Recurrent Amplification of the Heterochromatin Protein 1 (HP1) Gene Family across Diptera

Quentin Helleu, Mia T Levine

Recurrent Amplification of the Heterochromatin Protein 1 (HP1) Gene Family across Diptera

Quentin Helleu[1] and Mia T. Levine*,[1]

[1]Department of Biology, Epigenetics Institute, University of Pennsylvania, Philadelphia, PA

***Corresponding author:** E-mail: m.levine@sas.upenn.edu.

Associate editor: Stuart Newfeld

Abstract

The heterochromatic genome compartment mediates strictly conserved cellular processes such as chromosome segregation, telomere integrity, and genome stability. Paradoxically, heterochromatic DNA sequence is wildly unconserved. Recent reports that many hybrid incompatibility genes encode heterochromatin proteins, together with the observation that interspecies hybrids suffer aberrant heterochromatin-dependent processes, suggest that heterochromatic DNA packaging requires species-specific innovations. Testing this model of coevolution between fast-evolving heterochromatic DNA and its packaging proteins begins with defining the latter. Here we describe many such candidates encoded by the Heterochromatin Protein 1 (HP1) gene family across Diptera, an insect Order that encompasses dramatic episodes of heterochromatic sequence turnover. Using BLAST, synteny analysis, and phylogenetic tree building across 64 Diptera genomes, we discovered a staggering 121 HP1 duplication events. In contrast, we observed virtually no gene duplication in gene families that share a common "chromodomain" with HP1s, including *Polycomb* and *Su(var)3-9*. The remarkably high number of Dipteran HP1 paralogs arises from distant clades undergoing convergent HP1 family amplifications. These independently derived, young HP1s span diverse ages, domain structures, and rates of molecular evolution, including episodes of positive selection. Moreover, independently derived HP1s exhibit convergent expression evolution. While ancient HP1 parent genes are transcribed ubiquitously, young HP1 paralogs are transcribed primarily in male germline tissue, a pattern typical of young genes. Pervasive gene youth, rapid evolution, and germline specialization implicate heterochromatin-encoded selfish elements driving recurrent HP1 gene family expansions. The 121 young genes offer valuable experimental traction for elucidating the germline processes shaped by Diptera's many dramatic episodes of heterochromatin turnover.

Key words: **heterochromatin, Diptera, HP1, chromo, chromoshadow.**

Introduction

Heterochromatin is the gene-poor, repeat-rich genome compartment concentrated at telomeres and centromeres (Henikoff 2000; Smith et al. 2007). Despite its ostensible dearth of functional genetic elements, the heterochromatin compartment supports vital nuclear processes, including chromosome segregation, genome stability, and gene regulation (Dorer and Henikoff 1994; Allshire et al. 1995; Bernard et al. 2001; Le et al. 2004; Grewal and Jia 2007; Eissenberg and Reuter 2009). These conserved functions rely on the integrity of heterochromatic DNA complexed with specialized chromatin proteins that package and epigenetically delineate heterochromatin from its gene-rich euchromatin counterpart (Elgin 1996; Nowick et al. 2010). One major obstacle to understanding how this nucleoprotein complex supports conserved nuclear functions is the pervasive and puzzling observation that heterochromatic DNA accounts for the most rapidly evolving sequence in eukaryotic genomes. Indeed, even very closely related species exhibit radical sequence divergence of the satellite repeats and mobile elements that dominate the heterochromatin compartment (Kamm et al. 1995; Henikoff et al. 2001; Kidwell 2002; Gallach 2014; Jagannathan et al. 2017). One possible resolution to this paradox is that specialized heterochromatin packaging proteins evolve rapidly to maintain conserved functions. Under this model, heterochromatin proteins and heterochromatic sequence coevolve: recurrent innovation of heterochromatin proteins is required to maintain conserved, essential functions that depend on heterochromatin integrity.

Consistent with this model, several population genetic and molecular evolution analyses have uncovered fast-evolving heterochromatin proteins. This list includes heterochromatin-associated telomere protection proteins, sex chromosome packaging proteins, centromeric histone variants, and genome defense machinery (Barbash et al. 2004; Vermaak et al. 2005; Obbard et al. 2006; Anderson et al. 2009; Bayes and Malik 2009; Klattenhoff et al. 2009; Jacobs et al. 2014; Helleu et al. 2016; Levine et al. 2016; Kursel and Malik 2017; Lee et al. 2017; Maheshwari et al. 2017; Parhad et al. 2017). The nonrandom accumulation of amino acid-changing mutations in heterochromatin packaging proteins, that is, positive selection, implicates perpetual coevolution with the largely uncharacterized, heterochromatic repetitive sequence. Under this framework, disruption

Article

of any one of the two coevolving parties should have devastating cellular and/or developmental consequences. Consistent with this prediction, genetic incompatibilities uncovered in interspecies hybrids frequently involve heterochromatin mis-packaging, chromosome mis-segregation, and mobile element activation (Brideau et al. 2006; Bayes and Malik 2009; Ferree and Barbash 2009; Phadnis and Orr 2009; Kelleher et al. 2012; Dion-Cote et al. 2014; Lopez-Maestre et al. 2017). These data reveal that an "evolutionary mismatch" between heterochromatic sequence from one genome and fast evolving packaging proteins from another—brought together in an F_1 hybrid—imposes dire consequences on fundamental nuclear processes.

Fast-evolving heterochromatin proteins diverge between species not only via amino acid changing mutations across orthologs but also by recurrent births of paralogs. A short list of recent phylogenomic analyses on Drosophila genomes suggests that gene duplication represents a potent genetic mechanism of heterochromatin protein innovation. Lee et al. (2017) recently uncovered gene birth, gene loss, and putative replacement in gene families that encode telomere protection proteins (Lee et al. 2017). Kursel and Malik (2017) report proliferation of the gene family encoding the centromeric histone 3, Cen-H3 ("CID" in flies; Kursel and Malik 2017). Finally, two phylogenomic investigations of the heterochromatin protein 1 (HP1) gene family across 60 My of Drosophila evolution revealed 30 gene births and at least five gene deaths (Levine et al. 2012, 2016). Like gene families associated with telomere and centromere proteins, Drosophila HP1s are required for conserved heterochromatin-dependent nuclear processes, like chromosome transmission and genome integrity (Kellum and Alberts 1995; Klattenhoff et al. 2009; Levine et al. 2015; Helleu et al. 2016). The founding member, HP1A, was named for its association with the heterochromatin compartment in *Drosophila melanogaster* (James and Elgin 1986; James et al. 1989). Since the HP1A discovery, HP1 genes have been uncovered in all major clades of eukaryotic organisms (Lomberk et al. 2006; Vermaak and Malik 2009) and typically encode three domains: a N-terminal chromodomain ("CD" hereafter) that mediates interaction with modified histone tails, a C-terminal chromoshadow domain ("CSD" hereafter) that mediates protein–protein interactions (Paro and Hogness 1991; Assland and Stewart 1995; Smothers and Henikoff 2000), and a variable Hinge domain that binds RNA and/or DNA (Smothers and Henikoff 2001; Muchardt et al. 2002; Meehan et al. 2003), fig. 1*A*). This domain structure has successfully guided previous Drosophila-wide phylogenomic analyses that uncovered pervasive youth and rapid evolution (Levine et al. 2012). Furthermore, functional analyses of two female germline–specialized Drosophila HP1s implicate coevolution with heterochromatin-embedded transposable elements while two male germline HP1s implicate coevolution with sex chromosome heterochromatin (Klattenhoff et al. 2009; Levine et al. 2015, 2016; Helleu et al. 2016; Parhad et al. 2017). Importantly, both transposable elements and heterochromatic sex chromosomes turn over across the Drosophila genus (Lerat et al. 2011; Vicoso and Bachtrog 2013; Bargues and Lerat 2017), raising the possibility that recurrent HP1 birth tracks the fast evolution of heterochromatic DNA.

To explore this hypothesis, we set out to identify HP1 genes across a densely sampled clade that spans many dramatic episodes of heterochromatin divergence. The 250 My captured by the insect Order, Diptera, is ideal. For example, while virtually all sequenced Dipteran species harbor heterochromatic Y-chromosomes, the yellow fever mosquito (*Aedes aegypti*) and the hump-back fly (*Megaselia abdita*) have independently evolved cytologically homomorphic sex chromosomes, which results in reduction of heterochromatin content and complexity (Bhalla 1973; Traut 1994; Hall et al. 2015; Vicoso and Bachtrog 2015). At the other extreme, the sepsid fly, *Themira minor*, and *Drosophila miranda* have undergone wholesale heterochromatic sex chromosome turnover (Zhou and Bachtrog 2012, 2015; Vicoso and Bachtrog 2015; Mahajan and Bachtrog 2017). The tsetse fly (*Glossina* species) and bloodsucking black flies (*Simuliidae*) harbor supernumerary heterochromatic "B" chromosomes (Warnes and Maudlin 1992; Chubareva et al. 1996), while the fungus gnat (*Sciara coprophila*; Gerbi 1986) and hessian fly (*Mayetiola destructor*; Stuart et al. 2012) have evolved heterochromatic germline-limited chromosomes. Finally, genome size varies considerably across much of the Diptera tree; for example, the ~220-My-old mosquito clade (Culicidae) harbors genomes ranging in size 10-fold, from 174 to 1,967 Mb (Chen et al. 2015)—changes driven largely by heterochromatic repeat content. Despite this remarkable diversity of heterochromatin, the specialized proteins with which this elusive DNA sequence coevolves are unknown. Moreover, direct investigation of the evolutionary and functional significance of this heterochromatic sequence divergence challenges our computational and experimental toolkits—repetitive DNA assembly is problematic and mapping phenotypes to specific heterochromatic sequence is typically unfeasible (Henikoff 2000; Smith et al. 2007). However, we can readily manipulate heterochromatin proteins (encoded by euchromatic genes) that package this recalcitrant genome compartment (Vermaak et al. 2009).

