytosol (Figure 4.3B).

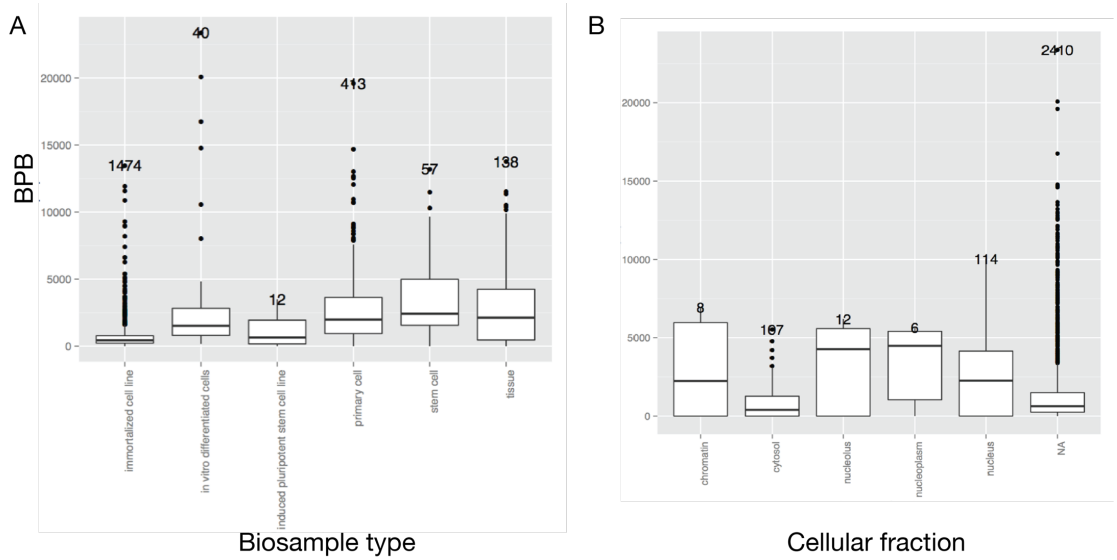

Figure 4.3. Number of unique branch points per billion reads (BPB) grouped by (A) biosample and (B) cellular fraction.

It is known that branchpoints are often proximal to the 3′ splice sites. The large number of branchpoints in this study provides an opportunity to examine the distribution of the distance. The median distance from a branchpoint to the closest downstream exon

is 26 nt, with 87.9% of branchpoints are within 50 nt upstream of the downstream exon introns and 91.1% of branchpoints are within 100 nt of the downstream exon, indicating that the majority of branchpoints are close to 3′ splice sites. Some branchpoints overlap with the 3′ splice sites (x = 1 in Figure 4.4), likely reflecting circular RNAs formed by 5′ introns directly ligated to 3′ introns.

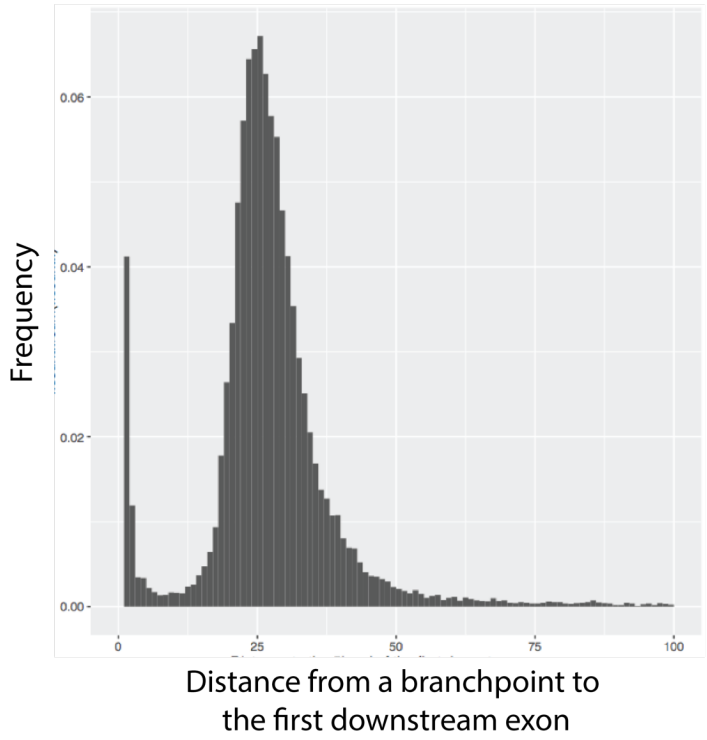

Figure 4.4. Distance from a branchpoint to the first downstream exon (i.e. distance from a branchpoint to the closest 3′ splice sites + 1)

Grouping branchpoints by the 5′ splice site sequence, we found that the most common 4 types of 5′ splice sites are GT, GC, AT, and GA. Branchpoints with GC, AT, or GA as the 5′ splice sites tend to stay closer to the next exon, which likely reflects that, compare to GT as the 5′ splice sites, these three types of 5′ splice sites are not as effective, and thus need to be closer to the 3′ splice site to allow splicing.

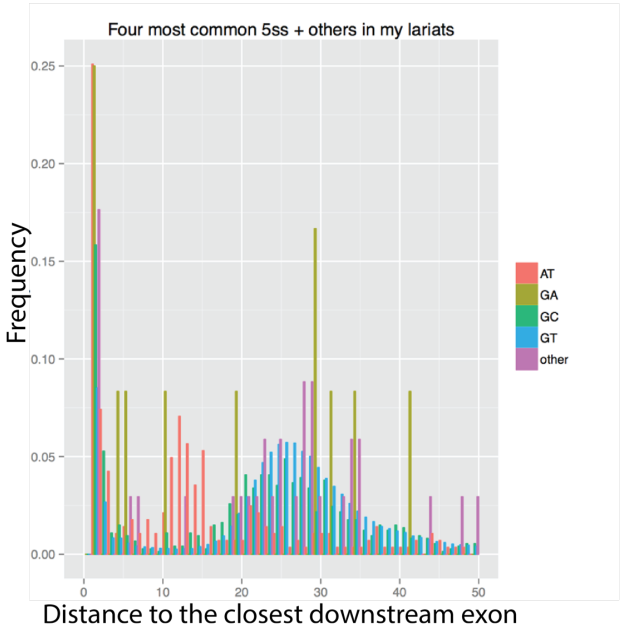

Figure 4.5. Distance from a branchpoint to the closest downstream exon, grouped by the 5′ splice site sequence.

Although AT rarely serves as the 5′ splice sites when all introns are taken into consideration, AT often is the 5′ splice for a group of introns (U12 introns) that uses the minor spliceosome for splicing. Thus, a set of known U12 introns were retrieved from U12DB (Alioto, 2007), and lifted over to the genome assemblies used in this study. Comparing to other introns, branchpoints of U12 introns are much closer to the next exon (Figure 4.6), suggesting that U12 introns have lower splicing efficiency and evolve their branchpoints to lie closer to 3′ splice sites.

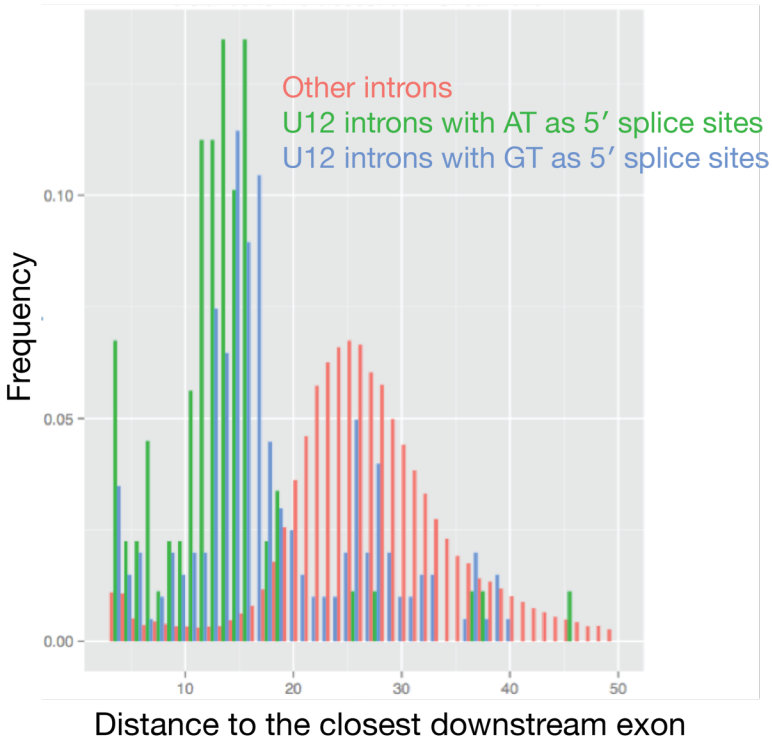

Figure 4.6. Distance from a branchpoint to the closest downstream exon, grouped by intron type (U2 and U12). U12 introns were further grouped into those with AT and GT as 5′ splice sites.

4.3.2 Branchpoints and alternative splicing

Four most common types of alternative splicing involve alternative donor sites, alternative acceptor sites, exon skipping and intron retention. We speculate that branchpoints may play a role in alternative splicing. To determine the relationship between alternative splicing and branchpoints, we extracted and compared branchpoints involved in the aforementioned 4 types of alternative splicing events.

At least 3 branchpoints (ES.12, ES.14 and ES.34) are used for exon skipping (Figure 4.7). To make the branchpoints comparable, we only consider exon skipping events where the trios of branchpoints were determined (complete cases). In total, 368 ES.12, 356 ES.14, and 463 ES.34 events were found. In term of distance to the closest

downstream exons, these 3 types of branchpoints have very similar distributions, suggesting that the distance to the closest downstream exon does not play a role in determining alternative splicing. However, in terms of splicing signals, these 3 types of branchpoints are different (Figure 4.7): branchpoints in ES.14 introns show the strongest TnA and PPT signals. The likely explanation is that the ES.14 intron—the longest intron in exon skipping—requires stronger signals at and downstream of branchpoints to compensate for the length of the intron. A further comparison of branchpoint strengths (measured as the frequency of the canonical TnA motif) for 3 groups of introns revealed that E14 does possess stronger branchpoints, compared to all branchpoints.

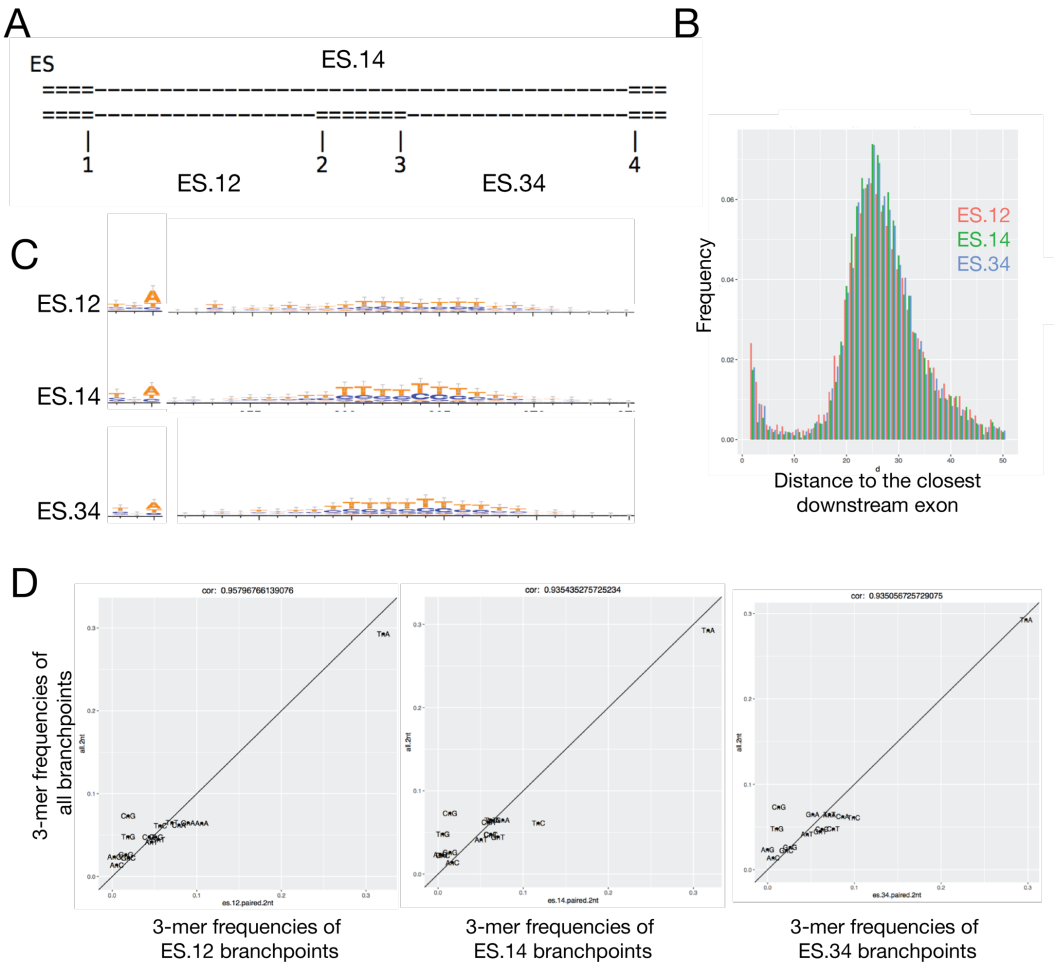

Figure 4.7. Comparison of branchpoints involved in exon skipping event. (A) A schematic of exon skipping. (B) Distance from branchpoints to the closest downstream exon, grouped by three types of introns in exon skipping. (C) Comparison of the branchpoint motif. (D) Comparison of 3-mer frequencies at and upstream of branchpoints in 3 types of branchpoints.

We then examined branchpoints in introns with alternative donor sites. Introns with alternative donor sites use at least two branchpoints (for AD.2 and AD.3 type introns, see figure 4.8). In total, we were able to determine 675 AD.2 type branchpoints and 737 AD.3 type branchpoints. Branchpoints in these two types of introns share similar distance distribution (Figure 4.8). However, the branchpoint signals for AD.2 is significantly

stronger than those for AD.3 ($p < 0.0001$), suggesting that longer introns (AD.2) requires stronger branchpoint signals to compete with AD.3 introns.

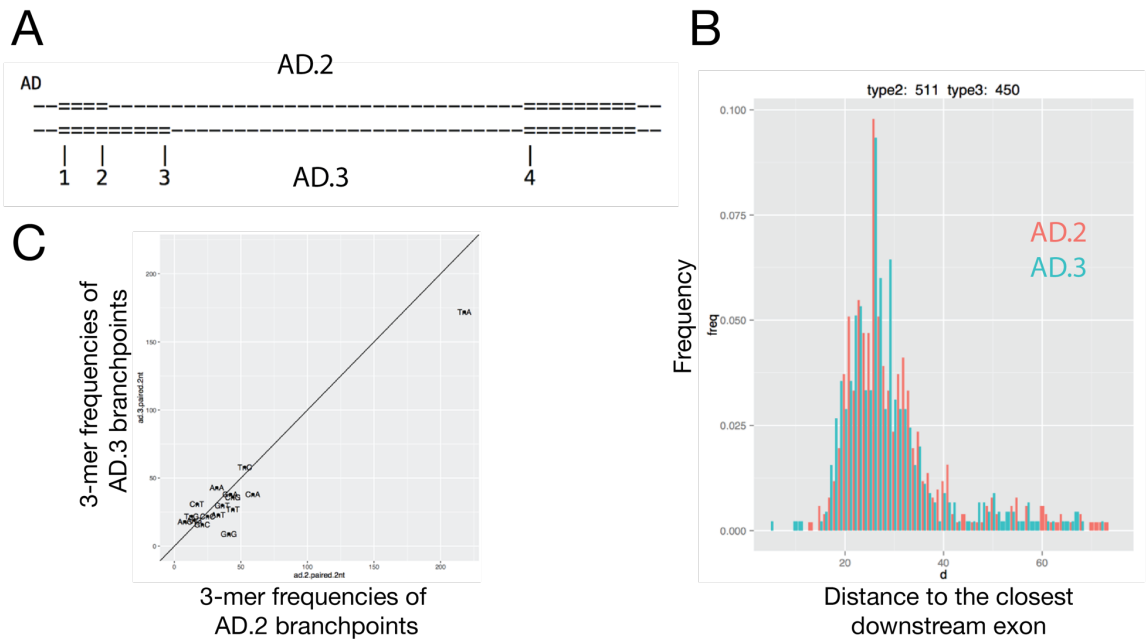

Figure 4.8. Comparison of branchpoints in introns with alternative donor sites. (A) A schematic of introns with alternative donor sites (AD.2 and AD.3). (B) Distance from branchpoints to the closest downstream exon, grouped by intron type (AD.2 and AD.3). (C) Comparison of 3-mer frequencies at and upstream of branchpoints in AD.2 and AD.3 types of branchpoints.

Next, we examined branchpoints in introns with alternative acceptor sites. Introns using alternative acceptor sites involve at least two branchpoints. Similar to the exon skipping analysis, only complete cases were examined. In total, 198 AC.2 and 211 AC.3 branchpoints were found. Branchpoints in AC.2 and AC.3 introns share similar distributions of distances to the closest downstream exons (Figure 4.9). Comparison of branchpoints in these types of introns versus all branchpoints revealed that shorter introns (i.e. AC.2 introns) possess stronger branchpoints. We currently do not understand how to explain this counterintuitive result.

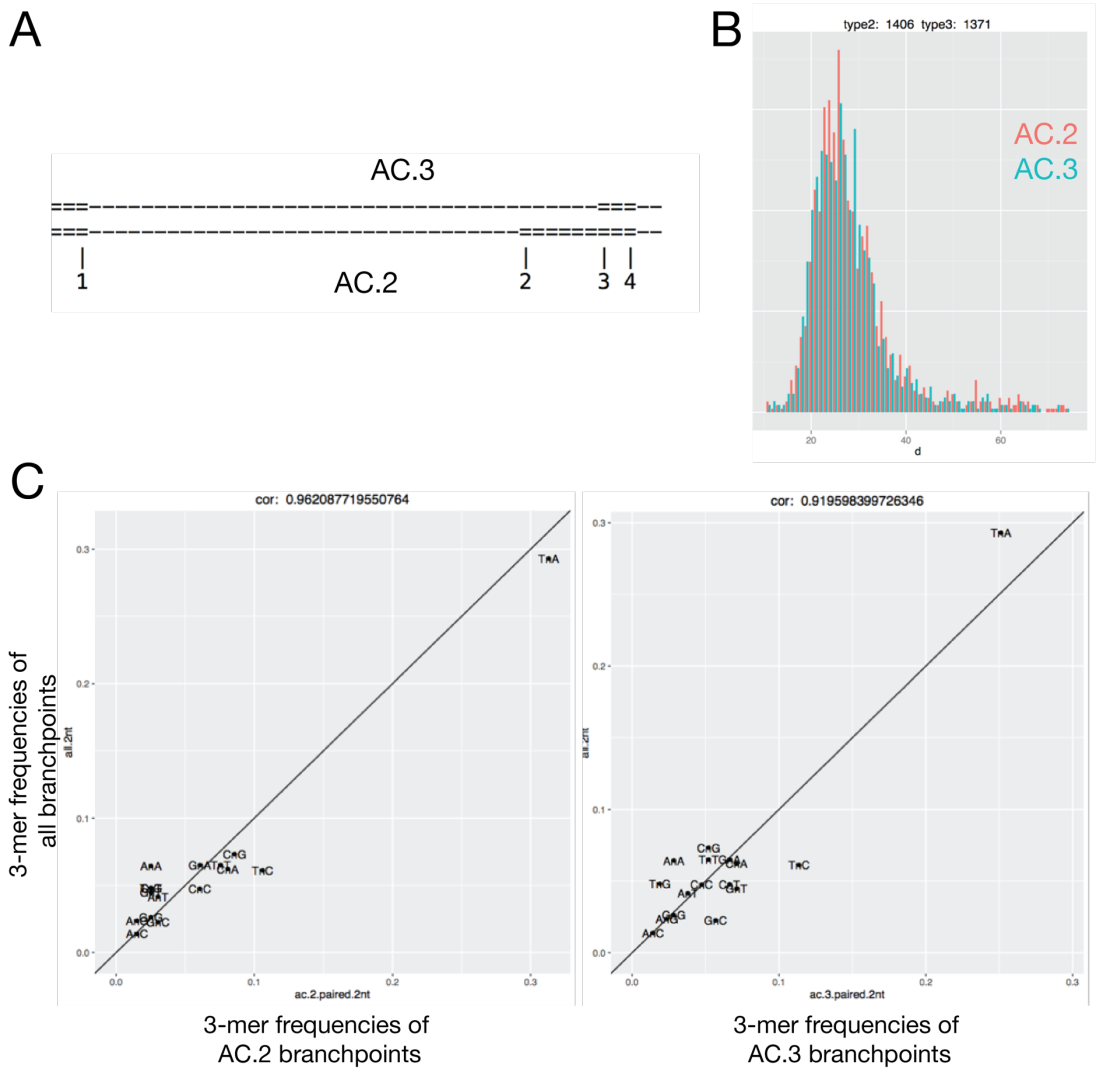

Figure 4.9. Comparison of branchpoints in introns with alternative acceptor sites. (A) A schematic of introns with alternative acceptor sites (AC.2 and AC.3). (B) Distance from branchpoints to the closest downstream exon, grouped by two types of introns (AC.2 and AC.3). (C) Comparison of 3-mer frequencies at and upstream of branchpoints in AC.2 and AC.3 types of branchpoints.

