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

Yu Fu,[1,2] Yujing Yang,[3†] Han Zhang,[3] Gwen Farley,[3] Junling Wang,[3] Kaycee A. Quarles,[3] Zhiping Weng,[2,4*] and Phillip D. Zamore[3*]

[1]Bioinformatics Program, Boston University, 44 Cummington Mall, Boston, MA 02215, USA

[2]Program in Bioinformatics and Integrative Biology, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605, USA

[3]RNA Therapeutics Institute and Howard Hughes Medical Institute, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605, USA

[4]Department of Biochemistry and Molecular Pharmacology, University of Massachusetts Medical School, 368 Plantation Street, Worcester, MA 01605, USA

***For correspondence:** zhiping.weng@umassmed.edu (ZW), phillip.zamore@umassmed.edu (PDZ)

†Present addresses: Yujing Yang, Laboratory of Pathology, State Key Laboratory of Biotherapy and Department of Pathology, West China Hospital, West China Medical School, Sichuan University, Chengdu, China

Abstract

We report a draft assembly of the genome of Hi5 cells from the lepidopteran insect pest, *Trichoplusia ni*, assigning 90.6% of bases to one of 28 chromosomes and predicting 14,037 protein-coding genes. Chemoreception and detoxification gene families reveal *T. ni*-specific gene expansions that may explain its widespread distribution and rapid adaptation to insecticides. Transcriptome and small RNA data from thorax, ovary, testis, and the germline-derived Hi5 cell line show distinct expression profiles for 295 microRNA- and >393 piRNA-producing loci, as well as 39 genes encoding small RNA pathway proteins. Nearly all of the W chromosome is devoted to piRNA production, and *T. ni* siRNAs are not 2´-*O*-methylated. To enable use of Hi5 cells as a model system, we have established genome editing and single-cell cloning protocols. The *T. ni* genome provides insights into pest control and allows Hi5 cells to become a new tool for studying small RNAs ex vivo.

Keywords: genome assembly; piRNA; siRNA; miRNA; sex determination; cultured cell; High Five; CRISPR; cabbage looper; *Trichoplusia ni;* Lepidoptera

Introduction

Lepidoptera (moths and butterflies), one of the most species-rich orders of insects, comprises more than 170,000 known species (*Mallet, J, 2007*; *Chapman, AD, 2009*), 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, JL et al., 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, J, 2001*). *T. ni* has evolved resistance to the chemical insecticide Dichlorodiphenyltrichloroethane (DDT; [*McEwen, FL, Hervey, GER, 1956*]) and the biological insecticide *Bacillus thuringiensis* toxin (*Janmaat, AF, Myers, J, 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, RR et al., 1986*; *Granados, RR et al., 1994*). Hi5 cells are a mainstay of recombinant protein production using baculoviral vectors (*Wickham, TJ et al., 1992*) and hold promise for the commercial-scale production of recombinant adeno-associated virus for human gene therapy (*Kotin, RM, 2011*; *van Oers, MM et al., 2015*). Hi5 cells produce abundant microRNAs (miRNAs), small interfering RNAs (siRNAs), and PIWI-interacting RNAs (*Kawaoka, S 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 (*Saito, K et al., 2006*; *Vagin, VV et al., 2006*; *Brennecke, J et al., 2007*; *Houwing, S et al., 2007*; *Aravin, AA et al., 2007*; *Kawaoka, S et al., 2008*). 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, WF, 1970*), are readily

transfected, and—unlike *B. mori* BmN4 cells (*Iwanaga, M 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.

To further understand this agricultural pest and its Hi5 cell line, we combined divers genomic sequencing data to assemble a chromosome-level, high-quality *T. ni* genome. Half the genome sequence resides in scaffolds >14.2 megabases (Mb), and >90% is assembled into 28 chromosome-length scaffolds. Automated gene prediction and subsequent manual curation, aided by extensive RNA-seq data, allowed us to examine gene orthology, gene families such as detoxification proteins, sex determination genes, and the miRNA, siRNA, and piRNA pathways. Our data allowed assembly of the gene-poor, repeat-rich W chromosome, which remarkably produces piRNAs across most of its length. To enable the use of cultured *T. ni* Hi5 cells as a novel insect model system, we established methods for efficient genome editing using the CRISPR/Cas9 system (*Ran, FA et al., 2013*) as well as single-cell cloning. With these new tools, *T. ni* promises to become a powerful companion to flies to study gene expression, small RNA biogenesis and function, and mechanisms of insecticide resistance in vivo and in cultured cells.

Results

Genome sequencing and assembly

We combined Pacific Biosciences long reads and Illumina short reads (Figure 1A, Table 1, and Materials and methods) to sequence genomic DNA from Hi5 cells and *T. ni* male and female pupae. The initial genome assembly from long reads (46.4× coverage with reads >5 kb) was polished using paired-end (172.7× coverage) and mate-pair reads (172.0× coverage) to generate 1,976 contigs spanning 368.2 megabases (Mb). Half of genomic bases reside in contigs >621.9 kb (N50). Hi-C long-range scaffolding (186.5×

coverage) produced 1,031 scaffolds (N50 = 14.2 Mb), with >90% of the sequences assembled into 28 major scaffolds. Karyotyping of metaphase Hi5 cells revealed that these cells have 112 ± 5 chromosomes (Figure 1B, Figure 1—figure supplement 1). Because lepidopteran cell lines are typically tetraploid (*Hink WF, 1972*), we conclude that the ~368.2 Mb *T. ni* genome comprises 28 chromosomes: 26 autosomes plus W and Z sex chromosomes (see below).

To evaluate the completeness of the assembled *T. ni* genome, we compared it to the Arthropoda data set of the Benchmark of Universal Single-Copy Orthologs (*Simão, FA et al., 2015*) (BUSCO v3). The *T. ni* genome assembly captures 97.5% of these gene orthologs, more than either the silkworm (95.5%) or monarch butterfly (*D. plexippus*; 97.0%) genomes (Supplementary file 1A). All 79 ribosomal proteins conserved between mammals and *D. melanogaster* (*Yoshihama, M et al., 2002*; *Marygold, SJ et al., 2007*) have orthologs in *T. ni*, further evidence of the completeness of the genome assembly (Supplementary file 1B). Finally, a search for genes in the highly conserved nuclear oxidative phosphorylation (OXPHOS) pathway (*Porcelli, D et al., 2007*) uncovered *T. ni* orthologs for all known *D. melanogaster* OXPHOS genes (Supplementary file 1C).

The genomes of wild insect populations are typically highly heterogeneous, which poses a significant impediment to assembly (*Keeling, CI et al., 2013*; *You, M et al., 2013*). We were unable to generate an isogenic *T. ni* strain by inbreeding. Therefore, our *T. ni* sequence reflects the genome of Hi5 cells, not cabbage looper itself. Hi5 cells presumably derive from a single immortalized, germline founder cell, which should reduce genomic variation among the cell line's four sets of chromosomes. To test this supposition, we identified the sequence variants in the Hi5 genome. In total, we called variants at 165,370 genomic positions (0.0449% of the genome assembly), with 2,710 in predicted coding regions (0.0132% of coding sequence), indicating that the genome of Hi5 cells is fairly homogenous. For the majority (88.8%) of these

genomic positions (covering 0.0399% of the genome), only one copy of the chromosome has the variant allele while the other three chromosomal copies match the reference genome. We can make three conclusions. First, Hi5 cells originated from a single founder cell or a homogenous population of cells. Second, the founder cells were haploid. Third, most sequence variants were acquired after the original derivation of the line from *T. ni* eggs.

We also assembled de novo *T. ni* genomes using paired-end DNA-seq data obtained from male and female pupae, but the resulting assemblies are fragmented (scaffold N50 ≤2.4 kb, Supplementary file 1D), likely due to the limitations of short-insert libraries and the high levels of heterozygosity commonly observed for genomes of wild insect populations (*Keeling, CI et al., 2013*; *You, M et al., 2013*). The animal genome contigs are highly concordant with the Hi5 genome , with ≤1.37% of animal contigs misassembled (Supplementary file 1D). Although we cannot determine scaffold-level differences between the animal and Hi5 cells, at the contig-level the Hi5 genome assembly is representative of the *T. ni* animal genome.

Gene orthology

We annotated 14,034 protein-coding genes in the *T. ni* genome (Supplementary file 1E), similar to other Lepidoptera (*Challis, RJ et al., 2016*). Analysis of the homology of *T. ni* genes to genes in 20 species that span the four common insect orders (Lepidoptera, Diptera, Coleoptera, Hymenoptera), non-insect arthropods, and mammals defines 30,448 orthology groups each containing orthologous proteins from two or more species (*Hirose, Y, Manley, JL, 1997*); 9,112 groups contain at least one *T. ni* gene. In all, 10,936 *T. ni* protein-coding genes are orthologous to at least one gene among the 20 reference species (Figure 1C, Figure 1—figure supplement 2).

T. ni contains 2,287 Lepidoptera-specific orthology groups (*T. ni*, *B. mori*, *D. plexippus*, and *P. xylostella* [diamondback moth]). Far fewer orthology groups are

unique to Diptera (404), Coleoptera (371), or Hymenoptera (1,344), suggesting that the lepidopteran lifestyle requires more order-specific genes. The *T. ni* genome additionally contains 3,098 orphan protein-coding genes for which we could detect no orthologous sequences in the 20 reference species. Of these orphan genes, 14.5% are present as two or more copies in the genome ("in-paralogs"), suggesting they evolved recently. Some of these in-paralogs may have arisen by gene duplication after the divergence of *T. ni* and *B. mori* ~111 mya (*Gaunt, MW, Miles, MA, 2002*; *Rota-Stabelli, O et al., 2013*; *Wheat, CW, Wahlberg, N, 2013*; *Rainford, JL et al., 2014*).

Opsins

The ability of insects to respond to light is crucial to their survival. Opsins, members of the G-protein-coupled receptor superfamily, play important roles in vision. Covalently bound to light-sensing chromophores, opsins absorb photons and activate the downstream visual transduction cascade (*Terakita, A, 2005*). The *T. ni* genome encodes ultraviolet, blue, and long-wavelength opsins. Thus, this nocturnal insect retains the full repertoire of insect opsins and has color vision (*Zimyanin, VL et al., 2008*) (Figure 1—figure supplement 3). *T. ni* also encodes an ortholog of the non-visual Rh7 opsin, which is found in a variety of insects (*Initiative, IGG, 2014*; *Futahashi, R et al., 2015*). In the *D. melanogaster* brain, Rh7 opsin participates in the entrainment of circadian rhythms by sensing violet light (*Ni, JD et al., 2017*). *T. ni* also encodes an ortholog of the vertebrate-like opsin, pterosin, which was first detected in the honeybee (*A. mellifera*) brain and is found widely among insects except for *Drosophilid* flies (*Velarde, RA et al., 2005*).

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, W et al., 2007*).

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 2A, Figure 2—figure supplement 1A). 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 (*Walters, JR et al., 2015*; *Gu, L et al., 2017*). 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 2B). 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, Figure 2—figure supplement 1B and 1C). 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, JR,*

Hardcastle, TJ, 2011), *Cydia pomonella* (*Gu, L et al., 2017*), and *Heliconius* butterflies (*Walters, JR et al., 2015*), 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, MR et al., 2016*) or *B. mori* (*The, ISG, 2008*), and earlier efforts to assemble the silkworm W resulted in fragmented sequences containing transposons (*Abe, H et al., 2005*; *Abe, H et al., 2008*; *Kawaoka, S et al., 2011*). The monarch genome scaffold continuity (N50 = 0.207 Mb versus N50 = 14.2 Mb for *T. ni*; [*Zhan, S 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 2B). Other lepidopteran W chromosomes are similarly enriched in repeats and depleted of genes (*Abe, H et al., 2005*; *Fuková, I et al., 2005*; *Traut, W et al., 2007*).

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 (Figure 2—figure supplement 1D). 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, S et al., 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 2C shows the multiple sequence alignment of the CCCH zinc finger domain of Masc proteins from several lepidopteran species.

Telomeres and centromeres

Like many non-dipteran insects, *T. ni* has a single telomerase gene and telomeres containing TTAGG repeats (*Sahara, K et al., 1999*). We found 40 $(TTAGG)_n$ stretches longer than 100 nt (mean ± S.D. = 600 ± 800 nt), nine at and 31 near contig boundaries (Supplementary file 1F; distance between $(TTAGG)_n$ and contig boundary = 5,000 ± 6,000 nt for the 40 stretches), indicating that our assembly captures the sequences of many telomeres. More than half (59%) of the sequences flanking the $(TTAGG)_n$ repeats are transposons, and ~49% of these belong to the non-long-terminal-repeat LINE/R1 family (Supplementary file 1G). These telomeric and subtelomeric characteristics of *T. ni* resemble those of *B. mori* (*Fujiwara, H et al., 2005*).

Lepidopteran chromosomes generally lack a coherent, monocentric centromere and are instead holocentric or diffuse (*Labbé, R et al., 2011*), and the silkworm, monarch butterfly, and diamondback moth genomes do not encode CenH3, a protein associated with monocentric chromosomes. The *T. ni* genome similarly does not contain a gene for CenH3, suggesting that its chromosomes are also holocentric.

CpG content and DNA methylation

The *T. ni* genome is 35.6% GC, slightly less than *B. mori* (37.3%). The distributions of observed/expected CpG ratios in genes and across the genome (Figure 2—supplement 2A) reveal that *T. ni* is similar to other lepidopterans (silkworm, monarch butterfly, diamondback moth) and a coleopteran species (red flour beetle, *T. castaneum*), but different from honeybee and fruit fly. The honeybee genome has a high CpG content in genes and exhibits a bimodal CpG distribution across the genome as a whole; the fruit fly genome is uniformly depleted of CpG dinucleotides. The differences in CpG patterns reflect the presence of both the DNMT1 and DNMT3 DNA methyltransferases in the honeybee, the absence of either in fruit fly, and the presence of only DNMT1 in *T. ni*, *B. mori*, *D. plexippus*, *P. xylostella*, and *T. castaneum*. Thus, like many other insects, the

T. ni genome likely has low levels of DNA methylation (*Xiang, H et al., 2010*; *Glastad, KM et al., 2011*).

Transposons and repeats

The *T. ni* genome contains 75.3 Mb of identifiable repeat elements (20.5% of the assembly), covering 458 repeat families (Figure 2—figure supplement 2B, Supplementary file 1H). With this level of repeat content, *T. ni* fits well with the positive correlation between genome size and repeat content among lepidopteran genomes (Figure 2—figure supplement 2C).

The DNA transposon piggyBac was originally isolated from a *T. ni* cell line (*Fraser, MJ et al., 1983*) and transposes effectively in a variety of species (*Lobo, N et al., 1999*; *Bonin, CP, Mann, RS, 2004*; *Wang, W et al., 2008*). We identified 262 copies of piggyBac in the Hi5 cell genome assembly. The family divergence rate of piggyBac is ~0.17%, substantially lower than other transposon families in the genome (Supplementary file 1I provides divergence rates for all transposon families). Among the individual piggyBac elements in the *T. ni* genome, 71 are specific to Hi5 cells. Compared to the 191 piggyBac insertions shared between *T. ni* and Hi5 cells (divergence rate = 0.22%), the Hi5 cell-specific elements are more highly conserved (divergence rate = 0.04%). We conclude that the piggyBac transposon entered the *T. ni* genome more recently than other transposons and, likely driven by the presence of many active piggyBac elements, expanded further during the immortalization of Hi5 cells in culture.

microRNAs

microRNAs (miRNAs) are ~22 nt non-coding RNAs that regulate mRNA stability and translation (*He, L, Hannon, GJ, 2004*; *Gao, G et al., 2005*). In insects, miRNA targets function in metamorphosis, reproduction, diapause, and other pathways of insect physiology and development (*Lucas, K, Raikhel, AS, 2013*). To characterize the *T. ni*

miRNA pathway, we sequenced RNA and small RNA from ovary, testis, thorax, and Hi5 cells. Then, we manually identified miRNA biogenesis genes such as *dcr-1*, *pasha*, *drosha*, and *ago2* (Supplementary file 2A) and computationally predicted 295 miRNA genes (Figure 3, Supplementary file 3A and Supplementary file 4), including 77 conserved, 31 Lepidoptera-specific, and 187 novel, *T. ni*-specific miRNAs.

In thorax, 222 of 270 miRNAs had comparable abundance in males and females (≤2-fold difference or false discovery rate [FDR] ≥0.1; Figure 3A). Of the 48 miRNAs having significantly different abundances in female and male thorax (>2-fold difference and FDR <0.1; Figure 3A), miR-1a, let-7, and miR-278 were highly abundant (>1000 parts per million [ppm]) in either female or male thorax. miR-1a, a miRNA thought to be expressed in all animal muscle, was the most abundant miRNA in thorax in both sexes, but was 2.2-fold more abundant in males. miR-1 was previously shown to regulate muscle development in fruit flies (*Sokol, NS, Ambros, V, 2005*) and to increase when locusts transition from solitary to swarming (*Wei, Y et al., 2009*). *T. ni* let-7, which has the same mature miRNA sequence as its *D. melanogaster*, *C. elegans*, and mammalian counterparts (*Lagos-Quintana, M et al., 2001*) was also more abundant in males, whereas miR-278 was 2.6-fold more abundant in females. let-7 may act in sex-specific pathways in metamorphosis (*Caygill, EE, Johnston, LA, 2008*), whereas miR-278 may play a sex-specific role in regulating energy homeostasis (*Teleman, AA et al., 2006*).

A subset of less well conserved miRNAs was also differentially expressed between male and female thorax. In general, poorly conserved miRNAs were less abundant: the median expression level for conserved miRNAs was 316 ppm, but only 161 ppm for Lepidoptera-specific and 4.22 ppm for *T. ni*-specific miRNAs. However, mir-2767, a Lepidoptera-specific miRNA, and three *T. ni*-specific miRNAs (mir-novel1, mir-novel4, mir-novel11) were both abundant (>1000 ppm) and differentially expressed in males and female thorax. We speculate that these recently evolved miRNAs may prove useful as targets for pest management.

Ovary, testis, and Hi5 cells have distinct miRNA expression profiles. We analyzed the expression patterns of the 44 most abundant miRNAs (Figure 3B and 3C), which explain 90% of miRNA reads in a tissue or cell line. Thirteen were expressed in ovaries, testes, and Hi5 cells. Of these 13, 11 were significantly more abundant in testis, 5 in ovary, and 3 in Hi5 cells (Figure 3B), suggesting that these miRNAs have important tissue- or cell-type-specific roles. miR-31 and miR-375, highly expressed in *T. ni* testis, are both mammalian tumor suppressors (*Creighton, CJ et al., 2010*; *Kinoshita, T et al., 2012*). miR-989, the most abundant miRNA in *T. ni* ovaries, plays an important role in border cell migration during *Drosophila* oogenesis (*Kugler, J-M et al., 2013*). miR-10, a miRNA in the Hox gene cluster, was preferentially expressed in Hi5 cells; its orthologs have been implicated in development and cancer (*Lund, AH, 2009*), suggesting miR-10 played a role in the immortalization of the germline cells from which Hi5 cells derive.

siRNAs

siRNAs, typically 20–22 nt long, regulate gene expression, defend against viral infection, and silence transposons (*Agrawal, N et al., 2003*; *van Rij, RP et al., 2006*; *Sanchez-Vargas, I et al., 2009*; *Tyler, DM et al., 2008*; *Tam, OH et al., 2008*; *Zambon, RA et al., 2006*; *Chung, WJ et al., 2008*; *Okamura, K et al., 2008b*; *Czech, B et al., 2008*; *Okamura, K et al., 2008b*; *Flynt, A et al., 2009*). They are processed by Dicer from double-stranded RNAs or hairpins into short double-stranded fragments bearing two-nucleotide, overhanging 3′ ends, which are subsequently loaded into Argonaute proteins (*Bernstein, E et al., 2001*; *Elbashir, SM et al., 2001*; *Siomi, H, Siomi, MC, 2009*). siRNAs require extensive sequence complementarity to their targets to elicit Argonaute-catalyzed target cleavage.

Endogenous siRNAs from transposons and *cis*-NATs

Endogenous siRNAs (endo-siRNAs) can derive from transposon RNAs, *cis*-natural antisense transcripts (*cis*-NATs), and long hairpin RNAs (*Czech, B et al., 2008*;

Ghildiyal, M et al., 2008; *Okamura, K et al., 2008a*; *Chung, WJ et al., 2008*; *Kawamura, Y et al., 2008*; *Okamura, K et al., 2008a*; *Tam, OH et al., 2008*; *Watanabe, T et al., 2008*) (hpRNAs). In *T. ni* ovary, testis, thorax, and Hi5 cells, 20.7–52.4% of siRNAs map to transposons, suggesting *T. ni* endogenous siRNAs suppress transposons in both the soma and the germline. Among the non-transposon siRNAs, <4.6% map to predicted hairpins, while 11.6–31.3% siRNAs map to *cis*-NATs (Supplementary file 3B).

Exogenous siRNAs against a virus

Hi5 cells are latently infected with a positive-sense, bipartite alphanodavirus, TNCL virus (Li et al., 2007, #97210; Miller and Ball, 2012, #84273) (Tn5 Cell Line virus). We asked if TNCL virus RNA is present in our *T. ni* samples and whether the RNAi pathway provides anti-viral defense via TNCL virus-derived siRNAs. We detected no viral RNA in the *T. ni* ovary, testis, or thorax transcriptome, but both TNCL virus RNA1 (5,010 fragments per kilobase of transcript per million mapped reads [FPKM]) and RNA2 (8,280 FPKM) were readily found in the Hi5 transcriptome (Figure 4A). To test whether Hi5 cells mount an RNAi defense to TNCL virus infection, we mapped small RNA-seq reads that were not mappable to the *T. ni* genome to the two TNCL virus genomic segments. TNCL virus-mapping small RNAs showed a median length of 21 nt (modal length = 20 nt; Figure 4A), typical for siRNAs, suggesting that the Hi5 RNAi pathway actively combats the virus. The TNCL virus-mapping small RNAs bear the two-nucleotide, 3′ overhanging ends that are the hallmark of siRNAs (Figure 4B) (*Elbashir, SM et al., 2001*; *Elbashir, SM et al., 2001*; *Elbashir, SM et al., 2001*). Moreover, the phased pattern of TNCL virus-mapping siRNAs suggests they are made one-after-another starting at the end of a dsRNA molecule: the distance between siRNA 5′ ends shows a periodicity of 20 nt, the length of a typical TNCL virus-mapping siRNA (Figure 4C). In *D. melanogaster*, Dicer-2 processively produces siRNAs, using ATP energy to translocate along a dsRNA molecule (*Cenik, ES et al., 2011*). The phasing of anti-viral

siRNAs in Hi5 cells suggests that *T. ni* Dicer-2 similarly generates multiple siRNAs from each molecule of dsRNA before dissociating.

In addition to siRNAs, the TNCL-mapping small RNAs include some 23–32 nt RNAs. These are unlikely to be anti-viral piRNAs, because they lack the characteristic first-nucleotide uridine bias and show no significant ping-pong signal (*Z*-score = −0.491). We conclude that Hi5 cells do not use piRNAs for viral defense.

Lepidopteran siRNAs are not 2′-*O*-methylated

The discovery that the 3′ ends of *D. melanogaster* siRNAs, but not miRNAs, are 2′-*O*-methylated (*Pelisson, A et al., 2007*) led to the idea that insects in general methylate both siRNAs and piRNAs. Resistance to oxidation by $NaIO_4$ is the hallmark of 3′ terminal, 2′-*O*-methylation, and the enrichment of a small RNA in a high-throughput sequencing library prepared from $NaIO_4$-treated RNA suggests 2′-*O*-methylation. Conversely, depletion of small RNAs, such as miRNAs, from such an oxidized RNA library is strong evidence for unmodified 2′,3′ vicinal hydroxyl groups. Surprisingly, TNCL virus-mapping siRNAs were 130-fold depleted from our oxidized small RNA-seq library (22.0 ppm) compared to the unoxidized library (2,870 ppm), suggesting that they are unmethylated. Sequencing of oxidized and unoxidized small RNA from *T. ni* ovary, testis, and thorax detected 20–22 nt peaks in unoxidized libraries; such peaks were absent from oxidized libraries (Figure 4D), suggesting that *T. ni* genome-mapping, endogenous siRNAs also lack 2′-O-methylation. We conclude that both *T. ni* exo- and endo-siRNAs are not 2′-*O*-methyl modified.

Are siRNAs unmethylated in other Lepidopteran species? We sequenced oxidized and unoxidized small RNAs from two additional Lepidoptera: *P. xylostella* and *B. mori*. Like *T. ni*, siRNAs from these Lepidoptera were abundant in libraries prepared from unoxidized small RNA but depleted from oxidized libraries (Figure 4—figure supplement 1A). The ratio of siRNAs in the oxidized library to siRNAs in the

corresponding unoxidized library (ox/unox) provides a measure of siRNA 2′,3′ modification. For *D. melanogaster* siRNAs, the median ox/unox ratio was 1.00, whereas the three Lepidoptera species had median ox/unox ratios between 0.17 and 0.22 (Figure 4E), indicating their siRNAs were depleted from oxidized libraries and therefore bear unmodified 2′,3′ hydroxyl groups. We conclude that the last common ancestor of *T. ni*, *B. mori*, and *P. xylostella,* which diverged 170 mya, lacked the ability to 2′-*O*-methylate siRNA 3′ ends. We do not currently know whether the last common ancestor of Lepidoptera lost the capacity to methylate siRNAs or if some or all members of Diptera, the sister order of Lepidoptera, acquired this function, which is catalyzed by the piRNA-methylating enzyme Hen1 (*Saito, K et al., 2007*; *Horwich, MD et al., 2007*; *Kirino, Y, Mourelatos, Z, 2007*).