Here, we define 121 young Heterochromatin Protein 1 (HP1) genes across Diptera. We leveraged the conserved domains of the HP1 gene family to conduct a comprehensive phylogenomic analysis of 64 publicly available genomes and/or transcriptomes that span this 250 My slice of evolutionary time. The exceptionally high numbers of HP1 paralogs represent almost exclusively young, lineage-restricted genes. Rampant gene birth and death confer both divergent and convergent HP1 gene numbers across distantly related species. Moreover, independently derived HP1s exhibit convergent expression evolution. Ancient HP1s are transcribed ubiquitously while young duplicate HP1s are transcribed primarily in male germline tissue, a pattern typical of young duplicate genes (Kaessmann 2010; Assis and Bachtrog 2013). Our data do not support the intuitive model that expansions and contractions of the sheer quantity of heterochromatic sequence selects for varying numbers of HP1s. Instead, our findings implicate distinct selfish, heterochromatin-embedded elements that hijack germline processes in shaping HP1 diversity across 250 My of Diptera evolution.

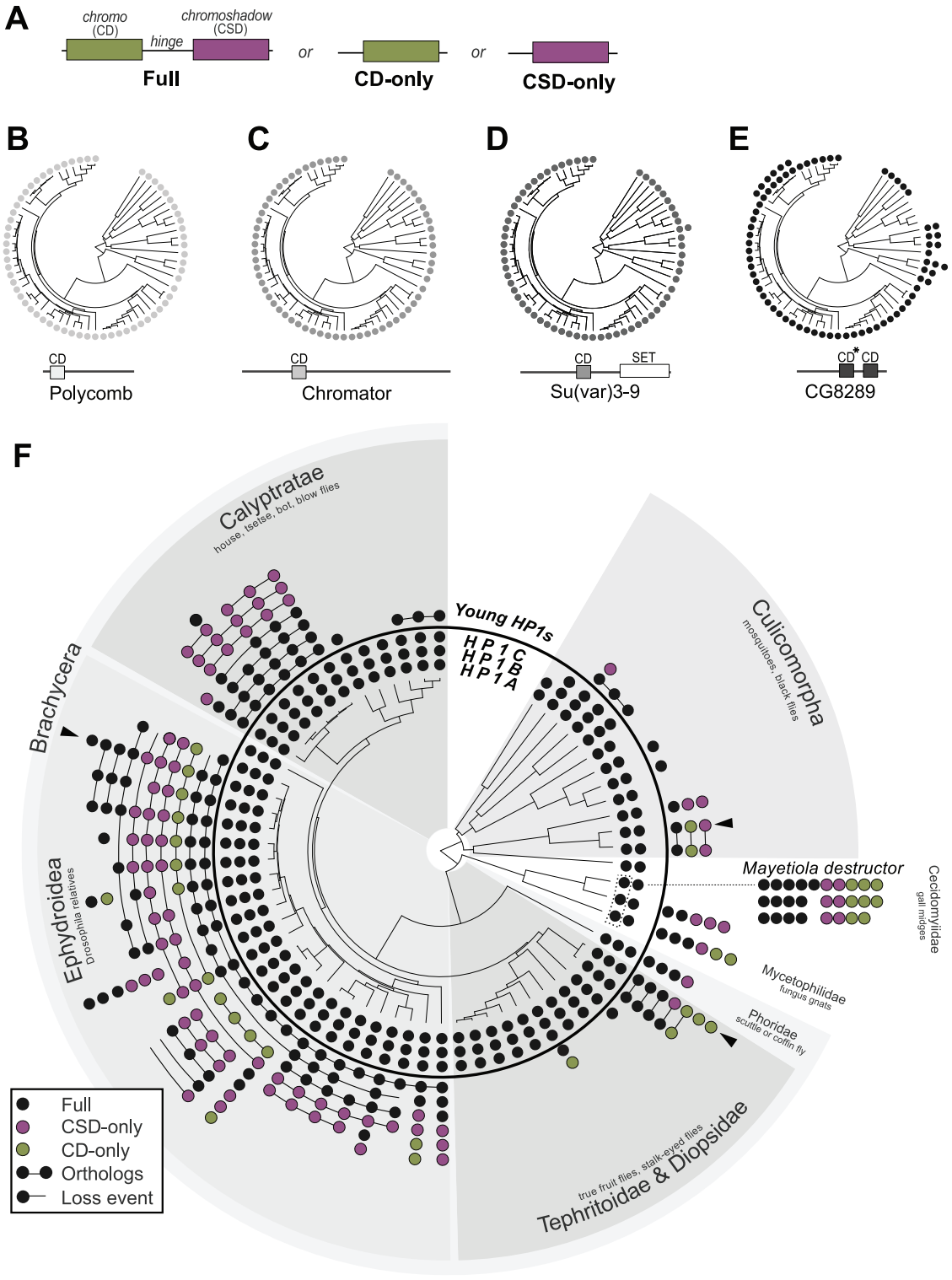

Fig. 1. Diptera-wide phylogenomic analysis of the HP1 family and families closely related to HP1s. (*A*) Domain structures of previously characterized HP1 genes. The canonical HP1 structure includes a chromodomain ("CD") and a chromoshadow domain ("CSD") separated by a hinge domain. (*B–E*) For all families of "HP1-relatives," a filled circle appears at the terminal end of each Diptera lineage where a given gene copy was discovered and phylogenetically validated. Terminal branches missing a circle (as in *E*) indicate an ancestral loss event. A second daughter copy (i.e., a paralog) is indicated with a second circle (as in *D*). (*F*) Diptera dendrogram indicating HP1 gene copies in each lineage. The orthologs of "ancient" HP1s (*HP1A*, *HP1B*, *HP1C*) appear in the inner circle while the relatively "young" HP1 orthologs (connected by a line, supported by synteny analysis) and paralogs (not connected by a line) appear outside the circle. Gene loss events (supported by evidence of a degenerated but recognizable sequence) are indicated by a line but no circle. Dashed box indicates *HP1A*-like genes for which we were unable to infer the ancient *HP1A* ortholog from young *HP1A*-like paralog. Arrowheads refer to harlequin fly (*Chironomus riparius*), the stalk-eyed fly (*Teleopsis dalmanni*) and an East Asian fruit fly (*Drosophila takahashii*) encode five, 10, and 13 *HP1*s, respectively.

Results

The HP1 Gene Family Recurrently (But Selectively) Amplifies across Diptera

To investigate the diversification of the HP1 gene family across Diptera, we performed a tBLASTn search of 64 curated genomes or transcriptomes that span 250 My of evolution (see Materials and Methods; supplementary tables S1 and S2 and fig. S1, Supplementary Material online). Canonical HP1s are characterized by the presence of both a chromodomain ("CD") and chromoshadow domain ("CSD"); however, abundant CD-only and CSD-only HP1s have been described across Drosophila (fig. 1A; Levine et al. 2012). While the CSD is diagnostic of HP1 gene family membership, many other euchromatin- and heterochromatin-packaging proteins encode a CD. We therefore considered true HP1 family members only those genes encoding a CD and a CSD ("full"), a CSD only, or a sole CD that forms a monophyletic clade with canonical "full" HP1s. A small subset of deeply diverging, CD-only HP1-like candidate genes required an additional reciprocal best BLAST analysis to confirm membership (see Materials and Methods). In any given Diptera genome, we inferred that our search for HP1 orthologs and paralogs was exhaustive when our *HP1-only* query (supplementary table S3, Supplementary Material online) recovered three non-HP1 genes that encode closely related CDs. These "HP1-relatives" are *Polycomb (Pc)*, *Chromator (Chro)*, and *Suppressor of variegation 3-9 (Su(var)3-9)*. We also identified in most Diptera genomes a fourth, more phylogenetically labile HP1-relative gene, *CG8289*, which we used only for phylogenetic delineation of true HP1 family members (supplementary fig. S2 and table S4, Supplementary Material online). The CDs of proteins that also have a helicase domain, like CHD1 and Mi-2 ("HP1-distant" genes), were too distantly related to offer additional HP1 phylogenetic resolution (supplementary fig. S2A, Supplementary Material online).

The four HP1-relatives offer baseline rates of gene family expansion and contraction in non-HP1 chromosomal proteins. Our tBLASTn followed by phylogenetic tree building and synteny analysis uncovered orthologous members of HP1-relative genes *Pc*, *Su(var)3-9*, and *Chro* in all 64 Diptera genomes (fig. 1B and C and supplementary table S4, Supplementary Material online). We found no evidence of *Pc* or *Chro* gene duplication or gene loss (fig. 1B and C; table 1). We discovered one additional *Su(var)3-9* paralog in the lineage leading to the Antarctic midge (*Belgica antarctica*, fig. 1D and supplementary fig. S3C and table S4, Supplementary Material online). We also discovered that *CG8289* was lost once—in the Malaria vector mosquitoes (Anophelinae, fig. 1E) and duplicated four times—once along the lineage leading to tsetse flies (Glossinidae) and at least three times along the lineages leading to the non-biting midges (e.g., *Belgica antarctica*) and the gall midges and fungus gnats (e.g., *Sitodiplosis mosellana*, fig. 1E and supplementary fig. S3C, D, and H, Supplementary Material online). In total, a Diptera-wide search across four CD-encoding, non-HP1 gene families uncovered five gain events and one loss event (fig. 1B–E and table 1).

Table 1. Gene Family Size-Changes across Diptera.

Gene	# Gains	# Losses
Polycomb	**0**	**0**
Chromator	**0**	**0**
Su(var)3-9	**1**	**0**
CG8289	**4**	**1**
***HP1-relative*: total**	**5**	**1**
***HP1*: Full**	**61**	**9**
***HP1*: CD-only**	**22**	**3**
***HP1*: CSD-only**	**38**	**7**
***HP1*: total**	**121**	**19**

NOTE.—Number of gene duplication events ("gains") and number of gene loss events detected across CD-encoding genes outside the HP1 family ("HP1-relative," upper panel in gray) and true HP1s genes parsed by those encoding both a CD and a CSD, a CD-only, or a CSD-only (lower panel).