Some introns can be retained in mature mRNAs. Each of such alternative splicing event only involve one optionally retained intron. We then compared branchpoints in retained introns versus all introns, which revealed that branchpoints in retained introns

have significantly weaker signals ($p < 0.0001$). We speculate that the reason of these introns being retained is that weaker splicing signals result in inefficient splicing.

4.3.3 The website

To facilitate queries and visualization, we built a web application with user-friendly interface. There are some considerations for this web application: a) the database backend should be lightweight and portable, allowing easy manipulation in the future; b) the frontend should present an intuitive and responsive interface; c) graphs should be rendered on the client-side to enable instant response and reduce the burden on the server. To meeting these, I built a SQLite database as the backend, used Flask (http://flask.pocoo.org/) to serve webpages using Bootstrap (https://getbootstrap.com/). For data visualization, I used D3.js, which provides easy-to-use and high customization function for plotting. The web app was developed and tested locally, and then deployed into a Docker container on a Weng Lab server.

The main functions include a dynamic table that, upon a users' query, returns branchpoint information, including the chromosome, coordinate, strand, base at the branchpoint, splicing donor site coordinate, and distance to the closest downstream acceptor site (Figure 4.10). Columns are customizable: branchpoint ID, distance to the splice donor site, downstream and upstream sequences, gene ID, transcript ID and intron ID. This table can be downloaded as JSON, CSV, XML, TXT, and EXCEL formats, allowing users to perform downstream analysis. The information panel on the right dynamically show detailed information about the one branchpoint clicked by the user. Currently, BPDB provides two graphs: nucleotide frequencies of bases at and flanking

branchpoints and distribution of distances from branchpoints to closest downstream 3′ splice sites. These two graphs are immediately updated once the selection of branchpoints has any changes (e.g. when the user queries branchpoints for a particular gene, or deselect a branchpoint in the table). In some cases, the user might not be interests in the graphs and information panel, so I added switches for these, which can provide a clean interface just containing the table. A user may be interested in just one particular gene, so I added search box to allow searches using partial and full matches of gene names.

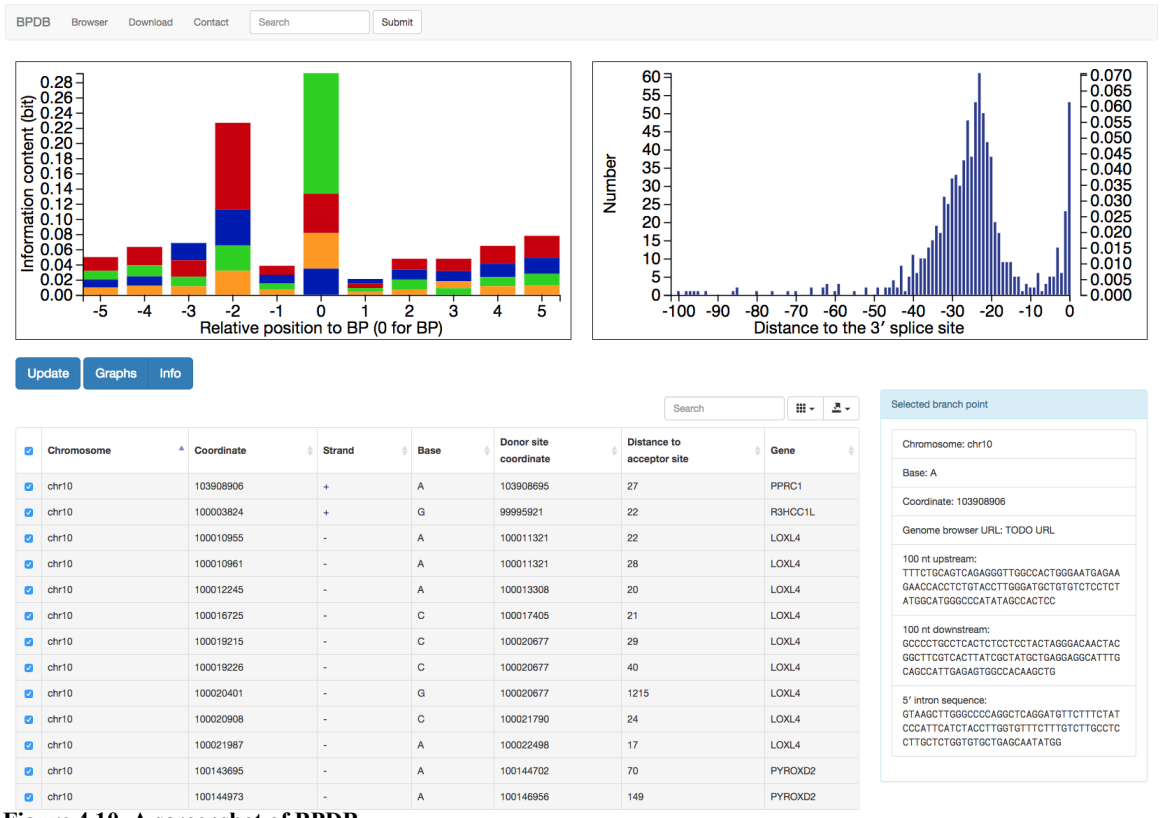

Figure 4.10. A screenshot of BPDB.

4.4 Discussion

Branchpoints are critical for RNA splicing, yet they are difficult to predict due to the high sequence degeneracy. Here, we mapped ~150,000 branchpoints for both human

and mouse genomes, which, our knowledge, is the most comprehensive catalog of branchpoints in human and mouse. We have built a highly efficient computational pipeline that screened >1.2 trillion reads from more than 40,000 RNA-seq datasets. Examination of these branchpoint reveals multiple branchpoint features, such as the proximity to their splice acceptor sites and canonical TnA motif. The large number of branchpoints also enable investigation of relationships between branchpoints and alternative splicing. In summary, longer introns in alternative splicing events require stronger branchpoint signals, whereas shorter introns and retained introns possess weaker branchpoint signals. To allowing easy queries and data visualization, I built the BPDB to provide a comprehensive branchpoint catalog. This resource should be valuable to biologists who need to manipulate introns by determining or mutating the branchpoints. Also, abnormal splicing can cause human disease (Singh & Cooper, 2012), including cancer (Yoshida et al., 2011). Mapping branchpoints is the first step towards a better understanding of these diseases.

Chapter 5. Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers

5.1 Introduction

High-throughput sequencing of RNA provides a quantitative measure of RNA abundance. However, library construction of RNA-seq and small RNA-seq can introduce bias at multiple steps, such as fragmentation of long RNAs, adapter ligation, PCR, and sequencing. Starting material is usually scarce, so PCR amplification is required during library construction to increase the number of cDNA molecules to an amount sufficient for sequencing. However, PCR randomly introduces errors that can propagate to later cycles (Cha & Thilly, 1993; Dohm, Lottaz, Borodina, & Himmelbauer, 2008). PCR also over- and under-amplifies certain molecules (Cha & Thilly, 1993). PCR duplicates are defined as reads made from the same original cDNA molecule via PCR.

A common method of PCR duplicate elimination is to remove all but one read of identical sequences, assuming that such identical reads have been created from the same cDNA molecule by PCR (e.g. samtools (H. Li et al., 2009)). This assumption may be flawed, especially with higher sequencing throughput, which increases the chance of observing reads with identical sequences originating from different cDNA molecules. The situation is even worse for small genomes (in which the genome coverage is substantially high) and for techniques that interrogate a subspace of the genome (e.g. small RNA-seq selects small RNAs, which are produced from very limited genomic loci (Brennecke et al., 2007; X. Z. Li et al., 2013)). The assumption is also systematically biased: in RNA-seq, shorter genes are more likely to produce identical reads than longer

genes with the same expression level, simply because the "genomic space" for RNA fragmentation is more limited for shorter genes. Finally, PCR duplicate identification also relies on mapping coordinates. (reads mapping to the exact same genomic location are considered to have identical sequences.) However, small RNAs from different loci (e.g. genomic repeats) can produce the same sequence; thus, strategies using genome coordinates to identify PCR duplicates result in biases for repeat-derived reads.

There are many variables for high-throughput library construction. Some are preset, e.g. PCR and sequencing error rates, but others are variable and depend on the parameters such as the amount of starting RNA used to generate a library, the number of reads sequenced (i.e., sequencing depth), and the PCR cycle number. While it is tempting to believe that more PCR cycles lead to more duplicate reads in high-throughput sequencing data, high PCR cycle numbers are often associated with scarce starting materials, which is another potential cause for PCR duplicate reads. Thus, PCR cycle numbers may be confounded with starting materials and sequence depth.

Unique molecular identifiers (UMIs) are often used to unambiguously and accurately detect PCR duplicates and improve transcript abundance quantification (Collins et al., 2015; G. K. Fu, Xu, et al., 2014; G. K. Fu, Hu, Wang, & Fodor, 2011; G. K. Fu, Wilhelmy, Stern, Fan, & Fodor, 2014; Islam et al., 2014; Kivioja et al., 2012; Shiroguchi, Jia, Sims, & Xie, 2012; T. Smith, Heger, & Sudbery, 2017). The idea is simple: if each molecule before PCR is tagged with a UMI, i.e., all molecules are unique (those molecules with identical sequences are ligated to different UMIs), then reads with the same sequence and the same UMI must be PCR duplicates.

One way to incorporate UMIs into reads is to introduce pre-defined sequences into the adapters. This avoids UMIs with suboptimal GC content and minimize complementarity between or within UMI sequences (Shiroguchi et al., 2012). Because UMI sequences are preset (and different UMIs have large edit distances), erroneous UMIs can be easily corrected to the pre-defined one by calculating edit distance. However, the drawback is that such pre-defined UMIs require a large number of costly, custom-synthesized oligonucleotides, perhaps prohibitive for many labs.

Another strategy uses adapters with random nucleotides at certain positions in the adapters. The length of random nucleotides leads to an exponential number of UMI combinations at almost no extra cost, because incorporating a random nucleotide costs the same as incorporating a specific nucleotide during DNA synthesis. UMIs bearing either 5 (4^5 = 1,024 unique UMIs) or 10 random nucleotides (4^{10} = 1,048,576 UMIs) were implemented cost-effectively and shown to improve PCR duplicate removal (Islam et al., 2014; Kivioja et al., 2012). A higher number of unique combinations can be achieved simply by increasing the length of random nucleotides. The number of UMI combinations must be sufficiently large because the chance that two cDNA molecules with identical sequences in the starting pool are tagged with the same UMI combination needs to be infinitesimally small.

Here, we describe novel experimental protocols and computational methods to unambiguously identify PCR duplicates in RNA-seq and small RNA-seq data. We show that removing PCR duplicates using UMIs is accurate, whereas removing PCR duplicates without UMIs is overly aggressive, eliminating many biologically meaningful reads,

worsening quantification. Finally, both the amount of starting materials and sequencing depth determine the level of PCR duplicates, but PCR amplification does not.

5.2 Methods

5.2.1 Simulation

Simulation procedure was performed similarly to (T. Smith et al., 2017). Briefly, 7 parameters were simulated: PCR and sequencing error rates, PCR amplification probability, UMI length, number of initial molecules, number of sequenced molecules, and number of PCR cycles, by varying one parameter and keeping other parameters constant. For each combination of the 7 parameters, 10,000 replicates were performed. UMI error correction for RNA-seq was implemented as described in (T. Smith et al., 2017). For small RNA-seq, we used read sequences instead of genomic coordinates when determining PCR duplicates. We used NetworkX (https://networkx.github.io/) for graph-related algorithms, and pysam (https://github.com/pysam-developers/pysam) for handling SAM/BAM files. Reads were mapped to the mouse mm10 genome as described in (Han et al., 2014). When reads were analyzed without UMIs, PCR duplicates were identified using Picard (https://github.com/broadinstitute/picard).

5.2.2 Availability

The tools developed for handling UMIs in our RNA-seq and small RNA-seq data can be found at https://github.com/weng-lab/umitools, and via PyPI (package: umitools). RNA-seq and small RNA-seq data have been deposited in the NCBI SRA under the

accession number PRJNA416930. For experimental procedures, see (Y. Fu, Wu, Beane, Zamore, & Weng, 2018).

5.3 Results

5.3.1 Adapting standard RNA-seq procedures to incorporate UMIs

We modified a published RNA-seq protocol in order to incorporate UMIs into strand-specific RNA-seq library construction protocol (Z. Zhang, Theurkauf, Weng, & Zamore, 2012). The original method has widely used for multiple species in multiple labs (X. Z. Li et al., 2013; Mohn et al., 2014; Z. Zhang et al., 2014). The standard protocol uses a single Y-shaped DNA adapter containing two partially complementary oligonucleotides and an unpaired 3′ thymidine that pairs with the single adenine tail added to both ends of the double-stranded cDNA fragments. We modified the adapters by inserting a five-nucleotide random UMI (Figure 5.1). Consequently, each cDNA fragment is ligated to an adapter with a UMI at each end, randomly choosing one out of 1,048,576 ($4^5 \times 4^5$) possible combinations provided by two UMIs.

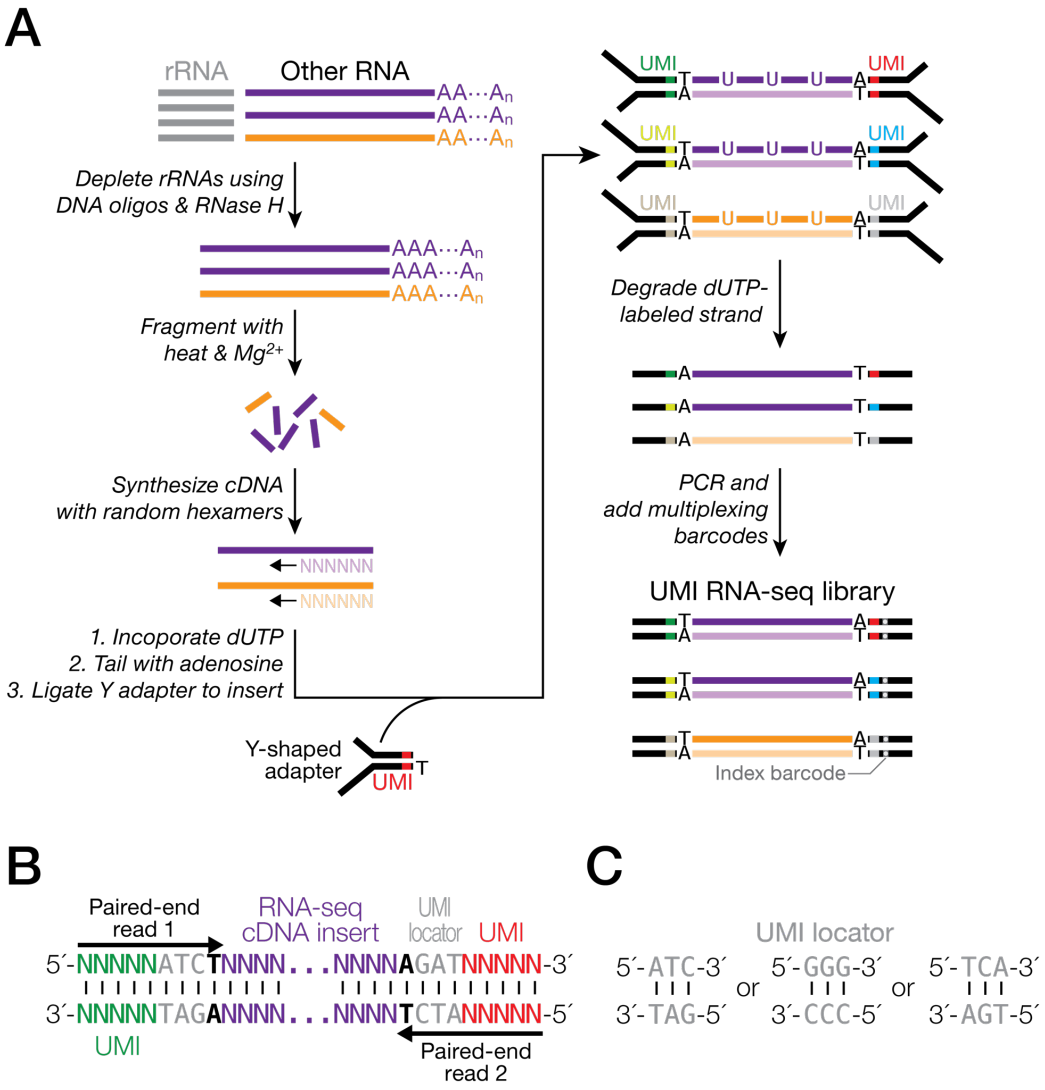

Figure 5.1. UMI incorporation into RNA-seq. (A) Overall workflow. Schematic of a read produced from RNA-seq with UMIs (B) and of UMI locators (C).

Our UMI RNA-seq adapters were designed so that the sequencing reaction begins at the very first nucleotide of the 5′ UMI (Figure 5.1), which guarantees the sequence diversity in the first five sequencing cycles. This is critical for commonly used Illumina sequencing platforms, such as HiSeq, MiSeq, and NextSeq, to accurately call bases (Mitra, Skrzypczak, Ginalski, & Rowicka, 2015). To avoid rare insertions or deletions

within or flanking a UMI from changing the UMI identity, we further designed a "UMI locator", a pre-defined trinucleotide 3′ to the UMI (e.g. 5′–NNNNNATC–3′). The three nucleotides serve as an anchor to allow unambiguous location of each UMI (Figure 5.1). Taking the properties of commonly used sequencing instrument into consideration, the 3 nt UMI locator sequence and the mandatory thymidine required for ligation that immediately follows (Figure 5.1) corresponded to the sequencing cycles 6–9, after the first five critical cycles required by the instrument for template generation. After we sequenced one lane of data using NextSeq, we found that NextSeq still considered these four invariant positions as low-complexity regions and reported N's or low qualities for these bases. Previously, this was solved by mixing the library with other samples or spike-in), or increasing the initial sequence diversity in the library (Mitra et al., 2015). In order to not comprise the sequencing depth, three UMI locator sequences were incorporated (Figure 5.1) and, by mixing 3 adapters with these sequences at equimolar amounts, the library complexity increases and the problem was solved. With this approach, we successfully generated RNA-seq libraries from total RNAs of multiple tissues. The libraries were comparable to libraries generated using the original protocol without UMIs, in terms of read depths, coverage, and qualities comparable to (see (Y. Fu, Wu, et al., 2018)). Thus, incorporating UMIs and UMI locators does not compromise library qualities and sequencing output.

5.3.2 Adapting standard small RNA-seq protocol to incorporate UMIs

Previously, the Zamore Lab has established a robust small RNA-seq protocol by modifying a published method (Lau et al., 2001). Compared to aforementioned RNA-seq

with UMI, UMI incorporation into the small RNA-seq requires some extra considerations. First, the number of distinct UMI combinations needs to be greater than that for RNA-seq, as some highly abundant small RNA species often have huge numbers of reads. For example, one single piRNA species (also the most abundant one) in this study produces 42,281 reads in one of our libraries. The situation is further exacerbated in the soma: the most abundant miRNA can take more than 40% of the total sequencing reads (tens of millions of reads in a typical sequencing run producing hundreds of millions of reads). That many reads have identical sequences but originate from different starting molecules demands a great number of UMI combinations to capture all distinct sequences. Second, the lengths of small RNAs (< 50 nt) plus a longer UMI (20 or even 30 nt) is still well within the limits of common sequencing instruments. Third, the length of a small RNA is a defining feature of its identity and thus, insertions or deletions could lead to misclassification of small RNAs. The latter two considerations also indicate that small RNA-seq is an ideal opportunity to test a large combination of UMIs.