Terminal 2′ methylation of *D. melanogaster* siRNAs is thought to protect them from non-templated nucleotide addition (tailing), 3′-to-5′ trimming, and wholesale degradation (*Ameres, SL et al., 2010*). Since *T. ni* siRNAs lack a 2′-*O*-methyl group at their 3′ ends, we first asked if we could observe frequent trimming by examining shorter TNCL-mapping siRNA (18–19 nt). These siRNAs account for 1.05% of all TNCL-mapping siRNAs. They did not possess the typical siRNA one-after-another pattern (Z_1 = -0.674, p = 0.500), yet more than 97.5% of these were prefixes of longer, phased siRNAs, indicating that these were trimmed siRNAs. We conclude that TNCL siRNA trimming is rare in Hi5 cells. We next asked whether *T. ni* and other lepidopteran siRNAs have higher frequencies of tailing. Despite the lack of 2′-*O*-methylation, most TNCL virus siRNAs were not tailed: just 6.69% of all virus-mapping small RNA reads contained 3′ non-templated nucleotides (Figure 4—figure supplement 1B). Among the 3′ non-templated nucleotides, the most frequent addition was one or more uridines (49.6%) as observed previously for miRNAs and siRNAs in other animals (*Ameres, SL et al., 2010*; *Chou, MT et al., 2015*). Endogenous siRNA tailing frequencies for the lepidopterans *T. ni* (10.2%, ovary), *B. mori* (5.97%, eggs), and *P. xylostella* (8.58%,

ovary) were also similar to *D. melanogaster* (6.71%, ovary). We speculate that lepidopterans have other mechanisms to maintain siRNA stability or that trimming and tailing in lepidopterans are less efficient than in flies.

siRNAs are non-randomly loaded into Argonaute proteins: the guide strand, the strand with the more weakly base paired 5′ end, is favored for loading (*Khvorova, A et al., 2003*; *Schwarz, DS et al., 2003*); the disfavored passenger strand is destroyed. Thus, loading skews the abundance of the two siRNA strands. To test if non-methylated siRNAs are loaded into Argonaute, we computationally paired single-stranded siRNAs that compose an siRNA duplex bearing two-nucleotide overhanging 3′ ends and calculated the relative abundance of the two siRNA strands. For TNCL-mapping siRNAs, 72.3% of siRNA duplexes had guide/passenger strand ratios ≥2 (median = 3.90; mean = 10.2; Figure 4—figure supplement 2). Among genome-mapping 20–22 nt small RNAs 78.5% of duplexes had guide/passenger strand ratios ≥2 (median 5.44; average 56.2). We conclude that the majority of exogenous and endogenous siRNAs are loaded, presumably into Ago2.

piRNAs

In animals, piRNAs, ~23–32 nt long, protect the germline genome by suppressing the transcription or accumulation of transposon and repetitive RNA (*Girard, A et al., 2006*; *Lau, NC et al., 2006*; *Vagin, VV et al., 2006*; *Brennecke, J et al., 2007*; *Aravin, AA et al., 2007*). In *D. melanogaster*, dedicated transposon-rich loci (piRNA clusters) give rise to piRNA precursor transcripts, which are processed into piRNAs loaded into one of three PIWI proteins, Piwi, Aubergine (Aub), or Argonaute3 (Ago3). Piwi acts in the nucleus to direct tri-methylation of histone H3 on lysine 9 on transposon and repetitive genomic sequences (*Sienski, G et al., 2012*; *Le Thomas, A et al., 2014*; *Le Thomas, A et al., 2014*). In fly cytoplasm, piRNAs guide the Piwi paralog Aub to cleave transposon mRNAs. The mRNA cleavage products can then produce more piRNAs, which are

loaded into Ago3. In turn, these sense piRNAs direct Ago3 to cleave transcripts from piRNA clusters, generating additional piRNAs bound to Aub. The resulting “Ping-Pong” feed-forward loop both amplifies piRNAs and represses transposon activity (*Brennecke, J et al., 2007*; *Gunawardane, LS et al., 2007*). Finally, Ago3 cleavage not only produces Aub-bound piRNAs, but also initiates the production of Piwi-bound, phased piRNAs that diversify the piRNA pool (*Mohn, F et al., 2015*; *Han, BW et al., 2015*).

piRNA pathway proteins

The *T. ni* genome contains a full repertoire of genes encoding piRNA pathway proteins (Supplementary file 2B). These genes were expressed in both germline and somatic tissues, but were higher in ovary, testis, and Hi5 cells compared to thorax (median ratios: ovary/thorax = 14.2, testis/thorax = 2.9, and Hi5/thorax = 4.9; Figure 5A). Expression of piRNA pathway genes in the Hi5 cell line suggests that it recapitulates the germline piRNA pathway. Although most *T. ni* piRNA pathway genes correspond directly to their *D. melanogaster* orthologs, *T. ni* encodes only two PIWI proteins, TnPiwi and TnAgo3. The fly proteins Aub and Piwi are paralogs that arose from a single ancestral PIWI protein after the divergence of flies and mosquitos (*Lewis, SH et al., 2016*). We do not yet know whether TnPiwi functions more like *Drosophila* Aub or Piwi. In *D. melanogaster*, piRNA clusters—the genomic sources of most transposon-silencing germline piRNAs—are marked by the proteins Rhino, Cutoff, and Deadlock, which allow transcription of these heterochromatic loci (*Klattenhoff, C et al., 2009*; *Pane, A et al., 2011*; *Mohn, F et al., 2014*; *Zhang, Z et al., 2014*). *T. ni* lacks detectable Rhino, Cutoff, and Deadlock orthologs. In fact, this trio of proteins is poorly conserved, and the mechanism by which they mark fly piRNA source loci may be unique to Drosophilids. In this regard, *T. ni* likely provides a more universal insect model for the mechanisms by which germ cells distinguish piRNA precursor RNAs from other protein-coding and non-coding transcripts.

piRNA cluster architecture

In both the germline and the soma, *T. ni* piRNAs originate from discrete genomic loci. To define these piRNA source loci, we employed an expectation-maximization algorithm that resolves piRNAs mapping to multiple genomic locations. Applying this method to multiple small RNA-seq datasets, we defined piRNA-producing loci comprising 10.7 Mb (348 clusters) in ovary, 3.1 Mb (79 clusters) in testis, 3.0 Mb (71 clusters) in Hi5 cells, and 2.4 Mb (65 clusters) in thorax (Figure 5B). For each tissue or cell-type, these 393 clusters explain >70% of uniquely mapped piRNAs and >70% of all piRNAs when using expectation-maximization mapping. A core set of piRNA-producing loci comprising 1.5 Mb is active in both germline and somatic tissues.

T. ni piRNA clusters vary substantially in size and expression level. In ovary, half the bases in piRNA clusters are in just 67 loci, with a median length of 53 kb. Among these, five span >200 kb, while the smallest is just 38 kb. The most productive piRNA source is a 264 kb locus on chromosome 13 (Figure 5—figure supplement 1); 7.8% of uniquely mapped piRNAs—50,000 distinct piRNA sequences—reside in this locus. Collectively, the top 20 ovary piRNA loci explain half the uniquely mapped piRNAs, yet constitute only 0.7% of the genome. Globally, 61.9% of bases in piRNA clusters are repetitive, and 74.5% transposon-mapping piRNAs are antisense, suggesting that *T. ni* uses antisense piRNAs to suppress transposon transcripts.

In the fly ovary germline, most piRNA clusters generate precursor RNAs from both DNA strands. These dual-strand clusters fuel the 'Ping-Pong' amplification cycle (*Brennecke, J et al., 2007*; *Gunawardane, LS et al., 2007*). Other fly piRNA clusters, such as the paradigmatic *flamenco* gene (*Prud'homme, N et al., 1995*; *Brennecke, J et al., 2007*; *Pelisson, A et al., 2007*; *Malone, CD et al., 2009*; *Goriaux, C et al., 2014*) are transcribed from one strand only and are organized to generate antisense piRNAs directly, without further Ping-Pong amplification (*Malone, CD et al., 2009*). These uni-strand clusters are the only sources of piRNAs in the follicle cells, somatic cells that

support fly oocyte development and express only a single PIWI protein, Piwi (*Malone, CD et al., 2009*).

The *T. ni* genome contains both dual- and uni-strand piRNA clusters. In ovary, 62 of 348 piRNA-producing loci are dual-strand (Watson/Crick >0.5 or Watson/Crick < 2). These loci produce 35.9% of uniquely mapped piRNAs and 22.8% of all piRNAs; 71.6% of transposon-mapping piRNA reads from these loci are antisense. The remaining 286 uni-strand loci account for 54.8% of uniquely mapped piRNAs and 36.7% of all piRNAs. Most piRNAs (74.8% of reads) from uni-strand clusters are antisense to transposons, the orientation required for repressing transposon mRNA accumulation. At least part of the piRNA antisense bias reflects positive selection for antisense insertions in uni-strand clusters: 57.1% of transposon insertions—79.7% of transposon-mapping nucleotides—are opposite the direction of piRNA precursor transcription, significantly different from dual-strand clusters, in which transposons are inserted randomly: 49.5% of transposon insertions in dual-strand clusters are in the antisense direction (Figure 5—figure supplement 2A). For one 77 kb uni-strand cluster on chromosome 20, 99.0% of piRNA reads (96% of piRNA sequences) that can be uniquely assigned are from the Crick strand, while 67.6% of transposon insertions and 79.7% of transposon-mapping nucleotides at this locus lie on the Watson strand.

Nearly the entire W chromosome produces piRNAs

The largest ovary cluster is a 462 kb W-linked region, consistent with our finding that the W chromosome is a major source of piRNAs (Figure 5B and 5C and Figure 5—figure supplement 2B). Our data likely underestimates the length of this large piRNA cluster, as it is difficult to resolve reads mapping to its flanking regions: 70.8% of bases in the flanking regions do not permit piRNAs to map uniquely to the genome. In fact, 85.1% of the sequences between clusters on the W chromosome are not uniquely mappable.

These gaps appear to reflect low mappability and not boundaries between discrete clusters. We propose that the W chromosome itself is a giant piRNA cluster.

To further test this idea, we identified piRNA reads that uniquely map to one location among all contigs and measured their abundance per kilobase of the genome. W-linked contigs had a median piRNA abundance of 14.4 RPKM in ovaries, 379-fold higher than the median of all autosomal and Z-linked contigs, consistent with the view that almost the entire W chromosome produces piRNAs. In *B. mori* females, a plurality of piRNAs come from the W chromosome: ovary-enriched piRNAs often map to W-linked sequences, but not autosomes (*Kawaoka, S et al., 2011*). Similarly, for *T. ni*, 27.2% of uniquely mapping ovary piRNAs derive from W-linked sequences, even though these contigs compose only 2.8% of the genome (Figure 5C). The W chromosome may produce more piRNAs than our estimate, as the unassembled repetitive portions of the W chromosome likely also produce piRNAs. Thus, the entire W chromosome is a major source of piRNAs in *T. ni* ovaries (Figure 5B). To our knowledge, the *T. ni* W chromosome is the first example of an entire chromosome devoted to piRNA production.

To determine if there are W-linked regions devoid of piRNAs, we mapped all piRNAs to the W-linked contigs and found that 11.0% of the W-linked bases were not covered by any piRNAs, indicating at least part of the W chromosome does not produce any piRNAs. Next, we manually inspected 74 putative W-linked protein-coding genes and nine putative W-linked miRNAs. All nine W-linked miRNAs (Figure 5B, Supplementary file 1J) are *T. ni*-specific, and small RNAs mapping to these predicted miRNA loci showed significant ping-pong signature (Z-score = 14.2, $p = 1.81\times10^{-45}$), suggesting that these are likely piRNAs, not authentic miRNAs. For the putative protein-coding genes, we categorized them into orphan genes (no homologs found), transposons (good homology to transposons), uncharacterized/hypothetical proteins, and potential protein-coding genes with homology to the NCBI non-redundant protein

sequences. We then asked whether piRNAs were produced from these genes (Figure 5—figure supplement 2C). Among W-linked genes, those with transposon homology on average produced the most piRNAs (44.9 median ppm) whereas those with homology to annotated genes produced the fewest (9.81 median ppm). Some putative genes (such as TNI001015 and TNI005339) produced no piRNAs at all. We conclude that although some W-linked loci do not produce piRNAs, nearly the entire W chromosome produces piRNAs.

In contrast to the W chromosome, *T. ni* autosomes and the Z chromosome produce piRNAs from discrete loci—63 autosomal and 11 Z-linked contigs had piRNA levels >10 rpkm. Few piRNAs are produced outside of these loci: for example, the median piRNA level across all autosomal and Z-linked contigs was ~0 in ovaries (Figure 5—figure supplement 2B).

Expression of piRNA clusters

In the *T. ni* germline, piRNA production from individual clusters varies widely, but the same five piRNA clusters produce the most piRNAs in ovary (34.9% of piRNAs), testis (49.3%), and Hi5 cells (44.0%), suggesting that they serve as master loci for germline transposon silencing. Other piRNA clusters show tissue-specific expression, with the W chromosome producing more piRNAs in ovary than in Hi5 cells, and three Z-linked clusters producing many more piRNAs in testis than in ovary (15.0–24.7 times more), even after accounting for the absence of dosage compensation in germline tissues (Figure 6—figure supplement 1A).

Hi5 cells are female, yet many piRNA-producing regions of the W chromosome that are active in the ovary produce few piRNAs in Hi5 cells (Figure 6—figure supplement 1A). We do not know whether this reflects a reorganization of cluster expression upon Hi5 cell immortalization or if Hi5 cells correspond to a specific germ cell type that is underrepresented in whole ovaries. At least 40 loci produce piRNAs in

Hi5 cells but not in ovaries. Comparison of DNA-seq data from *T. ni* and Hi5 identified 74 transposon insertions in 12 of the Hi5-specific piRNA clusters. Older transposons have more time to undergo sequence drift from the consensus sequence of the corresponding transposon family. The 74 Hi5-specific transposon insertions, which include both DNA and LTR transposons, had significantly lower divergence rates than those common to ovary and Hi5 cells (Figure 6A), consistent with the idea that recent transposition events generated the novel piRNA clusters in Hi5 cells. We conclude that the Hi5-specific piRNA-producing loci are quite young, suggesting that *T. ni* and perhaps other lepidopterans can readily generate novel piRNA clusters.

piRNA clusters active in thorax occupy ~0.57% of the genome and explain 86.8% of uniquely mapped somatic piRNAs in females and 89.5% in males. More than 90% of bases in clusters expressed in thorax are shared with clusters expressed in ovary (Figure 6—figure supplement 1B). Such broadly expressed clusters explain 83.7% of uniquely mapping piRNAs in female thorax and 86.1% in male thorax. Thus, the majority of piRNAs in the *T. ni* soma come from clusters that are also active in the germline. In general, autosomal piRNA cluster expression is similar between female and male thorax, but 12 clusters are differentially expressed between male and female thorax. Of these, nine are W-linked clusters that produce significantly more piRNAs in female than in male thorax (Figure 6B).

piRNA precursor transcripts are rarely spliced

In *D. melanogaster*, Rhino suppresses splicing of piRNA precursors transcribed from dual-strand piRNA clusters (*Mohn, F et al., 2014*; *Zhang, Z et al., 2014*). Fly uni-strand piRNA clusters do not bind Rhino and behave like canonical RNA polymerase II transcribed genes (*Brennecke, J et al., 2007*; *Goriaux, C et al., 2014*). Although *T. ni* has no *rhino* ortholog, its piRNA precursor RNAs are rarely spliced as observed for clusters in flies. We identified splicing events in our RNA-seq data, requiring ≥10 reads

596 that map across exon-exon junctions and a minimum splicing entropy of 2 to exclude
597 PCR duplicates (*Graveley, BR et al., 2011*). This approach detected just 27 splice sites
598 among all piRNA precursor transcripts from ovary, testis, thorax, and Hi5 piRNA
599 clusters (Figure 6C). Of these 27 splice sites, 19 fall in uni-strand piRNA clusters. We
600 conclude that, as in flies, transcripts from *T. ni* dual-strand piRNAs clusters are rarely if
601 ever spliced. Unlike flies (*Goriaux, C et al., 2014*), RNA from *T. ni* uni-strand piRNA
602 clusters also undergoes splicing infrequently.
603 The absence of piRNA precursor splicing in dual-strand piRNA clusters could
604 reflect an active suppression of the splicing machinery or a lack of splice sites. To
605 distinguish between these two mechanisms, we predicted gene models for piRNA-
606 producing loci, employing the same parameters used for protein-coding genes. For
607 piRNA clusters, this approach generated 1,332 gene models encoding polypeptides
608 >200 amino acids. These models comprise 2,544 introns with consensus splicing
609 signals (Figure 6—figure supplement 1C). Notably, ~90% of these predicted gene
610 models had high sequence similarity to transposon consensus sequences (BLAST e-
611 value<10^{-10}), indicating that many transposons in piRNA clusters have intact splice
612 sites. We conclude that piRNA precursors contain splice sites, but their use is actively
613 suppressed.
614 To measure splicing efficiency, we calculated the ratio of spliced to unspliced
615 reads for each predicted splice site in the piRNA clusters. High-confidence splice sites
616 in protein-coding genes outside piRNA clusters served as a control. Compared to the
617 control set of genes, splicing efficiency in piRNA loci was 9.67-fold lower in ovary, 2.41-
618 fold lower in testis, 3.23-fold lower in thorax, and 17.0-fold lower in Hi5 cells (Figure 6D),
619 showing that *T. ni* piRNA precursor transcripts are rarely and inefficiently spliced. To
620 test whether uni- and dual-strand piRNA cluster transcripts are differentially spliced in *T.*
621 *ni*, we evaluated the experimentally supported splice sites from Hi5, ovary, testis, and
622 thorax collectively. Dual-strand cluster transcripts had 1.71-fold lower splicing efficiency

compared to uni-strand clusters (Figure 6D). Thus, *T. ni* suppresses splicing of dual- and uni-strand piRNA cluster transcripts by a mechanism distinct from the Rhino-dependent pathway in *D. melanogaster*. That this novel splicing suppression pathway is active in Hi5 cells should facilitate its molecular dissection.

Genome-editing and single-cell cloning of Hi5 cells

The study of arthropod piRNAs has been limited both by a lack of suitable cultured cell models and by the dominance of *D. melanogaster* as a piRNA model for arthropods generally. Although Vasa-positive *D. melanogaster* ovarian cells have been isolated and cultured (*Niki, Y et al., 2006*), no dipteran germ cell line is currently available. *D. melanogaster* somatic OSS, OSC and Kc167 cells produce piRNAs, but lack key features of the canonical germline pathway (*Lau, NC et al., 2009*; *Saito, K et al., 2009*; *Vrettos, N et al., 2017*). In addition to Hi5 cells, lepidopteran cell lines from *Spodoptera frugiperda* (Sf9) and *B. mori* (BmN4) produce germline piRNAs (*Kawaoka, S et al., 2009*). The *S. frugiperda* genome remains a draft with 37,243 scaffolds and an N50 of 53.7 kb (*Kakumani, PK et al., 2014*). Currently, the BmN4 cell line is the only ex vivo model for invertebrate germline piRNA biogenesis and function. The *B. mori* genome sequence currently comprises 43,463 scaffolds with an N50 of 4.01 Mb (*The, ISG, 2008*). Unfortunately, BmN4 cells readily differentiate into two morphologically distinct cell types (*Iwanaga, M et al., 2014*). Although genome editing with Cas9 has been demonstrated in BmN4 cells (*Zhu, L et al., 2015*), no protocols for cloning individual, genome-modified BmN4 cells have been reported (*Mon, H et al., 2004*; *Kawaoka, S et al., 2009*; *Honda, S et al., 2013*). In contrast, Hi5 cells are cultured using commercially available media, readily transfected, and, we report here, efficiently engineered with Cas9 and grown from single cells into clonal lines.

The bacterial DNA nuclease Cas9, targeted by a single guide RNA (sgRNA), enables rapid and efficient genome editing in worms, flies, and mice, as well as in a

649 variety of cultured animal cell lines (*Jinek, M et al., 2012*; *Barrangou, R, Horvath, P,*
650 *2017*; *Komor, AC et al., 2017*). The site-specific double-strand DNA breaks catalyzed by
651 Cas9 can be repaired by error-prone non-homologous end joining (NHEJ), disrupting a
652 protein-coding sequence or, when two sgRNAs are used, deleting a region of genomic
653 DNA. Alternatively, homology-directed repair (HDR) using an exogenous DNA template
654 allows the introduction of novel sequences, including fluorescent proteins or epitope
655 tags, as well as point mutations in individual genes (*Cong, L et al., 2013*).
656 As a proof-of-concept, we used Cas9 and two sgRNAs to generate a deletion in
657 the piRNA pathway gene *TnPiwi*. The two sgRNAs, whose target sites lie 881 bp apart
658 (Figure 7A), were transcribed *in vitro*, loaded into purified, recombinant Cas9 protein,
659 and the resulting sgRNA/Cas9 ribonucleoprotein complexes (RNPs) transfected into Hi5
660 cells. PCR of genomic DNA isolated 48 h later was used to detect alterations in the
661 *TnPiwi* gene. A novel PCR product, ~900 bp smaller than the product amplified using
662 DNA from control cells, indicated that the desired deletion had been created (Figure
663 7B). Sanger sequencing of the PCR products confirmed deletion of 881–896 bp from
664 the *TnPiwi* gene. The presence of indels—short deletions and non-templated nucleotide
665 additions—at the deletion junction is consistent with a Cas9-mediated dsDNA break
666 having been repaired by NHEJ (Figure 7A). We note that these cells still contain at least
667 one wild-type copy of *TnPiwi*. We have not yet obtained cells in which all four copies of
668 *TnPiwi* are disrupted, perhaps because in the absence of Piwi, Hi5 cells are inviable.
669 To test whether an exogenous donor DNA could facilitate the site-specific
670 incorporation of protein tag sequences into Hi5 genome, we designed two sgRNAs with
671 target sites ~90 bp apart, flanking the *vasa* start codon (Figure 7C). As a donor, we
672 used a single-stranded DNA (ssDNA) encoding EGFP and an HA epitope tag flanked by
673 genomic sequences 787 bp upstream and 768 bp downstream of the *vasa* start codon
674 (Figure 7C). Cas9 and the two sgRNAs were cotransfected with the ssDNA donor, and,
675 one week later, EGFP-positive cells were detected by fluorescence microscopy. PCR

amplification of the targeted region using genomic DNA from EGFP-expressing cells confirmed integration of EGFP and the HA tag into the *vasa* gene (Figure 7D). Sanger sequencing further confirmed integration of EGFP and the HA tag in-frame with the *vasa* open reading frame (Supplemental file 9).

To establish a clonal line from the EGFP-HA-tagged Vasa-expressing cells, individual EGFP-positive cells were isolated by FACS and cultured on selectively permeable filters above a feeder layer of wild-type Hi5 cells (Figure 8A). Growth of the genome-modified single cells required live Hi5 feeder cells—conditioned media did not suffice—presumably because the feeder cells provide short-lived growth factors or other trophic molecules. Single EGFP-positive clones developed one month after seeding and could be further grown without feeder cells as a clonally derived cell line (Figure 8B).

Hi5 cell Vasa is present in a nuage-like, perinuclear structure

In the germline of *D. melanogaster* and other species, components of the piRNA biogenesis pathway, including Vasa, Aub, Ago3, and multiple Tudor-domain proteins, localize to a perinuclear structure called nuage (*Eddy, EM, 1976*; *Findley, SD et al., 2003*; *Lim, AK, Kai, T, 2007*; *Li, C et al., 2009*; *Liu, L et al., 2011*; *Webster, A et al., 2015*). Vasa, a germline-specific nuage component, is widely used as a marker for nuage. In BmN4 cells, transiently transfected Vasa localizes to a perinuclear structure resembling nuage (*Xiol, J et al., 2012*; *Patil, AA et al., 2017*). To determine whether nuage-like structures are present in Hi5 cells, we examined Vasa localization in the Hi5 cells in which the endogenous *vasa* gene was engineered to fuse EGFP and an HA epitope tag to the Vasa amino-terminus. We used two different immunostaining strategies to detect the EGFP-HA-Vasa fusion protein: a mouse monoclonal anti-GFP antibody and a rabbit monoclonal anti-HA antibody. GFP and HA colocalized in a perinuclear structure, consistent with Vasa localizing to nuage in Hi5 cells (Figure 8C).

Discussion

Using Hi5 cells, we have sequenced and assembled the genome of the cabbage looper, *T. ni*, a common and destructive agricultural pest that feeds on many plants of economic importance. Examination of the *T. ni* genome and transcriptome reveals the expansion of detoxification-related gene families (Table 1 and Supplementary file 6), many members of which are implicated in insecticide resistance and are potential targets of pest control. The *T. ni* genome should enable study of the genetic diversity and population structure of this generalist pest, which adapts to different environmental niches worldwide. Moreover, as the sister order of Diptera, Lepidoptera like *T. ni* provide a counterpoint for the well-studied insect model *D. melanogaster*.

The use of Hi-C sequencing was an essential step in assembling the final 368.2 Mb *T. ni* genome into high-quality, chromosome-length scaffolds. The integration of long reads, short reads, and Hi-C provides a rapid and efficient paradigm for generating chromosome-level assemblies of other animal genomes. This strategy assembled the gene-poor, repeat-rich *T. ni* W chromosome, which is, to our knowledge, the first chromosome-level sequence of a lepidopteran W chromosome. Our analysis of autosomal, Z-linked, and W-linked transcripts provides insights into lepidopteran dosage compensation and sex determination. Our data show that *T. ni* compensates for Z chromosome dosage in the soma by reducing transcription of both Z homologs in males, but Z dosage is uncompensated in the germline.

In addition to long RNAs, we characterized miRNAs, siRNAs, and piRNAs in *T. ni* gonads, soma, and cultured Hi5 cells. miRNAs are widely expressed in *T. ni* tissues, providing examples of germline-enriched and somatic miRNAs, as well as highly conserved, lepidopteran-specific, and novel *T. ni* miRNAs. Like flies, *T. ni* possess siRNAs that map to transposons, *cis*-NATs and hpRNAs. Unexpectedly, *T. ni* siRNAs—and likely all lepidopteran siRNAs—lack a 2′-*O*-methyl modification at their 3′ ends, unlike siRNAs in *D. melanogaster*. Consistent with siRNA production by a processive

728 Dicer-2 enzyme, Hi5 cells produce phased siRNAs from the RNA genome of a latent
729 alphanodavirus. The commonalities and differences between *T. ni* and *D. melanogaster*
730 small RNA pathways will help identify both deeply conserved and rapidly evolving
731 components.