In striking contrast to this rarity of gene duplication and gene loss across the HP1-relatives, we uncovered 121 HP1 duplication events and at least 19 gene loss events across 64 Diptera genomes (fig. 1F and table 1; supplementary figs. S2–S5 and tables S4–S6, Supplementary Material online). *HP1A* (encoded by *Su(var)205*) and *HP1B* are strictly retained, as observed for *Pc*, *Chro*, and *Su(var)3-9*, indicating that these two HP1 genes arose >250 Ma. *HP1C* appears in 52 of 64 genomes, also consistent with a pre-Diptera birth event but followed by multiple loss events along sublineages leading to mosquitos, fungus gnats, and midges (fig. 1F). Beyond these three ancient HP1s, the remaining 121 are all highly lineage-restricted, with the vast majority (~80%) emerging <70 Ma. Importantly, these reported HP1 gene numbers represent a minimum estimate. Although genome scaffold length (N50) does not correlate with HP1 number per species (supplementary fig. S6, Supplementary Material online), we expect additional HP1 paralogs (especially tandem duplicates) to emerge as genome assemblies improve.

The number of lineage-restricted HP1s encoded by a given species varies dramatically across Diptera lineages, with some species encoding only two HP1s and others encoding more than 10 (fig. 1F). This result contrasts with previous findings from Drosophila where each lineage encodes a relatively similar HP1 number (Levine et al. 2012). For example, several mosquitos and black flies ("Culicomorpha") encode only the ancient *HP1A* and *HP1B* (fig. 1F), while the harlequin fly (*Chironomus riparius*), the stalk-eyed fly (*Teleopsis dalmanni*) and an East Asian fruit fly (*D. takahashii*) encode five, 10, and 13, respectively (fig. 1F, arrowheads). The species with the highest number of lineage-restricted HP1s was the crop pest *Mayetiola destructor*, with an impressive 29 gene family members (fig. 1F and supplementary tables S4 and S6, Supplementary Material online). We ruled out the possibility that the 27 young paralogs include multiple alleles—each predicted gene was flanked by 1 kb of 5′ and 3′ unique sequence and maps to a unique position along the genome assembly based on contig coordinates (supplementary table S4, Supplementary Material online). The absence of an *M. destructor* close relative on our Diptera tree precluded our ability to ask how long these young genes have been retained. However, the vast majority of the young HP1s reported here

appear in many divergent clades, consistent with gene retention for 1–180 My. These data, along with evidence of expression (supplementary table S2, Supplementary Material online), functional constraint (see below), and previously defined functions of *D. melanogaster* HP1s of similar evolutionary age (Klattenhoff et al. 2009; Ross et al. 2013; Levine et al. 2015, 2016; Helleu et al. 2016), support the inference that the vast majority of HP1s reported here are functional.

These nonoverlapping HP1 repertoires across Diptera lineages implicate pervasive gene loss along with gene gains. Only in the densely sampled Drosophilids were we able to rigorously identify partially degenerated sequence consistent with recent loss. Here, we detected at least 2:1 duplication: loss events (supplementary table S5 and fig. S5, Supplementary Material online). Extrapolating from this clade, we infer that gene gain *and* loss accounts for the apparent gene replacement inferred from nonoverlapping HP1 paralogs across the many divergent Diptera species.

The abundant young HP1s of Diptera span diverse domain structures and parent genes-of-origin. Of the 121 gene birth events, 61 genes encode both a CD and CSD ("full"), 22 encode only a CD, and 38 encode only a CSD. While rapid evolution obscured the parent gene source of half of these young duplicates, of the well-resolved daughter genes, *HP1A* is the most prolific parent (at least 31 daughter/granddaughter genes), followed by *HP1B* (at least 23 daughter/granddaughter genes) and finally, *HP1C* with none (supplementary figs. S3 and S4 and table S4, Supplementary Material online). These strikingly dissimilar duplicate gene retention rates suggest that *HP1C* behaves like an HP1-relative gene (e.g., *Pc* or *Chro*), while *HP1A* and *HP1B* serve as primary sources of young HP1s across Diptera. Intriguingly, *HP1C* is the only "ancient" HP1 protein restricted to the euchromatin compartment (Smothers and Henikoff 2001), implicating specifically heterochromatin functions as the biological engine driving HP1 family proliferation (see Discussion). Across this 250-My snapshot, extreme gene family size differences, domain structure divergence, and pervasive lineage-restriction suggest that HP1s uniquely proliferate and diversify over both short and long stretches of evolutionary time relative to other DNA packaging proteins that encode a closely related CD.

Functionally Defined Residues Are Constrained across Young HP1s

To further evaluate whether young Dipteran HP1s encode functional gene products rather than persistent pseudogenes, we compared turnover of young and ancient HP1 residues defined previously by structural and biochemical analysis of CDs and CSDs (supplementary fig. S7, Supplementary Material online; Smothers and Henikoff 2000; Jacobs and Khorasanizadeh 2002; Nielsen et al. 2002). We reasoned that if young HP1s encode predominantly pseudogenes, critical residues for protein structure and chromatin recognition ("functionally defined residues") would turn over at rates similar to undefined sites. We calculated percent identity ("%ID") across 49 HP1s encoded by five different Diptera species. Each species encodes HP1A, HP1B, and HP1C ("ancient HP1s") and at least six young, species-restricted HP1s ("young HP1s"). HP1A, HP1B, and HP1C provide a baseline percent identity of defined sites: considered together, critical residues are conserved despite ancient gene births, divergent localization along chromosomes, and nonoverlapping functions (supplementary fig. S7, Supplementary Material online; Vermaak and Malik 2009). Not unexpectedly, percent identity estimates for all residues considered together is lower for young HP1s compared with ancient HP1s (Wilcoxon, $W = 9{,}250$, $P < 0.0001$, supplementary table S7, Supplementary Material online). However, between site classes percent identity is significantly elevated for functionally defined relative to undefined residues of young HP1s (fig. 2*A* and supplementary table S7, Supplementary Material online). Strikingly, the high percent identity of these functionally defined residues from young HP1s is indistinguishable from those of ancient HP1s (fig. 2*A*). These results suggest that HP1 duplicate genes maintain the canonical HP1 structure and recognition capacity but innovate at other, less constrained residues.

We visualized this elevated percent identity of functionally defined sites in amino acid LOGO plots of the CD and CSD of young and ancient proteins (Crooks et al. 2004). The young HP1 "histone binding domain" residues, which recognize either the methylated lysine 9 of histone H3 (a modification classically enriched in heterochromatin) or nearby histone tail sites, exhibit the highest conservation of the three site classes (median = 0.65, fig. 2*B*, red residues; Nielsen et al. 2002). Relatively high conservation was also apparent in the CSD's "Chromoshadow Fold" domain (median = 0.48, blue residues, fig. 2*B*; Jacobs and Khorasanizadeh 2002). The percent identity of the protein–protein interaction domain was lowest (median = 0.35, yellow residues, fig. 2*B*; Smothers and Henikoff 2000; Jacobs and Khorasanizadeh 2002). Only two of eight positions (arrows) have a high bitscore. Together these data suggest that young HP1s encode functional proteins that retain core structural and histone recognition features characteristic of the gene family, but protein interaction repertoires likely vary across young duplicates. Indeed, the two young, functionally characterized HP1s in *D. melanogaster* that diverge significantly at this protein: protein interaction domain, Umbrea and Rhino/HP1D, coimmunoprecipitated with protein-specific interaction partners (Ross et al. 2013; Mohn et al. 2014; Chen et al. 2016). The Hinge domain also determines paralog-specific localization among the ancient HP1s (Smothers and Henikoff 2001), but the dramatic structural divergence at this domain precludes our ability to conduct an equivalent percent identity analysis (supplementary fig. S8, Supplementary Material online). Nevertheless, we detected differential retention of the previously characterized nuclear localization signal that likely contributes, together with the CSD, to diverse subcellular localization of young HP1 proteins (Smothers and Henikoff 2001; Mishima et al. 2013; supplementary fig. S8, Supplementary Material online).

Young HP1s Rapidly Accumulate Amino Acid-Changing Mutations

Young genes are typically enriched for signatures of adaptive evolution compared with their parent genes (Han et al. 2009;