UMIs containing 10 consecutive random nucleotides were first tested. Although both the 3′ and 5′ adapters containing 10 nt UMIs ligated to small RNAs with nearly the same efficiency as the original adapters without UMIs, the resulting small RNA-seq libraries yielded unexpectedly short, variable-length reads that contained truncated insert and adapter sequences (data not shown). We speculate that long stretches of random nucleotides interfere with oligonucleotide annealing, a critical step in cDNA synthesis, PCR, and sequencing, by increasing the chance that a ′r anneals to a UMI instead of its

target sequences. Inter- and intramolecular annealing of 10 nt UMIs may also contribute to truncated reads.

To avoid a long stretch of random nucleotides, we used the UMI locator strategy described above to space out several short stretches of random nucleotides. For each adapter, we designed three trinucleotide UMI sequences, each separated from another by a trinucleotide UMI locator (e.g., 5′–NNN-CGA-NNN-TAC-NNN–3′; Figure 5.2). Two adapters with such UMIs can produce a trillion combinations, which should suffice all deep-sequencing applications. Similar to our RNA-seq strategy, we designed adapters with two different sets of UMI locator sequences at equimolar to increase the sequence complexity in the early sequencing cycles. This strategy allowed us to successfully generate and sequence the UMI small RNA-seq libraries, unambiguously locate UMIs, and computationally remove reads containing insertions or deletions in UMIs due to reverse transcription, PCR, and sequencing errors (Figure 5.2). We tested our method using total RNAs extracted from mouse testes isolated 17.5 days after birth. To assess the impact of the amount of starting materials on PCR duplicates, we prepared small RNA-seq libraries using a range of 39–5,000 ng RNAs made from serial dilution. To test the effect of PCR cycles, we gradually increased the PCR cycles for each library with a two-cycle increment. The resulting UMI small RNA-seq libraries yielded high-quality sequencing data, comparable to those generated with the original non-UMI protocol.

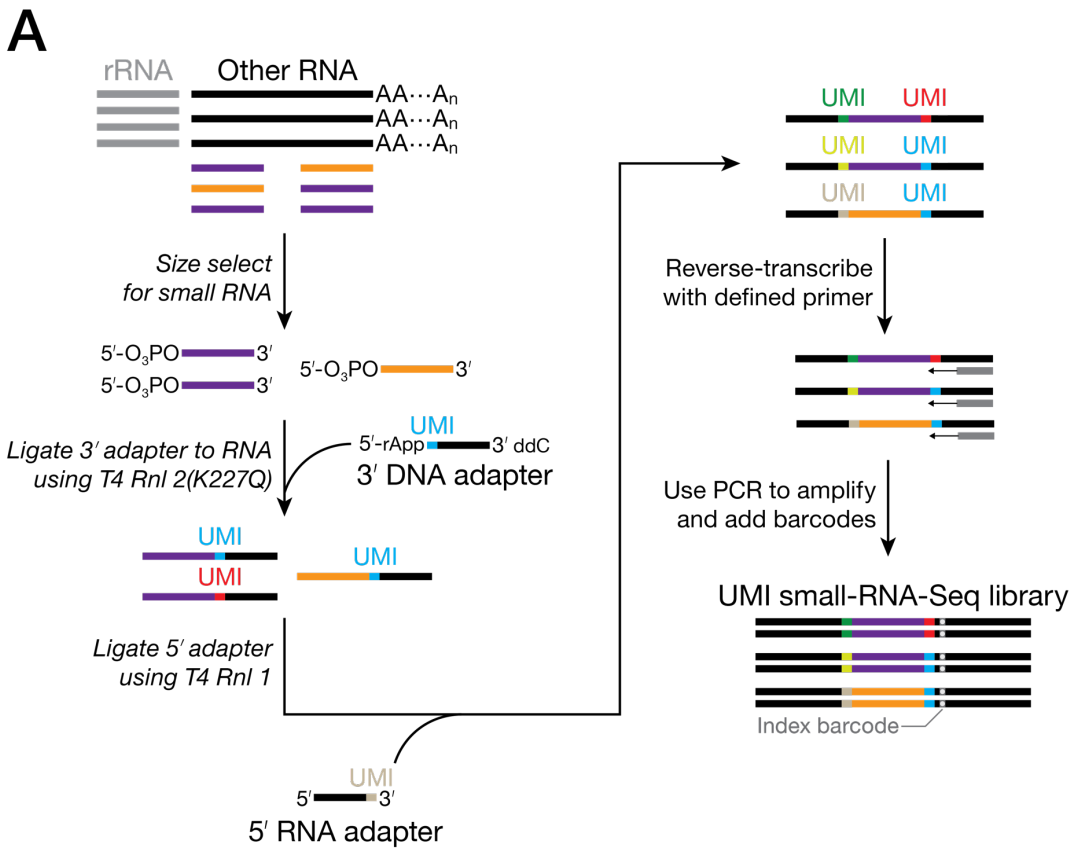

Figure 5.2. UMI incorporation into small RNA-seq. (A) Overall workflow. The method uses a 3′ adapter composed of DNA, except for a single, 5′ ribonucleotide (rA); the 5′ adapter is entirely RNA. A standard index barcode allows multiplexing. (B) Schematic of a read produced from small RNA-seq with UMIs.

5.2.3 Diverse UMIs capture all read species in RNA-seq and small RNA-seq

As mentioned above, to accurately identify PCR duplicates using UMIs, it is critical that the number of distinct UMIs far exceeds the maximal number of starting

molecules with identical sequences, such that these molecules have an infinitesimal probability of being ligated to adapters with the same UMI. Previous UMI methods were designed for sequencing single cells or an organism with a less complex transcriptome than mammals (G. K. Fu, Wilhelmy, et al., 2014; Shiroguchi et al., 2012). In particular, testis has a higher-complexity transcriptome than many other tissues such as muscle, liver, and even brain (Soumillon et al., 2013), demanding a large number of UMI combinations. Our UMI RNA-seq protocol theoretically provides ~1 million (4^{10}) distinct combinations. We then tested whether this diversity far exceeded the maximal number of reads with identical sequences in our libraries. Indeed, the transcripts derived from the 299-bp *7S RNA 1* gene produce 19,271 identical reads mapping to the same genomic coordinate, all of which are attached to distinct UMI sequences, indicating that all of these reads were from different starting RNA molecules. In conclusion, our UMI RNA-seq protocol is more than sufficient to disambiguate biologically identical reads from PCR duplicates. Our UMI small RNA-seq provides an even higher number of possible combinations with 18 nt UMIs—68.7 billion (4^{18})—much larger than the number of reads currently produced by a sequencing run. In terms of small RNA-seq, the most abundant small RNA species in our datasets is a piRNA with 42,281 reads, far fewer than the number of UMI combinations our protocol provides. We conclude that the UMI lengths used in the RNA-seq and small RNA-seq protocols contain a sufficient UMI diversity for current and, most likely, future sequencing experiments.

5.2.4 Error-correction for UMIs only slightly improves PCR duplicate identification

To test whether UMIs could help us accurately identify PCR duplicates, we first evaluated their performance using simulated data, where we know the ground truth. Assuming a library has sufficiently diverse UMI sequences, the simplest way to determine biologically identical reads is to look for reads with the same sequence but are tagged by different UMIs. This approach assumes that there is no error in the replication or reading of the UMI sequences, since such errors could render identical UMI sequences different and vice versa, causing misidentification of PCR duplicates. UMI errors could occur during PCR sequencing, and computationally correcting these errors has been shown to improve identification of PCR duplicates (Bose et al., 2015; Islam et al., 2014; Macosko et al., 2015; T. Smith et al., 2017, 2017; Yaari & Kleinstein, 2015).

We designed a strategy for correcting UMI errors by exploting the following assumptions. First, UMI errors are rare, with rates stipulated by the chemistry of PCR and sequencing ($\sim 10^{-5}$ and $\sim 10^{-3}$ errors per position respectively) (Flaman et al., 1994; Lundberg et al., 1991; Schirmer et al., 2016). Second, when two sufficiently long UMIs (for example, 10 and 18 nt in this study) that differ by just one base are connected to two reads with identical sequences, the probability that these are PCR duplicates of the same UMI with an error, albeit low ($p < 10^{-3}$) is still much higher than the probability that these are two distinct UMIs ($p = 4^{-10}$ for RNA-seq and 4^{-18} for small RNA-seq in this study). Adopting an error-correction method previously developed for RNA-seq (T. Smith et al., 2017), we built a UMI graph for each group of reads (Figure. 5.3). For RNA-seq, the reads that map to the same genomic position form a group. This approach does not work

for small RNAs, because they often originate from multiple genomic loci. Thus, we simply defined a group of small RNA reads as those with identical sequences. In both the RNA-seq and small RNA-seq UMI graphs, a node denotes a unique UMI and further holds the number of reads with that UMI (Figure 5.3). For each pair of UMIs (say, UMI *a* and UMI *b*) that differ by just one base (one edit distance apart), we connect their nodes if $n_a \geq 2 \times n_b - 1$, where n_a and n_b represent read counts for the two UMIs. We require a twofold difference between n_a and n_b, because as described above, the error rates for PCR and sequencing are low, and the twofold differences corresponds to the most extreme case whereby an error occurred during the first PCR cycle. However, a twofold difference is too stringent for pairs of UMIs with low read counts (e.g., 1 versus 2), for which the error predominantly arose from sequencing. We therefore added "−1" to ensure that these UMIs could be connected. All connected UMIs are then assumed to originate from the most abundant UMIs in the graph. This scheme allows correction of two or more errors in UMIs, provided that the intermediate UMIs are observed (for example, the intermediate UMI with one error and UMI with two errors in Figure 5.3A–B). One could relax the stringency of this method by adding direct connections between two nodes that differ in two or more positions.

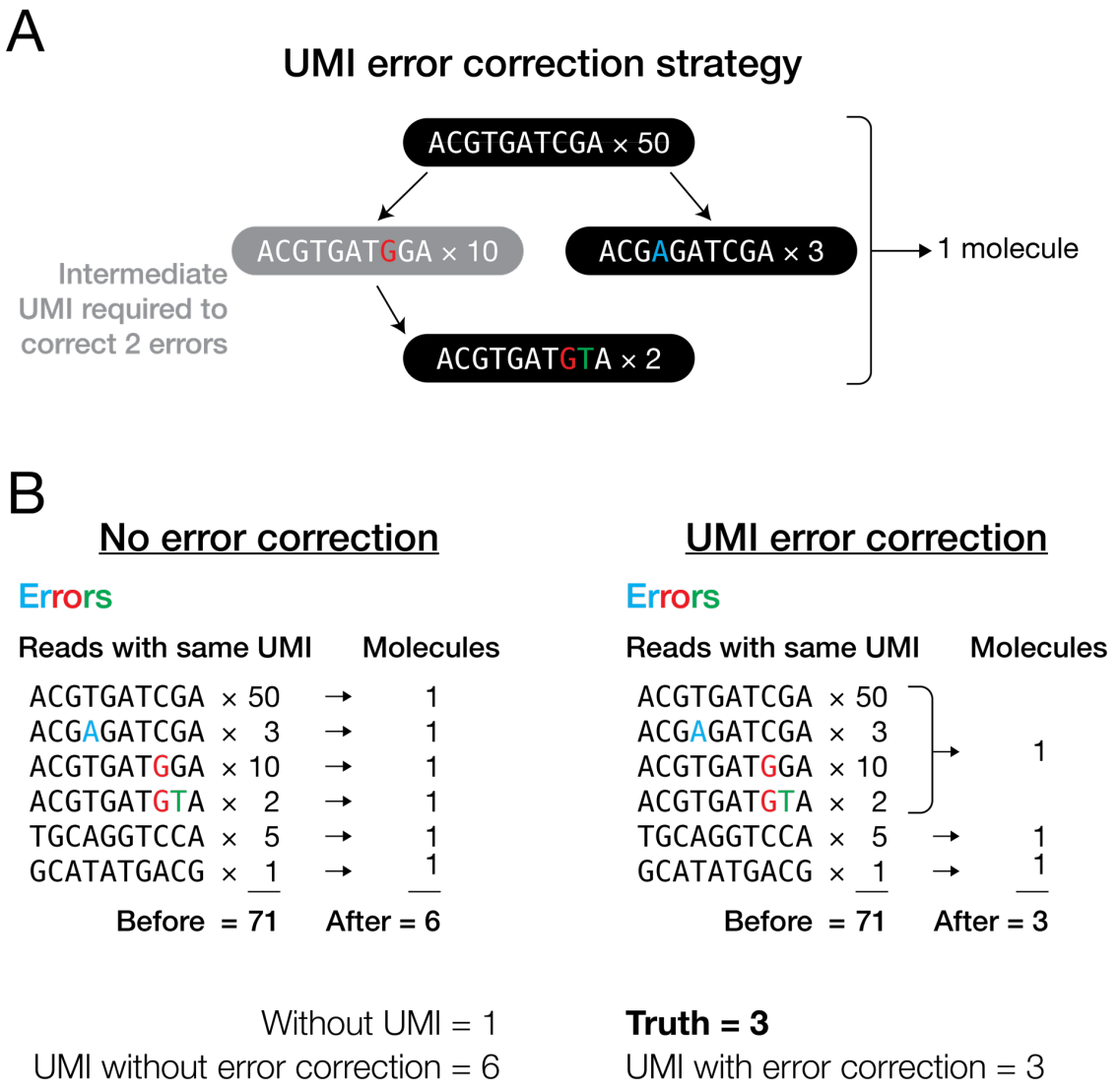

Figure 5.3. Identifying PCR duplicates. (A) Strategy for correcting errors in UMIs. (B) Illustration of how correcting errors in UMIs increases accuracy of PCR duplicate elimination.

The need for error-correction might depend on the experimental conditions, including the PCR amplification probability, PCR and sequencing error rates, UMI length, number of initial molecules, number of sequenced molecules, and number of PCR cycles. We performed computer simulations to investigate the effects of these seven experimental conditions on UMI error correction by systematically varying one variable

at a time while holding the other six constant. Each round of simulation produced a known number of PCR duplicates and therefore, unlike experimental data, the true fraction of all reads corresponding to PCR duplicates can be determined in the simulated data. To assess the accuracy of PCR duplicate identification using UMIs, we calculated the difference between the number of reads after PCR duplicate removal ("estimate") and the true value ("truth") relative to the true value: (estimate − truth) / truth. This metric reflects the extent to which UMIs over- or underestimate the truth as a fraction of the true value. We started the simulation with 100 initial molecules. We then performed PCR by randomly assigning a probability to each molecule (tagged with an 18 nt UMI) to be duplicated in each PCR cycle. The probability follows a uniform distribution between *m* and 1, where *m* denotes minimum amplification probability (it can be any value between 0 and 1 and is set to 0.8 in the baseline condition). Minimum amplification probability can be interpreted as PCR efficiency, because the efficiency (average probability) that a molecule is doubled during each PCR cycle is (1-*m*)/2. Ten cycles of PCR (PCR error rate set to 3×10^{-5}) (Flaman et al., 1994; Lundberg et al., 1991) generated a pool of 61,000 ± 1,000 (mean ± S.D.) molecules. To test the effect of sequencing depth, we randomly drew 100 molecules from the pool for sequencing (sequencing error rate set to 10^{-3}) (Schirmer et al., 2016) (Figure 5.4 and Figure. 5.5). We call this set of parameters "baseline condition", and it forms the base line from which we systematically varied each parameter. For each condition, we performed 10,000 trials.

We first assumed that there was no error in UMIs (Figure 5.3) and found that on average, (estimate – truth) / truth = 2.10% across 10,000 trials under the baseline

condition. Thus, without performing UMI error correction, we slightly overestimated the total number of biological molecules as an error in a UMI would artificially create an extra UMI, and in turn, we slightly underestimated the fraction of PCR duplicates (red vs gray lines in Figure 5.5 and Figure 5.6). Next, we used the UMI graph approach described above (Figure 5.3A, B) for correcting errors in UMIs, and the new average of (estimate – truth) / truth = 0.0388%. Even though correcting UMI errors consistently gives better (estimate – truth) / truth than not correcting the errors, the absolute difference in the fractions of PCR duplicates between the two approaches is small (Figure 5.4; Figure 5.5). For example, under the baseline condition, the true fraction of duplicates was 37.8 ± 3.2%; without correcting UMI errors yielded 36.5 ± 3.3%, and correcting UMI errors gave 37.8 ± 3.2%.

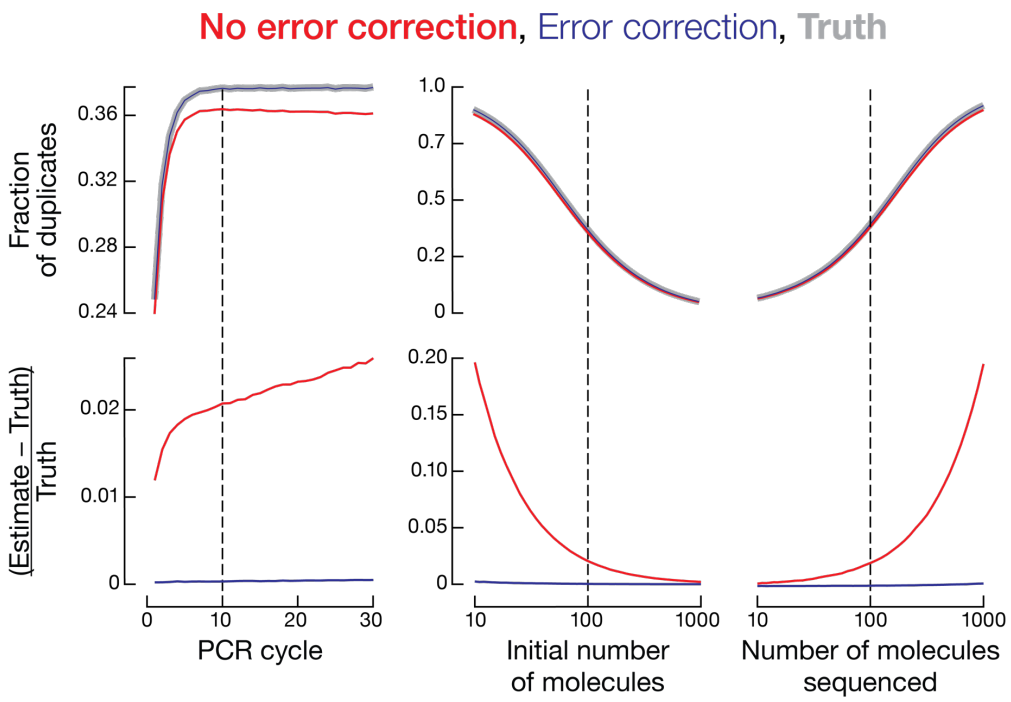

Figure 5.4. Simulation of PCR duplicate removal with or without error correction for UMIs. One parameter (PCR cycle number, starting material, or sequencing depth) was varied with the other parameters kept constant.

Upper plots show the fraction of duplicates, while lower plots show the accuracy of duplicate detection. Each dotted line indicates the value for this parameter used in other simulations.