732 A major motivation for sequencing the *T. ni* genome was the establishment of a
733 tractable cell culture model for studying small RNAs, especially piRNAs. We believe that
734 our genome assembly and gene-editing protocols will enable the use of *T. ni* Hi5 cells to
735 advance our understanding of how piRNA precursors are defined, made into piRNAs
736 and act to silence transposons in the germline. Hi5 cells express essentially all known
737 piRNA pathway genes except those specific to Drosophilids. Furthermore, *T. ni* Vasa
738 localizes to a perinuclear, nuage-like structure in Hi5 cells, making them suitable for
739 studying the assembly of the subcellular structures thought to organize piRNA
740 biogenesis. We have defined genomic piRNA-producing loci in Hi5 cells, as well as in
741 the soma, testis, and ovary. The most productive piRNA clusters are shared among
742 ovary, testis, and Hi5 cells. In addition, Hi5 cells contain novel piRNA clusters not found
743 in the moth itself, suggesting that the process of establishing new piRNA-producing loci
744 can be recapitulated by experimental manipulation of Hi5 cells.

745 As in *D. melanogaster*, splicing of *T. ni* piRNA precursor transcripts is efficiently
746 suppressed, yet *T. ni* lacks paralogs of the proteins implicated in splicing suppression in
747 flies. The ability to study the mechanisms by which piRNA clusters form and how
748 precursor RNAs are transcribed, exported, and marked for piRNA production in *T. ni*
749 promises to reveal both conserved and lepidopteran-specific features of this pathway.
750 Notably, the W chromosome not only is a major piRNA source, but also produces
751 piRNAs from almost its entirety. Future studies are needed to determine whether this is
752 a common feature of W chromosomes in Lepidoptera and other insects.

753 The establishment of procedures for genome editing and single-cell cloning of
754 Hi5 cells, combined with the *T. ni* genome sequence, make this germ cell line a

powerful tool to study RNA and protein function ex vivo. Our strategy combines transfection of pre-assembled Cas9/sgRNA complexes with single clone isolation using a selectable marker (e.g., EGFP) and feeder cells physically separated from the engineered cells. Compared with nucleic acid-based delivery of Cas9, transfection of Cas9 RNP minimizes the off-target mutations caused by prolonged Cas9 expression and eliminates the risk of integration of sgRNA or Cas9 sequences into the genome (*Lin, S et al., 2014*; *Kim, S et al., 2014*). Compared to plasmid donors (*Yu, Z et al., 2014*; *Ge, DT et al., 2016*), ssDNA homology donors similarly reduce the chance of introducing exogenous sequences at unintended genomic sites. Techniques for injecting the embryos of other lepidopteran species have already been established (*Wang, Y et al., 2013*; *Takasu, Y et al., 2014*; *Zhang, Z et al., 2015*). In principle, Cas9 RNP injected into cabbage looper embryos could be used to generate genetically modified *T. ni* strains both to explore lepidopteran biology and to implement novel strategies for safe and effective pest control.

Acknowledgements

We thank members of the Weng and Zamore laboratories for helpful discussions and comments on the manuscript; UMass Deep Sequencing Core for Pacific Biosciences sequencing; Zdenka Matijasevic for sharing the karyotyping protocol. This work was supported in part by National Institutes of Health grants R37GM062862 to PDZ and HD078253 to ZW.

References

Abe, H et al. 2008. Identification of the female-determining region of the W chromosome in Bombyx mori. *Genetica* ***133***: 269–282. 10.1007/s10709-007-9210-1

Abe, H, Mita, K, Yasukochi, Y, Oshiki, T, Shimada, T. 2005. Retrotransposable elements on the W chromosome of the silkworm, Bombyx mori. *Cytogenet*

780 *Genome Res* ***110***: 144–151. 10.1159/000084946
781 Agrawal, N, Dasaradhi, PVN, Mohmmed, A, Malhotra, P, Bhatnagar, RK, Mukherjee,
782 SK. 2003. RNA Interference: Biology, Mechanism, and Applications. *Microbiology*
783 *and Molecular Biology Reviews* ***67***: 657–685. 10.1128/MMBR.67.4.657-685.2003
784 Ai, J et al. 2011. Genome-wide analysis of cytochrome P450 monooxygenase genes in
785 the silkworm, Bombyx mori. *Gene* ***480***: 42–50. 10.1016/j.gene.2011.03.002
786 Altschul, SF, Gish, W, Miller, W, Myers, EW, Lipman, DJ. 1990. Basic local alignment
787 search tool. *Journal of Molecular Biology* ***215***: 403–410. 10.1016/S0022-
788 2836(05)80360-2
789 Ameres, SL et al. 2010. Target RNA-directed trimming and tailing of small silencing
790 RNAs. *Science* ***328***: 1534–1539. 10.1126/science.1187058
791 Aravin, AA, Sachidanandam, R, Girard, A, Fejes-Toth, K, Hannon, GJ. 2007.
792 Developmentally regulated piRNA clusters implicate MILI in transposon control.
793 *Science* ***316***: 744–747.
794 Attrill, H et al. 2016. FlyBase: establishing a Gene Group resource for Drosophila
795 melanogaster. *Nucleic Acids Res* ***44***: D786–92. 10.1093/nar/gkv1046
796 Barrangou, R, Horvath, P. 2017. A decade of discovery: CRISPR functions and
797 applications. *Nature microbiology* ***2***: 17092. 10.1038/nmicrobiol.2017.92
798 Belton, J-M, McCord, RP, Gibcus, JH, Naumova, N, Zhan, Y, Dekker, J. 2012. Hi–C: A
799 comprehensive technique to capture the conformation of genomes. *Methods* ***58***:
800 268–276. 10.1016/j.ymeth.2012.05.001
801 Benton, R, Vannice, KS, Gomez-Diaz, C, Vosshall, LB. 2009. Variant Ionotropic
802 Glutamate Receptors as Chemosensory Receptors in Drosophila. *Cell* ***136***: 149–
803 162. 10.1016/j.cell.2008.12.001
804 Bernstein, E, Caudy, AA, Hammond, SM, Hannon, GJ. 2001. Role for a bidentate
805 ribonuclease in the initiation step of RNA interference. *Nature* ***409***: 363–366.
806 10.1038/35053110

807 Bernt, M et al. 2013. MITOS: improved de novo metazoan mitochondrial genome
808 annotation. *Mol Phylogenet Evol* ***69***: 313–319. 10.1016/j.ympev.2012.08.023
809 Bonin, CP, Mann, RS. 2004. A piggyBac Transposon Gene Trap for the Analysis of
810 Gene Expression and Function in Drosophila. *Genetics* ***167***: 1801–1811.
811 10.1534/genetics.104.027557
812 Brennecke, J et al. 2007. Discrete Small RNA-Generating Loci as Master Regulators of
813 Transposon Activity in Drosophila. *Cell* ***128***: 1089–1103.
814 10.1016/j.cell.2007.01.043
815 Burton, JN, Adey, A, Patwardhan, RP, Qiu, R, Kitzman, JO, Shendure, J. 2013.
816 Chromosome-scale scaffolding of de novo genome assemblies based on
817 chromatin interactions. *Nature Biotechnology*
818 *Nat Biotech* ***31***: 1119–1125. 10.1038/nbt.2727
819 Campbell, MS, Holt, C, Moore, B, Yandell, M. 2014. Genome Annotation and Curation
820 Using MAKER and MAKER-P. *Current protocols in bioinformatics* ***48***: 4.11.1–
821 4.11.39. 10.1002/0471250953.bi0411s48
822 Capella-Gutiérrez, S, Silla-Martínez, JM, Gabaldón, T. 2009. trimAl: a tool for
823 automated alignment trimming in large-scale phylogenetic analyses.
824 *Bioinformatics* ***25***: 1972–1973. 10.1093/bioinformatics/btp348
825 Capinera, J. (2001). *Handbook of Vegetable Pests* (Gulf Professional Publishing).
826 Castresana, J. 2000. Selection of Conserved Blocks from Multiple Alignments for Their
827 Use in *Phylogenetic Analysis*. *Molecular Biology and Evolution*
828 *Mol Biol Evol* ***17***: 540–552. 10.1093/oxfordjournals.molbev.a026334
829 Caygill, EE, Johnston, LA. 2008. Temporal Regulation of Metamorphic Processes in
830 Drosophila by the let-7 and miR-125 Heterochronic MicroRNAs. *Current Biology*
831 ***18***: 943–950. 10.1016/j.cub.2008.06.020
832 Cenik, ES et al. 2011. Phosphate and R2D2 restrict the substrate specificity of Dicer-2,
833 an ATP-driven ribonuclease. *Molecular cell* ***42***: 172–184.

834 10.1016/j.molcel.2011.03.002
835 Challis, RJ, Kumar, S, Dasmahapatra, KKK, Jiggins, CD, Blaxter, M. 2016. Lepbase:
836 the Lepidopteran genome database. *bioRxiv* 056994. 10.1101/056994
837 Chapman, AD. (2009). Numbers of Living Species in Australia and the World: Report for
838 the Australian Biological Resources Study (Canberra: Department of the
839 Environment, Water, Heritage and the Arts).
840 Chou, MT, Han, BW, Hsiao, CP, Zamore, PD, Weng, Z, Hung, JH. 2015. Tailor: a
841 computational framework for detecting non-templated tailing of small silencing
842 RNAs. *Nucleic Acids Res* 10.1093/nar/gkv537
843 Chung, WJ, Okamura, K, Martin, R, Lai, EC. 2008. Endogenous RNA Interference
844 Provides a Somatic Defense against *Drosophila* Transposons. *Curr Biol* ***18***: 795–
845 802. 10.1016/j.cub.2008.05.006
846 Conesa, A, Götz, S, García-Gómez, JM, Terol, J, Talón, M, Robles, M. 2005. Blast2GO:
847 a universal tool for annotation, visualization and analysis in functional genomics
848 research. *Bioinformatics* ***21***: 3674–3676. 10.1093/bioinformatics/bti610
849 Cong, L et al. 2013. Multiplex Genome Engineering Using CRISPR/Cas Systems.
850 *Science* ***339***: 819–823. 10.1126/science.1231143
851 Creighton, CJ et al. 2010. Molecular profiling uncovers a p53-associated role for
852 microRNA-31 in inhibiting the proliferation of serous ovarian carcinomas and
853 other cancers. *Cancer research* ***70***: 1906–1915. 10.1158/0008-5472.CAN-09-
854 3875
855 Croset, V et al. 2010. Ancient Protostome Origin of Chemosensory Ionotropic
856 Glutamate Receptors and the Evolution of Insect Taste and Olfaction. *PLOS*
857 *Genetics* ***6***: e1001064. 10.1371/journal.pgen.1001064
858 Czech, B et al. 2008. An endogenous small interfering RNA pathway in Drosophila.
859 *Nature* ***453***: 798–802. 10.1038/nature07007
860 DePristo, MA, Banks, E, Poplin, R, Garimella…, KV. 2011. A framework for variation

discovery and genotyping using next-generation DNA sequencing data. *Nature …*

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. 10.1016/j.ibmb.2013.11.001

Eddy, EM. 1976. Germ plasm and the differentiation of the germ cell line. *International review of cytology*

Edgar, RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucl Acids Res* ***32***: 1792–1797. 10.1093/nar/gkh340

Elbashir, SM, Harborth, J, Lendeckel, W, Yalcin, A, Weber, K, Tuschl, T. 2001. Duplexes of 21-nucleotide RNAs mediate RNA interference in cultured mammalian cells. *Nature* ***411***: 494–498.

Findley, SD, Tamanaha, M, Clegg, NJ, Ruohola-Baker, H. 2003. Maelstrom, a *Drosophila* spindle-class gene, encodes a protein that colocalizes with Vasa and RDE1/AGO1 homolog, Aubergine, in nuage. *Development* ***130***: 859–871. 10.1242/dev.00310

Finn, RD et al. 2016. The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res* ***44***: D279–D285. 10.1093/nar/gkv1344

Flynt, A, Liu, N, Martin, R, Lai, EC. 2009. Dicing of viral replication intermediates during silencing of latent *Drosophila* viruses. *Proc Natl Acad Sci U S A* ***106***: 5270–5275. 10.1073/pnas.0813412106

Fraser, MJ, Smith, GE, Summers, MD. 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. *J Virol* ***47***: 287–300.

Friedländer, MR et al. 2008. Discovering microRNAs from deep sequencing data using miRDeep. *Nat Biotech* ***26***: 407–415. 10.1038/nbt1394

888 Friedlander, MR, Mackowiak, SD, Li, N, Chen, W, Rajewsky, N. 2012. miRDeep2
889 accurately identifies known and hundreds of novel microRNA genes in seven
890 animal clades. *Nucleic Acids Res* ***40***: 37–52. 10.1093/nar/gkr688
891 Fujiwara, H, Osanai, M, Matsumoto, T, Kojima, KK. 2005. Telomere-specific non-LTR
892 retrotransposons and telomere maintenance in the silkworm, Bombyx mori.
893 *Chromosome Res* ***13***: 455–467. 10.1007/s10577-005-0990-9
894 Fuková, I, Nguyen, P, Marec, F. 2005. Codling moth cytogenetics: karyotype,
895 chromosomal location of rDNA, and molecular differentiation of sex
896 chromosomes. *Genome* ***48***: 1083–1092. 10.1139/g05-063
897 Futahashi, R et al. 2015. Extraordinary diversity of visual opsin genes in dragonflies.
898 *PNAS* ***112***: E1247–E1256. 10.1073/pnas.1424670112
899 Gao, G, Vandenberghe, LH, Wilson, JM. 2005. New recombinant serotypes of AAV
900 vectors. *Curr Gene Ther* ***5***: 285–297. 10.2174/1566523054065057
901 Gaunt, MW, Miles, MA. 2002. An insect molecular clock dates the origin of the insects
902 and accords with palaeontological and biogeographic landmarks. *Molecular*
903 *biology and evolution* ***19***: 748–761. 10.1093/oxfordjournals.molbev.a004133
904 Ge, DT, Tipping, C, Brodsky, MH, Zamore, PD. 2016. Rapid Screening for CRISPR-
905 Directed Editing of the Drosophila Genome Using white Coconversion. *G3*
906 *(Bethesda)* ***6***: 3197–3206. 10.1534/g3.116.032557
907 Ghildiyal, M et al. 2008. Endogenous siRNAs derived from transposons and mRNAs in
908 Drosophila somatic cells. *Science* ***320***: 1077–1081.
909 Girard, A, Sachidanandam, R, Hannon, GJ, Carmell, MA. 2006. A germline-specific
910 class of small RNAs binds mammalian Piwi proteins. *Nature* ***442***: 199–202.
911 10.1038/nature04917
912 Glastad, KM, Hunt, BG, Yi, SV, Goodisman, MAD. 2011. DNA methylation in insects: on
913 the brink of the epigenomic era. *Insect Molecular Biology* ***20***: 553–565.
914 10.1111/j.1365-2583.2011.01092.x

915 Gong, D-P, Zhang, H-J, Zhao, P, Xia, Q-Y, Xiang, Z-H. 2009. The Odorant Binding
916 Protein Gene Family from the Genome of Silkworm, Bombyx mori. *BMC*
917 *Genomics* ***10***: 332. 10.1186/1471-2164-10-332
918 Goodman, WG, Granger, NA (2005). The Juvenile Hormones. Gilbert, LI, ed.
919 (Amsterdam: Elsevier), pp. 319-408.
920 Goriaux, C, Desset, S, Renaud, Y, Vaury, C, Brasset, E. 2014. Transcriptional
921 properties and splicing of the flamenco piRNA cluster. *EMBO Rep*
922 10.1002/embr.201337898
923 Granados, RR, Derksen, ACG, Dwyer, KG. 1986. Replication of the Trichoplusia ni
924 granulosis and nuclear polyhedrosis viruses in cell cultures. *Virology* ***152***: 472–
925 476. 10.1016/0042-6822(86)90150-9
926 Granados, RR, Guoxun, L, Derksen, ACG, McKenna, KA. 1994. A new insect cell line
927 from Trichoplusia ni (BTI-Tn-5B1-4) susceptible to Trichoplusia ni single
928 enveloped nuclear polyhedrosis virus. *Journal of Invertebrate Pathology* ***64***: 260–
929 266. 10.1016/S0022-2011(94)90400-6
930 Graveley, BR et al. 2011. The developmental transcriptome of Drosophila
931 melanogaster. *Nature* ***471***: 473–479. 10.1038/nature09715
932 Gu, L, Walters, JR, Knipple, DC. 2017. Conserved Patterns of Sex Chromosome
933 Dosage Compensation in the Lepidoptera (WZ/ZZ): Insights from a Moth Neo-Z
934 Chromosome. *Genome Biol Evol* ***9***: 802–816. 10.1093/gbe/evx039
935 Gunawardane, LS et al. 2007. A Slicer-Mediated Mechanism for Repeat-Associated
936 siRNA 5′ End Formation *in Drosophi*la. *Science* ***315***: 1587–1590.
937 10.1126/science.1140494
938 Hahn, C, Bachmann, L, Chevreux, B. 2013. Reconstructing mitochondrial genomes
939 directly from genomic next-generation sequencing reads—a baiting and iterative
940 mapping approach. *Nucl Acids Res* gkt371. 10.1093/nar/gkt371
941 Han, BW, Wang, W, Zamore, PD, 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. 10.1093/bioinformatics/btu647

Han, BW, Wang, W, Li, C, Weng, Z, Zamore, PD. 2015. Noncoding RNA. piRNA-guided transposon cleavage initiates Zucchini-dependent, phased piRNA production. *Science* ***348***: 817–821. 10.1126/science.aaa1264

He, L, Hannon, GJ. 2004. MicroRNAs: small RNAs with a big role in gene regulation. *Nat Rev Genet* ***5***: 522–531.

Hekmat-Scafe, DS, Scafe, CR, McKinney, AJ, Tanouye, MA. 2002. Genome-Wide Analysis of the Odorant-Binding Protein Gene Family in Drosophila melanogaster. *Genome Res* ***12***: 1357–1369. 10.1101/gr.239402

Hink, WF. 1972. A Catalog of Invertebrate Cell Lines. In Invertebrate Tissue Culture, Vol. 2, C., V, ed. (New York, NY: Academic Press), pp. 363-387.

Hink, WF. 1970. Established insect cell line from the cabbage looper, Trichoplusia ni. *Nature* ***226***: 466–467. 10.1038/226466b0

Hirose, Y, Manley, JL. 1997. Creatine phosphate, not ATP, is required for 3' end cleavage of mammalian pre-mRNA in vitro. *J Biol Chem* ***272***: 29636–29642. 10.1074/jbc.272.47.29636

Honda, S et al. 2013. Mitochondrial protein BmPAPI modulates the length of mature piRNAs. *RNA* ***19***: 1405–1418. 10.1261/rna.040428.113

Horwich, MD et al. 2007. The *Drosophila* RNA methyltransferase, DmHen1, modifies germline piRNAs and single-stranded siRNAs in RISC. *Curr Biol* ***17***: 1265–1272. 10.1016/j.cub.2007.06.030

Houwing, S et al. 2007. A role for Piwi and piRNAs in germ cell maintenance and transposon silencing in Zebrafish. *Cell* ***129***: 69–82.

Hsu, PD et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* ***31***: 827–832. 10.1038/nbt.2647

969 Initiative, IGG. 2014. Genome Sequence of the Tsetse Fly (Glossina morsitans): Vector
970 of African Trypanosomiasis. *Science* ***344***: 380–386. 10.1126/science.1249656
971 Iwanaga, M, Adachi, Y, Uchiyama, K, Tsukui, K, Katsuma, S, Kawasaki, H. 2014. Long-
972 term adaptation of the Bombyx mori BmN4 cell line to grow in serum-free culture.
973 *In Vitro CellDevBiol—Animal* ***50***: 792–796. 10.1007/s11626-014-9781-y
974 Janmaat, AF, Myers, J. 2003. Rapid evolution and the cost of resistance to Bacillus
975 thuringiensis in greenhouse populations of cabbage loopers, Trichoplusia ni.
976 *Proceedings of the Royal Society B* ***270***: 2263–2270. 10.1098/rspb.2003.2497
977 Jinek, M, Chylinski, K, Fonfara, I, Hauer, M, Doudna, JA, Charpentier, E. 2012. A
978 programmable dual-RNA-guided DNA endonuclease in adaptive bacterial
979 immunity. *Science* ***337***: 816–821. 10.1126/science.1225829
980 Jones, P et al. 2014. InterProScan 5: genome-scale protein function classification.
981 *Bioinformatics* ***30***: 1236–1240. 10.1093/bioinformatics/btu031
982 Kakumani, PK, Malhotra, P, Mukherjee, SK, Bhatnagar, RK. 2014. A draft genome
983 assembly of the army worm, Spodoptera frugiperda. *Genomics* ***104***: 134–143.
984 10.1016/j.ygeno.2014.06.005
985 Kanost, MR et al. 2016. Multifaceted biological insights from a draft genome sequence
986 of the tobacco hornworm moth, Manduca sexta. *Insect Biochem Mol Biol* ***76***:
987 118–147. 10.1016/j.ibmb.2016.07.005
988 Katsuma, S, Sugano, Y, Kiuchi, T, Shimada, T. 2015. Two Conserved Cysteine
989 Residues Are Required for the Masculinizing Activity of the Silkworm Masc
990 Protein. *J Biol Chem* ***290***: 26114–26124. 10.1074/jbc.M115.685362
991 Kawamura, Y et al. 2008. *Drosophila* endogenous small RNAs bind to Argonaute2 in
992 somatic cells. *Nature* ***453***: 793–797. 10.1038/nature06938
993 Kawaoka, S et al. 2008. Bombyx small RNAs: genomic defense system against
994 transposons in the silkworm, Bombyx mori. *Insect Biochem Mol Biol* ***38***: 1058–
995 1065. 10.1016/j.ibmb.2008.03.007

Kawaoka, S et al. 2009. The Bombyx ovary-derived cell line endogenously expresses PIWI/PIWI-interacting RNA complexes. *RNA* ***15***: 1258–1264. 10.1261/rna.1452209

Kawaoka, S et al. 2011. The silkworm W chromosome is a source of female-enriched piRNAs. *RNA* ***17***: 2144–2151. 10.1261/rna.027565.111

Keeling, CI et al. 2013. Draft genome of the mountain pine beetle, Dendroctonus ponderosae Hopkins, a major forest pest. *Genome Biology* ***14***: R27. 10.1186/gb-2013-14-3-r27

Khvorova, A, Reynolds, A, Jayasena, SD. 2003. Functional siRNAs and miRNAs exhibit strand bias. *Cell* ***115***: 209–216.

Kim, S, Kim, D, Cho, SW, Kim, J, Kim, JS. 2014. Highly efficient RNA-guided genome editing in human cells via delivery of purified Cas9 ribonucleoproteins. *Genome Res* ***24***: 1012–1019. 10.1101/gr.171322.113

Kinoshita, T, Hanazawa, T, Nohata, N, Okamoto, Y, Seki, N. 2012. The functional significance of microRNA-375 in human squamous cell carcinoma: Aberrant expression and effects on cancer pathways. *Journal of Human Genetics* ***57***: 556–563. 10.1038/jhg.2012.75

Kirino, Y, Mourelatos, Z. 2007. The mouse homolog of HEN1 is a potential methylase for Piwi-interacting RNAs. *RNA* ***13***: 1397–1401. 10.1261/rna.659307

Klattenhoff, C et al. 2009. The *Drosophila* HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. *Cell* ***138***: 1137–1149. 10.1016/j.cell.2009.07.014

Komor, AC, Badran, AH, Liu, DR. 2017. CRISPR-Based Technologies for the Manipulation of Eukaryotic Genomes. *Cell* ***168***: 20–36. 10.1016/j.cell.2016.10.044

Koren, S, Walenz, BP, Berlin, K, Miller, JR, Bergman, NH, Phillippy, AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and

repeat separation. *bioRxiv* 071282. 10.1101/071282

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

Kotin, RM. 2011. Large-scale recombinant adeno-associated virus production. *Hum Mol Genet* ***20***: R2–6. 10.1093/hmg/ddr141

Kugler, J-M, Chen, Y-W, Weng, R, Cohen, SM. 2013. miR-989 Is Required for Border Cell Migration in the Drosophila Ovary. *PLOS ONE* ***8***: e67075. 10.1371/journal.pone.0067075

Kurtz, S et al. 2004. Versatile and open software for comparing large genomes. *Genome Biol* ***5***: R12. 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***: 243–256. 10.1111/j.1365-2583.2010.01064.x

Lagesen, K, Hallin, P, Rødland, EA, Stærfeldt, H-H, Rognes, T, Ussery, DW. 2007. RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucl Acids Res* ***35***: 3100–3108. 10.1093/nar/gkm160

Lagos-Quintana, M, Rauhut, R, Lendeckel, W, Tuschl, T. 2001. Identification of Novel Genes Coding for Small Expressed RNAs. *Science* ***294***: 853–858.

Lau, NC et al. 2009. Abundant primary piRNAs, endo-siRNAs, and microRNAs in a *Drosophila* ovary cell line. *Genome Res* ***19***: 1776–1785. 10.1101/gr.094896.109

Lau, NC et al. 2006. Characterization of the piRNA complex from rat testes. *Science* ***313***: 363–367.