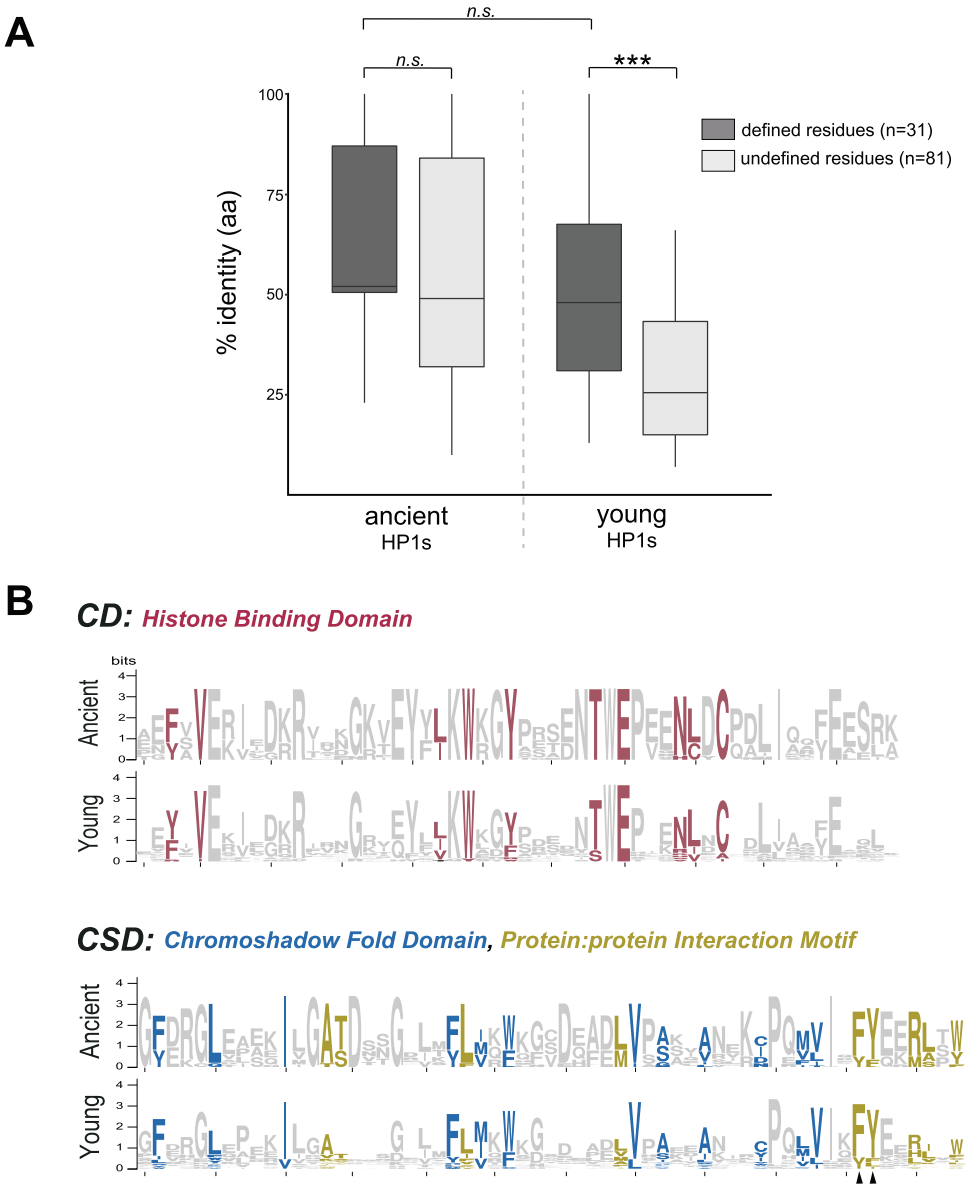

Fig. 2. Conservation of functionally defined residues in young HP1s. (*A*) Percent identity of functionally defined residues (supplementary fig. S7, Supplementary Material online) and undefined residues encoded by ancient (HP1A, HP1B, HP1C) and young HP1 paralogs compared across five Dipteran species: *Drosophila melanogaster*, *Glossina morsitans*, *Drosophila miranda*, *Teleopsis dalmanni*, *Scaptodrosophila lebanonensis*. Sample size "*n*" refers to number of sites compared per site class (for young HP1 comparison: $Z = 3.41$, "***": $P < 0.001$). (*B*) LOGO plot generated for ancient and young HP1s amino acid alignment with critical residues color-coded by domain/motif.

Assis and Bachtrog 2013; Jiang and Assis 2017). To investigate the possibility that turnover of undefined sites reflects adaptation of young HP1s, we quantified rates of molecular evolution between orthologs across seven pairs of closely related sister species that broadly sample the Diptera tree (fig. 3*A*). Specifically, we calculated the pairwise d*N*/d*S* (nonsynonymous divergence/synonymous divergence) of young and ancient HP1s. The pairwise d*N*/d*S* distributions across these CDs and CSDs vary significantly (fig. 3*B* and supplementary table S8, Supplementary Material online, Kruskal–Wallis = 34.41, $P < 0.0001$). For both the CDs and CSDs, young HP1s harbor elevated d*N*/d*S* relative to ancient HP1s, consistent with young HP1s rapidly accumulating amino acid-changing mutations under positive selection and/or loss of constraint.

To evaluate these alternatives, we investigated rates of evolution among HP1s in the tsetse fly (*Glossina*) clade using maximum likelihood methods. We focused on the *Glossina* clade because of its many closely related, publicly available genomes, a requirement for more in-depth molecular evolution analysis (supplementary table S2, Supplementary Material online). Virtually all lineage-restricted HP1s harbor elevated d*N*/d*S* estimates compared with ancient HP1s (table 2 and supplementary table S9, Supplementary Material online). Using a maximum likelihood framework implemented in PAML (see Materials and Methods), we detected significant signatures of positive selection in three genes, *HP1.Gm2*, *HP1.Gm3*, and *HP1.Gm7csd* ("*Gm*" = *Glossina morsitans*, see Materials and Methods for

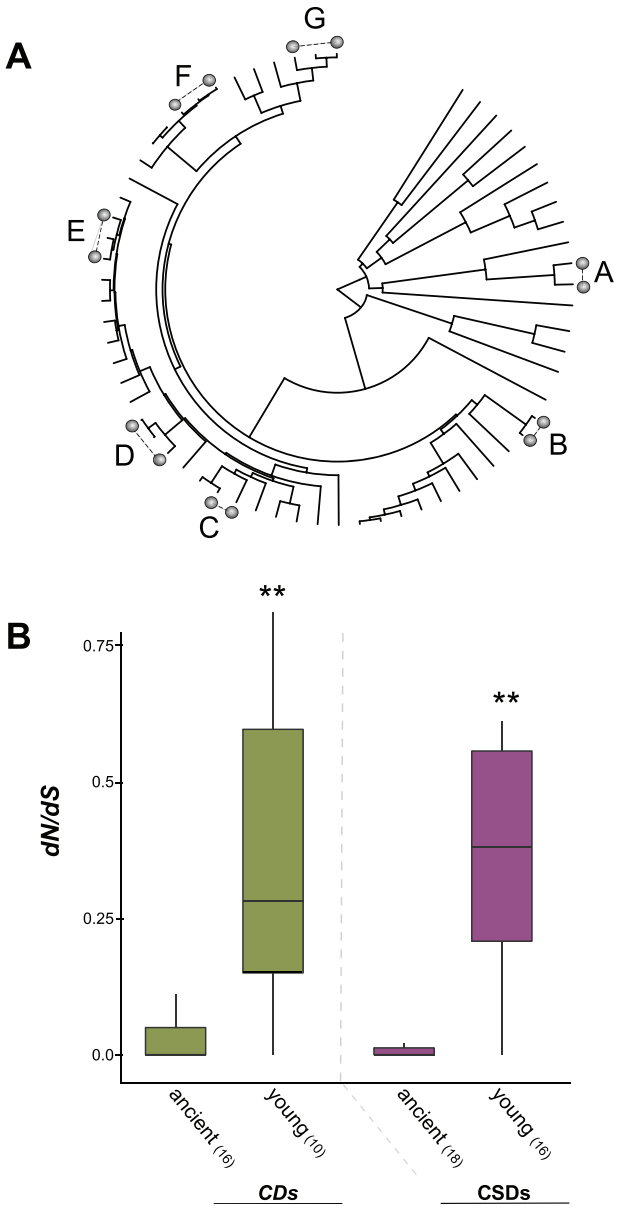

FIG. 3. Molecular evolution of CDs and CSDs. (*A*) Seven focal species pairs distributed across the Diptera dendrogram, where average dS (rate of synonymous substitution) is <0.5 for each pair per gene. The following species pairs are represented by each letter: A = *Chironomus tentans*, *Chironomus riparius*, B = *Teleopsis dalmanni*, *Teleopsis whitei*, C = *Bactrocera tryoni*, *Bactrocera oleae*, D = *Drosophila miranda*, *Drosophila pseudoobscura*, E = *Drosophila suzukii*, *Drosophila takahashii*, F = *Glossina morsitans*, *Glossina austeni*, G = *Calliphora vicina*, *Cuprina sericata*. (*B*) Comparisons of mean pairwise dN/dS values across old versus young CDs and CSDs in species pairs found in (*A*). The mean dN/dS of young CDs and CSDs is significantly larger than those of ancient HP1s ("**": $P < 0.01$).

nomenclature description). Only *HP1.Gm1* evolves extremely slowly, at a rate similar to the ancient HP1s. A comparison of one ratio models (M0) with either dN/dS set to 1 or estimated from the data uncovered evidence of functional constraint at *HP1.Gm1* (supplementary table S9, Supplementary Material online). The signatures of positive selection in the tsetse fly are reminiscent of several independently derived Drosophila HP1 genes reported previously (Vermaak et al. 2005; Levine et al. 2012). As anticipated earlier, the positively selected codons identified correspond to undefined residues (*HP1.Gm2*, *HP1.Gm3* in the CD, *HP1.Gm7csd* in the N-terminal extension, table 2). The Drosophila and Glossina data together suggest a general pattern of positive selection shaping many young HP1 genes.

Young HP1s Are Enriched in Germline Tissues

To begin elucidating the biology shaped by recurrent HP1 gene family amplifications, we conducted expression analysis in adult tissues. We first utilized available RNA-seq data sets prepared from somatic (head) and germline (testis or ovaries) tissue in lineages that encode at least six young HP1s (<40 My old; fig. 4 and supplementary table S10, Supplementary Material online). For three such species, *D. melanogaster*, *D. miranda*, and *Teleopsis dalmanni*, we observed ubiquitous expression for the relatively ancient *HP1A*, *HP1B*, and *HP1C* genes. Consistent with a previous RT-PCR-based report for *D. melanogaster* (Levine et al. 2012), we observed testis-biased expression for all young HP1s except *HP1D/rhino* and *oxpecker*, which are instead enriched in the female germline. *D. miranda* encodes six independently duplicated HP1 paralogs, all of which are predominantly testis-expressed. Finally, the distantly related *T. dalmanni* (stalk-eyed fly, 70 My diverged) harbors three ubiquitously expressed and four testis-restricted young HP1 genes. These data are consistent with convergent germline expression evolution across independently derived HP1s.

We also identified scores of lineage-specific HP1s in many Diptera species for which soma- versus germline-transcriptomes are not publicly available. To investigate gene- and tissue-specific expression patterns in two such species—*Glossina morsitans* and *Scaptodrosophila lebanonensis*—we performed RT-qPCR on dissected male and female heads and reproductive tissues (fig. 4). For both species, we detected expression in all tissue types for the ancient *HP1A*, *HP1B*, and *HP1C* genes. Five of the six young *S. lebanonensis* HP1 genes are restricted to the male testis while one is enriched in the ovaries (*Sl3cd*). In contrast, young HP1s from *G. morsitans* exhibited predominantly female germline enrichment. This unique ovary-biased pattern among the sampled species raises the possibility that the many young *G. morsitans* HP1s mediate its unique proliferation of young, heterochromatic B chromosomes (Warnes and Maudlin 1992) and/or silence female germline transposable elements (see below). The RT-qPCR and RNA-seq data together implicate rapidly evolving, heterochromatin germline biology in the retention of young, lineage-restricted HP1 genes.

Discussion

The discovery of *Drosophila melanogaster* HP1A, the founding HP1 family member, profoundly altered our grasp of heterochromatin's role in the organization and regulation of the eukaryotic nucleus (James and Elgin 1986; James et al. 1989; Eissenberg and Elgin 2014). For decades prior, the highly repetitive content and minimal recombination of the heterochromatic genome compartment obstructed insight and inspired benign neglect. Analysis of heterochromatin

Table 2. Molecular Evolution of HP1s across the tsetse Fly Genus (*Glossina*).