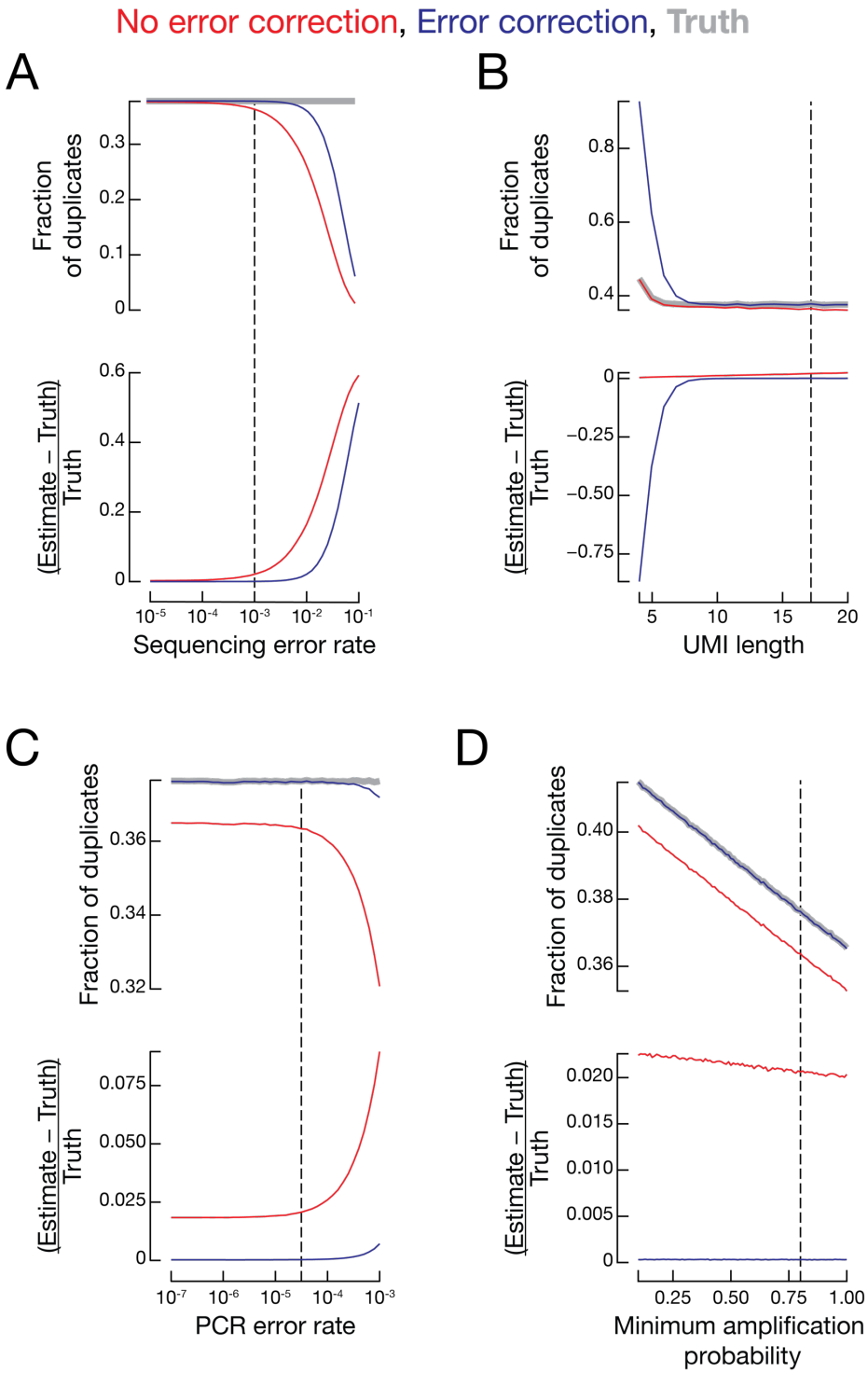

Figure 5.5. Accuracy and fraction of duplicates for simulated data varying (A) sequencing error rate, (B) UMI length, (C) PCR error rate, or (D) minimum amplification probability. Each dotted line indicates the value for this parameter used in other simulations.

Under some extreme conditions, correcting UMI errors yields substantially better results. For example, if we modify PCR error rate in the baseline condition from the default 3×10^{-5} to 10^{-3}, correcting UMI errors still yields a fraction of duplicates (37.2 ± 3.2%) very close to the truth (37.2 ± 3.1%), while not correcting the errors underestimates the fraction of duplicates (32.1 ± 3.5%). In conclusion, error-correction for UMIs consistently, albeit slightly, improves PCR duplicate identification. Therefore, we performed error correction for all following analyses.

5.2.5 Removing PCR duplicates without using UMIs is fundamentally flawed

Does the common practice of removing PCR duplicates without UMIs improve the quantification of both long and short transcripts and in particular, of small RNAs such as microRNAs or piRNAs, which collectively originate from a small portion of the genome? We compared PCR duplicate identification using UMIs together with mapping coordinates of the reads to the conventional approach of using coordinates alone. When only mapping coordinates were used (RNA-seq data from eight mouse tissues) (see (Y. Fu, Wu, et al., 2018)), 16.4%–44.5% RNA-seq reads were determined to be PCR duplicates, whereas using UMI information in conjunction with coordinates identified only 1.89%–10.67% as duplicates. That is, the majority of reads mapping to identical coordinates were in fact not PCR duplicates but rather from distinct starting molecules that should be counted for transcript abundance. The situation is even worse for small RNA-seq data, when only small RNA sequences were used, the majority (56.0%–76.8%)

of reads were flagged as PCR duplicates and therefore excluded from analysis. In contrast, when UMI information was used together with the sequences of reads, just 1.05%–13.6% of reads were determined to be duplicates. Thus, most of the identical reads in RNA-seq and small RNA-seq are biologically real and not PCR duplicates, consistent with the view that small RNAs, which tend to come from precisely the same small genomic regions, can easily be mistaken for PCR duplicates when UMI information is not used. Moreover, the assumption that common mapping coordinates indicate PCR duplicates becomes increasingly problematic as sequencing depth increases, because the chance of observing two identical reads that legitimately derive from different molecules before PCR also increases.

We further tested whether PCR duplicate removal using only mapping coordinates is appropriate for transcript quantification (Figure 5.6A). The conventional method underestimated the abundance of 119 transcripts by 1.25 fold or more: removing PCR duplicates based only on coordinates is too aggressive. These 119 transcripts are significantly shorter (median length = 602 nt) and more highly expressed (median abundance = 200 FPKM) than the other transcripts (median length = 1,620 nt; median abundance = 13.2 FPKM; Wilcoxon rank sum test p values = 2.22×10^{-44} and 1.80×10^{-59}, respectively) (Figure 6B). Thus, overestimation of PCR duplicates without UMIs reflects (1) a higher tendency of short transcripts to produce identical fragments due to more limited possibilities in fragmentation, and (2) a higher tendency of highly expressed genes to produce identical fragments. We conclude that removing PCR duplicates solely

by mapping coordinates introduces substantial bias and that UMIs allow more accurate quantification of PCR duplicates and transcript abundance.

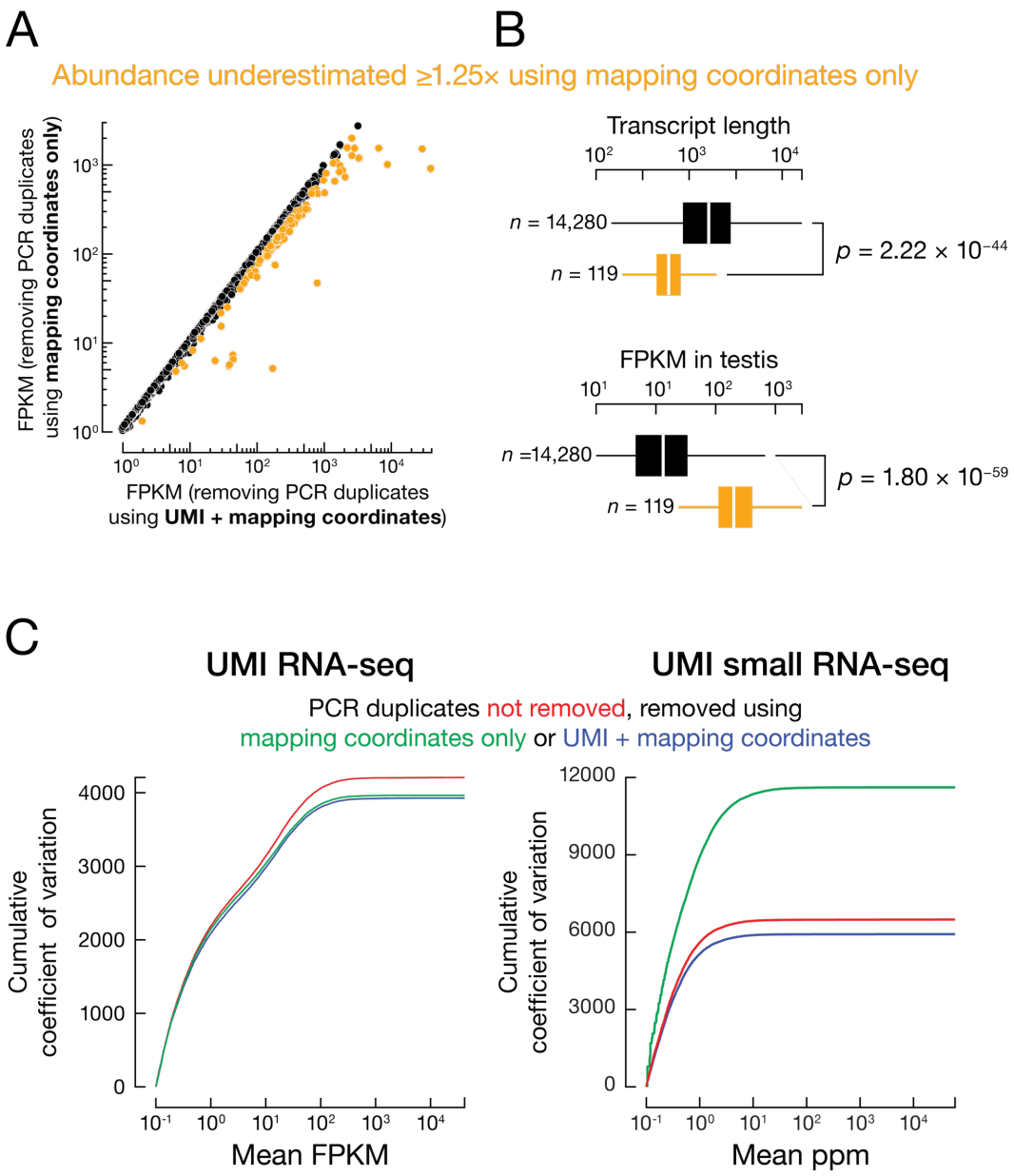

Figure 5.6. (A) Transcript abundance (FPKM) calculated by removing PCR duplicates using only mapping coordinates compared to using mapping coordinates and UMIs. (B) Using only mapping coordinates significantly biases against abundant and short genes. Outliers omitted. Wilcoxon rank sum test; n, number of genes in each group. (C) Relationship between cumulative coefficient of variation and transcript abundance.

5.2.6 UMIs improve data reproducibility

One metric for evaluating the quality of experimental data is the reproducibility between technical replicates. We evaluated how UMIs affect the reproducibility of transcript quantification using five libraries generated using the same sample of total mouse testis RNA, but with gradually decreasing amounts of starting RNA and correspondingly increasing numbers of PCR cycles: 4 μg (8 PCR cycles), 2 μg (9 PCR cycles), 1 μg (10 PCR cycles), 500 ng (11 PCR cycles), 125 ng (13 PCR cycles) (Supplemental Table S1A). We then analyzed the data sets treating PCR duplicates using one of three approaches: (1) no PCR duplicates were removed; (2) PCR duplicates were removed using the conventional approach of identical genomic locations; and (3) PCR duplicates were removed using UMIs together with mapping coordinates. We compared the three approaches by calculating coefficients of variation (CV = S.D. / mean) for transcript abundance across the five RNA-seq libraries. Compared to removing no duplicates, removing duplicates according to their mapping coordinates decreased the total CV by 5.80% (from 4,210 to 3,960), while using UMIs with mapping coordinates decreased the total CV by 6.67% (from 4,210 to 3,930) (Figure 5.6C). For example, when two RNA-seq libraries (125 ng with 12 PCR cycles and 1 μg with 10 PCR cycles) were compared, the number of transcripts whose abundance differed by ≥25% decreased when duplicates were removed (1,880 without duplicate removal, 1,503 removing duplicates by genomic coordinates, and 1,415 removing duplicates using UMIs). We conclude that removing PCR duplicates, using mapping coordinates alone or together with UMIs, improves the precision of transcript quantification.

Next, we evaluated the performance of these three approaches for a series of small RNA-seq libraries (starting material 39–5,000 ng). Compared to removing no duplicates, using UMIs to remove duplicates decreased the total CV by 8.72% (Figure 5.6C). Surprisingly, removing duplicates according to their mapping coordinates alone increased CV by 79.1% (from 6,490 to 11,620) (Figure 5.6C). For example, between two small RNA-seq libraries in this series, one generated from 150 ng and the other from 1 μg of the same total RNA sample, genomic loci (piRNA genes and GENCODE-annotated genes) whose small RNA abundance differed by ≥25% decreased 8.30% when duplicates were removed using UMIs (from 2,613 to 2,396 genes). In contrast, when duplicates were removed using solely mapping coordinates, the number of such irreproducible genes increased by 159% (6,762 genes). These results show that removing PCR duplicates with UMIs leads to more consistent quantification across libraries, whereas removing duplicates without UMIs is overly aggressive and decreases the reproducibility of small RNA-seq experiments.

5.2.7 PCR cycles alone do not determine the frequency of PCR duplicates

It is widely accepted that the number of PCR cycles used to amplify the initial cDNA is the major cause of PCR duplicates in sequencing libraries (Andrews, Good, Miller, Luikart, & Hohenlohe, 2016). We sought to test this assumption and to identify other experimental contributing factors. As described above, we performed computer simulations to test the impact of UMI error correction on PCR duplicate detection. We considered seven parameters that could impact the level of PCR duplicates during an RNA-seq or small RNA-seq experiment. Assuming that we have performed UMI error

correction, we now examine in detail these seven parameters for their impact on the level of PCR duplicates.

Four of the parameters—PCR amplification efficiency, PCR error rate, sequencing error rate, and UMI length—are specified by the experimental reagents and sequencing platform and typically not adjusted from experiment to experiment. Our simulation results indicate that varying the sequencing error rate, the PCR error rate, or the UMI length around their default values in the baseline condition (i.e., within the ranges stipulated by experimental settings) did not have a significant effect on the faction of PCR duplicates (the blue line is flat around the dashed vertical line in Figure 5.5A–C, top panels). In comparison, PCR efficiency had a measurable effect (the blue line in the top panel of Figure S1D reveals a negative correlation with PCR efficiency). This is because that at lower PCR efficiency, some molecules are less likely to be amplified and become underrepresented, causing a decrease in library complexity and correspondingly higher fractions of PCR duplicates.

The other three parameters—the number of initial molecules, the number of molecules sequenced (i.e., sequencing depth), and the number of PCR cycles—are often adjusted to meet specific experimental conditions. Our simulations revealed that a change in PCR cycle number alone only minimally affected the fraction of PCR duplicates (the blue line in the top-left panel of Figure 5.4 is nearly flat around the dashed vertical line), because the starting molecules of the original pool are proportionally propagated to the final library (Head et al., 2014). In contrast, decreasing the number of initial molecules or

increasing the number of molecules sequenced sharply raised the frequency of PCR duplicates (Figure 5.4, two top-right panels).

We further tested these findings using experimental datasets. We first analyzed a set of five UMI RNA-seq libraries made with gradually decreasing amounts of starting RNA and correspondingly increasing numbers of PCR cycles: 4 µg (8 cycles), 2 µg (9 cycles), 1 µg (10 cycles), 500 ng (11 cycles), 125 ng (13 cycles). We observed that less starting RNA and correspondingly more PCR amplification resulted in higher fractions of PCR duplicates (Figure 5.7A). For example, the 125 ng, 13-cycle library yielded 10.7% (median over 43,432 genes) PCR duplicates, while the 4 µg, 8-cycle library made by the same procedure contained only 1.79% PCR duplicates. Similarly, analysis of UMI small RNA-seq libraries generated from 39 ng (30 cycles) to 5 µg (16 cycles) total RNA revealed that starting with less RNA caused higher fractions of PCR duplicates (Figure 5.7A).

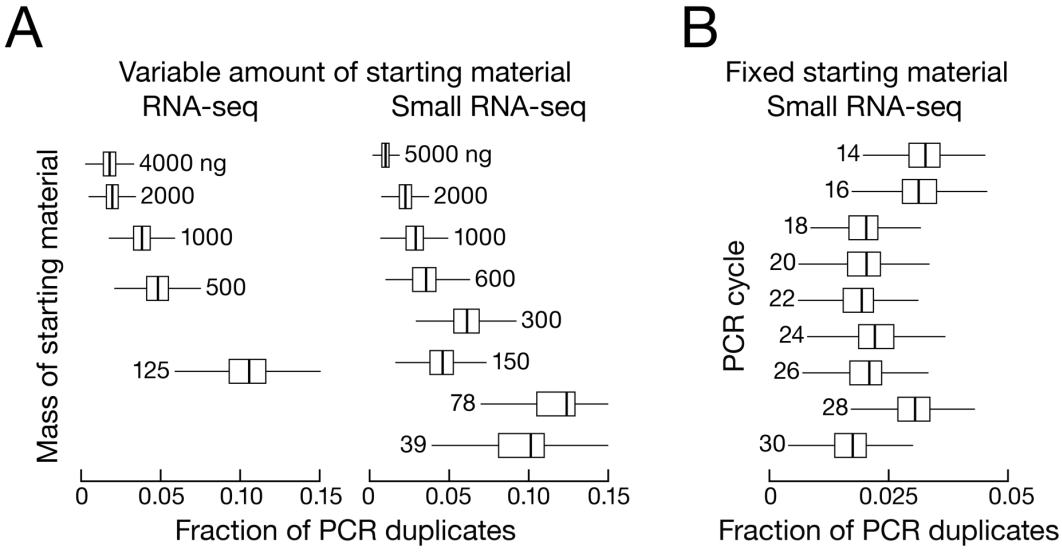

Figure 5.7. Fraction of PCR duplicates across genes for (A) a series of UMI RNA-seq and small RNA-seq libraries made with different amount of starting materials, and (B) a series of UMI small RNA-seq libraries all made with 5μg of total mouse testis RNA and with an increasing number of PCR cycles.

Simulations argue that the increase in PCR duplicates is not a consequence of greater PCR amplification but rather is caused by the use of lower starting material. To test this idea, we analyzed a second set of nine UMI small RNA-seq libraries, all generated from 5 μg total RNA from the same mouse testis, but amplified using 14 to 30 PCR cycles. Consistent with the simulations, these libraries did not show a discernable trend between fraction of PCR duplicates and the number of PCR cycles (Figure 5.7B). Thus, the higher fraction of PCR duplicates observed in libraries made from low amounts of RNA followed by high PCR cycle numbers more likely reflects the reduced complexity of the starting pool, rather than the increased number of PCR cycles. Together, our simulated and experimental data demonstrate that less starting RNA or higher sequencing depth, but not more PCR cycles *per se*, accounts for the frequency of PCR duplicates.

5.4 Discussion

We have described experimental protocols and computational methods that, by incorporating UMIs into standard procedures, allow accurate PCR duplicate removal from RNA-seq and small RNA-seq data. Our approach increases reproducibility and decreases noise in sequencing libraries generated using a broad range of starting RNA amount and number of PCR cycles, enabling accurate quantification of the abundance of both long and short RNAs. We tested the importance of a key aspect of data processing—error correction for UMIs—and showed that under typical experimental conditions for bulk sequencing (represented by dotted lines in Figure 5.4; Figure 5.5), correcting or not

correcting errors in the UMI sequences has little absolute effect on PCR duplicate quantification. However, sequencing libraries made from a small number of cells, amount of tissue, or amount of RNA, have become increasingly common (Stegle, Teichmann, & Marioni, 2015), and they are more severely affected by PCR duplicates. Single-cell sequencing poses three specific challenges for PCR duplicate removal. First, it uses a limited amount of starting RNA, causing too low library complexity. Second, the ongoing discovery of new species of non-coding RNAs, many poorly understood, increases the number of species being measured, requiring longer UMIs. Finally, the increasingly high sequencing depth provided by advances in technology increases both the number of species that can be detected and the background noise. Together, these three factors make PCR duplicate measurement without UMI error correction especially problematic for single-cell sequencing. Our UMI approach should be directly applicable to single-cell RNA-seq. Error correction for UMIs mitigates these challenges by improving PCR duplicate identification.

The two most widely used computational tools for PCR duplicate removal, Picard MarkDuplicates (http://broadinstitute.github.io/picard/) and SAMtools rmdup (H. Li et al., 2009) rely only on the mapping coordinates of sequencing reads. Our data suggest that most identical reads reflect biological reality. Thus, removing PCR duplicate reads using only mapping coordinates erroneously eliminates many usable reads, particularly those produced from short transcripts and small RNAs.

The eight mouse tissues we analyzed span a range of transcriptome complexity: previous analyses showed that the mouse testis transcriptome contains ~18,700

autosomal protein-coding transcripts, ~8,600 non-coding RNAs, and ~31.7 Mb of intergenic RNA, while the liver transcriptome contains only ~15,500 autosomal protein-coding transcripts, ~1,000 non-coding RNAs, and ~7.2 Mb of intergenic RNA (Soumillon et al., 2013). Among the eight mouse tissues we tested, removing duplicate reads based on only mapping coordinates eliminates many biologically meaningful reads even when the libraries were made using ample starting RNA and optimal experimental conditions. Given the anti-correlation between RNA complexity and PCR duplicate occurrence, UMIs will improve the accuracy of comparing long or small RNA abundance across different tissues or cell types. Short RNAs, such as miRNAs and piRNAs, as well as highly abundant transcripts are particularly susceptible to underestimation by the conventional mapping coordinate method of PCR duplicate removal.