Le Thomas, A et al. 2014. Transgenerationally inherited piRNAs trigger piRNA biogenesis by changing the chromatin of piRNA clusters and inducing precursor processing. *Genes & Development* ***28***: 1667–1680. 10.1101/gad.245514.114

Le Thomas, A, Marinov, G, Aravin, A. 2014. A Transgenerational Process Defines piRNA Biogenesis in Drosophila virilis. *Cell Reports* 10.1016/j.celrep.2014.08.013

1050 Lee, E et al. 2013. Web Apollo: a web-based genomic annotation editing platform.
1051 *Genome Biology* ***14***: R93. 10.1186/gb-2013-14-8-r93
1052 Lewis, SH, Salmela, H, Obbard, DJ. 2016. Duplication and Diversification of Dipteran
1053 Argonaute Genes, and the Evolutionary Divergence of Piwi and Aubergine.
1054 *Genome Biol Evol* ***8***: 507–518. 10.1093/gbe/evw018
1055 Li, C et al. 2009. Collapse of Germline piRNAs in the Absence of Argonaute3 Reveals
1056 Somatic piRNAs in Flies. *Cell* ***137***: 509–521. 10.1016/j.cell.2009.04.027
1057 Lim, AK, Kai, T. 2007. Unique germ-line organelle, nuage, functions to repress selfish
1058 genetic elements in *Drosophila melanogaster*. *Proc Natl Acad Sci U S A* ***104***:
1059 6714–6719. 10.1073/pnas.0701920104
1060 Lin, S, Staahl, BT, Alla, RK, Doudna, JA. 2014. Enhanced homology-directed human
1061 genome engineering by controlled timing of CRISPR/Cas9 delivery. *eLife* ***3***:
1062 e04766. 10.7554/eLife.04766
1063 Liu, L, Qi, H, Wang, J, Lin, H. 2011. PAPI, a novel TUDOR-domain protein, complexes
1064 with AGO3, ME31B and TRAL in the nuage to silence transposition.
1065 *Development* ***138***: 1863–1873. 10.1242/dev.059287
1066 Liu, S et al. 2011. Genome-wide identification and characterization of ATP-binding
1067 cassette transporters in the silkworm, Bombyx mori. *BMC Genomics* ***12***: 491.
1068 10.1186/1471-2164-12-491
1069 Lobo, N, Li, X, Fraser, MJ. 1999. Transposition of the piggyBac element in embryos of
1070 Drosophila melanogaster, Aedes aegypti and Trichoplusia ni. *Mol Gen Genet*
1071 ***261***: 803–810. 10.1007/s004380050024
1072 Lucas, K, Raikhel, AS. 2013. Insect MicroRNAs: Biogenesis, expression profiling and
1073 biological functions. *Insect Biochemistry and Molecular Biology* ***43***: 24–38.
1074 10.1016/j.ibmb.2012.10.009
1075 Lund, AH. 2009. miR-10 in development and cancer. *Cell Death Differ* ***17***: 209–214.
1076 10.1038/cdd.2009.58

1077 Luo, R et al. 2012. SOAPdenovo2: an empirically improved memory-efficient short-read
1078 de novo assembler. *Gigascience* ***1***: 18. 10.1186/2047-217X-1-18
1079 Mallet, J. 2007. Taxonomy of Lepidoptera: the scale of the problem. *The Lepidoptera*
1080 *Taxome Project*
1081 Malone, CD et al. 2009. Specialized piRNA Pathways Act in Germline and Somatic
1082 Tissues of the *Drosophila* Ovary. *Cell* ***137***: 522–535. 10.1016/j.cell.2009.03.040
1083 Marchler-Bauer, A et al. 2015. CDD: NCBI's conserved domain database. *Nucl Acids*
1084 *Res* ***43***: D222–D226. 10.1093/nar/gku1221
1085 Marygold, SJ et al. 2007. The ribosomal protein genes and Minute loci of Drosophila
1086 melanogaster. *Genome Biology* ***8***: R216. 10.1186/gb-2007-8-10-r216
1087 Matijasevic, Z, Krzywicka-Racka, A, Sluder, G, Jones, SN. 2008. MdmX regulates
1088 transformation and chromosomal stability in p53-deficient cells. *Cell Cycle* ***7***:
1089 2967–2973. 10.4161/cc.7.19.6797
1090 McEwen, FL, Hervey, GER. 1956. An Evaluation of Newer Insecticides for Control of
1091 DDT-Resistant Cabbage Loopers. *Journal of Economic Entomology* ***49***: 385–
1092 387. 10.1093/jee/49.3.385
1093 McKenna, A, Hanna, M, Banks, E, Sivachenko…, A. 2010. The Genome Analysis
1094 Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing
1095 data. *Genome …*
1096 Mitchell, A et al. 2015. The InterPro protein families database: the classification
1097 resource after 15 years. *Nucl Acids Res* ***43***: D213–D221. 10.1093/nar/gku1243
1098 Mohn, F, Handler, D, Brennecke, J. 2015. piRNA-guided slicing specifies transcripts for
1099 Zucchini-dependent, phased piRNA biogenesis. *Science* ***348***: 812–817.
1100 10.1126/science.aaa1039
1101 Mohn, F, Sienski, G, Handler, D, Brennecke, J. 2014. The Rhino-Deadlock-Cutoff
1102 Complex Licenses Noncanonical Transcription of Dual-Strand piRNA Clusters in
1103 Drosophila. *Cell* ***157***: 1364–1379. 10.1016/j.cell.2014.04.031

1104 Mon, H, Kusakabe, T, Lee, JM, Kawaguchi, Y, Koga, K. 2004. In vivo DNA double-
1105 strand breaks enhance gene targeting in cultured silkworm cells. *Comparative*
1106 *Biochemistry and Physiology Part B: Biochemistry and Molecular Biology* ***139***:
1107 99–106. 10.1016/j.cbpc.2004.06.013
1108 Nelson, DR. 2009. The Cytochrome P450 Homepage. *Human Genomics* ***4***: 59.
1109 10.1186/1479-7364-4-1-59
1110 Ni, JD, Baik, LS, Holmes, TC, Montell, C. 2017. A rhodopsin in the brain functions in
1111 circadian photoentrainment in Drosophila. *Nature* ***545***: 340–344.
1112 10.1038/nature22325
1113 Niki, Y, Yamaguchi, T, Mahowald, AP. 2006. Establishment of stable cell lines of
1114 *Drosophila* germ-line stem cells. *Proc Natl Acad Sci U S A* ***103***: 16325–16330.
1115 10.1073/pnas.0607435103
1116 Okamura, K, Balla, S, Martin, R, Liu, N, Lai, EC. 2008a. Two distinct mechanisms
1117 generate endogenous siRNAs from bidirectional transcription in *Drosophila*
1118 *melanogaster*. *Nat Struct Mol Biol* ***15***: 581–590. 10.1038/nsmb.1438
1119 Okamura, K, Chung, WJ, Ruby, JG, Guo, H, Bartel, DP, Lai, EC. 2008b. The *Drosophila*
1120 hairpin RNA pathway generates endogenous short interfering RNAs. *Nature* ***453***:
1121 803–806. 10.1038/nature07015
1122 Pace, JK, Feschotte, C. 2007. The evolutionary history of human DNA transposons:
1123 Evidence for intense activity in the primate lineage. *Genome Res* ***17***: 422–432.
1124 10.1101/gr.5826307
1125 Pane, A, Jiang, P, Zhao, DY, Singh, M, Schupbach, T. 2011. The Cutoff protein
1126 regulates piRNA cluster expression and piRNA production in the *Drosophila*
1127 germline. *EMBO J* 10.1038/emboj.2011.334
1128 Patil, AA et al. 2017. Characterization of Armitage and Yb containing granules and their
1129 relationship to nuage in ovary-derived cultured silkworm cell. *Biochem Biophys*
1130 *Res Commun* ***490***: 134–140. 10.1016/j.bbrc.2017.06.008

1131 Pelisson, A, Sarot, E, Payen-Groschene, G, Bucheton, A. 2007. A novel repeat-
1132 associated small interfering RNA-mediated silencing pathway downregulates
1133 complementary sense *gypsy* transcripts in somatic cells of the *Drosophila* ovary.
1134 *J Virol* ***81***: 1951–1960. 10.1128/JVI.01980-06
1135 Porcelli, D, Barsanti, P, Pesole, G, Caggese, C. 2007. The nuclear OXPHOS genes in
1136 insecta: a common evolutionary origin, a common cis-regulatory motif, a common
1137 destiny for gene duplicates. *BMC Evolutionary Biology* ***7***: 215. 10.1186/1471-
1138 2148-7-215
1139 Prud'homme, N, Gans, M, Masson, M, Terzian, C, Bucheton, A. 1995. *Flamenco*, a
1140 gene controlling the gypsy retrovirus of *Drosophila melanogaster*. *Genetics* ***139***:
1141 697–711.
1142 Rainford, JL, Hofreiter, M, Nicholson, DB, Mayhew, PJ. 2014. Phylogenetic Distribution
1143 of Extant Richness Suggests Metamorphosis Is a Key Innovation Driving
1144 Diversification in Insects. *PLOS ONE* ***9***: e109085. 10.1371/journal.pone.0109085
1145 Ran, FA, Hsu, PD, Wright, J, Agarwala, V, Scott, DA, Zhang, F. 2013. Genome
1146 engineering using the CRISPR-Cas9 system. *Nat Protocols* ***8***: 2281–2308.
1147 10.1038/nprot.2013.143
1148 Robertson, HM, Gordon, KHJ. 2006. Canonical TTAGG-repeat telomeres and
1149 telomerase in the honey bee, *Apis mellifera*. *Genome Res* ***16***: 1345–1351.
1150 10.1101/gr.5085606
1151 Rota-Stabelli, O, Daley, AC, Pisani, D. 2013. Molecular timetrees reveal a Cambrian
1152 colonization of land and a new scenario for ecdysozoan evolution. *Current*
1153 *Biology* ***23***: 392–398. 10.1016/j.cub.2013.01.026
1154 Sahara, K, Marec, F, Traut, W. 1999. TTAGG telomeric repeats in chromosomes of
1155 some insects and other arthropods. *Chromosome Res* ***7***: 449–460.
1156 10.1023/A:1009297729547
1157 Saito, K et al. 2009. A regulatory circuit for piwi by the large Maf gene *traffic jam* in

Drosophila. *Nature* ***461***: 1296–1299. 10.1038/nature08501

Saito, K et al. 2006. Specific association of Piwi with rasiRNAs derived from retrotransposon and heterochromatic regions in the *Drosophila* genome. *Genes Dev* ***20***: 2214–2222.

Saito, K, Sakaguchi, Y, Suzuki, T, Suzuki, T, Siomi, H, Siomi, MC. 2007. Pimet, the *Drosophila* homolog of HEN1, mediates 2′-O-methylation of Piwi-interacting RNAs at their 3′ ends. *Genes Dev* ***21***: 1603–1608. 10.1101/gad.1563607

Sanchez-Vargas, I et al. 2009. Dengue virus type 2 infections of *Aedes aegypti* are modulated by the mosquito's RNA interference pathway. *PLoS Pathog* ***5***: e1000299. 10.1371/journal.ppat.1000299

Schneider, I. 1979. Giemsa staining of insect chromosomes for karyotype analysis. *Methods in Cell Science* ***5***: 1027–1028.

Schwarz, DS, Hutvágner, G, Du, T, Xu, Z, Aronin, N, Zamore, PD. 2003. Asymmetry in the assembly of the RNAi enzyme complex. *Cell* ***115***: 199–208.

Shirayama, M et al. 2012. piRNAs Initiate an Epigenetic Memory of Nonself RNA in the *C. elegans* Germline. *Cell* ***150***: 65–77. 10.1016/j.cell.2012.06.015

Sienski, G, Donertas, D, Brennecke, J. 2012. Transcriptional silencing of transposons by Piwi and maelstrom and its impact on chromatin state and gene expression. *Cell* ***151***: 964–980. 10.1016/j.cell.2012.10.040

Simão, FA, Waterhouse, RM, Ioannidis, P, Kriventseva, EV, Zdobnov, EM. 2015. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* btv351. 10.1093/bioinformatics/btv351

Siomi, H, Siomi, MC. 2009. On the road to reading the RNA-interference code. *Nature* ***457***: 396–404. 10.1038/nature07754

Slater, GSC, Birney, E. 2005. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* ***6***: 31. 10.1186/1471-2105-6-31

Smit, AFA, Hubley, R, Green, P. 2017. RepeatMasker Open-3.0.

1185 Sokol, NS, Ambros, V. 2005. Mesodermally expressed *Drosophila* microRNA-1 is
1186 regulated by Twist and is required in muscles during larval growth. *Genes Dev*
1187 ***19***: 2343–2354.
1188 Stanke, M, Tzvetkova, A, Morgenstern, B. 2006. AUGUSTUS at EGASP: using EST,
1189 protein and genomic alignments for improved gene prediction in the human
1190 genome. *Genome Biology* ***7***: 1–8. 10.1186/gb-2006-7-s1-s11
1191 Suetsugu, Y et al. 2013. Large Scale Full-Length cDNA Sequencing Reveals a Unique
1192 Genomic Landscape in a Lepidopteran Model Insect, Bombyx mori. *G3* ***3***: 1481–
1193 1492. 10.1534/g3.113.006239
1194 Takasu, Y, Tamura, T, Sajwan, S, Kobayashi, I, Zurovec, M. 2014. The use of TALENs
1195 for nonhomologous end joining mutagenesis in silkworm and fruitfly. *Methods* ***69***:
1196 46–57. 10.1016/j.ymeth.2014.02.014
1197 Tam, OH et al. 2008. Pseudogene-derived small interfering RNAs regulate gene
1198 expression in mouse oocytes. *Nature* ***453***: 534–538. 10.1038/nature06904
1199 Teleman, AA, Maitra, S, Cohen, SM. 2006. *Drosophila* lacking microRNA miR-278 are
1200 defective in energy homeostasis. *Genes Dev* ***20***: 417–422. 10.1101/gad.374406
1201 Terakita, A. 2005. The opsins. *Genome Biology* ***6***: 213. 10.1186/gb-2005-6-3-213
1202 The, ISG. 2008. The genome of a lepidopteran model insect, the silkworm Bombyx
1203 mori. *Insect Biochemistry and Molecular Biology* ***38***: 1036–1045.
1204 10.1016/j.ibmb.2008.11.004
1205 Traut, W, Sahara, K, Marec, F. 2007. Sex chromosomes and sex determination in
1206 Lepidoptera. *Sex Dev* ***1***: 332–346. 10.1159/000111765
1207 Tyler, DM et al. 2008. Functionally distinct regulatory RNAs generated by bidirectional
1208 transcription and processing of microRNA loci. *Genes Dev* ***22***: 26–36.
1209 10.1101/gad.1615208
1210 Vagin, VV, Sigova, A, Li, C, Seitz, H, Gvozdev, V, Zamore, PD. 2006. A distinct small
1211 RNA pathway silences selfish genetic elements in the germline. *Science* ***313***:

320–324. 10.1126/science.1129333

Van der Auwera, GA et al. 2013. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics* ***43***: 11.10.1–33. 10.1002/0471250953.bi1110s43

van Oers, MM, Pijlman, GP, Vlak, JM. 2015. Thirty years of baculovirus-insect cell protein expression: from dark horse to mainstream technology. *J Gen Virol* ***96***: 6–23. 10.1099/vir.0.067108-0

van Rij, RP et al. 2006. The RNA silencing endonuclease Argonaute 2 mediates specific antiviral immunity in Drosophila melanogaster. *Genes Dev* ***20***: 2985–2995. 10.1101/gad.1482006

van Schooten, B, Jiggins, CD, Briscoe, AD, Papa, R. 2016. Genome-wide analysis of ionotropic receptors provides insight into their evolution in Heliconius butterflies. *BMC Genomics* ***17***: 254. 10.1186/s12864-016-2572-y

Vastenhouw, NL et al. 2010. Chromatin signature of embryonic pluripotency is established during genome activation. *Nature* 10.1038/nature08866

Velarde, RA, Sauer, CD, O. Walden, KK, Fahrbach, SE, Robertson, HM. 2005. Pteropsin: A vertebrate-like non-visual opsin expressed in the honey bee brain. *Insect Biochemistry and Molecular Biology* ***35***: 1367–1377. 10.1016/j.ibmb.2005.09.001

Vrettos, N, Maragkakis, M, Alexiou, P, Mourelatos, Z. 2017. Kc167, a widely used Drosophila cell line, contains an active primary piRNA pathway. *RNA* ***23***: 108–118. 10.1261/rna.059139.116

Walker, BJ et al. 2014. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLOS ONE* ***9***: e112963. 10.1371/journal.pone.0112963

Walters, JR, Hardcastle, TJ. 2011. Getting a Full Dose? Reconsidering Sex Chromosome Dosage Compensation in the Silkworm, Bombyx mori. *Genome*

Biol Evol ***3***: 491–504. 10.1093/gbe/evr036

Walters, JR, Hardcastle, TJ, Jiggins, CD. 2015. Sex Chromosome Dosage Compensation in Heliconius Butterflies: Global yet Still Incomplete. *Genome Biol Evol* ***7***: 2545–2559. 10.1093/gbe/evv156

Wang, W et al. 2008. Chromosomal transposition of PiggyBac in mouse embryonic stem cells. *PNAS* ***105***: 9290–9295. 10.1073/pnas.0801017105

Wang, Y et al. 2013. The CRISPR/Cas system mediates efficient genome engineering in Bombyx mori. *Cell research* ***23***: 1414. 10.1038/cr.2013.146

Wanner, KW, Robertson, HM. 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***: 621–629. 10.1111/j.1365-2583.2008.00836.x

Watanabe, T et al. 2008. Endogenous siRNAs from naturally formed dsRNAs regulate transcripts in mouse oocytes. *Nature* ***453***: 539–543. 10.1038/nature06908

Webster, A et al. 2015. Aub and Ago3 Are Recruited to Nuage through Two Mechanisms to Form a Ping-Pong Complex Assembled by Krimper. *Mol Cell* ***59***: 564–575. 10.1016/j.molcel.2015.07.017

Wei, Y, Chen, S, Yang, P, Ma, Z, Kang, L. 2009. Characterization and comparative profiling of the small RNA transcriptomes in two phases of locust. *Genome Biology* ***10***: R6. 10.1186/gb-2009-10-1-r6

Wheat, CW, Wahlberg, N. 2013. Phylogenomic insights into the cambrian explosion, the colonization of land and the evolution of flight in arthropoda. *Syst Biol* ***62***: 93–109. 10.1093/sysbio/sys074

Wickham, TJ, Davis, T, Granados, RR, Shuler, ML, Wood, HA. 1992. Screening of Insect Cell Lines for the Production of Recombinant Proteins and Infectious Virus in the Baculovirus Expression System. *Biotechnol Progress* ***8***: 391–396. 10.1021/bp00017a003

1266 Xavier, B, David, M, Piulachs, M-D. 2005. The Mevalonate Pathway and the Synthesis
1267 of Juvenile Hormone in Insects. *Annual Review of Entomology* ***50***: 181–199.
1268 10.1146/annurev.ento.50.071803.130356
1269 Xiang, H et al. 2010. Single base-resolution methylome of the silkworm reveals a sparse
1270 epigenomic map. *Nat Biotech* ***28***: 516–520. 10.1038/nbt.1626
1271 Xiol, J et al. 2012. A Role for Fkbp6 and the Chaperone Machinery in piRNA
1272 Amplification and Transposon Silencing. *Mol Cell* ***47***: 970–979.
1273 10.1016/j.molcel.2012.07.019
1274 Yoshihama, M et al. 2002. The Human Ribosomal Protein Genes: Sequencing and
1275 Comparative Analysis of 73 Genes. *Genome Res* ***12***: 379–390.
1276 10.1101/gr.214202
1277 You, M et al. 2013. A heterozygous moth genome provides insights into herbivory and
1278 detoxification. *Nat Genet* ***45***: 220–225. 10.1038/ng.2524
1279 Yu, Q-Y, Lu, C, Li, W-L, Xiang, Z-H, Zhang, Z. 2009. Annotation and expression of
1280 carboxylesterases in the silkworm, Bombyx mori. *BMC Genomics* ***10***: 553.
1281 10.1186/1471-2164-10-553
1282 Yu, Q et al. 2008. Identification, genomic organization and expression pattern of
1283 glutathione S-transferase in the silkworm, Bombyx mori. *Insect Biochemistry and*
1284 *Molecular Biology* ***38***: 1158–1164. 10.1016/j.ibmb.2008.08.002
1285 Yu, Z et al. 2014. Various applications of TALEN-and CRISPR/Cas9-mediated
1286 homologous recombination to modify the Drosophila genome. *Biology Open* ***3***:
1287 271–280. 10.1242/bio.20147682
1288 Zambon, RA, Vakharia, VN, Wu, LP. 2006. RNAi is an antiviral immune response
1289 against a dsRNA virus in *Drosophila melanogaster*. *Cell Microbiol* ***8***: 880–889.
1290 Zhan, S, Merlin, C, Boore, J, Reppert, S. 2011. The Monarch Butterfly Genome Yields
1291 Insights into Long-Distance Migration. *Cell* ***147***: 1171–1185.
1292 10.1016/j.cell.2011.09.052

Zhang, Z, Theurkauf, WE, Weng, Z, Zamore, PD. 2012. Strand-specific libraries for high throughput RNA sequencing (RNA-Seq) prepared without poly(A) selection. *Silence* ***3***: 9. 10.1186/1758-907X-3-9

Zhang, Z et al. 2014. The HP1 Homolog Rhino Anchors a Nuclear Complex that Suppresses piRNA Precursor Splicing. *Cell* ***157***: 1353–1363. 10.1016/j.cell.2014.04.030

Zhang, Z, Aslam, AFM, Liu, X, Li, M, Huang, Y, Tan, A. 2015. Functional analysis of Bombyx Wnt1 during embryogenesis using the CRISPR/Cas9 system. *Journal of Insect Physiology* ***79***: 73–79. 10.1016/j.jinsphys.2015.06.004

Zhu, L, Mon, H, Xu, J, Lee, JM, Kusakabe, T. 2015. CRISPR/Cas9-mediated knockout of factors in non-homologous end joining pathway enhances gene targeting in silkworm cells. *Sci Rep* ***5***: 18103. 10.1038/srep18103

Zhuang, J, Wang, J, Theurkauf, W, Weng, Z. 2014. TEMP: a computational method for analyzing transposable element polymorphism in populations. *Nucleic Acids Res* ***42***: 6826–6838. 10.1093/nar/gku323

Zimyanin, VL et al. 2008. In vivo imaging of oskar mRNA transport reveals the mechanism of posterior localization. *Cell* ***134***: 843–853. 10.1016/j.cell.2008.06.053

Table Legend

Table 1. Genome and gene set statistics for *T. ni* and *B. mori* (*The, ISG, 2008*). Cytochrome P450s, glutathione S-transferases, carboxylesterases, and ATP-binding cassette transporters for *B. mori* were retrieved from (*Yu, Q et al., 2008*; *Yu, Q-Y et al., 2009*; *Ai, J et al., 2011*; *Liu, S et al., 2011*).

	T. ni	***B. mori***
Genome Metrics		
Genome size (Mb)	368.2	431.7
Chromosome count	28	28
Scaffold N50 (Mb)	14.2	3.7
Contig N50 (kb)	621.9	15.5
Mitochondrial genome (kb)	15.8	15.7
Quality Control Metrics		
BUSCO complete (%)	97.5	95.5
CRP genes (%)	100%	100%
OXPHOS genes (%)	100%	100%
Genomic Features		
Repeat content (%)	20.5%	43.6%
GC content	35.6%	37.3%
CpG (O/E)	1.07	1.13
Coding (%)	5.58	4.11
Sex chromosomes	ZW	ZW
Gene Statistics		
Protein-coding genes	14,043	14,623
with Pfam matches	9,295	9,685
with GO terms	9,790	10,148
Cytochrome P450 proteins	108	83
Glutathione S-transferases	34	23
Carboxylesterases	87	76
ATP-binding cassette transporters	54	51
Universal orthologs lost	156	75
Species-specific genes	3,098	2,313

Figure Legends

Figure 1. Chromosomes and genes in the *T. ni* genome based on data from the Hi5 cell line. (A) Genome assembly and annotation workflow. (B) An example of a DAPI-stained spread of Hi5 cell mitotic chromosomes used to determine the karyotype. (C) Phylogenetic tree and orthology assignment of *T. ni* with 18 arthropod and two mammalian genomes. Colors denote gene categories. The category 1:1:1 represents universal single-copy orthologs, allowing absence and/or duplication in one genome. N:N:N orthologs include orthologs with variable copy numbers across species, allowing absence in one genome or two genomes from different orders. Lepidoptera-specific genes are present in at least three of the four lepidopteran genomes; Hymenoptera-specific genes are present in at least one wasp or bee genome and at least one ant genome. Coleoptera-specific genes are present in both coleopteran genomes; Diptera-specific genes present in at least one fly genome and one mosquito genome. Insect indicates other insect-specific genes. Mammal-specific genes are present in both mammalian genomes. The phylogenetic tree is based on the alignment of 1:1:1 orthologs.

Figure 1—figure supplement 1. Hi5 cell Karyotyping. Thirty images showing the numbers of chromosomes (N) in Hi5 cells. N ranged from 103 to 122; mean ± S.D. = 111.7 ± 5.45. Since lepidopteran cell lines are typically tetraploid, the haploid genome likely contains 28 (mean ± S.D. = 27.9 ± 1.36) pairs of chromosomes.

Figure 1—figure supplement 2. Phylogenetic tree of 21 species showing the scale, branch lengths and bootstrap support. Strict 1:1:1 orthologs were used to compute the phylogenetic tree using the maximum likelihood method. Black, branch length; red, bootstrap support.

Figure 1—figure supplement 3. Opsins in insects.

Figure 2. *T. ni* males are ZZ and females are ZW. (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.

Figure 2—figure supplement 1. (A) Genomic coverage comparison of Z-linked, W-linked and autosomal contigs. Contig coverage was shuffled 1,000,000 times to calculate the coverage ratio. Outliers are not shown. (B) Autosomal, Z-linked and W-linked transcript abundance in Hi5 cells and *T. ni* tissues. (C) Transcript abundance ratios of autosomal, Z-linked, and W-linked genes in Hi5 cells and *T. ni* tissues. Error bars represent 95% confidence interval estimated from 1,000 bootstrap replicates. (D). Sex-specific splicing of *T. ni doublesex* pre-mRNA.

Figure 2—figure supplement 2. CpG ratios and transposons. (A) Distribution of observed-to-expected CpG ratios in protein-coding genes (left panel) and in 500 bp genomic windows (right panel) in *A. mellifera*, *B. mori*, *D. plexippus*, *D. melanogaster*, *P. xylostella*, *T. castaneum*, and *T. ni*. (B) Proportion of the genome occupied by transposons versus transposon sequence divergence. Sequence divergence was calculated by comparing individual transposon copies with the corresponding consensus sequence (See Materials and methods). (C) Repeat content in lepidopteran genomes.

Figure 3. miRNA expression in *T. ni*. (A) Comparison of miRNA abundance in male and female *T. ni* thoraces. Solid circles, miRNAs with FDR < 0.1 and fold change > 2. Outlined circles, all other miRNAs. (B) Comparison of the tissue distribution of the 44 most abundant miRNAs among *T. ni* ovaries, testes, and Hi5. (C) Heat map showing the abundance of miRNAs in (B). miRNAs are ordered according to abundance in ovary. Conservation status uses the same color scheme in (A).