Gene	# Codons	# Species[a]	dN/dS	M8a versus M8 (X^2)	*P* value (df=1)	Significant BEB sites
HP1A	**210**	**5**	**0.040**	**0.00**	**0.99**	
HP1B	**172**	**5**	**0.022**	**0.00**	**0.99**	
HP1C	**230**	**5**	**0.030**	**0.00**	**0.99**	
HP1.Gm1	**170**	**5**	**0.063**	**0.03**	**0.84**	
HP1.Gm2	**139**	**5**	**0.417**	**4.07**	**0.044**	**12S, 56N**
HP1.Gm3	**169**	**5**	**0.906**	**4.24**	**0.039**	**9V, 33E**
HP1.Gm4	**165**	**4**	**0.408**	**0.00**	**0.98**	
HP1.Gm5csd	**70**	**5**	**0.555**	**0.26**	**0.61**	
HP1.Gm6csd	**63**	**5**	**0.979**	**0.28**	**0.60**	
HP1.Gm7csd	**91**	**3**	**0.593**	**10.17**	**0.0014**	**3A, 19E**

NOTE.—Results from codeML analysis in the PAML software package of ancient HP1s (*HP1A*, *HP1B*, *HP1C*) and young HP1s restricted to *Glossina*. *HP1.Gm1*, *HP1.Gm2*, *HP1.Gm3*, and *HP1.Gm4* represent full HP1s while *HP1.Gm5csd*, *HP1.Gm6csd*, and *HP1.Gm7csd* encode only a chromoshadow domain.
[a]*Glossina morsitans*, *G. pallidipes*, *G. austeni*, *G. palpalis*, *G. fuscipes* are the five species encoded by most of the HP1s analyzed here. *HP1.Gm4* is not present in *G. palpalis*. *HP1.Gm7csd* is not present in *G. palpalis and G. fuscipes*. "BEB sites" refers to sites that exceeded a posterior probability threshold estimated by Bayes Empirical Bayes of 90% and 95% (underlined).

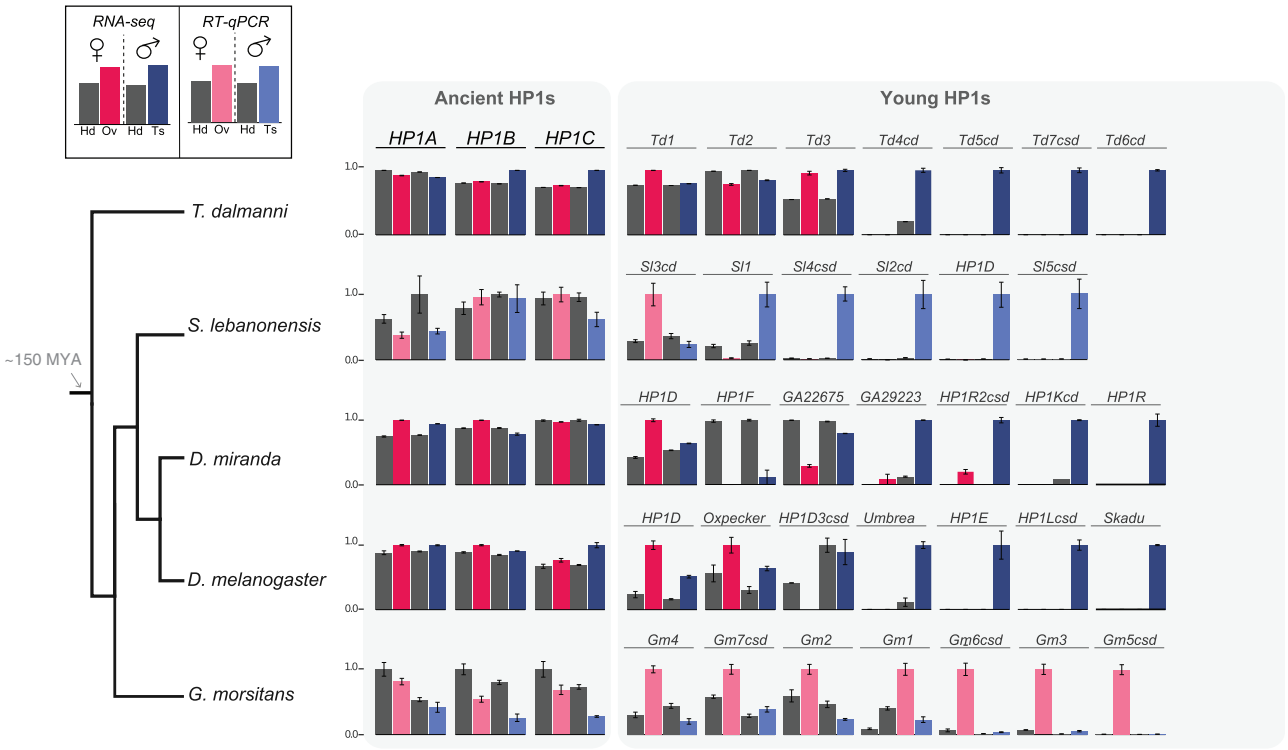

FIG. 4. Tissue-restricted gene expression of ancient and young HP1 orthologs and paralogs. Analysis of RNA-seq and RT-qPCR on samples prepared from adult tissues across six Diptera species that span the 150 My of evolution. "Hd" = head, "Ov" = ovaries, "Ts"=testes. Within each species, the young genes are sorted left-to-right by soma-biased expression to germline-biased expression and ovary to testis-biased expression, with the testis-restricted genes appearing at the far right for each species.

packaging proteins like HP1A, however, offered experimental traction that revealed this compartment's unexpected role in conserved, essential processes like chromosome segregation, telomere integrity, and genome defense (Kellum and Alberts 1995; Fanti et al. 1998). Here, we leverage this "surrogate approach" to expand our toolkit for probing the functional significance of heterochromatin and its rapid evolution (Vermaak et al. 2009). We report over 100 previously undefined members of the Heterochromatin Protein 1 (HP1) gene family across Diptera, an insect Order that spans remarkable heterochromatin diversity. The 121 young HP1s genes reported here likely represent an underestimate—the restricted number of codons per domain combined with rapid sequence evolution limits our power to infer HP1 membership of chromodomain-only HP1s. Many such unassigned genes branch as polytomies at the root of trees generated by MrBayes (supplementary fig. S3, Supplementary Material online) and PhyML (data not shown). The reported HP1 diversity spans 250 My of Diptera evolution, which brims with heterochromatic chromosome losses and gains, genome size expansions and contractions, and transposable element invasions (see Introduction). Like these recurrent heterochromatin turnover events, virtually all identified HP1 births are recent—between 0.5 and 70 Ma. This pervasive lineage restriction is consistent with recurrent gene turnover within and between Dipteran clades.

Widespread expansions and contractions of Diptera's HP1 gene family contrast sharply with the evolutionary dynamics of four other gene families that share a chromodomain (CD). *Pc*, *Chro*, and *Su(var)3-9* are strictly retained in all focal genomes. We detected only a single duplication event of *Su(var)3-9* in an Antarctic midge. *CG8289*, an uncharacterized CD-encoding gene, is the most dynamic non-HP1 gene family, with a single loss event and four duplication events. The strikingly different gene family dynamics of HP1s compared with four closely related families strongly support the idea that the HP1 family uniquely regulates and responds to fast evolving heterochromatic DNA packaging requirements over time. Consistent with heterochromatin function specifically driving HP1 gene family diversification, we did not detect even one young HP1 derived from the only euchromatin-localized HP1 protein, HP1C (Smothers and Henikoff 2001).

Variation in lineage-specific HP1 number is not restricted to the Order, Diptera: we also found lineages of non-Dipteran arthropods with divergent family sizes (supplementary fig. S9 and tables S11–S13, Supplementary Material online). For example, the bed bug (*Cimex lectularius*) encodes three young HP1 genes while the parasitoid wasp, *Nasonia vitripennis*, encodes only one (Fang et al. 2015). This 500-My snapshot also fortuitously revealed that the intensively studied, absolutely essential HP1A protein (encoded by *Su(var)205* in *D. melanogaster*) is itself surprisingly young, born ~300 Ma. An *HP1B*-like gene appears to be the ancestor of all arthropod HP1s (supplementary fig. S9, Supplementary Material online; Vermaak and Malik 2009). These data challenge our intuition that only ancient genes support conserved functions and vice versa. Ross et al. 2013 drew a similar inference from an extremely young *HP1* that, despite its recent birth, was essential for viability (Ross et al. 2013). The HP1 family demonstrates that essentiality and gene age are not necessarily correlated, joining a small but growing list of examples from Drosophila (Chen et al. 2010; Kondo et al. 2017), Lepidoptera (Drinnenberg et al. 2014), and Giardia (Paredez et al. 2011).

What is the biological significance of pervasive young HP1 gene retention shaping heterochromatin functions? Using gross characterizations of heterochromatin compartment evolution, including bulk repeat content and sex chromosome turnover, we found no clear correlations with raw HP1 number (supplementary figs. S10 and S11, Supplementary Material online). Instead, patterns of young HP1 gene expression offer some insight. Of the focal 34 young HP1 genes subjected to expression analysis, 28 (>80%) are transcribed primarily in germline tissue. Some are expressed in reproductive tissue only of males and females, others expressed primarily in ovaries only, and still others expressed primarily in testis only. Of all three classes, testis-enrichment accounts for the largest fraction of focal genes. These data suggest that while somatic functions may also select for young HP1s (~20%), germline processes drive the bulk of this innovation.