Our UMI approach builds on well-established protocols, requiring few changes in the procedures and little additional cost. We expect UMI analysis to be particularly useful when sequencing RNAs derived from a limited number of genomic loci, such as CaptureSeq (Mercer et al., 2014) and CAGE-seq (Carninci et al., 2006). Our approach can theoretically be adapted to any sequencing technique using synthetic oligonucleotide adapters. For example, sequencing immunoprecipitated chromatin (ChIP-seq) and the alternative CUT&RUN survey the genomic regions bound by proteins of interest (Park, 2009; Skene & Henikoff, 2017). The CUT&RUN method uses a nuclease to achieve more precise chromatin cleavage than the conventional ChIP-seq procedure, which utilizes sonication to randomly shear the DNA. Therefore, the likelihood of yielding identical reads also increases for CUT&RUN. By nature, protein-bound fragments also

map to a smaller portion of genomic positions than RNA-seq reads. UMIs can improve discovery of protein binding sites by minimizing noise. Similarly, degradome sequencing profiles the 5′ ends of 3′ cleaved RNA products (Addo-Quaye, Eshoo, Bartel, & Axtell, 2008); incorporating UMIs will enable precise quantification of cleaved RNA abundance.

Chapter 6. Conclusions, Prospective, and Future Work

The first study presented in this thesis provides the most comprehensive analysis of the *T. ni* genome to date, which lays the foundation for further characterizing the genome. The availability of its genome and annotation allows researchers to perform comparative studies, e.g. for detoxification genes, which would provide insights into the insecticide resistance of *T. ni*, a common and destructive agricultural pest. Since *T. ni* has been found in different environmental niches worldwide, the genome sequence is the first step to understand its genetic diversity and populations. The assembly strategy used in this study can be readily applied to other species, allowing rapid and low-cost genome assemblies.

The highly complete *T. ni* genome also offers a unique opportunity to examine some interesting features. Previously, efforts have been made to assemble Lepidopteran W chromosomes, but none achieved chromosome-level assemblies, likely due to the technical limitations of read lengths. Here, the *T. ni* genome contains the first chromosome-level assembly of the W chromosome in Lepidoptera. Examination of the W chromosome reveals that it is highly repetitive and is a major source of piRNAs.

Characterization of Hi5 cells also reveals that they are latently infected with an alphanodavirus and that Hi5 cells use RNAi to defend against this virus. Further investigation of such siRNAs revealed their phasing pattern, indicating that such siRNAs are produced in a one-after-another manner, consistent with Dicer processivity. Hi5 cells hold promises for recombinant proteins for therapeutics, and understanding the virus and the RNAi is the first step towards eliminating this virus to produce virus-free Hi5 cells. Unexpectedly, siRNAs in *T. ni* does not possess 2′-O-methylation at 3′ ends, unlike drosophilids and mammals. Furthermore, other lepidopterans similarly do not 2′-O-methylated their siRNAs. We currently do not understand why lepidopterans lack 2′-O-methylation in siRNAs.

One of the motivations to assembly this genome is that Hi5 cells have the active piRNA pathway, and can serve as a germline cell line to accelerate studies of small RNA pathways. We have annotated miRNAs and piRNA loci, allowing others to study them. Hi5 cells share the largest piRNA clusters with ovary and test, suggesting that it can recapitulate piRNA biogenesis. Additionally, Hi5 cells gained new piRNA clusters after its derivation of *T. ni* cells, suggesting that it can serve as a model to study piRNA evolution. *T. ni* lacks rhino, cutoff and deadlock, which are responsible for splicing suppression of piRNA cluster transcripts in multiple drosophilids. However, piRNA cluster transcripts are still rarely and inefficiently spliced, suggesting that *T. ni*—and likely many other insect species—has alternative mechanisms to suppress the splicing of piRNA cluster transcripts. Notably, almost the entirety of the *T. ni* W chromosome

produces piRNAs in ovary, which, to our knowledge, is the first example of a chromosome devoted to piRNA production.

The second study in this thesis investigates pachytene piRNAs during mouse spermatogenesis. These piRNAs, albeit abundant, are poorly understood. They are loaded into PIWI proteins and presumably can cleave mRNA targets. To test this, we integrated degradome-seq and small RNA data to obtain potential cleavage sites, which reveals that pachytene piRNAs can cleave RNAs when there is extensive complementarity. Interestingly, GU wobbles—compared to mismatches, such as A vs C and C vs T—promotes the target cleavage, suggesting that thermodynamically stable pairing leads to target cleavage. Further characterization is required to explain why some piRNA gene mutants do not have detectable phenotypes.

The third study presents the most comprehensive branchpoint annotation to date. The computational pipeline I developed for branchpoint discovery is highly efficient, and can be applied to more RNA-seq datasets, should more datasets become available. Such branchpoints are important resources to study splicing in human and mouse. The web application provides an easy-to-use interface that should facilitate future studies into RNA splicing.

The fourth study presents two new protocols with unique molecular identifiers (UMIs). Together with the companion Python package, UMIs are easy to implement and process both experimentally and computationally, rendering it very helpful for identifying PCR duplicates and improving transcript quantification. We also looked into what factors cause higher fraction of PCR duplicates in high-throughput sequencing

experiments, and found that scare starting material and high sequencing depth—but not the number of PCR cycles—cause the high duplicate fraction. This conclusion is profound in that experimentalists should focus on extracting more starting material in order to reduce the fraction of duplicates. Sequencing depth usually cannot be tweaked due to the number of reads required for quantification and other purposes; PCR cycle number often depends on the starting material, because sequencing machine requires certain abundant DNA to perform its function. Thus, to decrease the fraction of PCR duplicates, the only way is to increase the amount of starting material. When this is not possible—for example, in cases of single-cell sequencing or limitation of the material—UMIs should be used to accurately eliminate PCR duplicates.

For future work, some steps of genome assemblies lack specialized bioinformatics tools. For example, one could design and develop a package for analyzing sex determination and dosage compensation by generalizing the procedures I developed for this genome. Such a tool, when applied to a variety of species, will deepen the understanding of sex determination. Another important step during genome assembly is the estimation of the genome size. Developing a model to estimate the genome size based on some Illumina sequencing data—which can be easily acquired by using existing high-quality genome assemblies—would be very useful and potentially highly cited. Knowing the strategy for de novo assembling genomes, I am now working on assembling the genome of Western corn rootworm (*Diabrotica virgifera virgifera*), which causes $1 billion in lost revenue every year, and—according to flow cytometry—has a 2.58 Gb genome. All the analyses should be readily applicable to this genome and it is particularly

interesting to investigate its gene repertoire to understand its insecticide resistance. With the huge number of branchpoints made available, the next step would be to develop a machine learning model to help us unravel some previously unknown important features. These features, when aided by experimental validation, would be a critical step towards a better understanding of RNA splicing. The UMI approach presented in the last study will be useful for low starting material cases, such as single-cell sequencing. Future work may involve processing such datasets to improve quantification of single-cell sequencing. In summary, high-throughput sequencing has now made it trivial to answer some questions that were once unfathomable, but at the same time brings new problems that require sophisticated bioinformatics approaches.

Appendix A

Genomes used in the orthology analysis

Species	Name	Version	Reference
Acyrthosiphon pisum	Pea aphid	OGS 2.0	(T. I. A. G. Consortium, 2010)
Aedes aegypti	Yellow fever mosquito	AaegL3	(Nene et al., 2007)
Anopheles gambiae	African malaria mosquito	AgamP4.3	(Holt et al., 2002)
Apis mellifera	Western honey bee	OGSv3.2	(Leadership et al., 2006)
Atta cephalotes	Leafcutter ant	OGS1.2	(Suen et al., 2011)
Bombyx mori	Silk moth	v2.0	(The International Silkworm Genome, 2008)
Danaus plexippus	Monarch butterfly	OGS 2.0	(Zhan et al., 2011)
Drosophila melanogaster	Fruit fly	r6.12	(Adams et al., 2000)
Drosophila pseudobscura	Fruit fly	r3.04	(Richards et al., 2005)
Harpegnathos saltator	Jerdon's jumping ant	OGS v3.3	(Bonasio et al., 2010)
Homo sapiens	Human	GRCh37.62	(I. H. G. S. Consortium, 2004)
Linepithema humile	Argentine ant	OGS1.2	(C. D. Smith et al., 2011)
Mus musculus	House mouse	NCBIM37.62	(Mouse Genome Sequencing Consortium et al., 2002)
Nasonia vitripennis	Jewel wasp	OGS1.2	(Werren et al., 2010)
Pediculus humanus humanus	Body louse	PhumU2.1	(Kirkness et al., 2010)
Plutella xylostella	diamondback moth	v1.1	(You et al., 2013)
Pogonomyrmex barbatus	Red harvester ant	OGS1.2	(C. R. Smith et al., 2011)
Tetranychus urticae	Two-spotted spider mite	ASM23943v1	(Grbić et al., 2011)
Tribolium castaneum	Red flour beetle	v3.0	(Richards et al., 2008)
Dendroctonus ponderosae	Mountain pine beetle	v1.0	(Keeling et al., 2013)

Appendix B

piRNA pathway genes by sequence orthology					
T. ni gene name	*T. ni* gene ID	*T. ni* scaffold	*D. melanogaster* gene	Mouse gene	*B. mori* gene
TnAgo3	TNI000234	group0	ago3	-	bmAgo3
TnPiwi	TNI008009	group8	aub	Mili	siwi
armi	TNI007690	gropu7	armi	Mov10l1	armi
tdrd12	TNI013819	NA	boYb	Tdrd12	tdrd12
capsuleen	TNI003589	group3	capsuleen	Prmt5	capsuleen
gtsf1	-*	group2	dmGtsf-1 (arx)	Gtsf1	gtsf1
eggless	TNI012914	group16	eggless	Setdb1	eggless
gasz (asz1)	TNI001897	group1	gasz (asz1)	Gasz	gasz
hen1	TNI005148	group4	hen1	Hen1	hen1
hsp83	TNI006421	group6	hsp83	Hsp83	hsp83
krimper	TNI003105	group2	krimper	-	krimper
mael	TNI014445	group19	maelstrom	Maelstrom	maelstrom
papi	TNI016458	tig00003674	papi	Tdrd2 (Tdrdh)	papi
qin	TNI011883	group14	qin	Rnf17	Qin
shutdown	TNI011578	group14	shutdown	Fkbp6	shutdown-1
shutdown-like	TNI002558	group2	-	-	shutdown-2
spn-E	TNI009030	group10	spn-E	Tdrd9	spn-E
tejas	TNI015432	group24	tejas	Tdrd5	tejas
tudor	TNI008782	group9	tudor	Tdrd6	tudor
uap56	TNI000513	group0	uap56	Uap56	usp56
valois	TNI014546	group21	valois	Mep50	valois
vasa	TNI000568	group0	vasa	Mvh	vasa
vreteno	TNI007276	group7	vreteno	Tdrd1	vreteno
zuc	-**	group4	zucchini	MitoPLD	zucchini
-	-	-	piwi	-	-
-	-	-	cutoff	-	-

-	-	-	deadlock	-	-
-	-	-	oskar	-	-
-	-	-	rhino	-	-
-	-	-	soyb	-	-
-	-	-	squash	-	-
-	-	-	yb	-	-
-	-	-	panx	-	-
-	-	-	-	Miwi	-
-	-	-	-	A-Myb	-

* Genome coordinate: group2:19662468-19666842

** Genome coordinate: group4:12497170-12499557

Appendix C

```
-------------------------------------------------------------------------
-- BP Table
-------------------------------------------------------------------------
-- Drop the table if it already exists
DROP TABLE IF EXISTS bp_part1;
DROP TABLE IF EXISTS bp_part2;
DROP TABLE IF EXISTS bp_part3;
DROP TABLE IF EXISTS bp;
-- BP Part 1
create table bp_part1 (bpid text PRIMARY KEY, chr text, coor int,
strand text, dschr text, dscoor int, d2ds int, d2asest int);
.import ../Tables/BP.main_info.table bp_part1
-- BP Part 2 (some seq info)
CREATE TABLE bp_part2 (bpid text PRIMARY KEY, base text, up100 text,
down100 text, dsseq text);
.import ../Tables/BP.seq_info.table bp_part2
-- BP Part 3 (conservation info)
CREATE TABLE bp_part3 (bpid text PRIMARY KEY, phylop101 text);
.import ../Tables/BP.cons_info.table bp_part3

.print "Sanity check: do numbers of rows match?"
.print "Number of rows in Table bp_part1:"
select count(*) from bp_part1;
.print "Number of rows in Table bp_part2:"
select count(*) from bp_part2;
.print "Number of rows in Table bp_part3:"
select count(*) from bp_part3;
-- Join these tables to get the BP table
-- number of rows in the resulting table:
.print "Number of rows in bp_part1 JOIN bp_part2 JOIN bp_part3:"
SELECT COUNT(*) FROM bp_part1 JOIN bp_part2 on bp_part1.bpid =
bp_part2.bpid JOIN bp_part3 on bp_part1.bpid = bp_part3.bpid;
-- Now actually create the table
-- Specifying the columns I need. Otherwise, bpid will appear twice...
CREATE TABLE bp AS SELECT bp_part1.bpid as bpid, chr, coor, strand,
dschr, dscoor, d2ds, d2asest, base, up100, down100, dsseq
FROM bp_part1 JOIN bp_part2 ON bp_part1.bpid = bp_part2.bpid
JOIN bp_part3 ON bp_part1.bpid = bp_part3.bpid;
.print "Number of rows in Table BP"
SELECT COUNT(*) FROM bp;
---- Create index on bpid, which is unique
-- CREATE UNIQUE INDEX idx_bp_bpid ON bp (bpid);
---- Create some indices to speed things up
-- This is the most useful one for joining bp and intron tables
CREATE INDEX idx_bp_chr_strand_coor ON bp (chr, strand, coor);
CREATE INDEX idx_bp_coor ON bp (coor);
CREATE INDEX idx_bp_strand ON bp (strand);
CREATE INDEX idx_bp_dschr ON bp (dschr);
CREATE INDEX idx_bp_dscoor ON bp (dscoor);
CREATE INDEX idx_bp_bpid ON bp (bpid);
-- Drop the temporary tables:
DROP TABLE IF EXISTS bp_part1;
```

```
DROP TABLE IF EXISTS bp_part2;
DROP TABLE IF EXISTS bp_part3;
-- SELECT * FROM BP LIMIT 5;
.print "Done creating table: bp."

-----------------------------------------------------------------------
-- Other tables
-----------------------------------------------------------------------
--Species table
DROP TABLE IF EXISTS species;
CREATE TABLE species (speciesid text PRIMARY KEY, speciesname text,
genus text);
.import ../Tables/no_header/Species.table.no_header species
CREATE UNIQUE INDEX idx_species_speciesid ON species (speciesid);
CREATE INDEX idx_species_speciesname ON species (speciesname);
CREATE INDEX idx_species_genus ON species (genus);
.print "Species table: Number of rows imported:"
select count(*) from species;

-- Gene table
DROP TABLE IF EXISTS gene;
CREATE TABLE gene (gid text PRIMARY KEY, speciesid text, assembly text,
genename text, genetype text, havanaid text, chr text, start integer,
end integer, strand text, FOREIGN KEY(speciesid) REFERENCES
species(speciesid));
.import ../Tables/no_header/Gene.table.no_header gene
CREATE UNIQUE INDEX idx_gene_gid ON gene (gid);
CREATE INDEX idx_gene_assembly ON gene (assembly);
CREATE INDEX idx_gene_genename ON gene (genename);
CREATE INDEX idx_gene_chr ON gene (chr);
CREATE INDEX idx_gene_start ON gene (start);
CREATE INDEX idx_gene_end ON gene (end);
CREATE INDEX idx_gene_strand ON gene (strand);
.print "Table gene: Number of rows imported:"
select count(*) from gene;

--Transcript table
DROP TABLE IF EXISTS transcript;
CREATE TABLE transcript (tid text PRIMARY KEY, gid text, tname text,
chr text, start int, end int, strand text, FOREIGN KEY(gid) REFERENCES
gene(gid));
.import ../Tables/no_header/Transcript.table.no_header transcript
CREATE UNIQUE INDEX id_transcript_tid ON transcript (tid);
CREATE INDEX idx_transcript_gid ON transcript (gid);
CREATE INDEX idx_transcript_tname ON transcript (tname);
CREATE INDEX idx_transcript_chr ON transcript (chr);
CREATE INDEX idx_transcript_start ON transcript (start);
CREATE INDEX idx_transcript_end ON transcript (end);
CREATE INDEX idx_transcript_strand ON transcript (strand);

.print "Table transcript: Number of rows imported:"
select count(*) from transcript;

-- Intron table
DROP TABLE IF EXISTS intron;
```

```
CREATE TABLE intron (intronid text PRIMARY KEY, gid text, tid text,
intronnum int, chr text, start integer, end integer, strand text,
FOREIGN KEY(gid) REFERENCES gene(gid), FOREIGN KEY(tid) REFERENCES
transcript(tid));
.import ../Tables/no_header/Intron.table.no_header intron
CREATE INDEX idx_intron_chr_strand_start_end ON intron (chr, strand,
start, end);
CREATE UNIQUE INDEX idx_intron_intronid ON intron (intronid);
CREATE INDEX idx_intron_chr ON intron (chr);
CREATE INDEX idx_intron_start ON intron (start);
CREATE INDEX idx_intron_end ON intron (end);
CREATE INDEX idx_intron_strand ON intron (strand);
CREATE INDEX idx_intron_gid ON intron (gid);
.print "Table intron: Number of rows imported:"
select count(*) from intron;

-- Exon table
-- Note that the GENCODE exonid cannot be used as the primary key
-- GENCODE reuses the same exon id for different transcripts if the two
-- transcripts have that same exon
DROP TABLE IF EXISTS exon;
CREATE TABLE exon (exonid text, gid text, tid text, exonnum int, chr
text, start integer, end integer, strand text, FOREIGN KEY(gid)
REFERENCES gene(gid), FOREIGN KEY(tid) REFERENCES transcript(tid));
.import ../Tables/no_header/Exon.table.no_header exon
-- Note that exon id's are used and these are not guaranteed to be
unique: two transcripts may have the same exon.
CREATE INDEX idx_exon_exonid ON exon (exonid);
CREATE INDEX idx_exon_gid ON exon (gid);
CREATE INDEX idx_exon_chr ON exon (chr);
CREATE INDEX idx_exon_start ON exon (start);
CREATE INDEX idx_exon_end ON exon (end);
CREATE INDEX idx_exon_strand ON exon (strand);

.print "Table exon: Number of rows imported:"
select count(*) from exon;

--Table of BP source
DROP TABLE IF EXISTS bp_src;
CREATE TABLE bp_src (bpid text, accession text, readnum int, speciesnum
int, FOREIGN KEY(bpid) REFERENCES bp(bpid));
.import ../Tables/bp_src.table bp_src

CREATE INDEX idx_bp_src_bpid ON bp_src (bpid);
CREATE INDEX idx_bp_src_accession ON bp_src (accession);

.print "bp_src table: Number of rows imported:"
select count(*) from bp_src;
```

BIBLIOGRAPHY

Abe, H., Fujii, T., Tanaka, N., Yokoyama, T., Kakehashi, H., Ajimura, M., … Shimada, T. (2008). Identification of the female-determining region of the W chromosome in Bombyxmori. *Genetica*, *133*(3), 269–282. https://doi.org/10.1007/s10709-007-9210-1

Abe, H., Seki, M., Ohbayashi, F., Tanaka, N., Yamashita, J., Fujii, T., … Shimada, T. (2005). Partial deletions of the W chromosome due to reciprocal translocation in the silkworm Bombyx mori. *Insect Molecular Biology*, *14*(4), 339–352. https://doi.org/10.1111/j.1365-2583.2005.00565.x

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., … Venter, J. C. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, *287*(5461), 2185–2195. https://doi.org/10.1126/science.287.5461.2185

Addo-Quaye, C., Eshoo, T. W., Bartel, D. P., & Axtell, M. J. (2008). Endogenous siRNA and miRNA Targets Identified by Sequencing of the Arabidopsis Degradome. *Current Biology*, *18*(10), 758–762. https://doi.org/10.1016/j.cub.2008.04.042

Agarwal, V., Bell, G. W., Nam, J.-W., & Bartel, D. P. (2015). Predicting effective microRNA target sites in mammalian mRNAs. *ELife*, *4*, e05005. https://doi.org/10.7554/eLife.05005