Figure 4. siRNA. (A) Distribution of siRNAs mapping to TNCL virus in the genomic (blue) and anti-genomic orientation (red). Inset: length distribution of TNCL virus-mapping small RNAs. (B) Distance between the 3′ and 5′ ends of siRNAs on opposite viral strands. (C) Distance between the 3′ and 5′ ends of siRNAs on the same viral strand. (D) Length distribution of small RNAs from unoxidized and oxidized small RNA-seq libraries. (E) Lepidopteran siRNAs are not 2′-*O*-methylated. The box plots display the ratio of abundance (as a fraction of all small RNAs sequenced) for each siRNA in oxidized versus unoxidized small RNA-seq libraries. The tree shows the phylogenetic relationships of the analyzed insects. Outliers are not shown.

Figure 4—figure supplement 1. (A) siRNA length distributions for multiple insects in oxidized and unoxidized small RNA-seq libraries. (B) Length distribution of fully matched and tailed TNCL virus-siRNAs.

Figure 4—figure supplement 2. Loading asymmetry of siRNAs mapping to TNCL RNA1 (A) and RNA2 (B). For each single-stranded siRNA species, we searched for siRNAs on the other strand that when paired produce a typical siRNA duplex with two-nucleotide overhanging 3′ ends.

Figure 5. piRNAs and miRNAs in the *T. ni* genome. (A) Abundance of mRNAs encoding piRNA pathway proteins) in Hi5 cells, ovary, testis, and thorax. (B) Ideogram displaying the positions of miRNA genes (arrowheads) and piRNA clusters in the *T. ni* genome. Color-coding reports tissue expression for Hi5 cells, ovaries, testis, and thorax. Contigs that cannot be placed onto chromosome-length scaffolds are arbitrarily concatenated and are marked 'Un.' (C) Distribution of piRNAs among the autosomes, Z, and W chromosomes in Hi5 cells, ovary, testis, and female and male thorax, compared with the fraction of the genome corresponding to autosomes, W, and Z chromosomes.

Figure 5—figure supplement 1. piRNA abundance (ppm) along the most productive piRNA cluster. Top, fixed scale (some data clipped); bottom, auto-scaled.

Figure 5—figure supplement 2. (A) piRNA clusters tend to produce piRNAs that are antisense to transposons. The *x*-axis represents the ratio of piRNAs from the plus strand to piRNAs from the minus strand, with the dotted lines indicating twofold difference. The *y*-axis indicates the ratio of transposons lengths on the plus strand over transposon length on the minus strand. The solid line indicates regression line and shading indicates 95% confidence interval by LOWESS. Boxplot shows fractions of antisense transposons (i.e. transposons inserted opposite to the direction of piRNAs precursor transcription) in dual- and uni-strand clusters. Outliers are not shown. Wilcoxon rank-sum test. (B) piRNA densities on autosomal, Z-linked and W-linked contigs in Hi5 cells, ovary, testis, and female and male thorax. (C) Abundance of piRNAs from putative W-linked genes.

Figure 6. (A) Hi5-specific piRNA clusters contain younger transposon copies. RC, rolling-circle transposons; LINE, Long interspersed nuclear elements; LTR, long terminal repeat retrotransposon; DNA, DNA transposon. (B) Comparison of piRNA abundance per cluster in female and male thorax. (C) piRNA precursors are rarely

spliced. The number of introns supported by exon-exon junction-mapping reads is shown for protein-coding genes and for piRNA clusters for each tissue or cell type. (D) piRNA precursors are inefficiently spliced. Splicing efficiency is defined as the ratio of spliced over unspliced reads. Splice sites were categorized into those inside and outside piRNA clusters. Outliers are not shown.

Figure 6—figure supplement 1. (A) Comparison of piRNA abundance (ppm) from ovary and Hi5 piRNA-producing loci and from ovary and testis piRNA-producing loci. (B) piRNA cluster lengths in *T. ni* ovary, testis, thorax, and Hi5 cells. (C) Motifs around intron boundaries of predicted protein-coding gene models within and outside of piRNA clusters.

Figure 7. Genome editing in Hi5 cells. (A) Strategy for using Cas9/sgRNA RNPs to generate a loss-of-function *TnPiwi* deletion allele. Red, protospacer-adjacent motif (PAM); blue, protospacer sequence. Arrows indicate the diagnostic forward and reverse primers used in PCR to detect genomic deletions (Δ). Sanger sequencing of the ~1700 bp PCR products validated the *TnPiwi* deletions. (B) An example of PCR analysis of a *TnPiwi* deletion event. (C) Strategy for using Cas9/sgRNA RNPs and a single-stranded DNA homology donor to insert EGFP and an HA-tag in-frame with the *vasa* open reading frame. (D) An example of PCR analysis of a successful HDR event. DNA isolated from wild type (WT) and FACS-sorted, EGFP-expressing Hi5 cells (HDR) were used as templates.

Figure 8. Hi5 cells contain nuage. (A) Schematic of single-clone selection of genome-edited Hi5 cells using the strategy described in Figure 7C. (B) A representative field of Hi5 cells edited to express EGFP-HA-Vasa from the endogenous locus. (C) A representative image of a fixed, EGFP-HA-Vasa-expressing Hi5 cell stained with DAPI,

anti-EGFP and anti-HA antibodies. EGFP and HA staining colocalize in a perinuclear structure consistent with Vasa localizing to nuage.

Supplementary Files

Supplementary file 1. *T. ni* genome statistics. (A) BUSCO assessments of *T. ni* and six other genomes. (B) CRP genes. (C) Genes in the OXPHOS pathway. (D) Genome comparisons. Genomes assembled using paired-end DNA-seq data from male and female *T. ni* pupae are compared with the Hi5 genome as the reference. The dot plots show genome alignments for contigs ≥1 kb. (E) Numbers of genes in lepidopteran genomes. (F) Positions of telomeric repeats: position of $(TTAGG)_n$ longer than 100 nt. (G) Transposons in *T. ni* subtelomeric regions. (H) Repeat statistics for the *T. ni* genome. (I) Transposon family divergence rates. (J) Manual curation of W-linked protein-coding genes and miRNAs.

Supplementary file 2. Genes encoding small RNA pathway proteins. (A) Genes encoding miRNA and siRNA pathway proteins. (B) Genes encoding piRNA pathway proteins (grouped by sequence orthology).

Supplementary file 3. *T. ni* miRNAs, siRNAs and piRNAs. (A) miRNA annotation. (B) Mapping statistics for endogenous siRNAs in *T. ni* and *D. melanogaster*. (C) piRNA cluster lengths. piRNA cluster coordinates in Hi5 (D), ovary (E), testis (F), female thorax (G), and male thorax (H).

Supplementary file 4. mirDeep2 output for *T. ni* miRNAs

Supplementary file 5. Genomes used in this study.

Supplementary file 6. *T. ni* detoxification-related genes. (A) P450 gene counts by clade in *T. ni* and *B. mori*. (B) Sequences of P450 proteins. (C) Sequences of glutathione-*S*-transferase proteins. (D) Carboxylesterase gene counts by clade in *T. ni* and *B. mori*. (E) Sequences of carboxylesterase proteins. (F) ATP-binding cassette transporter gene counts by clade in *T. ni* and *B. mori*. (G) Sequences of ATP-binding cassette transporter proteins.

Supplementary file 7. *T. ni* chemoreception genes. (A) Sequences of olfactory receptor proteins. (B) Sequences of gustatory receptor proteins. (C) Sequences of ionotropic receptor proteins.

Supplementary file 8. Genes in the juvenile hormone biosynthesis and degradation pathways.

Supplementary file 9. Genome-modified sequences.

Supplementary file 10. Single-stranded DNA donor purification

Materials and methods

Genomic DNA libraries

Hi5 cells (ThermoFisher, Waltham, MA, USA) were cultured at 27℃ in Express Five Serum Free Medium (ThermoFisher) following the manufacturer's protocol. Thorax were dissected from four-day-old female or male *T. ni* pupa (Benzon Research, Carlisle, PA, USA). Cells or tissues were lysed in 2× PK buffer (200 mM Tris-HCl [pH7.5], 300 mM NaCl, 25 mM EDTA, 2% w/v SDS) containing 200 µg/ml proteinase K at 65℃ for 1 h, extracted with phenol:chloroform:isoamyl alcohol (25:24:1; Sigma, St. Louis, MO, USA), and genomic DNA collected by ethanol precipitation. The precipitate was dissolved in 10 mM Tris-HCl (pH 8.0), 0.1mM EDTA, treated with 20 µg/ml RNase A at 37℃ for 30 min, extracted with phenol:chloroform:isoamyl alcohol (25:24:1), and collected by ethanol precipitation. DNA concentration was determined (Qubit dsDNA HS Assay, ThermoFisher). Genomic DNA libraries were prepared from 1 µg genomic DNA (Illumina TruSeq LT kit, NextSeq 500, Illumina, San Diego, CA, USA).

Long-read genome sequencing with a 23 kb average insert range was constructed from 16 µg genomic DNA using the SMRTbell Template Prep Kit 1.0 SPv3 (Pacific Biosciences, Menlo Park, CA, USA) according to manufacturer's protocol. Sequence analysis was performed using P6/C4 chemistry, 240 min data collection per SMRTcell on an RS II instrument (Pacific Biosciences). Mate pair libraries with 2 kb and 8 kb insert sizes were constructed (Nextera Mate Pair Library Prep Kit, Illumina) according to manufacturer's protocol from 1 µg Hi5 cell genomic DNA. Libraries were sequenced to obtain 79 nt paired-end reads (NextSeq500, Illumina).

Hi-C

Hi-C libraries were generated from Hi5 cells as described (*Belton, J-M et al., 2012*), except that 50 million cells were used. Hi-C Libraries were sequenced using the NextSeq500 platform (Illumina) to obtain 79 nt, paired-end reads.

Karyotyping

Hi5 cells were first incubated in Express Five medium containing 1 µg/ml colcemid at 27℃ for 8 h (*Schneider, I, 1979*), then in 4 ml 0.075 M KCl for 30 min at 37℃, and fixed with freshly prepared methanol:acetic acid (3:1, v/v) precooled to −20°C. Mitotic chromosomes were spread, mounted by incubation in ProLong Gold Antifade Mountant with DAPI (4',6'-diamidino-2-phenylindole; ThermoFisher) overnight in the dark, and imaged using a DMi8 fluorescence microscope equipped with an 63× 1.40 N.A. oil immersion objective (HCX PL APO CS2, Leica Microsystems, Buffalo Grove, IL, USA) as described (*Matijasevic, Z et al., 2008*).

Small RNA libraries

Ovaries, testes, and thoraces were dissected from cabbage looper adults 24–48 h after emerging. Total RNA (30 µg) was isolated (mirVana miRNA isolation kit, Ambion, Austin, TX, USA) and sequenced using the NextSeq500 platform (Illumina) to obtain 59 nt single-end reads as previously described (*Han, BW et al., 2015*).

RNA-seq

Adult ovaries, testes, or thoraces were dissected from cabbage looper adults 24 to 48 h after emerging. Total RNA (3 µg) was purified (mirVana miRNA isolation kit, Ambion) and sequenced as described (*Zhang, Z et al., 2012*) using the NextSeq500 platform (Illumina) to obtain 79 nt, paired-end reads.

Genome assembly

Canu v1.3 (*Koren, S et al., 2017*) was used to assemble long reads into contigs, followed by Quiver (github.com/PacificBiosciences/GenomicConsensus) to polish the contigs using the same set of reads. Pilon (*Walker, BJ et al., 2014*) was used to further polish the assembly using Illumina paired-end reads. Finally, to assemble the genome into chromosome-length scaffolds, we joined the contigs using Hi-C reads and

LACHESIS (*Burton, JN et al., 2013*). The mitochondrial genome was assembled separately using MITObim (six iterations, *D. melanogaster* mitochondrial genome as bait; [*Hahn, C et al., 2013*]).

To evaluate the quality of the genome assembly, we ran BUSCO v3 (*Simão, FA et al., 2015*) using the arthropod profile and default parameters to identify universal single-copy orthologs. We further evaluated genome quality using conserved gene sets: OXPHOS and CRP genes. *B. mori* and *D. melanogaster* OXPHOS and CRP protein sequences were retrieved (*Marygold, SJ et al., 2007*; *Porcelli, D et al., 2007*) and BLASTp was used to search for their *T. ni* homologs, which were further validated by querying using InterPro (*Jones, P et al., 2014*; *Mitchell, A et al., 2015*). We also assembled *T. ni* genomes from male and female animals respectively using SOAPdenovo2 (kmer size 69; [*Luo, R et al., 2012*]. We then compared the animal genomes with the *T. ni* genome assembled from Hi5 cells using QUAST (`-m 500`) (Gurevich et al., 2013, #56036;) and the nucmer and mummerplot (`--layout --filter`) functions from MUMmer 3.23 (*Kurtz, S et al., 2004*). To determine the genomic variants, we used HaplotypeCaller from GATK (*McKenna, A et al., 2010*; *DePristo, MA et al., 2011*; *Van der Auwera, GA et al., 2013*) (`-ploidy 4 -genotyping_mode DISCOVERY`').

Genome annotation

To annotate the *T. ni* genome, we first masked repetitive sequences and then integrated multiple sources of evidence to predict gene models. We used RepeatModeler to define repeat consensus sequences and RepeatMasker (`-s -e ncbi`) to mask repetitive regions (*Smit, AFA et al., 2017*). We used RNAmmer (*Lagesen, K et al., 2007*) to predict 8S, 18S, 28S rRNA genes, and Barrnap (https://github.com/tseemann/barrnap) to predict 5.8S rRNA genes. We used Augustus v3.2.2 (*Stanke, M et al., 2006*) and SNAP (*Korf, I, 2004*) to computationally predicted gene models. Predicted gene models

were compiled by running six iterations of MAKER (*Campbell, MS et al., 2014*), aided with homology evidence of well annotated genes (UniProtKB/Swiss-Prot and Ensembl) and of transcripts from related species (*B. mori* [*Suetsugu, Y et al., 2013*] and *D. melanogaster* [*Attrill, H et al., 2016*]). We used BLAST2GO (*Conesa, A et al., 2005*) to integrate results from BLAST, and InterPro (*Mitchell, A et al., 2015*) to assign GO terms to each gene. We used MITOS (*Bernt, M et al., 2013*) web server to predict mitochondrial genes and WebApollo (*Lee, E et al., 2013*) for manual curation of genes of interest. To characterize telomeres, we used $(TTAGG)_{200}$ (*Robertson, HM, Gordon, KHJ, 2006*) as the query to search the *T. ni* genome using BLASTn with the option '`-dust no`' and kept hits longer than 100 nt. The genomic coordinates of these hits were extended by 10 kb to obtain the subtelomeric region.

Orthology and evolution

To place genes into ortholog groups, we compared the predicted proteomes from 21 species (Supplementary file 5). Orthology assignment was determined using OrthoMCL (*Hirose, Y, Manley, JL, 1997*) with default parameters. MUSCLE v3.8.31 (*Edgar, RC, 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, J, 2000*) with default parameters, and fed into PhyML 3.0 (*Vastenhouw, NL 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 (*Shirayama, M et al., 2012*).

Sex determination and sex chromosomes

To identify sex-linked contigs, we mapped genomic sequence reads from males and females to the contigs. Reads with MAPQ scores ≥20 were used to calculate contig coverage, which was then normalized by the median coverage. The distribution of

normalized contig coverage ratios (male:female ratios, M:F ratios) was manually checked to empirically determine the thresholds for Z-linked and W-linked contigs (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, RJ et al., 2016*). Z/AA ratio was calculated according to (*Gu, L et al., 2017*).

Gene families for detoxification and chemoreception

To curate genes related to detoxification and chemoreception, we obtained seed alignments from Pfam (*Finn, RD et al., 2016*) and ran hmmbuild 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, (Supplementary file 6, 7 and 8). We then used these HMM profiles to search for gene models in the predicted *T. ni* proteome (hmmsearch, e-value cutoff: 1×10^{-5}). We also retrieved reference sequences of P450, GST, COE, ABC, OR, GR, IR, OBP, and juvenile hormone pathway genes from the literature (*Hekmat-Scafe, DS et al., 2002*; ; *Xavier, B et al., 2005*; *Wanner, KW, Robertson, HM, 2008*; *Yu, Q et al., 2008*; *Benton, R et al., 2009*; *Gong, D-P et al., 2009*; *Yu, Q-Y et al., 2009*; *Croset, V et al., 2010*; *Ai, J et al., 2011*; *Liu, S et al., 2011*; *Dermauw, W, Van Leeuwen, T, 2014*; *van Schooten, B et al., 2016*). These were aligned to the *T. ni* genome using tBLASTx (*Altschul, SF et al., 1990*) and Exonerate (*Slater, GSC, Birney, E, 2005*) to search for homologs. Hits were manually inspected to ensure compatibility with RNA-seq data, predicted gene models, known protein domains (using CDD [*Marchler-Bauer, A et al., 2015*]) and homologs from other species. P450 genes were submitted to David Nelson's Cytochrome P450 Homepage (*Nelson, DR, 2009*) for nomenclature and classification. Sequences and statistics of these genes are in Supplementary files 6, 7 and 8.

To determine the phylogeny of these gene families, we aligned the putative protein sequences from *T. ni* and *B. mori* genomes using MUSCLE (*Edgar, RC, 2004*), trimmed the multiple sequence alignments using TrimAl (*Capella-Gutiérrez, S et al., 2009*) (with the option `-automated1`), and performed phylogenetic analysis (PhyML 3.0 [*Vastenhouw, NL 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, we used opsin mRNA and peptide sequences from other species (*Zimyanin, VL et al., 2008*; *Futahashi, R et al., 2015*) to search for homologs in *T. ni*. To discriminate opsin genes from other G-protein-coupled receptors, we required that the top hit in the NCBI non-redundant database and UniProt were opsins.

Transposon analysis

To determine transposon age, we calculated the average percent divergence for each transposon family: the percent divergence (RepeatMasker) of each transposon copy was multiplied by its length, and the sum of all copies were divided by the sum of lengths of all copies in the family (*Pace, JK, Feschotte, C, 2007*). We used TEMP (*Zhuang, J et al., 2014*) to identify transposon insertions in the Hi5 genome.

miRNA and siRNA analysis

mirDeep2 (*Friedländer, MR et al., 2008*; *Friedlander, MR et al., 2012*) with default parameters predicted miRNA genes. Predicted miRNA hairpins were required to have homology (exact seed matches and BLASTn e-value < 1×10^{-5}) to known miRNAs and/or miRDeep2 scores ≥10. miRNAs were named according to exact seed matches and high sequence identities (BLASTn e-value < 1×10^{-5}) 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.

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

To search for viral transcripts in *T. ni*, we downloaded viral protein sequences from NCBI (http://www.ncbi.nlm.nih.gov/genome/viruses/) and used using tBLASTn to map them to the *T. ni* genome and to the transcriptomes of Hi5 cells and five *T. ni* tissues. We filtered hits (percent identity ≥ 0.80, e-val ≤ 1×10^{-20}, and alignment length ≥100) and mapped small RNA-seq reads to the identified viral transcripts.

Candidate genomic hairpins were defined according to (*Okamura, K et al., 2008b*). Candidate *cis*-NATs were defined according to (*Ghildiyal, M et al., 2008*).

piRNA analysis

To determine the genomic coordinates of piRNA-producing loci, we mapped small RNAs to the genome as described (*Han, BW et al., 2014*). We then calculated the abundance of piRNAs in 5 kb genomic windows. For each window, we counted the number of uniquely mapped reads and the number of reads mapped to multiple loci (multimappers) 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 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.

To identify differentially expressed piRNA loci, we used the ppm and rpkm values, normalized to the total number of uniquely mapped reads, to measure piRNA abundance. For analyses including all mapped reads (uniquely mapped reads and multimappers), reads were apportioned by the number of times that they were mapped to the genome. To make piRNA loci comparable across tissues, we merged piRNA loci from ovary, testis, female and male thorax, and Hi5 cells. For the comparison between female and male thoraces, the cluster on tig00001980 was removed as this cluster likely corresponds to a mis-assembly. We used Spearman correlations to calculate the pairwise correlations of piRNA abundances. As for defining sex-linked contigs, we calculated M:F ratios and used the same thresholds 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. We used AUGUSTUS (*Stanke, M et al., 2006*), with the model trained for *T. ni* genome-wide gene prediction, to predict gene models and their splice sites in *T. ni* piRNA clusters.

β-elimination

Total RNAs were extracted from Hi5 cells using mirVana kit as described previously. We then incubated 100 µg total RNA with 25 mM $NaIO_4$ in borate buffer (148 mM Borax,148 mM Boric acid, pH 8.6) for 30 min at room temperature, beta-elimination was performed in 50 mM NaOH at 45℃ for 90 min (*Horwich, MD et al., 2007*). The resultant RNA was collected by ethanol precipitation.

sgRNA design

sgRNAs for the target loci (5′-end of *TnPiwi* and 5′-end of *vasa*) were designed using crispr.mit.edu (*Hsu, PD et al., 2013*) to retrieve all possible guide sequences, and guide sequences adjacent to deletion or insertion targets were chosen. Supplementary file 9 lists guide sequences.

ssDNA donor purification

Donor template sequence was produced as a gBlock (Integrated DNA Technologies, San Diego, CA, USA). A biotinylated forward primer and a standard reverse primer were used in PCR to generate a double-stranded, biotinylated DNA donor. The biotinylated DNA was captured on M-280 streptavidin Dynabeads (ThermoFisher), and the biotinylated strand was separated from the non-biotinylated strand essentially as described in the manufacturer's protocol. Supplemental file 10 provides a detailed protocol.

Transfection of Hi5 cells

sgRNAs were transcribed using T7 RNA polymerase, gel purified, then incubated with Cas9 in serum-free Hi5 culture medium supplemented with 18 mM L-glutamine. The resulting sgRNA/Cas9 RNPs were incubated with Trans-IT insect reagent (Mirus Bio, Madison, WI, USA) for 15 min at room temperature, then evenly distributed onto 90% confluent Hi5 cells. Culture medium was replaced with fresh medium 12 h later. Genomic DNA was isolated and analyzed by PCR 48 h later.

PCR to validate genomic editing in transfected cells

Forty eight hours after transfection, Hi5 cells from one 90% confluent well of a six-well plate (Corning, Corning, NY, USA) were collected, washed once with PBS (ThermoFisher) and lysed in 2× PK buffer containing 200 µg/ml proteinase K, extracted with phenol:chloroform:isoamyl alcohol (25:24:1), and then genomic DNA collected by

ethanol precipitation. Deletions in *TnPiwi* were detected by PCR using primers flanking the deleted region (Supplementary file 9). To confirm deletions by sequencing PCR products were resolved by agarose gel electrophoresis, purified (QIAquick Gel Extraction Kit, QIAGEN, Germantown, MD, USA), and cloned into pCR-Blunt II-Topo vector (ThermoFisher). The recombinant plasmid was transformed into Top10 competent *E.coli* (ThermoFisher) following supplier's protocol. PCR products amplified using M13 (−20) forward and M13 reverse primers from a sample of a single bacterial colony were sequenced by GENEWIZ (South Plainfield, NJ, USA).

Single clone selection

Wild-type Hi5 cells were seeded into a 96-well Transwell permeable support receiver plate (Corning, Corning, NY, USA) at 30% confluence and incubated overnight in serum free medium with 100 U/ml penicillin and 100 µg/ml streptomycin. A Transwell permeable support insert plate with media in each well was inserted into the receiver plate, and a single EGFP-positive cell was sorted into each insert well by FACS. After 14 days incubation at 27℃, wells were examined for EGFP-positive cell clones using a DMi8 fluorescent microscope (Leica).

Immunostaining

EGFP-HA-Vasa-expressing Hi5 cells were seeded on 22 × 22 mm cover slips (Fisher Scientific, Pittsburgh, PA, USA) in a well of a six-well plate (Corning). After cells had attached to the coverslip, the medium was removed and cells were washed three times with PBS (Gibco). Cells were fixed in 4% (w/v) methanol-free formaldehyde (ThermoFisher) in PBS at room temperature for 15 min, washed three times with PBS, permeabilized with 0.1% (w/v) Triton X-100 in PBS for 15 min at room temperature, and then washed three times with PBS. For antibody labeling, cells were incubated in 0.4% (v/v) Photo-Flo in 1× PBS for 10 min at room temperature, then 10 min in 0.1% (w/v) Triton X-100 in PBS and 10 min in 1× ADB-PBS (3 mg/ml bovine serum albumen, 1%

(v/v) donkey serum, 0.005% (w/v) Triton X-100 in 1× PBS). Next, cells were incubated with primary antibodies (mouse anti-GFP antibody (GFP-1D2, Developmental Studies Hybridoma Bank, Iowa City, IA, USA) and rabbit anti-HA Tag antibody (C29F4, Cell Signaling, Danvers, MA, USA), diluted 1:200 in ADB (30 mg/ml BSA, 10% (v/v) donkey serum, 0.05% (w/v) Triton X-100 in 1× PBS) at 4℃ overnight. After three washes in PBS, cells were incubated sequentially in 0.4% (v/v) Photo-Flo in 1× PBS, 0.1% (w/v) Triton X-100 in PBS, and 1× ADB-PBS, each for 10 min at room temperature. Cells were then incubated with secondary Alexa Fluor 488-labeled donkey anti-mouse (ThermoFisher) and Alexa Fluor 680-labeled donkey anti-rabbit (ThermoFisher) antibodies, diluted 1:500 in ADB at room temperature for one hour. After washing three times with 0.4% (v/v) Photo-Flo in 1× PBS and once with 0.4% (v/v) Photo-Flo in water, coverslips were air dried in the dark at room temperature. Slides were mounted in ProLong Gold Antifade Mountant with DAPI and examined by confocal microscopy (TCS SP5 II Laser Scanning Confocal, Leica).

Data deposition

The *T. ni* Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession NKQN00000000. The version described here is version NKQN01000000. All sequencing data are available through the NCBI Sequence Read Archive under the accession number PRJNA336361. Further details are available at the Cabbage Looper Database (http://cabbagelooper.org/).