To appreciate the biological significance of this pervasive germline enrichment of young proteins that potentially interact with repetitive, heterochromatic DNA sequence, we turn to the evolutionary dynamics of two heterochromatin-interacting, non-HP1 gene families and a handful of functionally characterized, germline-specialized young HP1s. The antiviral/antitransposon Argonaute gene family also proliferates rapidly across the Diptera tree (Lewis, Salmela et al. 2016). A detailed phylogenomic analysis of Argonautes in the Drosophila *obscura* clade revealed pervasive male germline restriction of young Argonaute2 (Ago2) paralogs (Lewis, Webster, et al. 2016). The authors put forward the hypothesis that intragenomic conflict with selfish elements, which gain access to the next generation only in the germline, drives Ago2 proliferation and ongoing evolution at least in this Diptera subclade. Another gene family with striking similarities to HP1s outside of Diptera is the vertebrate-restricted KRAB-ZNF gene family, which also diversifies rapidly (Nowick et al. 2010; Thomas and Schneider 2011). Germline restriction again characterizes these young, DNA-binding proteins that target transposable elements. Jacobs et al. (2014) reported functional evidence of tit-for-tat coevolution between human KRAB-ZNF genes and human retrotransposons, a discovery supported more broadly by Imbeault et al. (2017) across vertebrates (Jacobs et al. 2014; Imbeault et al. 2017). Consistent with the possibility that TE evolution may also drive HP1 gene family diversification, two young, female germline HP1s from *D. melanogaster* suppress transposable elements. *HP1D/rhino* evolves under strong positive selection and regulates genome defense by promoting transcription of the primary Piwi-RNA (piRNA) transcript (Vermaak et al. 2005; Klattenhoff et al. 2009; Parhad et al. 2017). *Oxpecker*, a young duplicate of *HP1D/rhino*, also suppresses several transposable elements incompletely silenced in the female germline by its parent gene in *D. melanogaster* (Levine et al. 2016). These data suggest that intragenomic conflict between the host genome and its transposable elements drive at least some HP1 gene evolution documented here, and specifically conflict negotiated in the ovaries.

We predict that genome defense function shapes the evolution of many young, ovary-enriched HP1s encoded by the tsetse fly clade (Glossina). All seven young HP1 genes are ovary-expressed, with four Glossina-restricted HP1s expressed primarily in the ovaries and three expressed exclusively in the ovaries. Intriguingly, this clade has also undergone a unique duplication of Ago3, a key player in the female germline, TE-silencing Piwi-interacting RNA pathway (Lewis, Salmela et al. 2016). Given that young genes rarely evolve ovary-biased expression (Kaessmann 2010; Assis and Bachtrog 2013), this pervasive transcriptional pattern (and evidence of positive selection) implicates dynamically evolving heterochromatin biology in the ovaries of this human disease vector. Intriguingly, this unusual ovary-biased signature in young genes is reminiscent of recent findings from another human disease vector, mosquito (Anopheles spp.; Papa et al. 2017).

The majority of young, germline-biased HP1s in Diptera, however, are expressed primarily in male germline tissue. Testis-biased expression is a hallmark of young duplicate genes (Kaessmann 2010). The broadly reported pattern suggests that both the transcriptional environment and uniquely dynamic evolutionary pressures on male reproductive biology shape proteins required for sperm competitiveness and sperm–egg interactions, among others (Swanson and Vacquier 2002;

Kaessmann 2010). Our data implicate male germline heterochromatin biology as no exception. While transposable element evolution may drive at least some of this diversification, two young, male germline-specialized HP1s implicate instead selfish sex chromosomes. The ~20-My-old, X-linked *HP1D2* gene in *Drosophila simulans* encodes a spermatogonia-restricted, *Y*-chromosome packaging protein (Helleu et al. 2016). The "*Sex Ratio*" or "*SR*" allele of *HP1D2*, in combination with a second allele 110 kb away, causes the heterochromatic Y chromosome to missegregate during male meiosis II (Cazemajor et al. 2000; Montchamp-Moreau et al. 2006). These "*SR*" males produce virtually no *Y*-bearing sperm and so father only daughters—a classic symptom of non-Mendelian segregation driven by a "selfish" *X*-linked element. Intriguingly, *HP1D2* is not the only testis-restricted family member that uniquely regulates sex chromosome segregation. The autosome-linked, *HP1E* gene encodes a paternal DNA packaging protein that when depleted from the male germline, causes paternal chromosome mis-segregation in the first zygotic division of embryos "fathered" by these mutant males (Levine et al. 2015). Levine et al. (2015) went on to show that the paternal *X* chromosome (in female embryos) and paternal *Y* chromosome (in male embryos) are uniquely vulnerable to the loss of the HP1E during sperm development. These two independent studies of testis-specialized HP1 functions implicate sex chromosome evolution and possibly *X*–*Y* chromosome conflict as drivers of the abundant young, testis-specialized HP1s both inside and outside of Drosophila.

Importantly, *Oxpecker*, *HP1D2*, and *HP1E* were identified initially by the same phylogenomic methods employed here (Vermaak et al. 2005; Levine et al. 2012). In this light, the reported 121 young Diptera HP1s offer a powerful toolkit for empirically investigating in molecular detail the still mysterious evolutionary and functional significance of the fast-evolving heterochromatic DNA. Given the enrichment of genome parasites and their targets in heterochromatin (Dimitri and Junakovic 1999; Kanizay et al. 2013; Larracuente 2014; Helleu et al. 2016; Li et al. 2017), young, germline HP1s have the unique power to elucidate how intragenomic conflict shapes fundamental biological processes like intergenerational chromosome transmission and genome integrity. Based on studies in Drosophila (Brideau et al. 2006; Bayes and Malik 2009; Ferree and Barbash 2009; Kelleher et al. 2012; Parhad et al. 2017), we also anticipate that the 121 HP1s reported here include genes involved in hybrid infertility and/or inviability (supplementary fig. S12, Supplementary Material online). Finally, rampant HP1 gene birth and death suggests that a complete picture of heterochromatin function demands investigation of lineage-specific biology. This HP1 toolkit offers new traction to address the traditionally evasive genetic and epigenetic determinants of heterochromatin integrity.

Materials and Methods

Selection of Publicly Available Diptera Genomes for Phylogenomic Analysis

To select Diptera species to include in our phylogenomic analysis, we generated a list of publicly available Diptera genome assemblies (supplementary table S1, Supplementary Material online). To assess suitability, we selected six genes from the 1200 "core insect genes" (Rosenfeld et al. 2016) based on a high conservation score, a restricted gene length, and the absence of gene duplication events. These six genes—*mago nashi*, *Ercc1*, *Proliferating cell nuclear antigen*, *spindle A*, *Catecholamines up*, and *eukaryotic translation initiation factor 2D*—range in size from 150 to 550 codons, substantially longer than the 50-codon chromodomains and chromoshadow domains used in our downstream analysis. We retained any genome assembly for which at least five of the six genes could be detected on a single contig. Based on these conservative criteria, we retained 64 species of the original 78 (supplementary tables S1 and S2 and fig. S1, Supplementary Material online) for our BLAST search.

BLAST Search for HP1 and HP1-like Gene Families across Diptera

To identify Heterochromatin Protein 1 (HP1) orthologs and paralogs across Diptera, we conducted a tBLASTn search of the 64 "curated" genome assemblies described earlier. We seeded this iterative search with 17 annotated chromodomains (CD) and chromoshadow domains (CSD) encoded by the three oldest HP1s (*HP1A*, *HP1B*, and *HP1C*) annotated in six Diptera species: *Aedes aegypti*, *Anopheles gambiae*, *Teleopsis dalmanni*, *Glossina morsitans*, *Musca domestica*, and *Drosophila melanogaster* (supplementary table S3, Supplementary Material online). These six species span major clades in Diptera and have high-quality assembled genomes and transcriptomes available (supplementary table S2, Supplementary Material online). Using these 34 queries, we investigated assembled genomes, assembled transcriptomes, or both (when available). For each investigated Diptera species, we extracted all hits with an e-value < 0.001 along with one kilobase of flanking upstream and downstream sequence. We leveraged these flanking sequences to later discriminate based on synteny paralogs from orthologs in the same genomes and orthologs across different genomes (for all species except the seven with transcriptome-only data sets; supplementary table S2, Supplementary Material online). We generated a list of unique loci (many hits often corresponded to the same locus) using the de novo assembler implemented in Geneious 9.1.5 (Biomatters; http://www.geneious.com; Kearse et al. 2012). We retained only open reading frames (ORFs) encoding >50 codons.

We identified CDs and CSDs in all unique loci based on BLAST alignment to query sequence. We then verified domain identification with NCBI Conserved Domain Search (Marchler-Bauer et al. 2015). We inferred that our HP1 search was exhaustive in a given genome when we identified a complete CD sequence of "HP1 relative" genes: *Pc*, *Su(var)3-9*, and *Chromator*. These three genes encode CDs that are closely related to the HP1 CD, in contrast to the CD and helicase domain-encoding genes, like CHD1 and Mi-2 (supplementary fig. S2A and table S4, Supplementary Material online). When available, we defined intron/exon boundaries based on transcriptomic sequence assembly. If a transcriptome was not available for gene model prediction, we aligned the predicted

HP1A, *HP1B*, *HP1C, or* HP1 relatives (see below) to orthologous exons identified in closely related species to determine intron/exon boundaries using MAFFT (Katoh and Standley 2013). Finally, we inferred loss events of HP1s by identifying syntenic degenerated sequences relative to an orthologous HP1 by aligning flanking regions (10 kb on either side) of the focal orthologous HP1 (supplementary table S5 and fig. S5, Supplementary Material online). This approach was only possible with high-quality genome assemblies of sister species, and so is restricted to the densely sampled Drosophila genomes. Finally, using the same methods, we conducted an HP1 search in seven non-Dipteran arthropods with well-annotated, high-quality genomes (assemblies with NCBI-designated "Full" genome representation, supplementary table S11, Supplementary Material online).

Nomenclature

For ease of recognition of young HP1 domain structure and Diptera clade of origin, we use the following nomenclature for the duplication events uncovered by our phylogenomic analysis. After "HP1" we refer to the domain structure. If the gene encodes both a CD and a CSD ("full"), no designation is made. If the HP1 encodes only a chromodomain, we used "cd" and if only a chromoshadow domain, we used "csd." A "." follows domain designation and then the first letters of the genus and species of first discovery (e.g., "Td" for *Teleopsis dalmanni*). If the species encodes more than one young HP1, we assigned a number to each gene, where "full" HP1s receive the lowest number(s) followed by CD-only and finally CSD-only HP1s. We retained the naming scheme for all Drosophilids described in (Levine et al. 2012, 2016).