Agrawal, N., Dasaradhi, P. V. N., Mohmmed, A., Malhotra, P., Bhatnagar, R. K., & Mukherjee, S. K. (2003). RNA Interference: Biology, Mechanism, and Applications. *Microbiology and Molecular Biology Reviews*, *67*(4), 657–685. https://doi.org/10.1128/MMBR.67.4.657-685.2003

Ai, J., Zhu, Y., Duan, J., Yu, Q., Zhang, G., Wan, F., & Xiang, Z. (2011). Genome-wide analysis of cytochrome P450 monooxygenase genes in the silkworm, Bombyx mori. *Gene*, *480*(1–2), 42–50. https://doi.org/10.1016/j.gene.2011.03.002

Alioto, T. S. (2007). U12DB: a database of orthologous U12-type spliceosomal introns. *Nucleic Acids Research*, *35*(suppl_1), D110–D115. https://doi.org/10.1093/nar/gkl796

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2

Ambros, V. (2004). The functions of animal microRNAs. *Nature*, *431*(7006), 350–355. https://doi.org/10.1038/nature02871

Andersen, J. F., Utermohlen, J. G., & Feyereisen, R. (1994). Expression of Housefly CYP6A1 and NADPH-Cytochrome P450 Reductase in Escherichia coli and Reconstitution of an Insecticide-Metabolizing P450 System. *Biochemistry*, *33*(8), 2171–2177. https://doi.org/10.1021/bi00174a025

Andrews, K. R., Good, J. M., Miller, M. R., Luikart, G., & Hohenlohe, P. A. (2016). Harnessing the power of RADseq for ecological and evolutionary genomics. *Nature Reviews Genetics*, *17*(2), 81–92. https://doi.org/10.1038/nrg.2015.28

Aravin, A. A., Hannon, G. J., & Brennecke, J. (2007). The Piwi-piRNA Pathway Provides an Adaptive Defense in the Transposon Arms Race. *Science*, *318*(5851), 761–764. https://doi.org/10.1126/science.1146484

Aravin, Alexei A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K. F., … Hannon, G. J. (2008). A piRNA Pathway Primed by Individual Transposons Is Linked to De Novo DNA Methylation in Mice. *Molecular Cell*, *31*(6), 785–799. https://doi.org/10.1016/j.molcel.2008.09.003

Attrill, H., Falls, K., Goodman, J. L., Millburn, G. H., Antonazzo, G., Rey, A. J., … Consortium, the F. (2016). FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *Nucleic Acids Research*, *44*(D1), D786–D792. https://doi.org/10.1093/nar/gkv1046

Bartel, D. P. (2004). MicroRNAs. *Cell*, *116*(2), 281–297. https://doi.org/10.1016/S0092-8674(04)00045-5

Bartel, D. P. (2009). MicroRNAs: Target Recognition and Regulatory Functions. *Cell*, *136*(2), 215–233. https://doi.org/10.1016/j.cell.2009.01.002

Baxter, S. W., Badenes-Pérez, F. R., Morrison, A., Vogel, H., Crickmore, N., Kain, W., … Jiggins, C. D. (2011). Parallel Evolution of Bacillus thuringiensis Toxin Resistance in Lepidoptera. *Genetics*, *189*(2), 675–679. https://doi.org/10.1534/genetics.111.130971

Benson, G. (1999). Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Research*, *27*(2), 573–580. https://doi.org/10.1093/nar/27.2.573

Benton, R., Vannice, K. S., Gomez-Diaz, C., & Vosshall, L. B. (2009). Variant Ionotropic Glutamate Receptors as Chemosensory Receptors in Drosophila. *Cell*, *136*(1), 149–162. https://doi.org/10.1016/j.cell.2008.12.001

Bernt, M., Donath, A., Jühling, F., Externbrink, F., Florentz, C., Fritzsch, G., … Stadler, P. F. (2013). MITOS: Improved de novo metazoan mitochondrial genome annotation. *Molecular Phylogenetics and Evolution*, *69*(2), 313–319. https://doi.org/10.1016/j.ympev.2012.08.023

Biology analysis group. (2004). A Draft Sequence for the Genome of the Domesticated Silkworm (Bombyx mori). *Science*, *306*(5703), 1937–1940. https://doi.org/10.1126/science.1102210

Bonasio, R., Zhang, G., Ye, C., Mutti, N. S., Fang, X., Qin, N., … Liebig, J. (2010). Genomic Comparison of the Ants Camponotus floridanus and Harpegnathos saltator. *Science*, *329*(5995), 1068–1071. https://doi.org/10.1126/science.1192428

Bonin, C. P., & Mann, R. S. (2004). A piggyBac Transposon Gene Trap for the Analysis of Gene Expression and Function in Drosophila. *Genetics*, *167*(4), 1801–1811. https://doi.org/10.1534/genetics.104.027557

Bose, S., Wan, Z., Carr, A., Rizvi, A. H., Vieira, G., Pe'er, D., & Sims, P. A. (2015). Scalable microfluidics for single-cell RNA printing and sequencing. *Genome Biology*, *16*, 120. https://doi.org/10.1186/s13059-015-0684-3

Brennecke, J., Aravin, A. A., Stark, A., Dus, M., Kellis, M., Sachidanandam, R., & Hannon, G. J. (2007). Discrete Small RNA-Generating Loci as Master Regulators of Transposon Activity in Drosophila. *Cell*, *128*(6), 1089–1103. https://doi.org/10.1016/j.cell.2007.01.043

Brown, J. B., Boley, N., Eisman, R., May, G. E., Stoiber, M. H., Duff, M. O., … Celniker, S. E. (2014). Diversity and dynamics of the Drosophila transcriptome. *Nature*, *512*(7515), 393–399. https://doi.org/10.1038/nature12962

Burton, J. N., Adey, A., Patwardhan, R. P., Qiu, R., Kitzman, J. O., & Shendure, J. (2013). Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. *Nature Biotechnology*, *31*(12), 1119–1125. https://doi.org/10.1038/nbt.2727

Campbell, M. S., Holt, C., Moore, B., & Yandell, M. (2014). Genome Annotation and Curation Using MAKER and MAKER-P. *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]*, *48*, 4.11.1-4.11.39. https://doi.org/10.1002/0471250953.bi0411s48

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., … Yandell, M. (2008). MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research*, *18*(1), 188–196. https://doi.org/10.1101/gr.6743907

Capella-Gutiérrez, S., Silla-Martínez, J. M., & Gabaldón, T. (2009). trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics*, *25*(15), 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Capinera, J. (2001). *Handbook of vegetable pests*. Gulf Professional Publishing. Retrieved from http://books.google.com/books?hl=en&lr=&id=8j7kOlaLhSwC&oi=fnd&pg=PP1&dq=%22Standard+Book+Number:%22+%22Are+the+Major+Vegetable%22+%

22Identi%C2%AEcation+is+so%22+%22Weevil,+Acanthoscelides+obtectus%22+%22(Chaudoir)+and+Slender%22+%22Flea+Beetle,+Phyllotreta%22+%22Western+Potato+Flea+Beetle,%22+%22Flea+Beetle,+Phyllotreta%22+&ots=NQqyCAynzt&sig=zoLZKB44YTZNdAt43t-Znz-3W_Y

Carninci, P., Sandelin, A., Lenhard, B., Katayama, S., Shimokawa, K., Ponjavic, J., … Hayashizaki, Y. (2006). Genome-wide analysis of mammalian promoter architecture and evolution. *Nature Genetics*, *38*(6), 626–635. https://doi.org/10.1038/ng1789

Castresana, J. (2000). Selection of Conserved Blocks from Multiple Alignments for Their Use in Phylogenetic Analysis. *Molecular Biology and Evolution*, *17*(4), 540–552. https://doi.org/10.1093/oxfordjournals.molbev.a026334

Cha, R. S., & Thilly, W. G. (1993). Specificity, efficiency, and fidelity of PCR. *PCR Methods Appl*, *3*(3), 18–29.

Challis, R. J., Kumar, S., Dasmahapatra, K. K. K., Jiggins, C. D., & Blaxter, M. (2016). Lepbase: the Lepidopteran genome database. *BioRxiv*, 056994. https://doi.org/10.1101/056994

Chinwalla, A. T., Cook, L. L., Delehaunty, K. D., Fewell, G. A., Fulton, L. A., Fulton, R. S., … Green, E. D. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. https://doi.org/10.1038/nature01262

Chung, W.-J., Okamura, K., Martin, R., & Lai, E. C. (2008). Endogenous RNA Interference Provides a Somatic Defense against Drosophila Transposons. *Current Biology*, *18*(11), 795–802. https://doi.org/10.1016/j.cub.2008.05.006

Collins, J. E., Wali, N., Sealy, I. M., Morris, J. A., White, R. J., Leonard, S. R., … Busch-Nentwich, E. M. (2015). High-throughput and quantitative genome-wide messenger RNA sequencing for molecular phenotyping. *BMC Genomics*, *16*(1). https://doi.org/10.1186/s12864-015-1788-6

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J., Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*, *21*(18), 3674–3676. https://doi.org/10.1093/bioinformatics/bti610

Consortium, I. H. G. S. (2004). Finishing the euchromatic sequence of the human genome. *Nature*, *431*(7011), 931. https://doi.org/10.1038/nature03001

Consortium, T. I. A. G. (2010). Genome Sequence of the Pea Aphid Acyrthosiphon pisum. *PLOS Biol*, *8*(2), e1000313. https://doi.org/10.1371/journal.pbio.1000313

Croset, V., Rytz, R., Cummins, S. F., Budd, A., Brawand, D., Kaessmann, H., … Benton, R. (2010). Ancient Protostome Origin of Chemosensory Ionotropic Glutamate Receptors and the Evolution of Insect Taste and Olfaction. *PLOS Genetics*, *6*(8), e1001064. https://doi.org/10.1371/journal.pgen.1001064

Czech, B., Malone, C. D., Zhou, R., Stark, A., Schlingeheyde, C., Dus, M., … Brennecke, J. (2008). An endogenous small interfering RNA pathway in Drosophila. *Nature*, *453*(7196), 798–802. https://doi.org/10.1038/nature07007

Daborn, P. J., Yen, J. L., Bogwitz, M. R., Goff, G. L., Feil, E., Jeffers, S., … ffrench-Constant, R. H. (2002). A Single P450 Allele Associated with Insecticide

Resistance in Drosophila. *Science*, *297*(5590), 2253–2256. https://doi.org/10.1126/science.1074170

Dermauw, W., & Van Leeuwen, T. (2014). The ABC gene family in arthropods: Comparative genomics and role in insecticide transport and resistance. *Insect Biochemistry and Molecular Biology*, *45*, 89–110. https://doi.org/10.1016/j.ibmb.2013.11.001

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, *36*(16), e105. https://doi.org/10.1093/nar/gkn425

Drinnenberg, I. A., deYoung, D., Henikoff, S., & Malik, H. S. (2014). Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *ELife*, *3*, e03676. https://doi.org/10.7554/eLife.03676

Dumesic, P. A., Natarajan, P., Chen, C., Drinnenberg, I. A., Schiller, B. J., Thompson, J., … Madhani, H. D. (2013). Stalled Spliceosomes Are a Signal for RNAi-Mediated Genome Defense. *Cell*, *152*(5), 957–968. https://doi.org/10.1016/j.cell.2013.01.046

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, *32*(5), 1792–1797. https://doi.org/10.1093/nar/gkh340

Enayati, A. A., Ranson, H., & Hemingway, J. (2005). Insect glutathione transferases and insecticide resistance. *Insect Molecular Biology*, *14*(1), 3–8. https://doi.org/10.1111/j.1365-2583.2004.00529.x

Feuda, R., Marlétaz, F., Bentley, M. A., & Holland, P. W. H. (2016). Conservation, Duplication, and Divergence of Five Opsin Genes in Insect Evolution. *Genome Biology and Evolution*, *8*(3), 579–587. https://doi.org/10.1093/gbe/evw015

Feyereisen, R. (2006). Evolution of insect P450. *Biochemical Society Transactions*, *34*(6), 1252–1255. https://doi.org/10.1042/BST0341252

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., … Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, *44*(D1), D279–D285. https://doi.org/10.1093/nar/gkv1344

Flaman, J. M., Frebourg, T., Moreau, V., Charbonnier, F., Martin, C., Ishioka, C., … Iggo, R. (1994). A rapid PCR fidelity assay. *Nucleic Acids Research*, *22*(15), 3259–3260.

Franklin, M. T., Ritland, C. E., & Myers, J. H. (2011). Genetic analysis of cabbage loopers, Trichoplusia ni (Lepidoptera: Noctuidae), a seasonal migrant in western North America. *Evolutionary Applications*, *4*(1), 89–99. https://doi.org/10.1111/j.1752-4571.2010.00135.x

Fraser, M. J., Smith, G. E., & Summers, M. D. (1983). Acquisition of Host Cell DNA Sequences by Baculoviruses: Relationship Between Host DNA Insertions and FP Mutants of Autographa californica and Galleria mellonella Nuclear Polyhedrosis Viruses. *Journal of Virology*, *47*(2), 287–300.

Friedländer, M. R., Chen, W., Adamidi, C., Maaskola, J., Einspanier, R., Knespel, S., & Rajewsky, N. (2008). Discovering microRNAs from deep sequencing data using

miRDeep. *Nature Biotechnology*, *26*(4), 407–415. https://doi.org/10.1038/nbt1394

Friedländer, M. R., Mackowiak, S. D., Li, N., Chen, W., & Rajewsky, N. (2012). miRDeep2 accurately identifies known and hundreds of novel microRNA genes in seven animal clades. *Nucleic Acids Research*, *40*(1), 37–52. https://doi.org/10.1093/nar/gkr688

Friedman, R. C., Farh, K. K.-H., Burge, C. B., & Bartel, D. P. (2008). Most mammalian mRNAs are conserved targets of microRNAs. *Genome Research*, *19*(1), 92–105. https://doi.org/10.1101/gr.082701.108

Fu, G. K., Hu, J., Wang, P.-H., & Fodor, S. P. A. (2011). Counting individual DNA molecules by the stochastic attachment of diverse labels. *Proceedings of the National Academy of Sciences*, *108*(22), 9026–9031. https://doi.org/10.1073/pnas.1017621108

Fu, G. K., Wilhelmy, J., Stern, D., Fan, H. C., & Fodor, S. P. A. (2014). Digital Encoding of Cellular mRNAs Enabling Precise and Absolute Gene Expression Measurement by Single-Molecule Counting. *Analytical Chemistry*, *86*(6), 2867–2870. https://doi.org/10.1021/ac500459p

Fu, G. K., Xu, W., Wilhelmy, J., Mindrinos, M. N., Davis, R. W., Xiao, W., & Fodor, S. P. A. (2014). Molecular indexing enables quantitative targeted RNA sequencing and reveals poor efficiencies in standard library preparations. *Proceedings of the National Academy of Sciences*, *111*(5), 1891–1896. https://doi.org/10.1073/pnas.1323732111

Fu, Y., Wu, P.-H., Beane, T., Zamore, P. D., & Weng, Z. (2018). Elimination of PCR duplicates in RNA-seq and small RNA-seq using unique molecular identifiers. *BioRxiv*, 251892. https://doi.org/10.1101/251892

Fu, Y., Yang, Y., Zhang, H., Farley, G., Wang, J., Quarles, K. A., … Zamore, P. D. (2018). The genome of the Hi5 germ cell line from Trichoplusia ni, an agricultural pest and novel model for small RNA biology. *ELife*, *7*, e31628. https://doi.org/10.7554/eLife.31628

Fujiwara, H., Osanai, M., Matsumoto, T., & Kojima, K. K. (2005). Telomere-specific non-LTR retrotransposons and telomere maintenance in the silkworm, Bombyx mori. *Chromosome Research*, *13*(5), 455–467. https://doi.org/10.1007/s10577-005-0990-9

Fuková, I., Nguyen, P., & Marec, F. (2005). Codling moth cytogenetics: karyotype, chromosomal location of rDNA, and molecular differentiation of sex chromosomes. *Genome*, *48*(6), 1083–1092. https://doi.org/10.1139/g05-063

Futahashi, R., Kawahara-Miki, R., Kinoshita, M., Yoshitake, K., Yajima, S., Arikawa, K., & Fukatsu, T. (2015). Extraordinary diversity of visual opsin genes in dragonflies. *Proceedings of the National Academy of Sciences*, *112*(11), E1247–E1256. https://doi.org/10.1073/pnas.1424670112

Gao, K., Masuda, A., Matsuura, T., & Ohno, K. (2008). Human branch point consensus sequence is yUnAy. *Nucleic Acids Research*, *36*(7), 2257–2267. https://doi.org/10.1093/nar/gkn073

Garcia, D. M., Baek, D., Shin, C., Bell, G. W., Grimson, A., & Bartel, D. P. (2011). Weak seed-pairing stability and high target-site abundance decrease the proficiency of lsy-6 and other microRNAs. *Nature Structural & Molecular Biology*, *18*(10), 1139–1146. https://doi.org/10.1038/nsmb.2115

Ghildiyal, M., Seitz, H., Horwich, M. D., Li, C., Du, T., Lee, S., … Zamore, P. D. (2008). Endogenous siRNAs Derived from Transposons and mRNAs in *Drosophila* Somatic Cells. *Science*, *320*(5879), 1077–1081. https://doi.org/10.1126/science.1157396

Girard, A., Sachidanandam, R., Hannon, G. J., & Carmell, M. A. (2006). A germline-specific class of small RNAs binds mammalian Piwi proteins. *Nature*, *442*(7099), 199–202. https://doi.org/10.1038/nature04917

Goh, W. S. S., Falciatori, I., Tam, O. H., Burgess, R., Meikar, O., Kotaja, N., … Hannon, G. J. (2015). piRNA-directed cleavage of meiotic transcripts regulates spermatogenesis. *Genes & Development*, *29*(10), 1032–1044. https://doi.org/10.1101/gad.260455.115

Gong, D.-P., Zhang, H.-J., Zhao, P., Xia, Q.-Y., & Xiang, Z.-H. (2009). The Odorant Binding Protein Gene Family from the Genome of Silkworm, Bombyx mori. *BMC Genomics*, *10*, 332. https://doi.org/10.1186/1471-2164-10-332

Goodman, W. G., & Granger, N. A. (2005). The Juvenile Hormones. In L. I. Gilbert (Ed.), *Comprehensive Molecular Insect Science* (pp. 319–408). Amsterdam: Elsevier. https://doi.org/10.1016/B0-44-451924-6/00039-9

Gou, L.-T., Dai, P., Yang, J.-H., Xue, Y., Hu, Y.-P., Zhou, Y., … Liu, M.-F. (2014). Pachytene piRNAs instruct massive mRNA elimination during late spermiogenesis. *Cell Research*, *24*(6), 680–700. https://doi.org/10.1038/cr.2014.41

Granados, R. R., Guoxun, L., Derksen, A. C. G., & McKenna, K. A. (1994). A new insect cell line from Trichoplusia ni (BTI-Tn-5B1-4) susceptible to Trichoplusia ni single enveloped nuclear polyhedrosis virus. *Journal of Invertebrate Pathology*, *64*(3), 260–266. https://doi.org/10.1016/S0022-2011(94)90400-6

Graveley, B. R., Brooks, A. N., Carlson, J. W., Duff, M. O., Landolin, J. M., Yang, L., … Celniker, S. E. (2011). The developmental transcriptome of Drosophila melanogaster. *Nature*, *471*(7339), 473–479. https://doi.org/10.1038/nature09715

Grbić, M., Van Leeuwen, T., Clark, R. M., Rombauts, S., Rouzé, P., Grbić, V., … Van de Peer, Y. (2011). The genome of Tetranychus urticae reveals herbivorous pest adaptations. *Nature*, *479*(7374), 487–492. https://doi.org/10.1038/nature10640

Grimson, A., Farh, K. K.-H., Johnston, W. K., Garrett-Engele, P., Lim, L. P., & Bartel, D. P. (2007). MicroRNA Targeting Specificity in Mammals: Determinants beyond Seed Pairing. *Molecular Cell*, *27*(1), 91–105. https://doi.org/10.1016/j.molcel.2007.06.017

Grimson, A., Srivastava, M., Fahey, B., Woodcroft, B. J., Chiang, H. R., King, N., … Bartel, D. P. (2008). Early origins and evolution of microRNAs and Piwi-interacting RNAs in animals. *Nature*, *455*(7217), 1193–1197. https://doi.org/10.1038/nature07415