A

7.1 million pairs of reads
28,760.2 Mb, 78.1× coverage
46.4× coverage for reads >5 kb

De novo genome assembly using long reads (Canu)

7.1 million pairs of reads
28,760.2 Mb, 78.1× coverage
46.4× coverage for reads >5 kb

Assembly polishing using long reads (Quiver)

317.8 million pairs of reads
63,550.1 Mb, 172.6× coverage

Assembly Polishing using short reads (Pilon)

316.5 million pairs of reads
63,292.4 Mb, 172.0× coverage

Scaffolding using mate pair reads (SOAPdenovo)

489.2 million pairs of reads
68,619.1 Mb, 186.4× coverage

Scaffolding using Hi-C reads (LACHESIS)

rRNA prediction (RNAmmer and barrnap)

Genome

Assemble transcriptome de novo using RNA-seq reads (Trinity)

Find protein sequences (UniProtKB/Swiss-Prot)

Build repeat consensus sequences (RepeatModeler)

Transcripts from other species (Ensembl)

Genome annotation pipeline (MAKER)

Mask repeats (RepeatMasker)

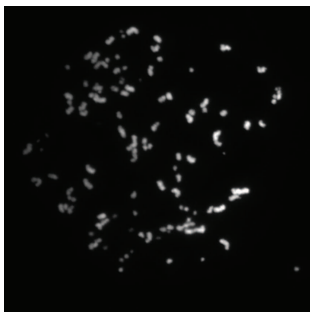
Predict genes (AUGUSTUS and SNAP)

Map repeat sequences to genome (tblastx, blastn)

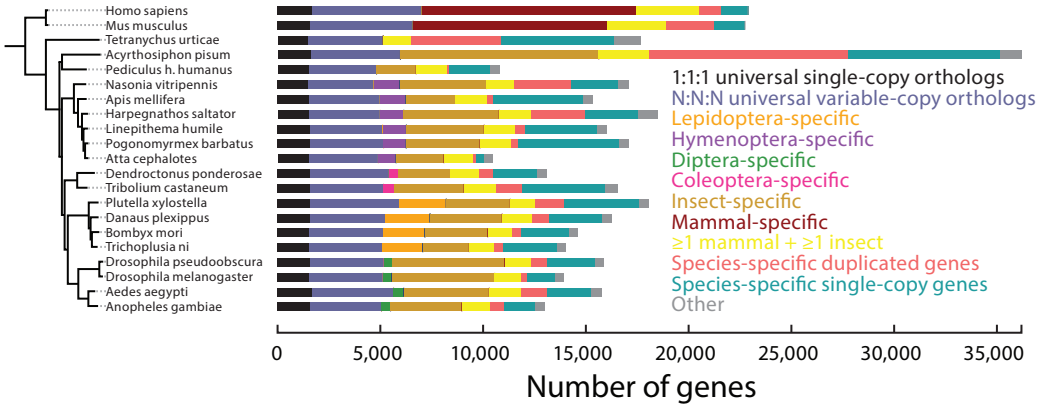
Identify tRNAs (tRNAscan-SE)

Annotation

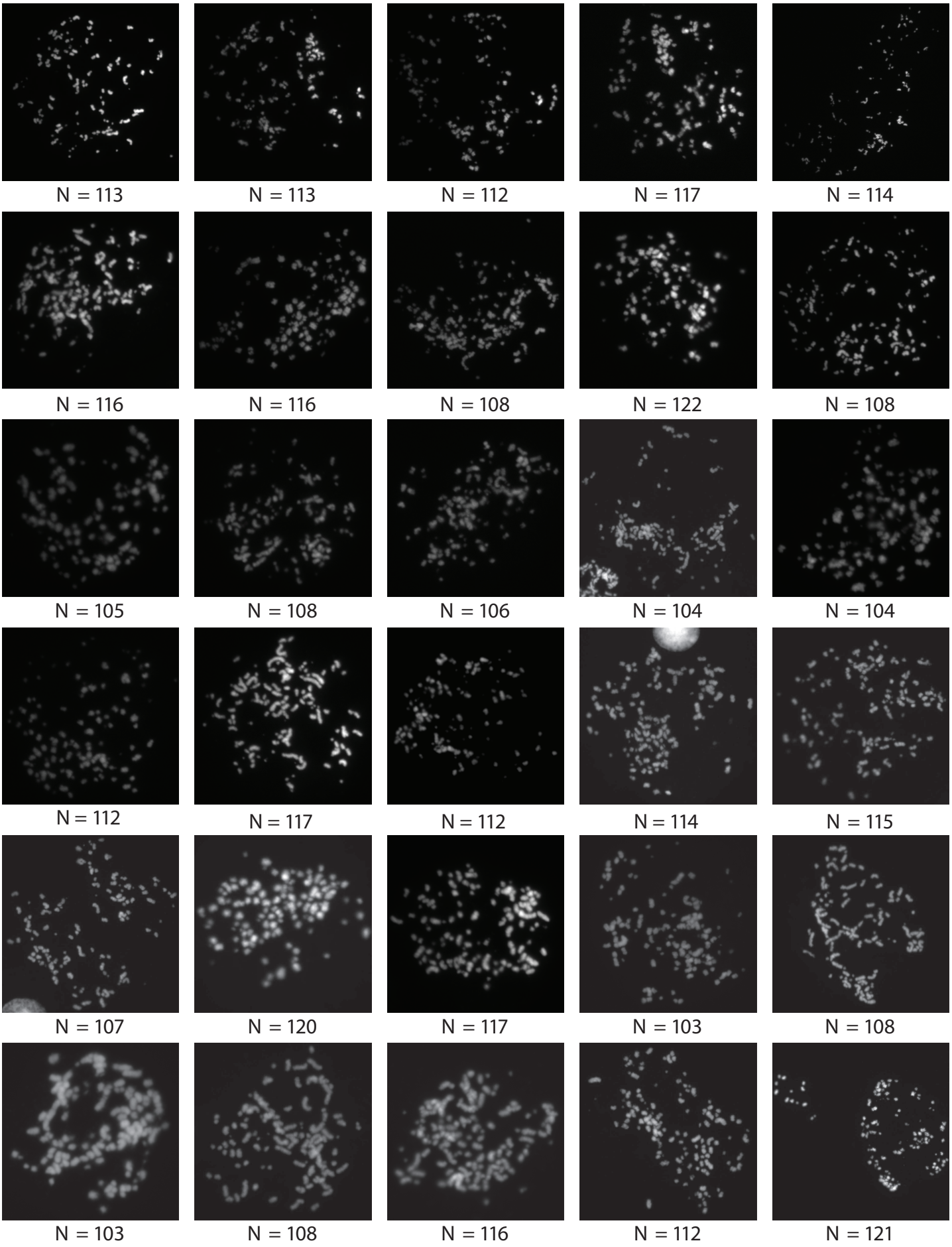
B

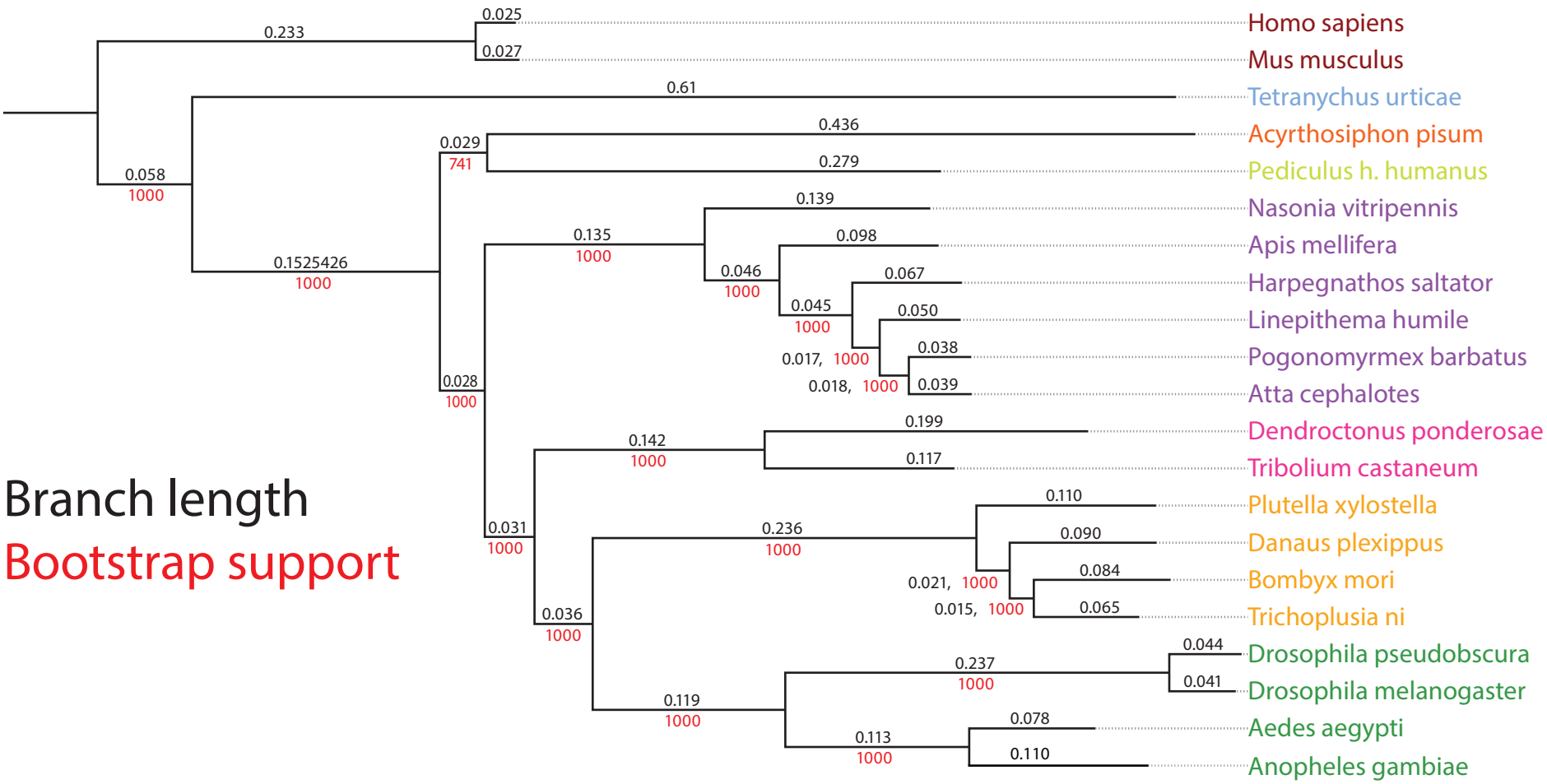


C

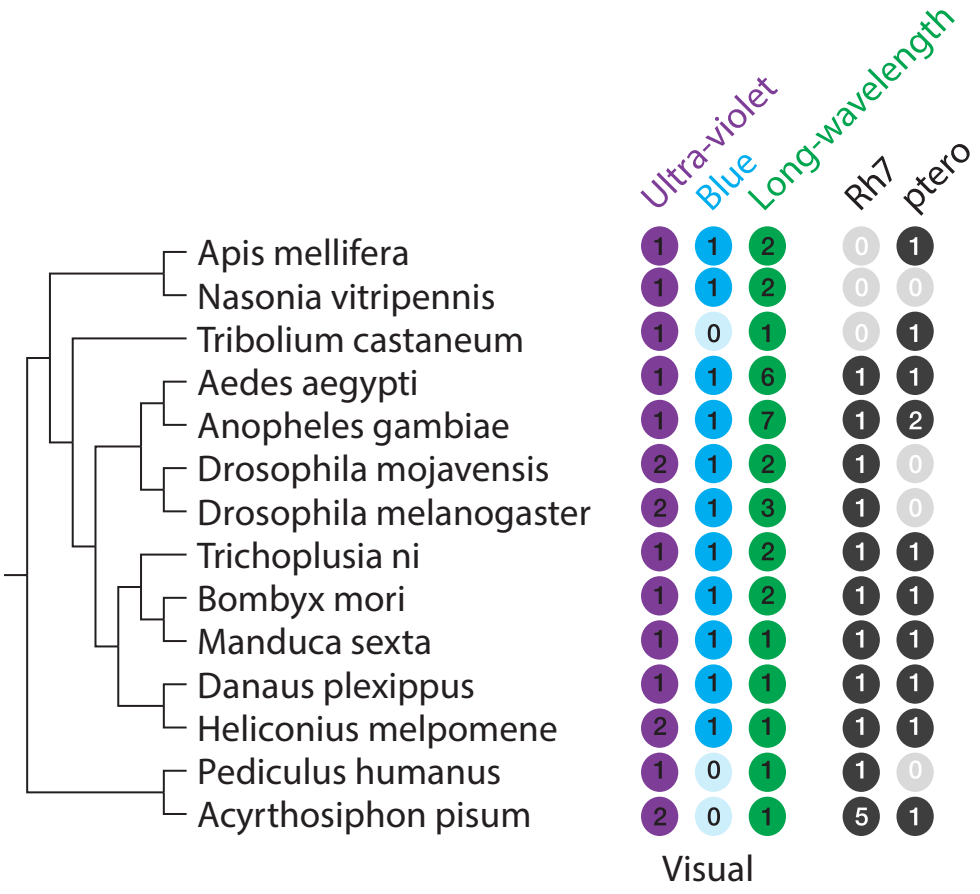


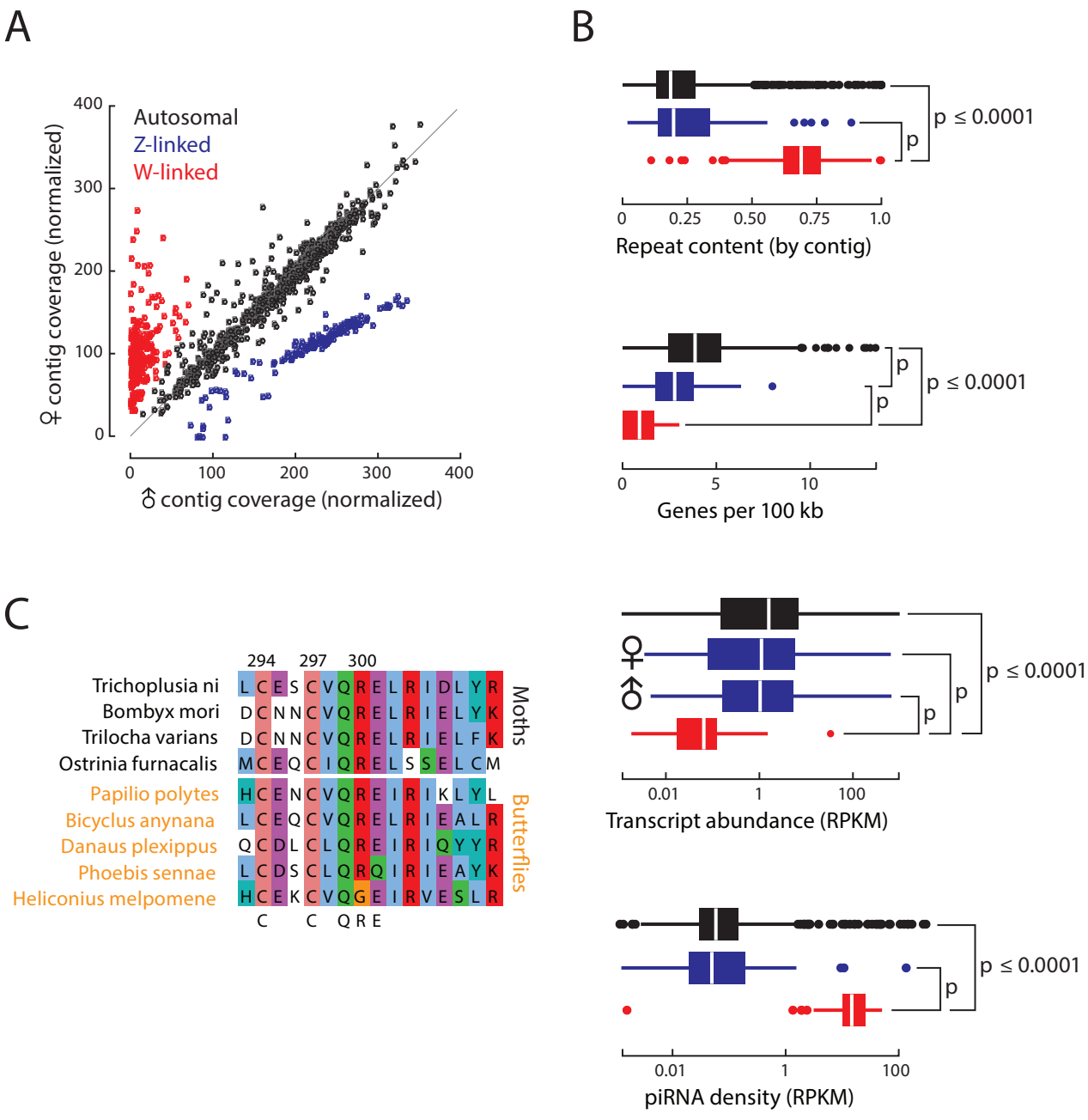
Fu et al., Figure 1—figure supplement 1

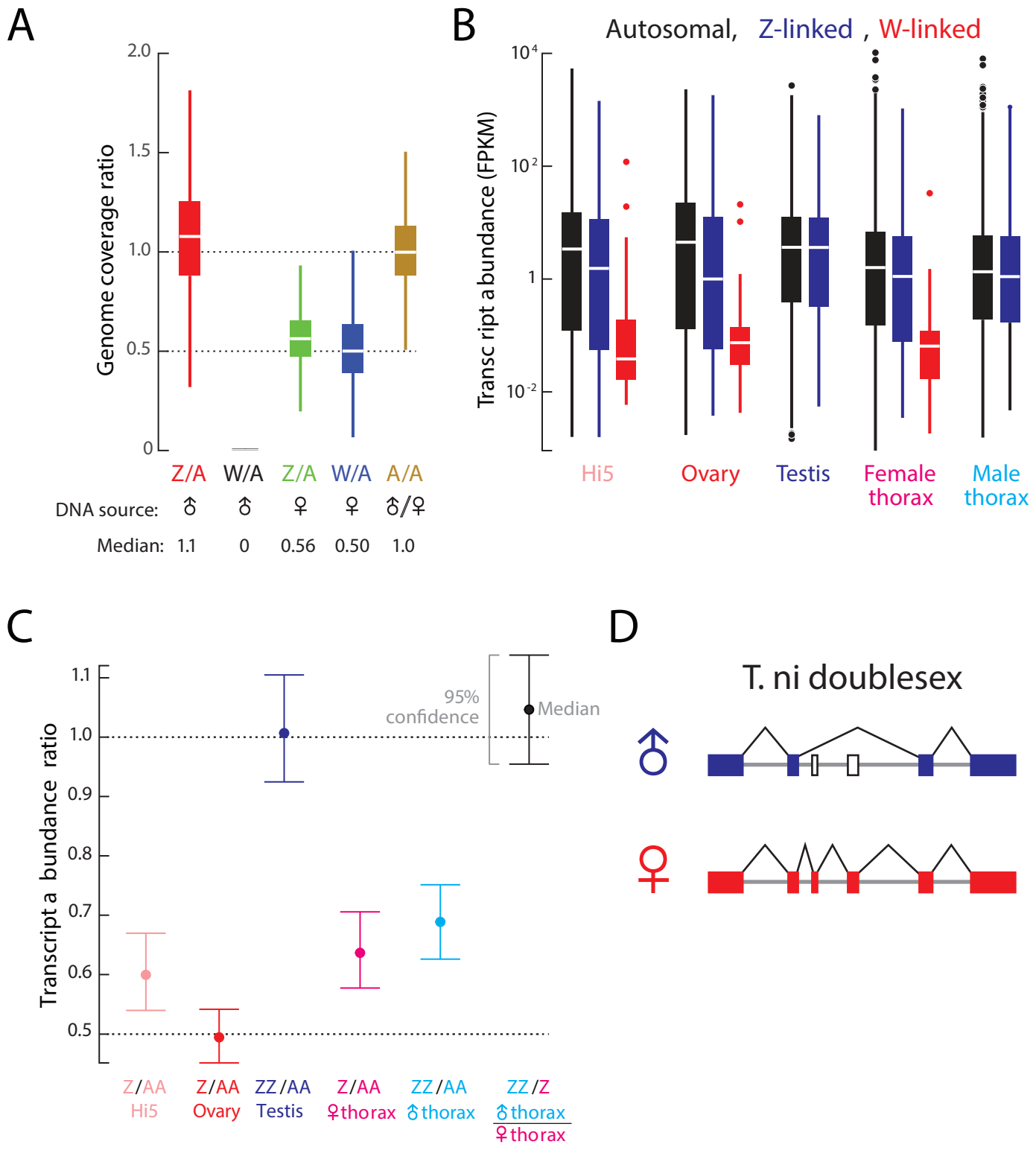



0.233
0.025
0.027
0.61
0.436
0.029
741
0.279
0.058
1000
0.139
0.135
1000
0.098
0.1525426
1000
0.046
1000
0.067
0.045
1000
0.050
0.017, 1000
0.038
0.028
1000
0.018, 1000
0.039
0.199
0.142
1000
0.117
0.110
0.031
1000
0.236
1000
0.090
0.021, 1000
0.084
0.036
1000
0.015, 1000
0.065
0.044
0.237
1000
0.041
0.119
1000
0.078
0.113
1000
0.110
Homo sapiens
Mus musculus
Tetranychus urticae
Acyrthosiphon pisum
Pediculus h. humanus
Nasonia vitripennis
Apis mellifera
Harpegnathos saltator
Linepithema humile
Pogonomyrmex barbatus
Atta cephalotes
Dendroctonus ponderosae
Tribolium castaneum
Plutella xylostella
Danaus plexippus
Bombyx mori
Trichoplusia ni
Drosophila pseudobscura
Drosophila melanogaster
Aedes aegypti
Anopheles gambiae
Branch length
Bootstrap support
Insects: Diptera , Lepidoptera , Coleoptera , Hymenoptera , Phthiraptera , Homoptera
Arachnida
Mammals