HP1 Gene Phylogeny and Family Membership Delineation

We aligned paralogs and orthologs using MAFFT nucleotide alignment v7.222 with default parameters, then realigned with MAFFT as a translated alignment with default parameters in Geneious (v.9.1.5). We checked the final alignment by eye and we removed all codon positions from the alignment that contain a gap in >50% of the sequences. We built CD and CSD gene trees with MrBayes v3.2 implemented in Geneious (> 2, 500, 000 trees, 500 sampling frequency, and 25% burn-in). We generated trees until the average SD of split frequencies was <0.025 for all trees but the two extremely large data sets in supplementary figure S2*B* and *C*, Supplementary Material online, for which we used a 0.05 cut-off (Guindon et al. 2010; Ronquist et al. 2012). For comparison, we also built gene trees using maximum-likelihood methods implemented in PhyML using the LG+G substitution model (1000 bootstraps to determine node support). Observing no conflicts between well-supported branches across the two programs, we report the MrBayes trees only, which offered stronger resolution.

For each species, only genes that met one of two criteria were considered true HP1 family members. First, any gene that encoded the diagnostic CSD is an HP1. Second, for those genes that encoded a CD-only, we relied on a combination of monophyly with "full" (CD and CSD) HP1s and/or reciprocal best BLAST analysis. Fourteen CD-only HP1s form a monophyletic clade with at least one full HP1 (supplementary fig. S3, Supplementary Material online). Fifty-three of the CD-only candidate genes branched deeply on the HP1 gene trees, consistent with rapid evolution that obscures family of origin (supplementary fig. S3 and table S6, Supplementary Material online). Our inability to confidently resolve these candidate HP1 genes from HP1-relatives motivated a tBLASTn analysis using the focal gene as the query against three of our highest quality genomes (*D. melanogaster*, *Glossina morsitans*, *Aedes aegypti*). We retained for further scrutiny only those genes whose best BLAST hit was an annotated HP1 in all three genomes. For a given target gene, we selected the most closely related species to its host genome and conducted a reciprocal best BLAST analysis. Noting that a classic reciprocal best BLAST test is not formally possible for lineage-specific paralogs, we inferred HP1 membership when the best BLAST hit in the sister species is used as a query and returns a lower e-value for the focal gene than any HP1-relative (*Polycomb* and *Chromator*). We report in supplementary table S6, Supplementary Material online, that only 10 of the initial 53 CD-only hits satisfied these criteria.

HP1A, *HP1C*, and *HP1A*-like genes formed well-supported monophyletic clades. *HP1B* genes from outside Brachycera (supplementary fig. S1, Supplementary Material online) frequently branched from the root (supplementary fig. S4*B* and *C* and table S12, Supplementary Material online). To confirm that these full HP1s are *HP1B* orthologs, we conducted a reciprocal best BLAST analysis with three well-annotated genomes that encode unambiguous *HP1B* genes. All 17 predicted *HP1B* genes met the reciprocal best BLAST criteria (supplementary table S12, Supplementary Material online).

For the non-Diptera arthropod HP1 phylogenomic analysis referenced in the Discussion, many HP1 genes did not form monophyletic clades with the well-supported *HP1A* and *HP1C* clades and instead form a polytomy on both the CD and CSD trees. We evaluated the possibility that these genes represent the more basal *HP1B*-like family by conducting a reciprocal best BLAST analysis using the confirmed *HP1B* genes from *D. melanogaster* on all other eight genomes represented on the non-Diptera arthropod species tree in supplementary figure S9, Supplementary Material online. The most significant hit for all *HP1B* candidate genes from these eight species was *HP1B* in *D. melanogaster* and vice versa (e-values < 0.001, supplementary table S13, Supplementary Material online).

Species Phylogeny

We built a dated species phylogeny by combining multiple published trees (Wiegmann et al. 2011; Misof et al. 2014). We used clade-specific publications to obtain phylogenetic relationship and estimates of evolutionary timescales inside Diptera for Culicidae (Reidenbach et al. 2009; Freitas et al. 2015; Neafsey et al. 2015), for the Bactrocera (Zhang et al. 2010), for Drosophilidae (van der Linde et al. 2010; Morales-Hojas et al. 2011; Whiteman et al. 2012; Yang et al. 2012; Ometto et al. 2013; Zhou and Bachtrog 2015), and for Chironomus species (Milner 1997). To date nodes that

differed between publications, we used the median node time proposed on TimeTree.org (Hedges et al. 2015).

For focal species in Calliphoridae and Diopsidae clades, we did not find any dated nodes in the published literature. We used instead the software "r8s" (v.1.81) that allows divergence time estimation without a molecular clock but a fixed node time and assumed a constant rate of evolution (Sanderson 2003). We applied r8s software on orthologous sequences of the following highly conserved genes: *aats1*, *cad*, *pgd*, and *tpi* (Wiegmann et al. 2011).

Evolutionary Rate Estimation

To investigate domain-specific rates of evolution within HP1s, we calculated pairwise dN/dS using CODEML for ten selected sister species across the tree (supplementary table S8, Supplementary Material online). We retained only gene comparisons with limited synonymous divergence ($dS < 0.5$). We used the nonparametric Kruskal–Wallis test to assess significance between group distributions. We used post hoc test Dunn test (Zar 2010) to calculate adjusted *P* values between conditions following Benjamini–Hochberg adjustment (Benjamini and Hochberg 2000).

We used the CODEML program in PAML 4.1 (Yang 2007) to test for positive selection on each HP1 ortholog shared by at least three species in the Glossina clade. We first generated nucleotide alignments (described earlier) for each HP1 gene. We found no evidence of recombination for any two genes using the program GARD. Specifically, for full HP1s GARD investigated 1411 models in 4: 19 wall-clock time. The alignment contained 408 potential breakpoints, translating into a search space of 83, 436 models with up to two breakpoints, of which 1.69% was explored. For CSD-only HP1s, GARD examined 2, 079 models in 1: 19 wall-clock time. The alignment contained 133 potential breakpoints, translating into a search space of 8911 models with up to two breakpoints, of which 23.33% was explored by the algorithm.

We then estimated dN/dS and tested for evidence of positive selection by conducting model comparisons of M8 (allows dN/dS values >1) and M8a (dN/dS values from 0 to 1). To evaluate evidence of constraint acting on the ancient and young HP1s, we compared one ratio models (M0) where dN/dS was empirically estimated or dN/dS set to one. We report only maximum likelihood estimates based on an F3x4 model of codon frequencies (supplementary table S9, Supplementary Material online).

Transcriptomic Analysis and RT-qPCR

To investigate tissue-restricted expression patterns of newly defined HP1 genes, we selected a panel of species based on multiple criteria. First, we retained only those that encode at least six lineage-restricted HP1s. Second, we selected species that broadly sample Diptera evolution. Based on our criteria, we analyzed RNA-Seq data sets for three species: *Drosophila melanogaster*, *Drosophila miranda*, and *Teleopsis dalmanni* (supplementary table S10, Supplementary Material online). We included *D. melanogaster* as a positive control—previously, Levine et al. (2012) reported RT-PCR analysis of adult tissue-specific expression for HP1s encoded by this species. We mapped reads with HISAT2 (Kim et al. 2015) to the available respective genome assemblies (*D. melanogaster* Flybase r6.12, *D. miranda* DroMir 2.2 NCBI, *T. dalmanni* NCBI TSA GBBP00000000) then sorted and counted mapped reads with featureCounts (Liao et al. 2014). Finally, we calculated normalized expression for each HP1 gene with DESeq2 (Love et al. 2014). We compared expression across four tissues: female head, male head, testis, and ovary.

To investigate additional species with no publicly available germline and somatic transcriptomic data set but many lineage-restricted, divergent HP1 genes, we selected *Glossina morsitans* (gift of Aksoy Lab, Yale University) and *Scaptodrosophila lebanonensis* (from Cornell Drosophila Species Stock Center). We extracted RNA from dissected heads and reproductive tissues of males and females for each species. We dissected the tissues in PBS and stored them in RNAlater (Invitrogen) before extracting the RNA using mirVana miRNA Isolation Kit (Invitrogen) according to the manufacturer's instructions. After DNase-treating (TURBO DNase, Invitrogen) all samples, we prepared cDNA (SuperScript III, Invitrogen). We performed qPCR reactions using SYBR Green (Invitrogen) on an ABI 7000 Real Time PCR System, using following cycle parameters: 50°C for 2 min, 95°C for 2 min, 40 cycles of 95°C for 15s, 60°C for 30s. We also included a "Reverse Transcriptase-minus (RT-)" control of each sample to rule out genomic contamination. We confirmed that all primer pairs (supplementary table S14, Supplementary Material online) had similar amplification efficiencies using a dilution series of genomic DNA (data not shown). We also amplified multiple candidate reference genes (supplementary table S14, Supplementary Material online) and evaluated their expression stability across tissues using four different methods (Vandesompele et al. 2002; Andersen et al. 2004; Pfaffl et al. 2004; Silver et al. 2006) on the *RefFinder* website (Cotton EST Database, http://150.216.56.64/referencegene.php). We normalized the transcript level of each genes by the reference genes using the Delta-Cq method on Qbase+ (https://www.qbaseplus.com/; Biogazelle). To facilitate comparisons between tissues and genes, we scaled the values to the tissue type of highest expression for each gene.

Supplementary Material

Supplementary data are available at *Molecular Biology and Evolution* online.

Acknowledgments

The authors thank C. Leek for assistance with RT-qPCR and L. Kursel, G. Lee, and four anonymous reviewers for their comments on earlier versions of the manuscript. The authors are also grateful to the Aksoy lab (Yale University), especially G. Attardo and B. Weiss, for their generosity of time, tissue, and input. This work was supported by the National Institutes of Health (R00 GM107351 and R35 GM124684 to M.T.L.).

References

Allshire RC, Nimmo ER, Ekwall K, Javerzat JP, Cranston G. 1995. Mutations derepressing silent centromeric domains in fission yeast disrupt chromosome segregation. *Genes Dev*. 9(2):218–233.

Andersen CL, Jensen JL, Ørntoft TF. 2004. Normalization of real-time quantitative reverse transcription-PCR data: a model-based variance estimation approach to identify genes suited for normalization, applied to bladder and colon cancer data sets. *Cancer Res*. 64(15):5245–5250.

Anderson JA, Gilliland WD, Langley CH. 2009. Molecular population genetics and evolution of *Drosophila meiosis* genes. *Genetics* 181(1):177–185.