Gu, L., Walters, J. R., & Knipple, D. C. (2017). Conserved Patterns of Sex Chromosome Dosage Compensation in the Lepidoptera (WZ/ZZ): Insights from a Moth Neo-Z Chromosome. *Genome Biology and Evolution*, *9*(3), 802–816. https://doi.org/10.1093/gbe/evx039

Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, *59*(3), 307–321. https://doi.org/10.1093/sysbio/syq010

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Ha, H., Song, J., Wang, S., Kapusta, A., Feschotte, C., Chen, K. C., & Xing, J. (2014). A comprehensive analysis of piRNAs from adult human testis and their relationship with genes and mobile elements. *BMC Genomics*, *15*, 545. https://doi.org/10.1186/1471-2164-15-545

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., … Regev, A. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols*, *8*(8), 1494–1512. https://doi.org/10.1038/nprot.2013.084

Hahn, C., Bachmann, L., & Chevreux, B. (2013). Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative

mapping approach. *Nucleic Acids Research*, gkt371. https://doi.org/10.1093/nar/gkt371

Han, B. W., Wang, W., Li, C., Weng, Z., & Zamore, P. D. (2015). piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. *Science*, *348*(6236), 817–821. https://doi.org/10.1126/science.aaa1264

Han, B. W., Wang, W., Zamore, P. D., & Weng, Z. (2014). piPipes: a set of pipelines for piRNA and transposon analysis via small RNA-seq, RNA-seq, degradome- and CAGE-seq, ChIP-seq and genomic DNA sequencing. *Bioinformatics*, btu647. https://doi.org/10.1093/bioinformatics/btu647

He, L., & Hannon, G. J. (2004). MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, *5*(7), 522–531. https://doi.org/10.1038/nrg1379

Head, S. R., Komori, H. K., LaMere, S. A., Whisenant, T., Van Nieuwerburgh, F., Salomon, D. R., & Ordoukhanian, P. (2014). Library construction for next-generation sequencing: Overviews and challenges. *BioTechniques*, *56*(2), 61-passim. https://doi.org/10.2144/000114133

Hekmat-Scafe, D. S., Scafe, C. R., McKinney, A. J., & Tanouye, M. A. (2002). Genome-Wide Analysis of the Odorant-Binding Protein Gene Family in Drosophila melanogaster. *Genome Research*, *12*(9), 1357–1369. https://doi.org/10.1101/gr.239402

Hink, W. F. (1970). Established Insect Cell Line from the Cabbage Looper, Trichoplusia ni. *Nature*, *226*(5244), 466–467. https://doi.org/10.1038/226466b0

Holt, R. A., Subramanian, G. M., Halpern, A., Sutton, G. G., Charlab, R., Nusskern, D. R., … Hoffman, S. L. (2002). The Genome Sequence of the Malaria Mosquito Anopheles gambiae. *Science*, *298*(5591), 129–149. https://doi.org/10.1126/science.1076181

Horwich, M. D., Li, C., Matranga, C., Vagin, V., Farley, G., Wang, P., & Zamore, P. D. (2007). The Drosophila RNA Methyltransferase, DmHen1, Modifies Germline piRNAs and Single-Stranded siRNAs in RISC. *Current Biology*, *17*(14), 1265–1272. https://doi.org/10.1016/j.cub.2007.06.030

Houwing, S., Kamminga, L. M., Berezikov, E., Cronembold, D., Girard, A., Elst, H. van den, … Ketting, R. F. (2007). A Role for Piwi and piRNAs in Germ Cell Maintenance and Transposon Silencing in Zebrafish. *Cell*, *129*(1), 69–82. https://doi.org/10.1016/j.cell.2007.03.026

Hutvágner, G., McLachlan, J., Pasquinelli, A. E., Bálint, É., Tuschl, T., & Zamore, P. D. (2001). A Cellular Function for the RNA-Interference Enzyme Dicer in the Maturation of the *let-7* Small Temporal RNA. *Science*, *293*(5531), 834–838. https://doi.org/10.1126/science.1062961

Iga, M., & Kataoka, H. (2012). Recent Studies on Insect Hormone Metabolic Pathways Mediated by Cytochrome P450 Enzymes. *Biological and Pharmaceutical Bulletin*, *35*(6), 838–843. https://doi.org/10.1248/bpb.35.838

Initiative, I. G. G. (2014). Genome Sequence of the Tsetse Fly (Glossina morsitans): Vector of African Trypanosomiasis. *Science*, *344*(6182), 380–386. https://doi.org/10.1126/science.1249656

International Human Genome Sequencing Consortium. (2001). Initial sequencing and analysis of the human genome. *Nature*, *409*(6822), 860–921. https://doi.org/10.1038/35057062

Islam, S., Zeisel, A., Joost, S., La Manno, G., Zajac, P., Kasper, M., … Linnarsson, S. (2014). Quantitative single-cell RNA-seq with unique molecular identifiers. *Nature Methods*, *11*(2), 163–166. https://doi.org/10.1038/nmeth.2772

Iwanaga, M., Adachi, Y., Uchiyama, K., Tsukui, K., Katsuma, S., & Kawasaki, H. (2014). Long-term adaptation of the Bombyx mori BmN4 cell line to grow in serum-free culture. *In Vitro Cellular & Developmental Biology - Animal*, 1–5. https://doi.org/10.1007/s11626-014-9781-y

Janmaat, A. F., & Myers, J. (2003). Rapid evolution and the cost of resistance to Bacillus thuringiensis in greenhouse populations of cabbage loopers, Trichoplusia ni. *Proceedings of the Royal Society B: Biological Sciences*, *270*(1530), 2263–2270. https://doi.org/10.1098/rspb.2003.2497

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., … Hunter, S. (2014). InterProScan 5: genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Jones, W. D., Cayirlioglu, P., Grunwald Kadow, I., & Vosshall, L. B. (2007). Two chemosensory receptors together mediate carbon dioxide detection in Drosophila. *Nature*, *445*(7123), 86–90. https://doi.org/10.1038/nature05466

Juliano, C. E., Reich, A., Liu, N., Götzfried, J., Zhong, M., Uman, S., … Lin, H. (2014). PIWI proteins and PIWI-interacting RNAs function in Hydra somatic stem cells.

Proceedings of the National Academy of Sciences, *111*(1), 337–342. https://doi.org/10.1073/pnas.1320965111

Kanost, M. R., Arrese, E. L., Cao, X., Chen, Y.-R., Chellapilla, S., Goldsmith, M. R., … Blissard, G. W. (2016). Multifaceted biological insights from a draft genome sequence of the tobacco hornworm moth, Manduca sexta. *Insect Biochemistry and Molecular Biology*, *76*, 118–147. https://doi.org/10.1016/j.ibmb.2016.07.005

Kasai, S., & Scott, J. G. (2000). Overexpression of Cytochrome P450 CYP6D1 Is Associated with Monooxygenase-Mediated Pyrethroid Resistance in House Flies from Georgia. *Pesticide Biochemistry and Physiology*, *68*(1), 34–41. https://doi.org/10.1006/pest.2000.2492

Katsuma, S., Sugano, Y., Kiuchi, T., & Shimada, T. (2015). Two Conserved Cysteine Residues Are Required for the Masculinizing Activity of the Silkworm Masc Protein. *Journal of Biological Chemistry*, *290*(43), 26114–26124. https://doi.org/10.1074/jbc.M115.685362

Kawamura, Y., Saito, K., Kin, T., Ono, Y., Asai, K., Sunohara, T., … Siomi, H. (2008). Drosophila endogenous small RNAs bind to Argonaute 2 in somatic cells. *Nature*, *453*(7196), 793–797. https://doi.org/10.1038/nature06938

Kawaoka, S., Hayashi, N., Suzuki, Y., Abe, H., Sugano, S., Tomari, Y., … Katsuma, S. (2009). The Bombyx ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. *RNA*, *15*(7), 1258–1264. https://doi.org/10.1261/rna.1452209

Kawaoka, Shinpei, Hayashi, N., Katsuma, S., Kishino, H., Kohara, Y., Mita, K., & Shimada, T. (2008). Bombyx small RNAs: Genomic defense system against transposons in the silkworm, Bombyx mori. *Insect Biochemistry and Molecular Biology*, *38*(12), 1058–1065. https://doi.org/10.1016/j.ibmb.2008.03.007

Kawaoka, Shinpei, Kadota, K., Arai, Y., Suzuki, Y., Fujii, T., Abe, H., … Katsuma, S. (2011). The silkworm W chromosome is a source of female-enriched piRNAs. *RNA*, *17*(12), 2144–2151. https://doi.org/10.1261/rna.027565.111

Keeling, C. I., Yuen, M. M., Liao, N. Y., Roderick Docking, T., Chan, S. K., Taylor, G. A., … Bohlmann, J. (2013). Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. *Genome Biology*, *14*, R27. https://doi.org/10.1186/gb-2013-14-3-r27

Khan, S. G., Metin, A., Gozukara, E., Inui, H., Shahlavi, T., Muniz-Medina, V., … Kraemer, K. H. (2004). Two essential splice lariat branchpoint sequences in one intron in a xeroderma pigmentosum DNA repair gene: mutations result in reduced XPC mRNA levels that correlate with cancer risk. *Human Molecular Genetics*, *13*(3), 343–352. https://doi.org/10.1093/hmg/ddh026

Kirkness, E. F., Haas, B. J., Sun, W., Braig, H. R., Perotti, M. A., Clark, J. M., … Pittendrigh, B. R. (2010). Genome sequences of the human body louse and its primary endosymbiont provide insights into the permanent parasitic lifestyle. *Proceedings of the National Academy of Sciences*, *107*(27), 12168–12173. https://doi.org/10.1073/pnas.1003379107

Kivioja, T., Vähärautio, A., Karlsson, K., Bonke, M., Enge, M., Linnarsson, S., & Taipale, J. (2012). Counting absolute numbers of molecules using unique molecular identifiers. *Nature Methods*, *9*(1), 72–74. https://doi.org/10.1038/nmeth.1778

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. https://doi.org/10.1101/gr.215087.116

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics*, *5*, 59. https://doi.org/10.1186/1471-2105-5-59

Kotin, R. M. (2011). Large-scale recombinant adeno-associated virus production. *Human Molecular Genetics*, *20*(R1), R2–R6. https://doi.org/10.1093/hmg/ddr141

Kurtz, S., Phillippy, A., Delcher, A. L., Smoot, M., Shumway, M., Antonescu, C., & Salzberg, S. L. (2004). Versatile and open software for comparing large genomes. *Genome Biology*, *5*, R12. https://doi.org/10.1186/gb-2004-5-2-r12

Labbé, R., Caveney, S., & Donly, C. (2011). Genetic analysis of the xenobiotic resistance-associated ABC gene subfamilies of the Lepidoptera. *Insect Molecular Biology*, *20*(2), 243–256. https://doi.org/10.1111/j.1365-2583.2010.01064.x

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. https://doi.org/10.1093/nar/gkm160

Laissue, P. P., & Vosshall, L. B. (2008). The olfactory sensory map in Drosophila. In *Brain development in Drosophila melanogaster* (pp. 102–114). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-0-387-78261-4_7

Lanning, C. L., Fine, R. L., Corcoran, J. J., Ayad, H. M., Rose, R. L., & Abou-Donia, M. B. (1996). Tobacco budworm P-glycoprotein: biochemical characterization and its involvement in pesticide resistance. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1291*(2), 155–162. https://doi.org/10.1016/0304-4165(96)00060-8

Lau, N. C., Lim, L. P., Weinstein, E. G., & Bartel, D. P. (2001). An Abundant Class of Tiny RNAs with Probable Regulatory Roles in Caenorhabditis elegans. *Science*, *294*(5543), 858–862. https://doi.org/10.1126/science.1065062

Lau, N. C., Seto, A. G., Kim, J., Kuramochi-Miyagawa, S., Nakano, T., Bartel, D. P., & Kingston, R. E. (2006). Characterization of the piRNA Complex from Rat Testes. *Science*, *313*(5785), 363–367. https://doi.org/10.1126/science.1130164

Le Thomas, A., Stuwe, E., Li, S., Du, J., Marinov, G., Rozhkov, N., … Aravin, A. A. (2014). Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes & Development*, *28*(15), 1667–1680. https://doi.org/10.1101/gad.245514.114

Leadership, O. project, Weinstock, G. M., Robinson, G. E., Investigators, P., Gibbs, R. A., Coordination, C., … Wright, R. (2006). Insights into social insects from the

genome of the honeybee Apis mellifera. *Nature*, *443*(7114), 931–949. https://doi.org/10.1038/nature05260

Leal, W. S. (2013). Odorant Reception in Insects: Roles of Receptors, Binding Proteins, and Degrading Enzymes. *Annual Review of Entomology*, *58*(1), 373–391. https://doi.org/10.1146/annurev-ento-120811-153635

Lee, E., Helt, G. A., Reese, J. T., Munoz-Torres, M. C., Childers, C. P., Buels, R. M., … Lewis, S. E. (2013). Web Apollo: a web-based genomic annotation editing platform. *Genome Biology*, *14*, R93. https://doi.org/10.1186/gb-2013-14-8-r93

Lee, H.-C., Gu, W., Shirayama, M., Youngman, E., Conte, D., & Mello, C. C. (2012). C. elegans piRNAs Mediate the Genome-wide Surveillance of Germline Transcripts. *Cell*, *150*(1), 78–87. https://doi.org/10.1016/j.cell.2012.06.016

Lee, Y., Ahn, C., Han, J., Choi, H., Kim, J., Yim, J., … Kim, V. N. (2003). The nuclear RNase III Drosha initiates microRNA processing. *Nature*, *425*(6956), 415–419. https://doi.org/10.1038/nature01957

Letunic, I., & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, *44*(W1), W242–W245. https://doi.org/10.1093/nar/gkw290

Lewis, B. P., Burge, C. B., & Bartel, D. P. (2005). Conserved Seed Pairing, Often Flanked by Adenosines, Indicates that Thousands of Human Genes are MicroRNA Targets. *Cell*, *120*(1), 15–20. https://doi.org/10.1016/j.cell.2004.12.035

Lewis, B. P., Shih, I. -hun., Jones-Rhoades, M. W., Bartel, D. P., Burge, C. B., & others. (2003). Prediction of mammalian microRNA targets. *Cell*, *115*(7), 787–798.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, L., Stoeckert, C. J., & Roos, D. S. (2003). OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, *13*(9), 2178–2189. https://doi.org/10.1101/gr.1224503

Li, M., Kuivenhoven, J. A., Ayyobi, A. F., & Pritchard, P. H. (1998). T→G or T→A mutation introduced in the branchpoint consensus sequence of intron 4 of lecithin:cholesterol acyltransferase (LCAT) gene: intron retention causing LCAT deficiency. *Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism*, *1391*(2), 256–264. https://doi.org/10.1016/S0005-2760(97)00198-7

Li, T.-C., Scotti, P. D., Miyamura, T., & Takeda, N. (2007). Latent Infection of a New Alphanodavirus in an Insect Cell Line. *Journal of Virology*, *81*(20), 10890–10896. https://doi.org/10.1128/JVI.00807-07

Li, X., Schuler, M. A., & Berenbaum, M. R. (2007). Molecular Mechanisms of Metabolic Resistance to Synthetic and Natural Xenobiotics. *Annual Review of Entomology*, *52*(1), 231–253. https://doi.org/10.1146/annurev.ento.51.110104.151104

Li, X. Z., Roy, C. K., Dong, X., Bolcun-Filas, E., Wang, J., Han, B. W., … Zamore, P. D. (2013). An Ancient Transcription Factor Initiates the Burst of piRNA Production

during Early Meiosis in Mouse Testes. *Molecular Cell*, *50*(1), 67–81. https://doi.org/10.1016/j.molcel.2013.02.016

Liu, S., Zhou, S., Tian, L., Guo, E., Luan, Y., Zhang, J., & Li, S. (2011). Genome-wide identification and characterization of ATP-binding cassette transporters in the silkworm, Bombyx mori. *BMC Genomics*, *12*, 491. https://doi.org/10.1186/1471-2164-12-491

Lobo, N., Li, X., & Fraser, M. J. (1999). Transposition of the piggyBac element in embryos of Drosophila melanogaster, Aedes aegypti and Trichoplusia ni. *Molecular & General Genetics: MGG*, *261*(4–5), 803–810.

Loman, N. J., Misra, R. V., Dallman, T. J., Constantinidou, C., Gharbia, S. E., Wain, J., & Pallen, M. J. (2012). Performance comparison of benchtop high-throughput sequencing platforms. *Nature Biotechnology*, *30*(5), 434–439. https://doi.org/10.1038/nbt.2198

Lucas, K., & Raikhel, A. S. (2013). Insect MicroRNAs: Biogenesis, expression profiling and biological functions. *Insect Biochemistry and Molecular Biology*, *43*(1), 24–38. https://doi.org/10.1016/j.ibmb.2012.10.009

Lundberg, K. S., Shoemaker, D. D., Adams, M. W. W., Short, J. M., Sorge, J. A., & Mathur, E. J. (1991). High-fidelity amplification using a thermostable DNA polymerase isolated from Pyrococcus furiosus. *Gene*, *108*(1), 1–6. https://doi.org/10.1016/0378-1119(91)90480-Y

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., … others. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, *1*(1), 18.

Macosko, E. Z., Basu, A., Satija, R., Nemesh, J., Shekhar, K., Goldman, M., … McCarroll, S. A. (2015). Highly Parallel Genome-wide Expression Profiling of Individual Cells Using Nanoliter Droplets. *Cell*, *161*(5), 1202–1214. https://doi.org/10.1016/j.cell.2015.05.002

Malone, C. D., Brennecke, J., Dus, M., Stark, A., McCombie, W. R., Sachidanandam, R., & Hannon, G. J. (2009). Specialized piRNA Pathways Act in Germline and Somatic Tissues of the Drosophila Ovary. *Cell*, *137*(3), 522–535. https://doi.org/10.1016/j.cell.2009.03.040

Marchler-Bauer, A., Derbyshire, M. K., Gonzales, N. R., Lu, S., Chitsaz, F., Geer, L. Y., … Bryant, S. H. (2015). CDD: NCBI's conserved domain database. *Nucleic Acids Research*, *43*(D1), D222–D226. https://doi.org/10.1093/nar/gku1221

Marygold, S. J., Roote, J., Reuter, G., Lambertsson, A., Ashburner, M., Millburn, G. H., … Cook, K. R. (2007). The ribosomal protein genes and Minute loci of Drosophila melanogaster. *Genome Biology*, *8*, R216. https://doi.org/10.1186/gb-2007-8-10-r216

Matsumoto, N., Nishimasu, H., Sakakibara, K., Nishida, K. M., Hirano, T., Ishitani, R., … Nureki, O. (n.d.). Crystal Structure of Silkworm PIWI-Clade Argonaute Siwi Bound to piRNA. *Cell*. https://doi.org/10.1016/j.cell.2016.09.002

McEwen, F. L., & Hervey, G. E. R. (1956). An Evaluation of Newer Insecticides for Control of DDT-Resistant Cabbage Loopers. *Journal of Economic Entomology*, *49*(3), 385–387. https://doi.org/10.1093/jee/49.3.385

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., … DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

McKenna, A., Kernytsky, A. M., Sivachenko, A. Y., Philippakis, A. A., Hartl, C., Altshuler, D., … Fennell, T. J. (2011). A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nature Genetics*, *43*(5), ng.806. https://doi.org/10.1038/ng.806

Mercer, T. R., Clark, M. B., Andersen, S. B., Brunck, M. E., Haerty, W., Crawford, J., … Mattick, J. S. (2015). Genome-wide discovery of human splicing branchpoints. *Genome Research*, gr.182899.114. https://doi.org/10.1101/gr.182899.114

Mercer, T. R., Clark, M. B., Crawford, J., Brunck, M. E., Gerhardt, D. J., Taft, R. J., … Mattick, J. S. (2014). Targeted sequencing for gene discovery and quantification using RNA CaptureSeq. *Nature Protocols*, *9*(5), 989–1009. https://doi.org/10.1038/nprot.2014.058

Miesen, P., Girardi, E., & van Rij, R. P. (2015). Distinct sets of PIWI proteins produce arbovirus and transposon-derived piRNAs in Aedes aegypti mosquito cells. *Nucleic Acids Research*, gkv590. https://doi.org/10.1093/nar/gkv590

Miller, L. K., & Ball, L. A. (2012). *The Insect Viruses*. Springer Science & Business Media.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., … Finn, R. D. (2015). The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Research*, *43*(D1), D213–D221. https://doi.org/10.1093/nar/gku1243

Mitra, A., Skrzypczak, M., Ginalski, K., & Rowicka, M. (2015). Strategies for Achieving High Sequencing Accuracy for Low Diversity Samples and Avoiding Sample Bleeding Using Illumina Platform. *PLoS ONE*, *10*(4). https://doi.org/10.1371/journal.pone.0120520