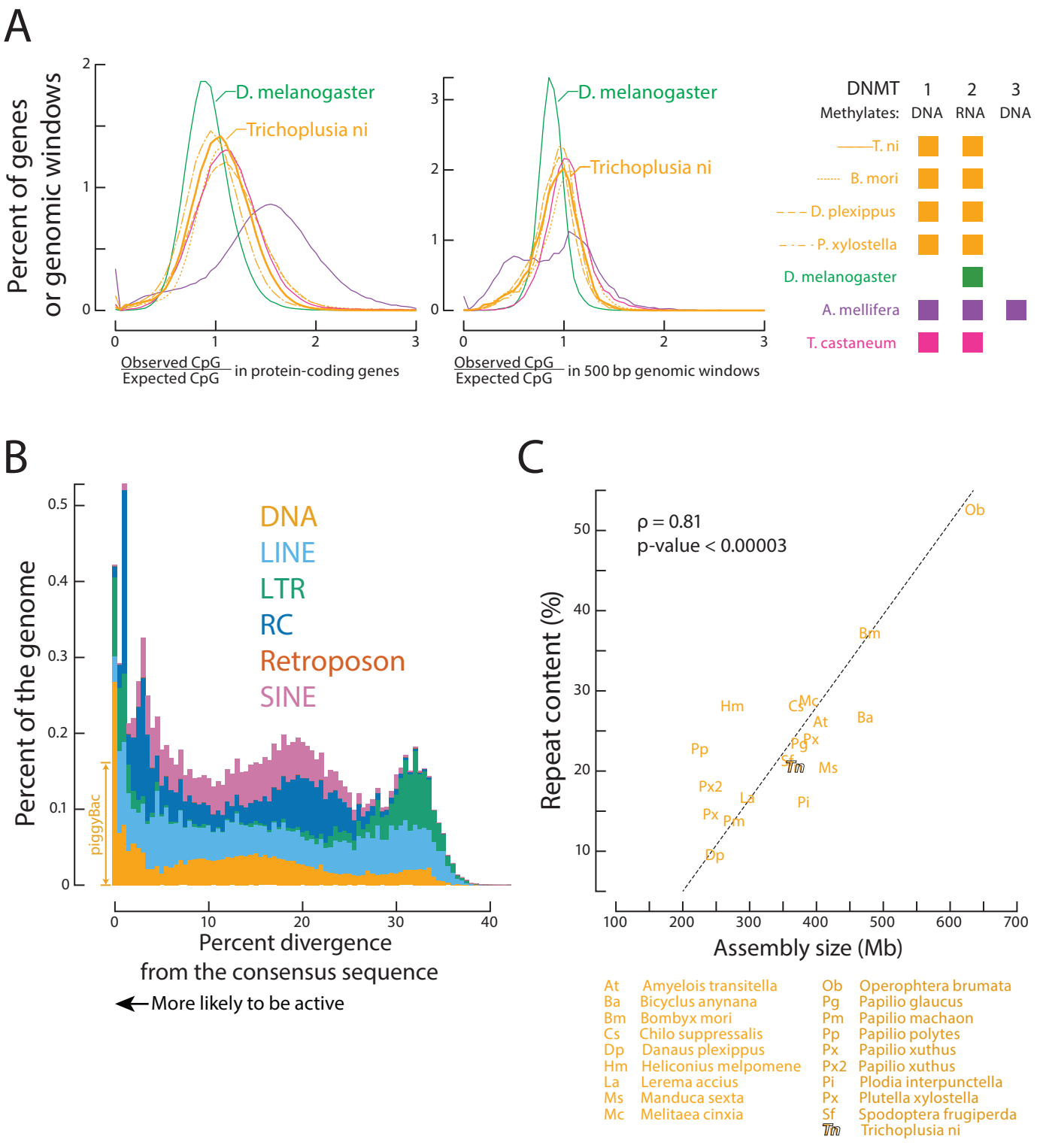
Fu et al., Figure 1—figure supplement 3

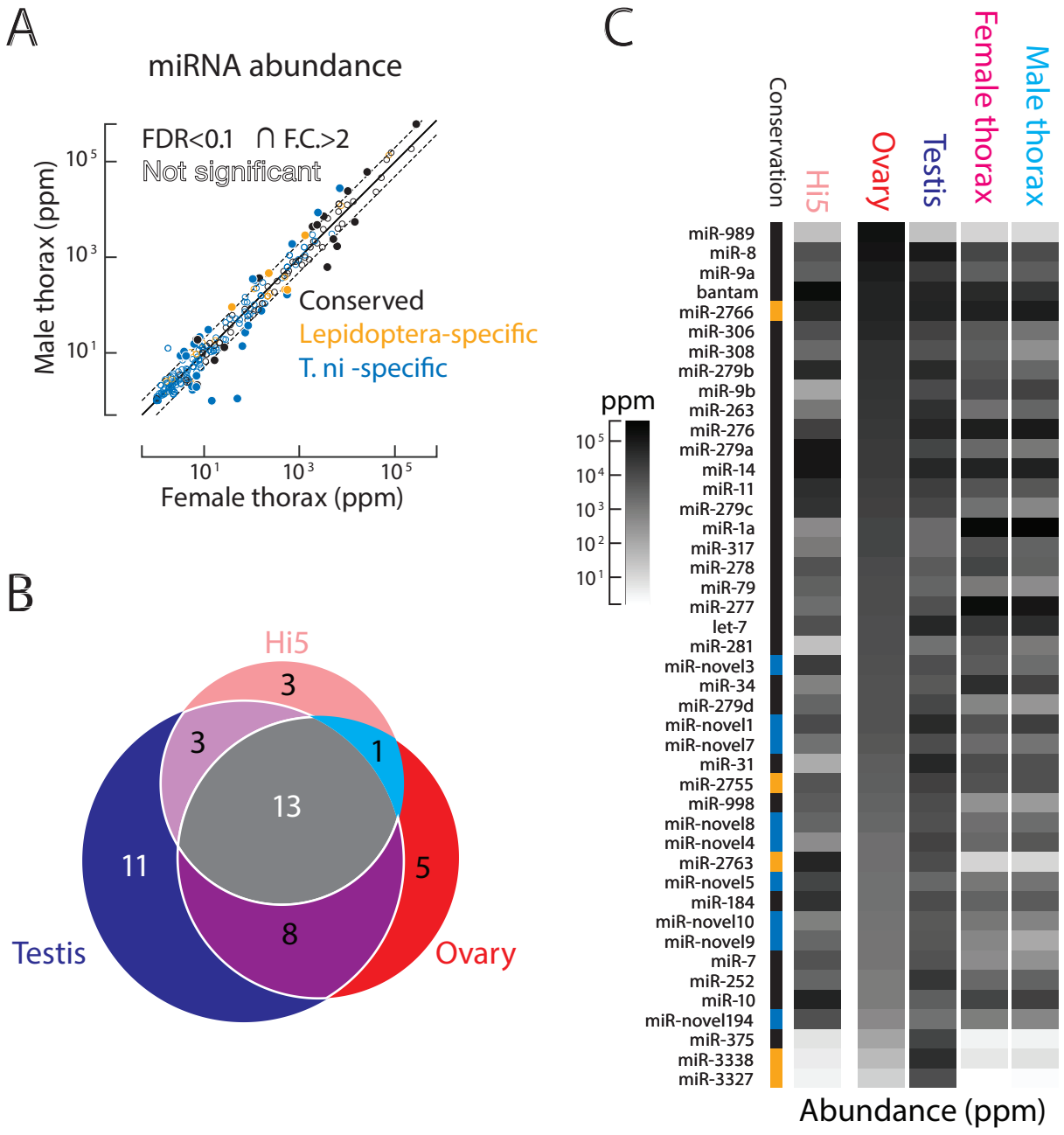


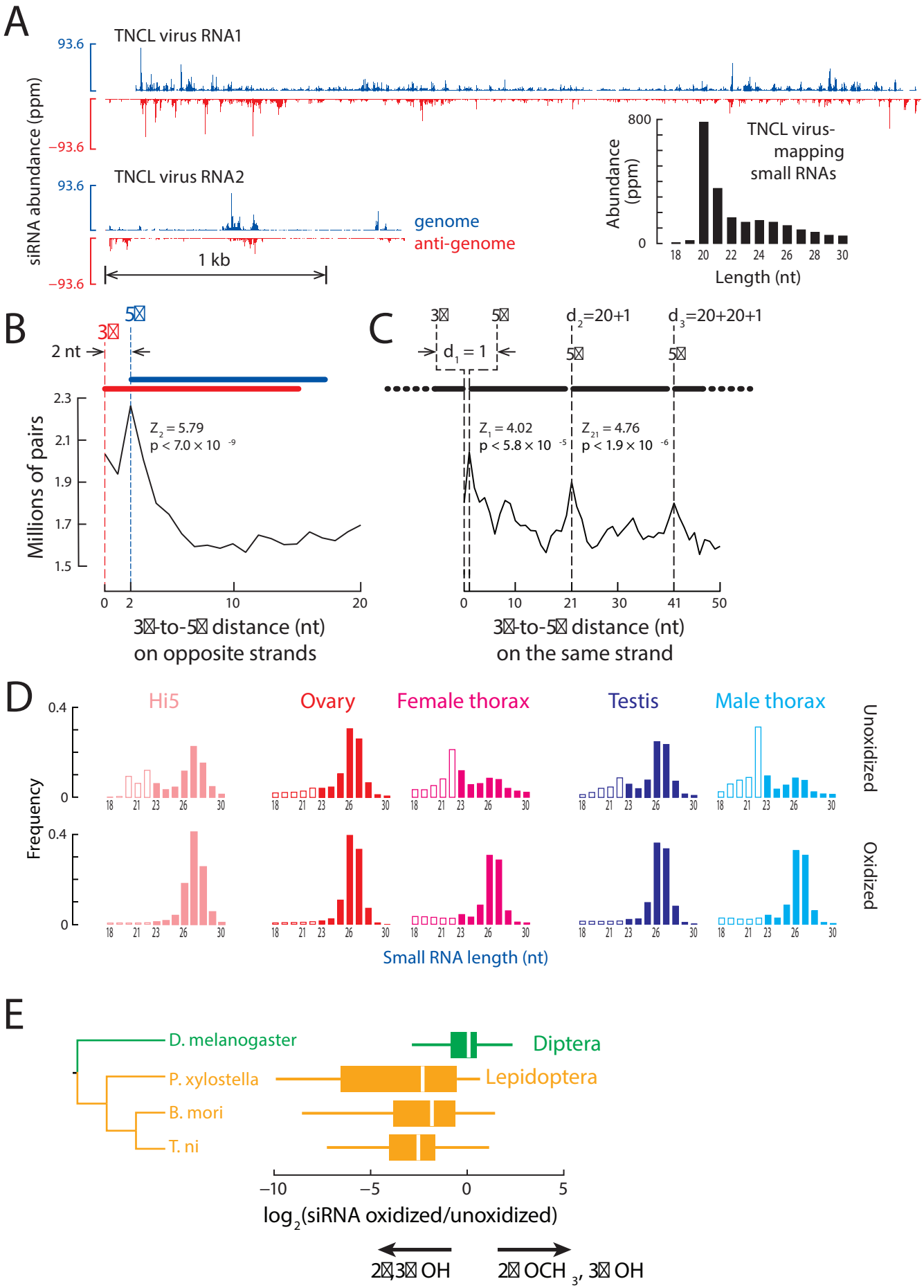

A
400
300
200
100
0
♀ contig coverage (normalized)
Autosomal
Z-linked
W-linked
0
100
200
300
400
♂ contig coverage (normalized)
B
p ≤ 0.0001
p
0
0.25
0.50
0.75
1.0
Repeat content (by contig)
p
p
p ≤ 0.0001
0
5
10
Genes per 100 kb
C
294
297
300
Trichoplusia ni
Bombyx mori
Trilocha varians
Ostrinia furnacalis
Papilio polytes
Bicyclus anynana
Danaus plexippus
Phoebis sennae
Heliconius melpomene
Moths
Butterflies
C C Q R E
p ≤ 0.0001
p
p
0.01
1
100
Transcript abundance (RPKM)
p ≤ 0.0001
p
0.01
1
100
piRNA density (RPKM)

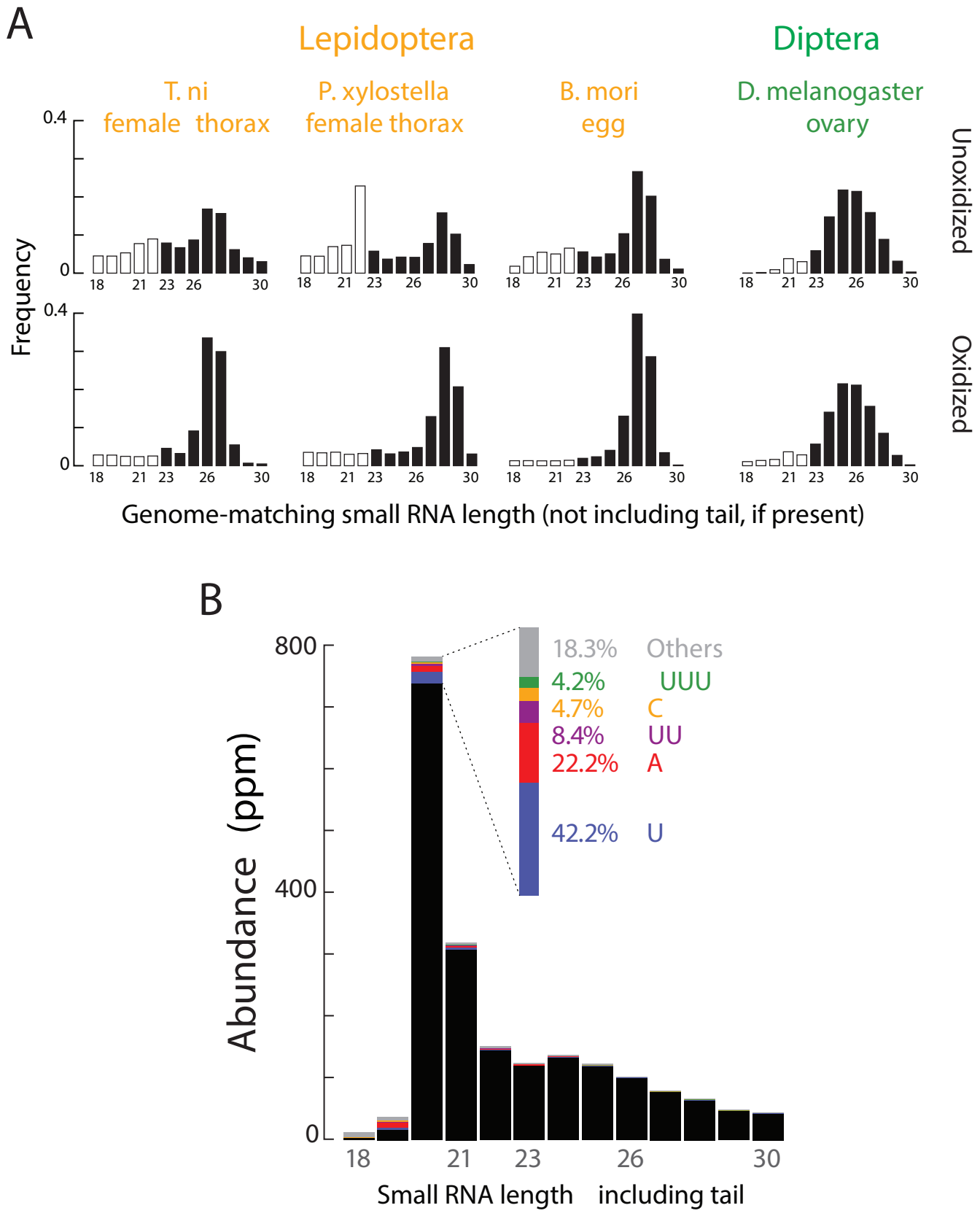

A
Genome coverage ratio
2.0
1.5
1.0
0.5
0
Z/A
W/A
Z/A
W/A
A/A
DNA source:
♂
♂
♀
♀
♂/♀
Median:
1.1
0
0.56
0.50
1.0
B
Autosomal, Z-linked , W-linked
Transc ript a bundance (FPKM)
10⁴
10²
1
10⁻²
Hi5
Ovary
Testis
Female thorax
Male thorax
C
Transcript a bundance ratio
1.1
1.0
0.9
0.8
0.7
0.6
0.5
95% confidence
Median
Z/AA Hi5
Z/AA Ovary
ZZ/AA Testis
Z/AA ♀thorax
ZZ/AA ♂thorax
ZZ/Z ♂thorax/♀thorax
D
T. ni doublesex
♂
♀

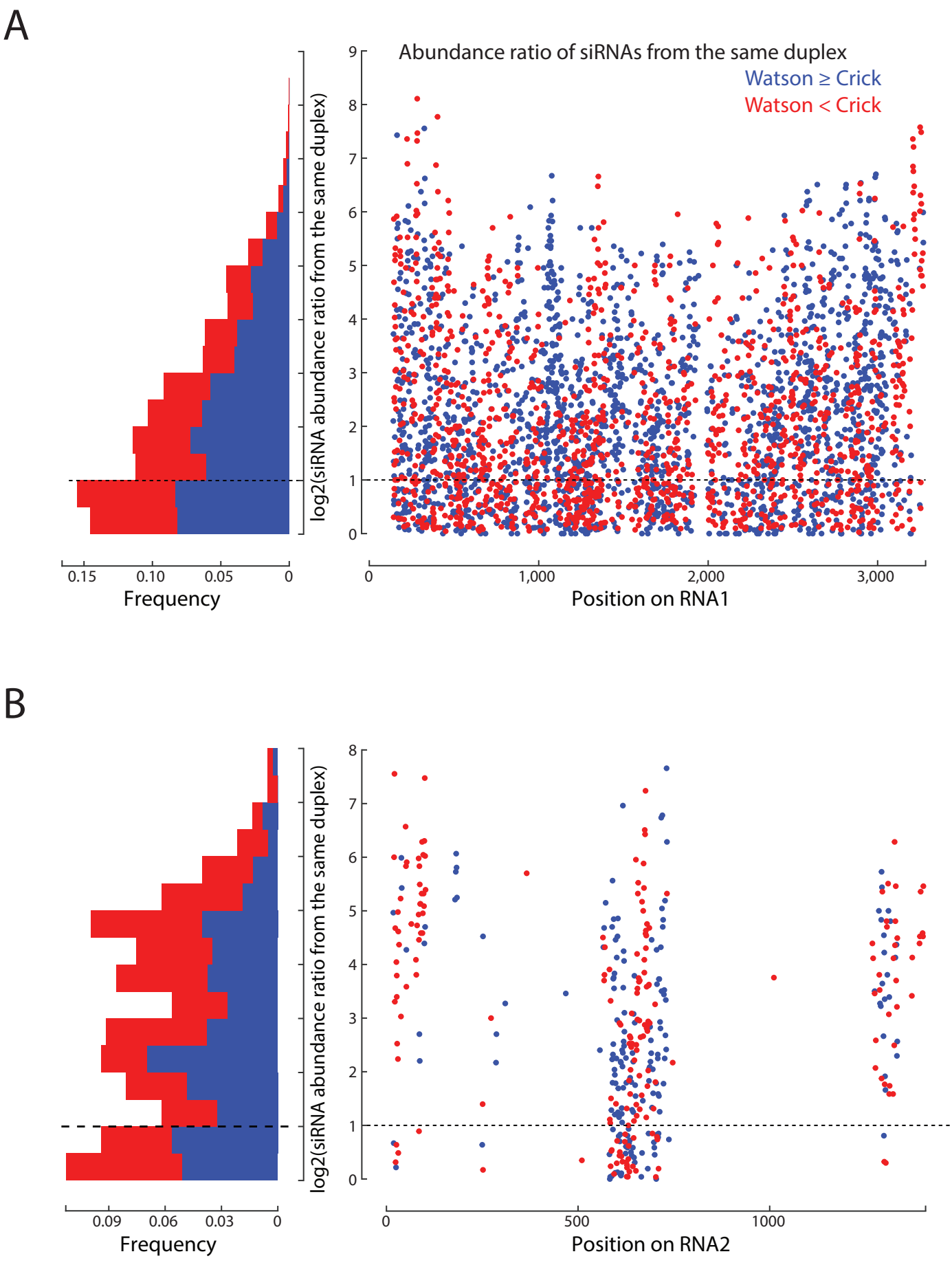
Fu et al., Figure 2—figure supplement 2

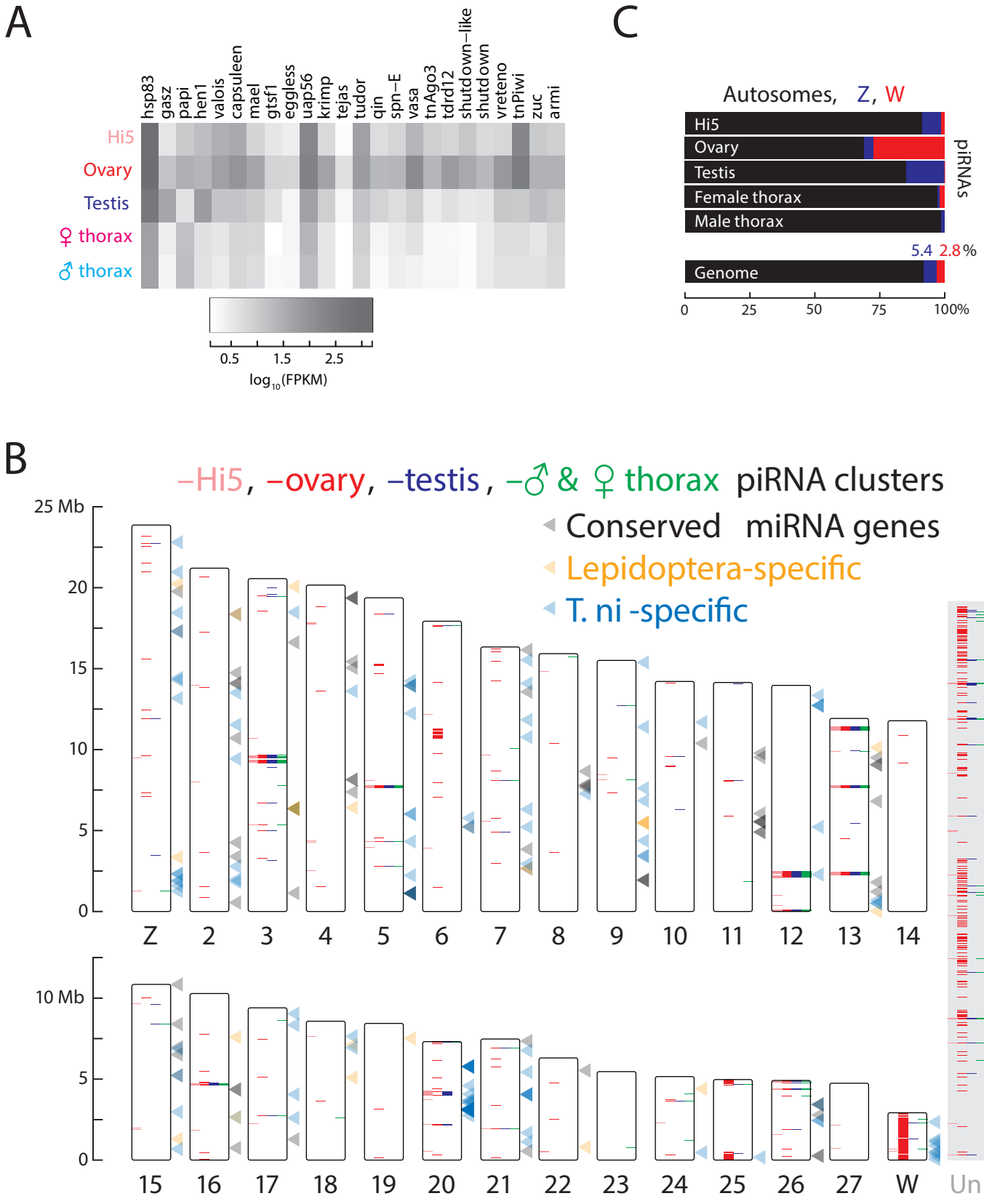



A
miRNA abundance
FDR<0.1 ∩ F.C.>2
Not significant
Conserved
Lepidoptera-specific
T. ni -specific
Male thorax (ppm)
Female thorax (ppm)
B
Hi5
Testis
Ovary
3
3
1
13
11
5
8
C
Conservation
Hi5
Ovary
Testis
Female thorax
Male thorax
ppm
miR-989
miR-8
miR-9a
bantam
miR-2766
miR-306
miR-308
miR-279b
miR-9b
miR-263
miR-276
miR-279a
miR-14
miR-11
miR-279c
miR-1a
miR-317
miR-278
miR-79
miR-277
let-7
miR-281
miR-novel3
miR-34
miR-279d
miR-novel1
miR-novel7
miR-31
miR-2755
miR-998
miR-novel8
miR-novel4
miR-2763
miR-novel5
miR-184
miR-novel10
miR-novel9
miR-7
miR-252
miR-10
miR-novel194
miR-375
miR-3338
miR-3327
Abundance (ppm)

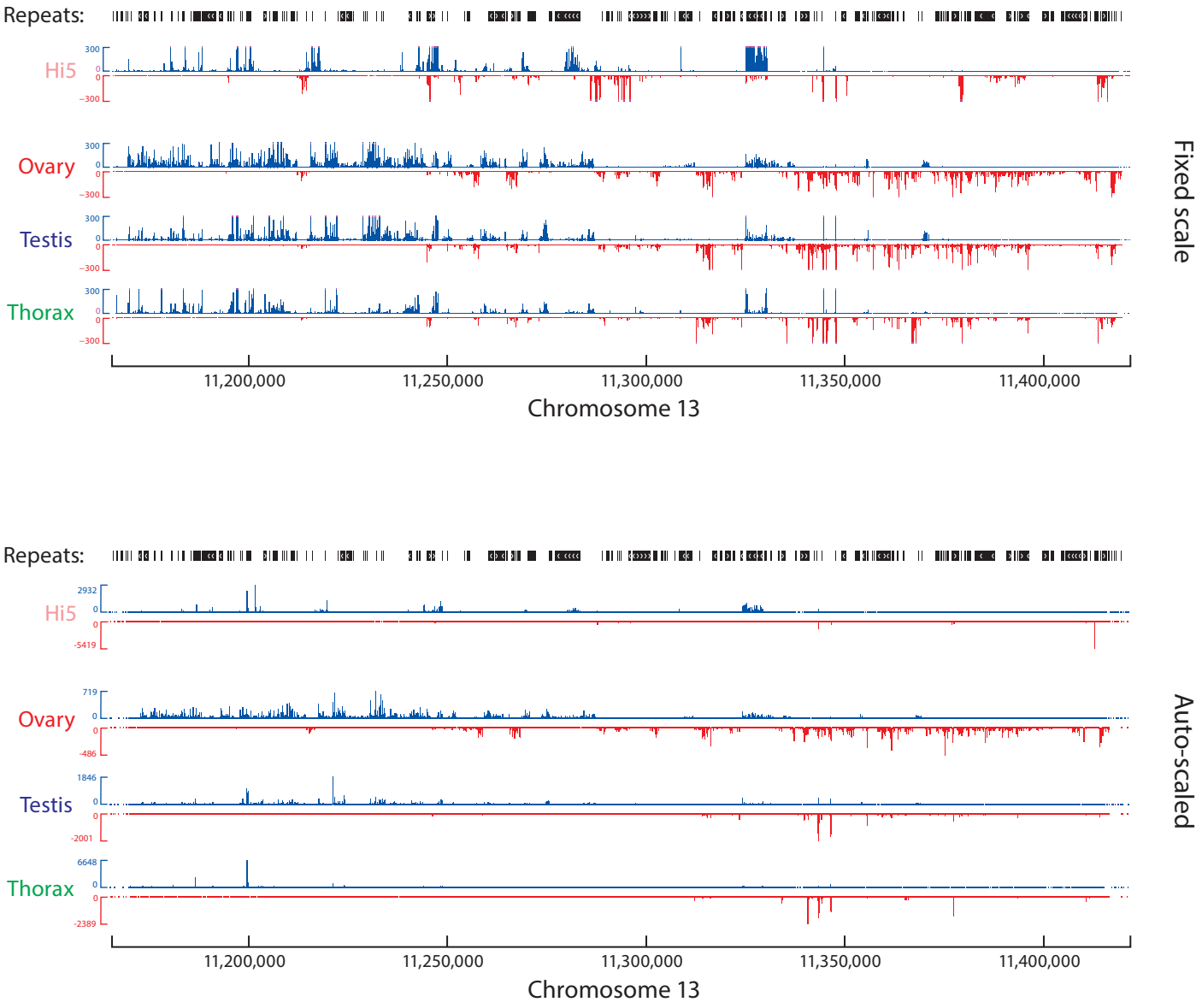

A
TNCL virus RNA1
93.6
−93.6
siRNA abundance (ppm)
TNCL virus RNA2
93.6
−93.6
genome
anti-genome
1 kb
800
Abundance (ppm)
0
TNCL virus-mapping small RNAs
18 20 22 24 26 28 30
Length (nt)
B
5′
3′
2 nt
2.3
2.1
1.9
1.7
1.5
Millions of pairs
Z2 = 5.79
p < 7.0 × 10 -9
0 2 10 20
3′-to-5′ distance (nt)
on opposite strands
C
3′
5′
d1 = 1
d2=20+1
d3=20+20+1
5′
5′
Z1 = 4.02
p < 5.8 × 10 -5
Z21 = 4.76
p < 1.9 × 10 -6
0 10 21 30 41 50
3′-to-5′ distance (nt)
on the same strand
D
Hi5
Ovary
Female thorax
Testis
Male thorax
Unoxidized
Oxidized
0.4
0
Frequency
18 21 23 26 30
Small RNA length (nt)
E
D. melanogaster
P. xylostella
B. mori
T. ni
Diptera
Lepidoptera
−10 −5 0 5
log2(siRNA oxidized/unoxidized)
2′,3′ OH
2′ OCH3, 3′ OH