Mohn, F., Handler, D., & Brennecke, J. (2015). piRNA-guided slicing specifies transcripts for Zucchini-dependent, phased piRNA biogenesis. *Science*, *348*(6236), 812–817. https://doi.org/10.1126/science.aaa1039

Mohn, F., Sienski, G., Handler, D., & Brennecke, J. (2014). The Rhino-Deadlock-Cutoff Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in Drosophila. *Cell*, *157*(6), 1364–1379. https://doi.org/10.1016/j.cell.2014.04.031

Montell, C. (2009). A taste of the Drosophila gustatory receptors. *Current Opinion in Neurobiology*, *19*(4), 345–353. https://doi.org/10.1016/j.conb.2009.07.001

Morin, G. B. (1989). The human telomere terminal transferase enzyme is a ribonucleoprotein that synthesizes TTAGGG repeats. *Cell*, *59*(3), 521–529. https://doi.org/10.1016/0092-8674(89)90035-4

Mouse Genome Sequencing Consortium, Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., … Lander, E. S. (2002). Initial sequencing and comparative analysis of the mouse genome. *Nature*, *420*(6915), 520–562. https://doi.org/10.1038/nature01262

Murray, C. L., Quaglia, M., Arnason, J. T., & Morris, C. E. (1994). A putative nicotine pump at the metabolic blood–brain barrier of the tobacco hornworm. *Journal of Neurobiology*, *25*(1), 23–34. https://doi.org/10.1002/neu.480250103

Nakagawa, T., Sakurai, T., Nishioka, T., & Touhara, K. (2005). Insect Sex-Pheromone Signals Mediated by Specific Combinations of Olfactory Receptors. *Science*, *307*(5715), 1638–1642. https://doi.org/10.1126/science.1106267

Nelson, D. R. (2009). The Cytochrome P450 Homepage. *Human Genomics*, *4*, 59. https://doi.org/10.1186/1479-7364-4-1-59

Nene, V., Wortman, J. R., Lawson, D., Haas, B., Kodira, C., Tu, Z. (Jake), … Severson, D. W. (2007). Genome Sequence of Aedes aegypti, a Major Arbovirus Vector. *Science*, *316*(5832), 1718–1723. https://doi.org/10.1126/science.1138878

Nikou, D., Ranson, H., & Hemingway, J. (2003). An adult-specific CYP6 P450 gene is overexpressed in a pyrethroid-resistant strain of the malaria vector, Anopheles gambiae. *Gene*, *318*, 91–102. https://doi.org/10.1016/S0378-1119(03)00763-7

Okamura, K., Ladewig, E., Zhou, L., & Lai, E. C. (2013). Functional small RNAs are generated from select miRNA hairpin loops in flies and mammals. *Genes & Development*, *27*(7), 778–792. https://doi.org/10.1101/gad.211698.112

Okamura, Katsutomo, Chung, W.-J., Ruby, J. G., Guo, H., Bartel, D. P., & Lai, E. C. (2008). The Drosophila hairpin RNA pathway generates endogenous short interfering RNAs. *Nature*, *453*(7196), 803–806. https://doi.org/10.1038/nature07015

Okamura, Katsutomo, & Lai, E. C. (2008). Endogenous small interfering RNAs in animals. *Nature Reviews Molecular Cell Biology*, *9*(9), 673–678. https://doi.org/10.1038/nrm2479

Padgett, R. A. (2012). New connections between splicing and human disease. *Trends in Genetics*, *28*(4), 147–154. https://doi.org/10.1016/j.tig.2012.01.001

Park, P. J. (2009). ChIP–seq: advantages and challenges of a maturing technology. *Nature Reviews Genetics*, *10*(10), 669–680. https://doi.org/10.1038/nrg2641

Porcelli, D., Barsanti, P., Pesole, G., & Caggese, C. (2007). The nuclear OXPHOS genes in insecta: a common evolutionary origin, a common cis-regulatory motif, a common destiny for gene duplicates. *BMC Evolutionary Biology*, *7*, 215. https://doi.org/10.1186/1471-2148-7-215

Rainford, J. L., Hofreiter, M., Nicholson, D. B., & Mayhew, P. J. (2014). Phylogenetic Distribution of Extant Richness Suggests Metamorphosis Is a Key Innovation Driving Diversification in Insects. *PLOS ONE*, *9*(10), e109085. https://doi.org/10.1371/journal.pone.0109085

Ranson, H., Claudianos, C., Ortelli, F., Abgrall, C., Hemingway, J., Sharakhova, M. V., … Feyereisen, R. (2002). Evolution of Supergene Families Associated with

Insecticide Resistance. *Science*, *298*(5591), 179–181. https://doi.org/10.1126/science.1076781

Reuter, M., Berninger, P., Chuma, S., Shah, H., Hosokawa, M., Funaya, C., … Pillai, R. S. (2011). Miwi catalysis is required for piRNA amplification-independent LINE1 transposon silencing. *Nature*, *480*(7376), 264–267. https://doi.org/10.1038/nature10672

Richards, S., Gibbs, R. A., Weinstock, G. M., Brown, S. J., Denell, R., Beeman, R. W., … Grossmann, D. (2008). The genome of the model beetle and pest Tribolium castaneum. *Nature*, *452*(7190), 949–955. https://doi.org/10.1038/nature06784

Richards, S., Liu, Y., Bettencourt, B. R., Hradecky, P., Letovsky, S., Nielsen, R., … Gibbs, R. A. (2005). Comparative genome sequencing of Drosophila pseudoobscura: Chromosomal, gene, and cis-element evolution. *Genome Research*, *15*(1), 1–18. https://doi.org/10.1101/gr.3059305

Robertson, H. M., & Gordon, K. H. J. (2006). Canonical TTAGG-repeat telomeres and telomerase in the honey bee, Apis mellifera. *Genome Research*, *16*(11), 1345–1351. https://doi.org/10.1101/gr.5085606

Sathyanarayanan, S., Zheng, X., Kumar, S., Chen, C.-H., Chen, D., Hay, B., & Sehgal, A. (2008). Identification of novel genes involved in light-dependent CRY degradation through a genome-wide RNAi screen. *Genes & Development*, *22*(11), 1522–1533. https://doi.org/10.1101/gad.1652308

Sato, K., Tanaka, K., & Touhara, K. (2011). Sugar-regulated cation channel formed by an insect gustatory receptor. *Proceedings of the National Academy of Sciences*, *108*(28), 11680–11685. https://doi.org/10.1073/pnas.1019622108

Schirmer, M., Smeekens, S. P., Vlamakis, H., Jaeger, M., Oosting, M., Franzosa, E. A., … Xavier, R. J. (2016). Linking the Human Gut Microbiome to Inflammatory Cytokine Production Capacity. *Cell*, *167*(4), 1125-1136.e8. https://doi.org/10.1016/j.cell.2016.10.020

Scott, J. G. (1999). Cytochromes P450 and insecticide resistance. *Insect Biochemistry and Molecular Biology*, *29*(9), 757–777. https://doi.org/10.1016/S0965-1748(99)00038-7

Shichida, Y., & Matsuyama, T. (2009). Evolution of opsins and phototransduction. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *364*(1531), 2881–2895. https://doi.org/10.1098/rstb.2009.0051

Shin, C., Nam, J.-W., Farh, K. K.-H., Chiang, H. R., Shkumatava, A., & Bartel, D. P. (2010). Expanding the MicroRNA Targeting Code: Functional Sites with Centered Pairing. *Molecular Cell*, *38*(6), 789–802. https://doi.org/10.1016/j.molcel.2010.06.005

Shiroguchi, K., Jia, T. Z., Sims, P. A., & Xie, X. S. (2012). Digital RNA sequencing minimizes sequence-dependent bias and amplification noise with optimized single-molecule barcodes. *Proceedings of the National Academy of Sciences*, *109*(4), 1347–1352. https://doi.org/10.1073/pnas.1118018109

Sienski, G., Dönertas, D., & Brennecke, J. (2012). Transcriptional Silencing of Transposons by Piwi and Maelstrom and Its Impact on Chromatin State and Gene Expression. *Cell*, *151*(5), 964–980. https://doi.org/10.1016/j.cell.2012.10.040

Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., & Zdobnov, E. M. (2015). BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, btv351. https://doi.org/10.1093/bioinformatics/btv351

Singh, R. K., & Cooper, T. A. (2012). Pre-mRNA splicing in disease and therapeutics. *Trends in Molecular Medicine*, *18*(8), 472–482. https://doi.org/10.1016/j.molmed.2012.06.006

Skene, P. J., & Henikoff, S. (2017). An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *ELife*, *6*, e21856. https://doi.org/10.7554/eLife.21856

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31. https://doi.org/10.1186/1471-2105-6-31

Slone, J., Daniels, J., & Amrein, H. (2007). Sugar Receptors in Drosophila. *Current Biology*, *17*(20), 1809–1816. https://doi.org/10.1016/j.cub.2007.09.027

Smith, C. D., Zimin, A., Holt, C., Abouheif, E., Benton, R., Cash, E., … Tsutsui, N. D. (2011). Draft genome of the globally widespread and invasive Argentine ant (Linepithema humile). *Proceedings of the National Academy of Sciences*, *108*(14), 5673–5678. https://doi.org/10.1073/pnas.1008617108

Smith, C. R., Smith, C. D., Robertson, H. M., Helmkampf, M., Zimin, A., Yandell, M., … Gadau, J. (2011). Draft genome of the red harvester ant Pogonomyrmex barbatus. *Proceedings of the National Academy of Sciences*, *108*(14), 5667–5672. https://doi.org/10.1073/pnas.1007901108

Smith, T., Heger, A., & Sudbery, I. (2017). UMI-tools: modeling sequencing errors in Unique Molecular Identifiers to improve quantification accuracy. *Genome Research*, *27*(3), 491–499. https://doi.org/10.1101/gr.209601.116

Sokol, N. S., & Ambros, V. (2005). Mesodermally expressed Drosophila microRNA-1 is regulated by Twist and is required in muscles during larval growth. *Genes & Development*, *19*(19), 2343–2354. https://doi.org/10.1101/gad.1356105

Soumillon, M., Necsulea, A., Weier, M., Brawand, D., Zhang, X., Gu, H., … Kaessmann, H. (2013). Cellular Source and Mechanisms of High Transcriptome Complexity in the Mammalian Testis. *Cell Reports*, *3*(6), 2179–2190. https://doi.org/10.1016/j.celrep.2013.05.031

Stanke, M., Tzvetkova, A., & Morgenstern, B. (2006). AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome Biology*, *7*(1), 1–8. https://doi.org/10.1186/gb-2006-7-s1-s11

Stegle, O., Teichmann, S. A., & Marioni, J. C. (2015). Computational and analytical challenges in single-cell transcriptomics. *Nature Reviews Genetics*, *16*(3), 133–145. https://doi.org/10.1038/nrg3833

Strange, R. C., Spiteri, M. A., Ramachandran, S., & Fryer, A. A. (2001). Glutathione-S-transferase family of enzymes. *Mutation Research/Fundamental and Molecular*

Mechanisms of Mutagenesis, *482*(1–2), 21–26. https://doi.org/10.1016/S0027-5107(01)00206-8

Suen, G., Teiling, C., Li, L., Holt, C., Abouheif, E., Bornberg-Bauer, E., … Currie, C. R. (2011). The Genome Sequence of the Leaf-Cutter Ant Atta cephalotes Reveals Insights into Its Obligate Symbiotic Lifestyle. *PLOS Genetics*, *7*(2), e1002007. https://doi.org/10.1371/journal.pgen.1002007

Suetsugu, Y., Futahashi, R., Kanamori, H., Kadono-Okuda, K., Sasanuma, S., Narukawa, J., … Mita, K. (2013). Large Scale Full-Length cDNA Sequencing Reveals a Unique Genomic Landscape in a Lepidopteran Model Insect, Bombyx mori. *G3: Genes|Genomes|Genetics*, *3*(9), 1481–1492. https://doi.org/10.1534/g3.113.006239

Sztal, T., Chung, H., Berger, S., Currie, P. D., Batterham, P., & Daborn, P. J. (2012). A Cytochrome P450 Conserved in Insects Is Involved in Cuticle Formation. *PLOS ONE*, *7*(5), e36544. https://doi.org/10.1371/journal.pone.0036544

Taggart, A. J., DeSimone, A. M., Shih, J. S., Filloux, M. E., & Fairbrother, W. G. (2012). Large-scale mapping of branchpoints in human pre-mRNA transcripts in vivo. *Nature Structural & Molecular Biology*, *19*(7), 719–721. https://doi.org/10.1038/nsmb.2327

Taggart, A. J., Lin, C.-L., Shrestha, B., Heintzelman, C., Kim, S., & Fairbrother, W. G. (2017). Large-scale analysis of branchpoint usage across species and cell lines. *Genome Research*, *27*(4), 639–649. https://doi.org/10.1101/gr.202820.115

Tam, O. H., Aravin, A. A., Stein, P., Girard, A., Murchison, E. P., Cheloufi, S., … Hannon, G. J. (2008). Pseudogene-derived small interfering RNAs regulate gene expression in mouse oocytes. *Nature*, *453*(7194), 534–538. https://doi.org/10.1038/nature06904

Tan, F. L., & Yin, J. Q. (2004). RNAi, a new therapeutic strategy against viral infection. *Cell Research*, *14*(6), 460–466. https://doi.org/10.1038/sj.cr.7290248

Terakita, A. (2005). The opsins. *Genome Biology*, *6*, 213. https://doi.org/10.1186/gb-2005-6-3-213

The International Silkworm Genome. (2008). The genome of a lepidopteran model insect, the silkworm Bombyx mori. *Insect Biochemistry and Molecular Biology*, *38*(12), 1036–1045. https://doi.org/10.1016/j.ibmb.2008.11.004

Traut, W., Sahara, K., & Marec, F. (2008). Sex Chromosomes and Sex Determination in Lepidoptera. *Sexual Development*, *1*(6), 332–346. https://doi.org/10.1159/000111765

Vagin, V. V., Sigova, A., Li, C., Seitz, H., Gvozdev, V., & Zamore, P. D. (2006). A Distinct Small RNA Pathway Silences Selfish Genetic Elements in the Germline. *Science*, *313*(5785), 320–324. https://doi.org/10.1126/science.1129333

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … DePristo, M. A. (2002). From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline. In *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc. https://doi.org/10.1002/0471250953.bi1110s43

van Schooten, B., Jiggins, C. D., Briscoe, A. D., & Papa, R. (2016). Genome-wide analysis of ionotropic receptors provides insight into their evolution in Heliconius butterflies. *BMC Genomics*, *17*, 254. https://doi.org/10.1186/s12864-016-2572-y

Varoquaux, N., Liachko, I., Ay, F., Burton, J. N., Shendure, J., Dunham, M. J., … Noble, W. S. (2015). Accurate identification of centromere locations in yeast genomes using Hi-C. *Nucleic Acids Research*, *43*(11), 5331–5339. https://doi.org/10.1093/nar/gkv424

Velarde, R. A., Sauer, C. D., O. Walden, K. K., Fahrbach, S. E., & Robertson, H. M. (2005). Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. *Insect Biochemistry and Molecular Biology*, *35*(12), 1367–1377. https://doi.org/10.1016/j.ibmb.2005.09.001

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., … Zhu, X. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Vourekas, A., Alexiou, P., Vrettos, N., Maragkakis, M., & Mourelatos, Z. (2016). Sequence-dependent but not sequence-specific piRNA adhesion traps mRNAs to the germ plasm. *Nature*, *advance online publication*. https://doi.org/10.1038/nature17150

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE*, *9*(11), e112963. https://doi.org/10.1371/journal.pone.0112963

Walters, J. R., & Hardcastle, T. J. (2011). Getting a Full Dose? Reconsidering Sex Chromosome Dosage Compensation in the Silkworm, Bombyx mori. *Genome Biology and Evolution*, *3*, 491–504. https://doi.org/10.1093/gbe/evr036

Wang, W., Lin, C., Lu, D., Ning, Z., Cox, T., Melvin, D., … Liu, P. (2008). Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *Proceedings of the National Academy of Sciences*, *105*(27), 9290–9295. https://doi.org/10.1073/pnas.0801017105

Wang, Z., Gerstein, M., & Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, *10*(1), 57–63. https://doi.org/10.1038/nrg2484

Wanner, K. W., & Robertson, H. M. (2008). The gustatory receptor family in the silkworm moth Bombyx mori is characterized by a large expansion of a single lineage of putative bitter receptors. *Insect Molecular Biology*, *17*(6), 621–629. https://doi.org/10.1111/j.1365-2583.2008.00836.x

Watanabe, T., Totoki, Y., Toyoda, A., Kaneda, M., Kuramochi-Miyagawa, S., Obata, Y., … Sasaki, H. (2008). Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature*, *453*(7194), 539–543. https://doi.org/10.1038/nature06908

Werren, J. H., Richards, S., Desjardins, C. A., Niehuis, O., Gadau, J., Colbourne, J. K., & Group, T. N. G. W. (2010). Functional and Evolutionary Insights from the Genomes of Three Parasitoid Nasonia Species. *Science*, *327*(5963), 343–348. https://doi.org/10.1126/science.1178028

Wickham, T. J., Davis, T., Granados, R. R., Shuler, M. L., & Wood, H. A. (1992). Screening of Insect Cell Lines for the Production of Recombinant Proteins and Infectious Virus in the Baculovirus Expression System. *Biotechnology Progress*, *8*(5), 391–396. https://doi.org/10.1021/bp00017a003

Xavier Bellés, David Martín, & Piulachs, M.-D. (2005). The Mevalonate Pathway and the Synthesis of Juvenile Hormone in Insects. *Annual Review of Entomology*, *50*(1), 181–199. https://doi.org/10.1146/annurev.ento.50.071803.130356

Yaari, G., & Kleinstein, S. H. (2015). Practical guidelines for B-cell receptor repertoire sequencing analysis. *Genome Medicine*, *7*, 121. https://doi.org/10.1186/s13073-015-0243-2

Yoshida, K., Sanada, M., Shiraishi, Y., Nowak, D., Nagata, Y., Yamamoto, R., … Ogawa, S. (2011). Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature*, *478*(7367), 64–69. https://doi.org/10.1038/nature10496

You, M., Yue, Z., He, W., Yang, X., Yang, G., Xie, M., … Wang, J. (2013). A heterozygous moth genome provides insights into herbivory and detoxification. *Nature Genetics*, *45*(2), 220–225. https://doi.org/10.1038/ng.2524

Yu, Q., Lu, C., Li, B., Fang, S., Zuo, W., Dai, F., … Xiang, Z. (2008). Identification, genomic organization and expression pattern of glutathione S-transferase in the silkworm, Bombyx mori. *Insect Biochemistry and Molecular Biology*, *38*(12), 1158–1164. https://doi.org/10.1016/j.ibmb.2008.08.002

Yu, Q.-Y., Lu, C., Li, W.-L., Xiang, Z.-H., & Zhang, Z. (2009). Annotation and expression of carboxylesterases in the silkworm, Bombyx mori. *BMC Genomics*, *10*, 553. https://doi.org/10.1186/1471-2164-10-553

Yusa, K. (2015). piggyBac Transposon. *Microbiology Spectrum*, *3*(2). https://doi.org/10.1128/microbiolspec.MDNA3-0028-2014

Zhan, S., Merlin, C., Boore, J. L., & Reppert, S. M. (2011). The Monarch Butterfly Genome Yields Insights into Long-Distance Migration. *Cell*, *147*(5), 1171–1185. https://doi.org/10.1016/j.cell.2011.09.052

Zhang, X., Tiewsiri, K., Kain, W., Huang, L., & Wang, P. (2012). Resistance of Trichoplusia ni to Bacillus thuringiensis Toxin Cry1Ac Is Independent of Alteration of the Cadherin-Like Receptor for Cry Toxins. *PLOS ONE*, *7*(5), e35991. https://doi.org/10.1371/journal.pone.0035991

Zhang, Z., Theurkauf, W. E., Weng, Z., & Zamore, P. D. (2012). Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. *Silence*, *3*(1), 9. https://doi.org/10.1186/1758-907X-3-9

Zhang, Z., Wang, J., Schultz, N., Zhang, F., Parhad, S. S., Tu, S., … Theurkauf, W. E. (2014). The HP1 Homolog Rhino Anchors a Nuclear Complex that Suppresses piRNA Precursor Splicing. *Cell*, *157*(6), 1353–1363. https://doi.org/10.1016/j.cell.2014.04.030

CURRICULUM VITAE