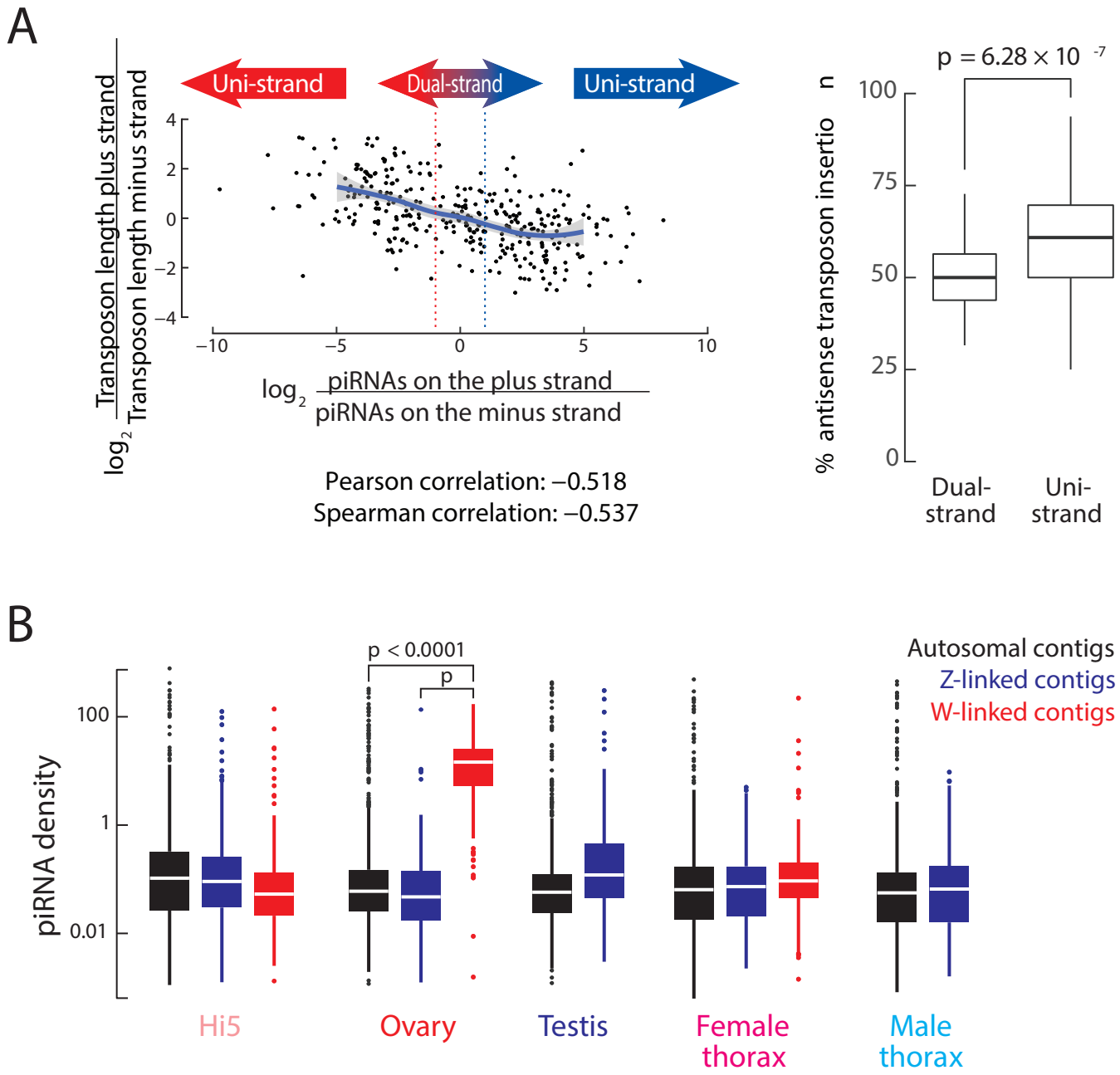

A
Lepidoptera
Diptera
T. ni
female thorax
P. xylostella
female thorax
B. mori
egg
D. melanogaster
ovary
Unoxidized
Oxidized
Frequency
0.4
0
18 21 23 26 30
Genome-matching small RNA length (not including tail, if present)
B
800
400
0
Abundance (ppm)
18.3% Others
4.2% UUU
4.7% C
8.4% UU
22.2% A
42.2% U
18 21 23 26 30
Small RNA length including tail

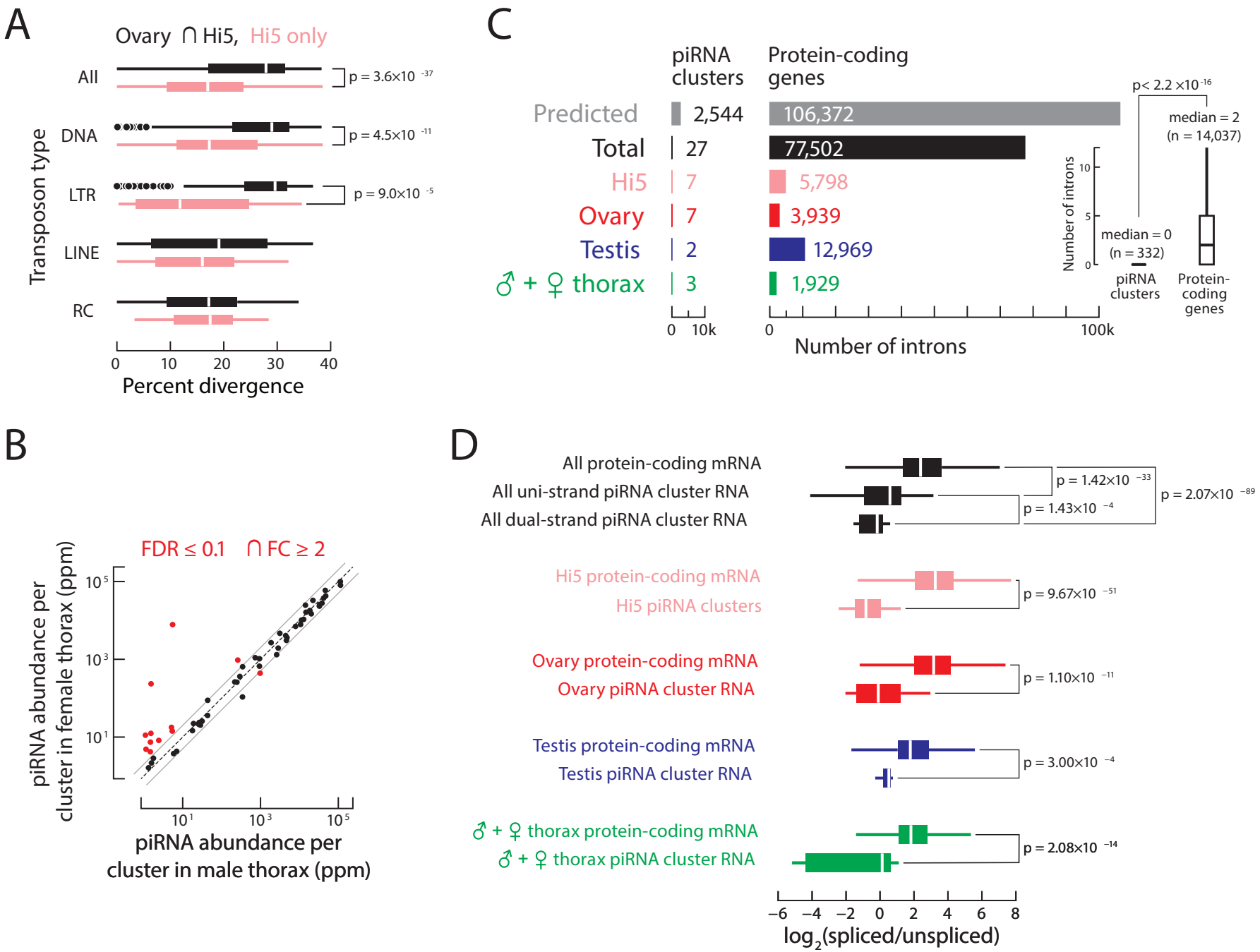

A
Abundance ratio of siRNAs from the same duplex
Watson ≥ Crick
Watson < Crick
log2(siRNA abundance ratio from the same duplex)
0
1
2
3
4
5
6
7
8
9
0.15
0.10
0.05
0
Frequency
0
1,000
2,000
3,000
Position on RNA1
B
log2(siRNA abundance ratio from the same duplex)
0
1
2
3
4
5
6
7
8
0.09
0.06
0.03
0
Frequency
0
500
1000
Position on RNA2

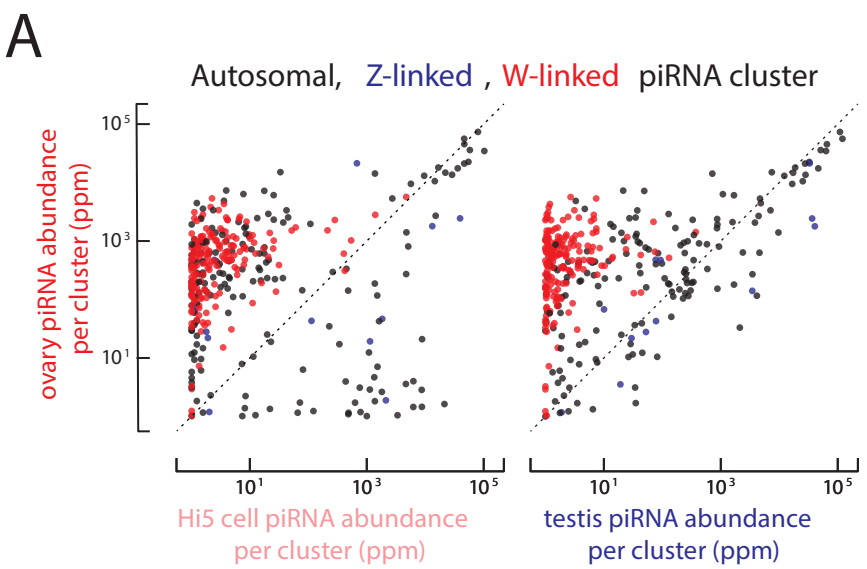

A
hsp83
gasz
papi
hen1
valois
capsuleen
mael
gtsf1
eggless
uap56
krimp
tejas
tudor
qin
spn–E
vasa
tnAgo3
tdrd12
shutdown–like
shutdown
vreteno
tnPiwi
zuc
armi
Hi5
Ovary
Testis
♀ thorax
♂ thorax
0.5
1.5
2.5
log10(FPKM)
C
Autosomes, Z, W
Hi5
Ovary
Testis
Female thorax
Male thorax
piRNAs
5.4 2.8%
Genome
0
25
50
75
100%
B
–Hi5, –ovary, –testis, –♂ & ♀ thorax piRNA clusters
Conserved miRNA genes
Lepidoptera-specific
T. ni -specific
25 Mb
20
15
10
5
0
Z
2
3
4
5
6
7
8
9
10
11
12
13
14
10 Mb
5
0
15
16
17
18
19
20
21
22
23
24
25
26
27
W
Un

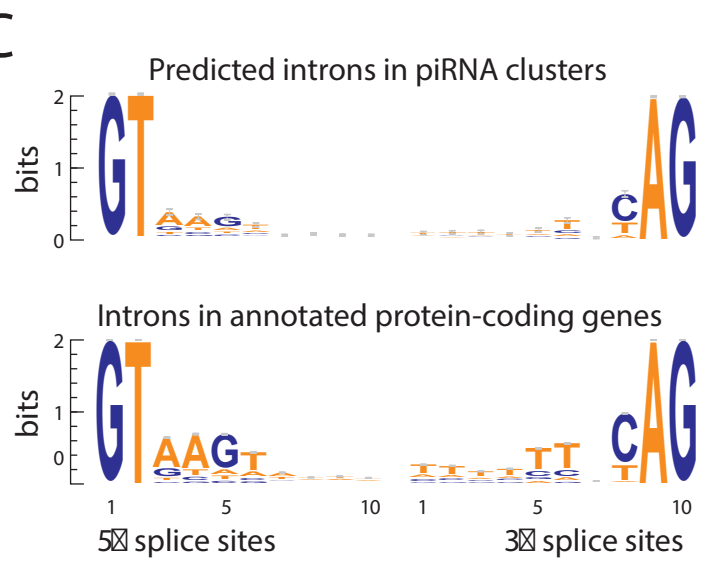
Fu et al., Figure 5—figure supplement 1

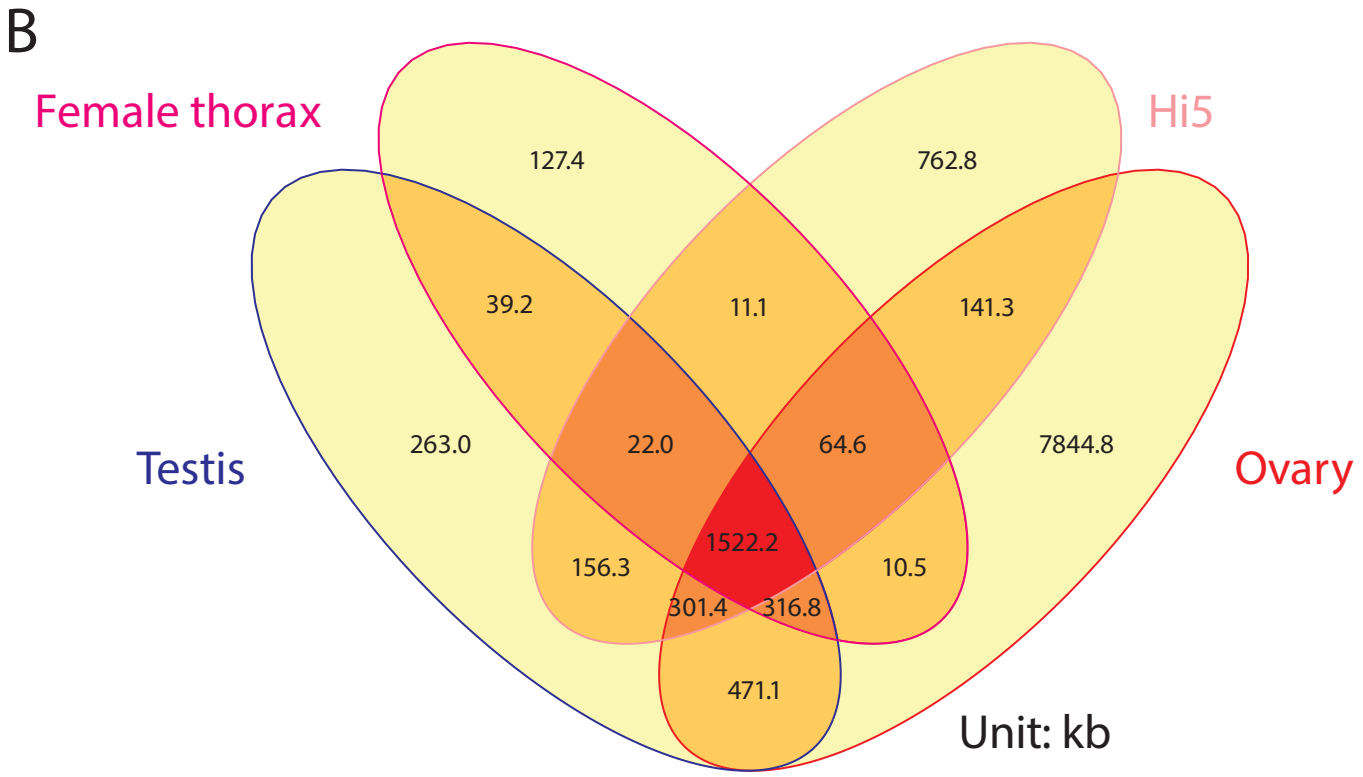


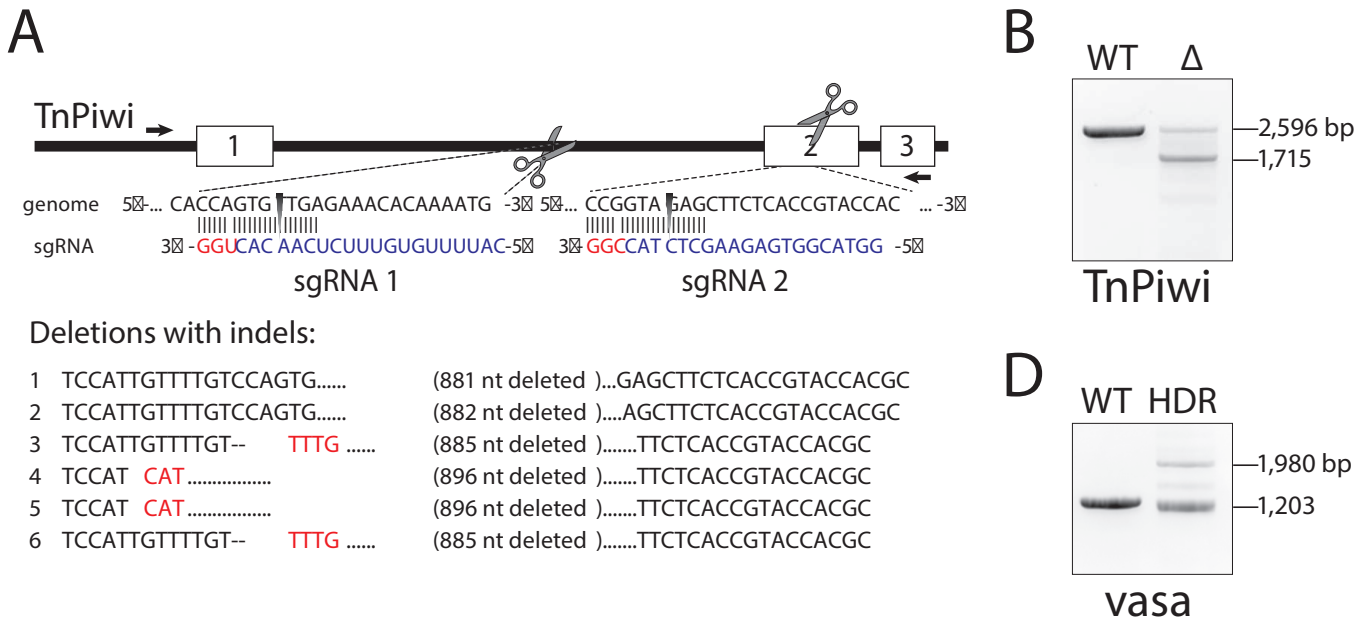

A
Uni-strand
Dual-strand
Uni-strand
Transposon length plus strand
Transposon length minus strand
log2
4
2
0
−2
−4
−10
−5
0
5
10
log2 piRNAs on the plus strand / piRNAs on the minus strand
Pearson correlation: −0.518
Spearman correlation: −0.537
p = 6.28 × 10 -7
% antisense transposon insertio n
100
75
50
25
0
Dual-strand
Uni-strand
B
p < 0.0001
p
Autosomal contigs
Z-linked contigs
W-linked contigs
piRNA density
100
1
0.01
Hi5
Ovary
Testis
Female thorax
Male thorax
C
piRNA abundance (ppm)
0
50
100
150
200
250
Transposon homology (n = 8)
Uncharacterized or hypothetical (n = 21)
Potential protein-coding (n = 12)
No homologs (n = 33)
p=2.3 × 10 -3
p = 3.2 × 10 -5
p = 0.0018

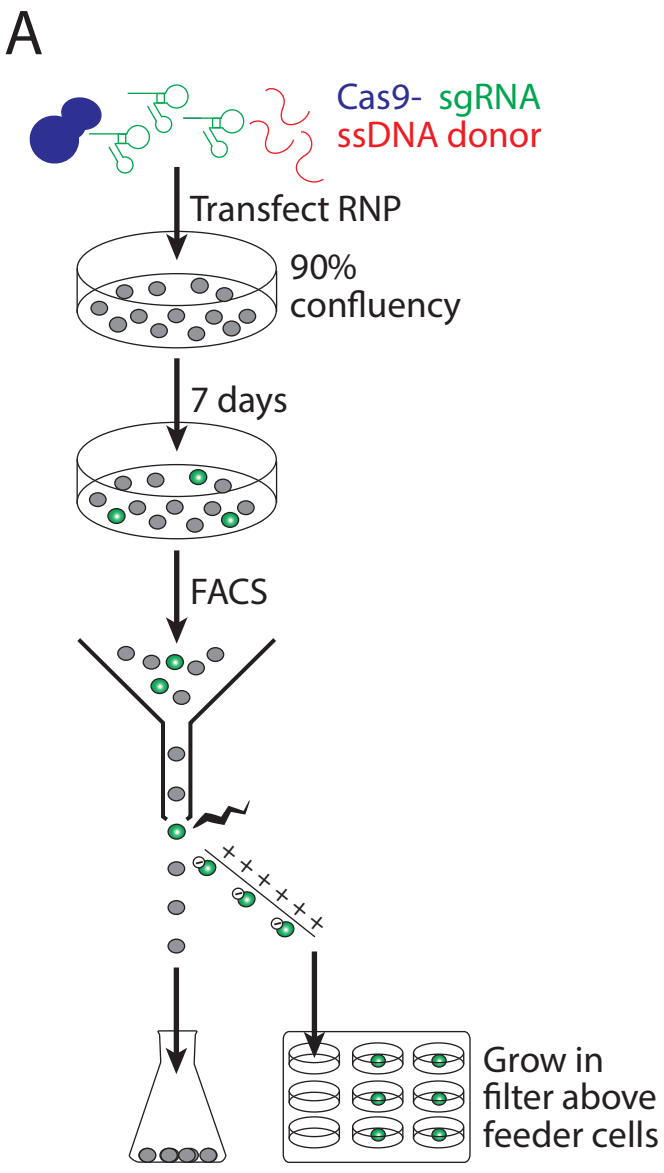

A
Ovary ∩ Hi5, Hi5 only
All
DNA
LTR
LINE
RC
Transposon type
p = 3.6×10 -37
p = 4.5×10 -11
p = 9.0×10 -5
0 10 20 30 40
Percent divergence
B
FDR ≤ 0.1 ∩ FC ≥ 2
piRNA abundance per cluster in female thorax (ppm)
10 5
10 3
10 1
10 1 10 3 10 5
piRNA abundance per cluster in male thorax (ppm)
C
piRNA clusters
Protein-coding genes
Predicted 2,544 106,372
Total 27 77,502
Hi5 7 5,798
Ovary 7 3,939
Testis 2 12,969
♂ + ♀ thorax 3 1,929
0 10k
0 100k
Number of introns
p< 2.2 ×10-16
median = 2 (n = 14,037)
median = 0 (n = 332)
Number of introns
10
5
0
piRNA clusters
Protein-coding genes
D
All protein-coding mRNA
All uni-strand piRNA cluster RNA
All dual-strand piRNA cluster RNA
p = 1.42×10 -33
p = 1.43×10 -4
p = 2.07×10 -89
Hi5 protein-coding mRNA
Hi5 piRNA clusters
p = 9.67×10 -51
Ovary protein-coding mRNA
Ovary piRNA cluster RNA
p = 1.10×10 -11
Testis protein-coding mRNA
Testis piRNA cluster RNA
p = 3.00×10 -4
♂ + ♀ thorax protein-coding mRNA
♂ + ♀ thorax piRNA cluster RNA
p = 2.08×10 -14
−6 −4 −2 0 2 4 6 8
log2(spliced/unspliced)

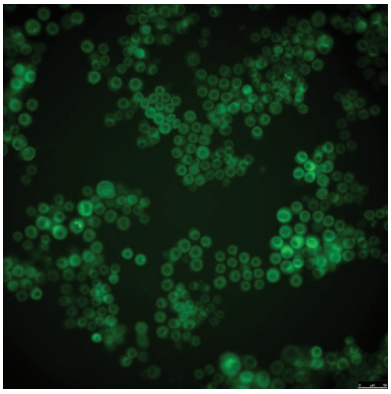

A
Autosomal, Z-linked , W-linked piRNA cluster
ovary piRNA abundance per cluster (ppm)
Hi5 cell piRNA abundance per cluster (ppm)
testis piRNA abundance per cluster (ppm)


C
Predicted introns in piRNA clusters
Introns in annotated protein-coding genes
bits
5′ splice sites
3′ splice sites

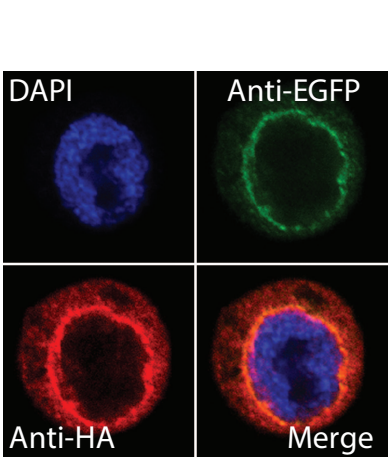

B
Female thorax
Hi5
Testis
Ovary
127.4
762.8
39.2
11.1
141.3
263.0
22.0
64.6
7844.8
1522.2
156.3
10.5
301.4
316.8
471.1
Unit: kb


A
TnPiwi
1
2
3
genome 5'... CACCAGTG TTGAGAAACACAAAATG -3' 5'... CCGGTA GAGCTTCTCACCGTACCAC ... -3'
sgRNA 3' -GGUCAC AACUCUUUGUGUUUUAC-5' 3' GGCCAT CTCGAAGAGTGGCATGG -5'
sgRNA 1
sgRNA 2
Deletions with indels:
1 TCCATTGTTTTGTCCAGTG..... (881 nt deleted)...GAGCTTCTCACCGTACCACGC
2 TCCATTGTTTTGTCCAGTG..... (882 nt deleted)....AGCTTCTCACCGTACCACGC
3 TCCATTGTTTTGT-- TTTG ..... (885 nt deleted).....TTCTCACCGTACCACGC
4 TCCAT CAT..... (896 nt deleted).....TTCTCACCGTACCACGC
5 TCCAT CAT..... (896 nt deleted).....TTCTCACCGTACCACGC
6 TCCATTGTTTTGT-- TTTG ..... (885 nt deleted).....TTCTCACCGTACCACGC
B
WT Δ
2,596 bp
1,715
TnPiwi
D
WT HDR
1,980 bp
1,203
vasa
C
start codon
sgRNA4
5'CCATAA ATCTATTACTACTGAGCATT ... ACATGGATGATGATTG GGATGATGGTGTCGGTG ATG TGG-3'
3'GGTATT TAGATAATGATGACTCG TAA... TGTACCTACTACTAACCCTACTACCACAGCCAC TACAC C-5'
sgRNA3
vasa
1
2
3
4
mutated PAM
ssDNA donor
homology arm
EGFP + HA insertion
homology arm
Homology-dependent repair:
1
EGFP + HA
2
3
4





Clonal EGFP-HA-Vasa-expressing Hi5 cells (20×)



EGFP-HA-Vasa-expressing Hi5 cell (63×)