Assis R, Bachtrog D. 2013. Neofunctionalization of young duplicate genes in Drosophila. *Proc Natl Acad Sci U S A*. 110(43):17409–17414.

Assland R, Stewart F. 1995. The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. *Nucleic Acids Res*. 23(16):3168–3173.

Barbash DA, Awadalla P, Tarone AM. 2004. Functional divergence caused by ancient positive selection of a Drosophila hybrid incompatibility locus. *PLoS Biol*. 2(6):e142.

Bargues N, Lerat E. 2017. Evolutionary history of LTR-retrotransposons among 20 Drosophila species. *Mob DNA* 8:7.

Bayes JJ, Malik HS. 2009. Altered heterochromatin binding by a hybrid sterility protein in Drosophila sibling species. *Science* 326(5959):1538–1541.

Benjamini Y, Hochberg Y. 2000. On the adaptive control of the false discovery rate in multiple testing with independent statistics. *J Educ Behav Stat*. 25(1):60.

Bernard P, Maure JF, Partridge JF, Genier S, Javerzat JP, Allshire RC. 2001. Requirement of heterochromatin for cohesion at centromeres. *Science* 294(5551):2539–2542.

Bhalla SC. 1973. Sex-linked translocations, semisterility and linkage alterations in the mosquito *Aedes aegypti*. *Can J Genet Cytol*. 15(1):9–20.

Brideau NJ, Flores HA, Wang J, Maheshwari S, Wang X, Barbash DA. 2006. Two Dobzhansky-Muller genes interact to cause hybrid lethality in Drosophila. *Science* 314(5803):1292–1295.

Cazemajor M, Joly D, Montchamp-Moreau C. 2000. Sex-ratio meiotic drive in *Drosophila simulans* is related to equational nondisjunction of the Y chromosome. *Genetics* 154:229–236.

Chen S, Zhang YE, Long M. 2010. New genes in Drosophila quickly become essential. *Science* 330(6011):1682–1685.

Chen X-G, Jiang X, Gu J, Xu M, Wu Y, Deng Y, Zhang C, Bonizzoni M, Dermauw W, Vontas J. 2015. Genome sequence of the Asian Tiger mosquito, *Aedes albopictus*, reveals insights into its biology, genetics, and evolution. *Proc Natl Acad Sci U S A*. 112(44):E5907–E5915.

Chen YA, Stuwe E, Luo Y, Ninova M, Le Thomas A, Rozhavskaya E, Li S, Vempati S, Laver JD, Patel DJ, et al. 2016. Cutoff suppresses RNA polymerase II termination to ensure expression of piRNA precursors. *Mol Cell* 63(1):97–109.

Chubareva LA, Petrova NA, Kachvorian EA. 1996. The morphokaryotypic traits of 4 species of black flies (Diptera: simuliidae). *Parazitologiia* 30(1):3–12.

Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res*. 14(6):1188–1190.

Dimitri P, Junakovic N. 1999. Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Trends Genet*. 15(4):123–124.

Dion-Cote AM, Renaut S, Normandeau E, Bernatchez L. 2014. RNA-seq reveals transcriptomic shock involving transposable elements reactivation in hybrids of young lake whitefish species. *Mol Biol Evol*. 31(6):1640–1640.

Dorer DR, Henikoff S. 1994. Expansions of transgene repeats cause heterochromatin formation and gene silencing in Drosophila. *Cell* 77(7):993–1002.

Drinnenberg IA, deYoung D, Henikoff S, Malik HS. 2014. Recurrent loss of CenH3 is associated with independent transitions to holocentricity in insects. *Elife* 3:

Eissenberg JC, Elgin SCR. 2014. HP1a: a structural chromosomal protein regulating transcription. *Trends Genet*. 30(3):103–110.

Eissenberg JC, Reuter G. 2009. Cellular mechanism for targeting heterochromatin formation in Drosophila. *Int Rev Cell Mol Biol*. 273:1–47.

Elgin SCR. 1996. Heterochromatin and gene regulation in Drosophila. *Curr Opin Genet Dev*. 6(2):193–202.

Fang C, Schmitz L, Ferree PM. 2015. An unusually simple HP1 gene set in Hymenopteran insects. *Biochem Cell Biol*. 93(6):596–603.

Fanti L, Giovinazzo G, Berloco M, Pimpinelli S. 1998. The heterochromatin protein 1 prevents telomere fusions in Drosophila. *Mol Cell* 2(5):527–538.

Ferree PM, Barbash DA. 2009. Species-specific heterochromatin prevents mitotic chromosome segregation to cause hybrid lethality in Drosophila. *PLoS Biol*. 7(10):e1000234.

Freitas LA, Russo CAM, Voloch CM, Mutaquiha OCF, Marques LP, Schrago CG. 2015. Diversification of the genus Anopheles and a neotropical clade from the late Cretaceous. *PLoS One* 10(8):e0134462.

Gallach M. 2014. Recurrent turnover of chromosome-specific satellites in Drosophila. *Genome Biol Evol*. 6(6):1279–1286.

Gerbi SA. 1986. Unusual chromosome movements in sciarid flies. *Results Probl. Cell Differ*. 13:71–104.

Grewal SIS, Jia S. 2007. Heterochromatin revisited. *Nat Rev Genet*. 8(1):35–46.

Guindon S, Dufayard JF, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol*. 59(3):307–321.

Hall AB, Basu S, Jiang X, Qi Y, Timoshevskiy VA, Biedler JK, Sharakhova MV, Elahi R, Anderson MAE, Chen X-G, et al. 2015. Sex determination. A male-determining factor in the mosquito *Aedes aegypti*. *Science* 348(6240):1268–1270.

Han MV, Demuth JP, McGrath CL, Casola C, Hahn MW. 2009. Adaptive evolution of young gene duplicates in mammals. *Genome Res*. 19(5):859–867.

Hedges SB, Marin J, Suleski M, Paymer M, Kumar S. 2015. Tree of life reveals clock-like speciation and diversification. *Mol Biol Evol*. 32(4):835–845.

Helleu Q, Gérard PR, Dubruille R, Ogereau D, Prud'homme B, Loppin B, Montchamp-Moreau C. 2016. Rapid evolution of a Y-chromosome heterochromatin protein underlies sex chromosome meiotic drive. *Proc Natl Acad Sci U S A*. 113(15):4110–4115.

Henikoff S. 2000. Heterochromatin function in complex genomes. *Biochim Biophys Acta* 1470(1):O1–O8.

Henikoff S, Ahmad K, Malik HS. 2001. The centromere paradox: stable inheritance with rapidly evolving DNA. *Science* 293(5532):1098–1102.

Imbeault M, Helleboid PY, Trono D. 2017. KRAB zinc-finger proteins contribute to the evolution of gene regulatory networks. *Nature* 543(7646):550–554.

Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. *Nature* 516(7530):242–245.

Jacobs SA, Khorasanizadeh S. 2002. Structure of HP1 chromodomain bound to a lysine 9-methylated histone H3 tail. *Science* 295(5562):2080–2083.

Jagannathan M, Warsinger-Pepe N, Watase GJ, Yamashita YM. 2017. Comparative analysis of satellite DNA in the *Drosophila melanogaster* species complex. *G3 (Bethesda)* 7(2):693–704.

James TC, Eissenberg JC, Craig C, Dietrich V, Hobson A, Elgin SC. 1989. Distribution patterns of HP1, a heterochromatin-associated nonhistone chromosomal protein of Drosophila. *Eur J Cell Biol*. 50(1):170–180.

James TC, Elgin SC. 1986. Identification of a nonhistone chromosomal protein associated with heterochromatin in *Drosophila melanogaster* and its gene. *Mol Cell Biol*. 6(11):3862–3872.

Jiang X, Assis R. 2017. Natural selection drives rapid functional evolution of young Drosophila duplicate genes. *Mol Biol Evol.* 34(12):3089–3098.
Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res.* 20(10):1313–1326.
Kamm A, Galasso I, Schmidt T, Heslop-Harrison JS. 1995. Analysis of a repetitive DNA family from Arabidopsis-Arenosa and relationships between Arabidopsis species. *Plant Mol Biol.* 27(5):853–862.
Kanizay LB, Albert PS, Birchler JA, Dawe RK. 2013. Intragenomic conflict between the two major knob repeats of maize. *Genetics* 194(1):81.
Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol*. 30(4):772–780.
Kearse M, Moir R, Wilson A, Stones-Havas S, Cheung M, Sturrock S, Buxton S, Cooper A, Markowitz S, Duran C, et al. 2012. Geneious basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* 28(12):1647–1649.
Kelleher ES, Edelman NB, Barbash DA. 2012. Drosophila interspecific hybrids phenocopy piRNA-pathway mutants. *PLoS Biol*. 10(11):e1001428.
Kelley JL, Peyton JT, Fiston-Lavier AS, Teets NM, Yee MC, Johnston JS, Bustamante CD, Lee RE, Denlinger DL. 2014. Compact genome of the Antarctic midge is likely an adaptation to an extreme environment. *Nat Commun.* 5:4611.
Kellum R, Alberts BM. 1995. Heterochromatin protein-1 is required for correct chromosome segregation in Drosophila embryos. *J Cell Sci*. 108:1419–1431.
Kidwell MG. 2002. Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115(1):49–63.
Kim D, Langmead B, Salzberg SL. 2015. HISAT: a fast spliced aligner with low memory requirements. *Nat Methods* 12(4):357–360.
Klattenhoff C, Xi H, Li C, Lee S, Xu J, Khurana JS, Zhang F, Schultz N, Koppetsch BS, Nowosielska A, et al. 2009. The Drosophila HP1 homolog Rhino is required for transposon silencing and piRNA production by dual-strand clusters. *Cell* 138(6):1137–1149.
Kondo S, Vedanayagam J, Mohammed J, Eizadshenass S, Kan L, Pang N, Aradhya R, Siepel A, Steinhauer J, Lai EC. 2017. New genes often acquire male-specific functions but rarely become essential in Drosophila. *Genes Dev.* 31(18):1841–1846.
Kosugi S, Hasebe M, Matsumura N, Takashima H, Miyamoto-Sato E, Tomita M, Yanagawa H. 2009. Six classes of nuclear localization signals specific to different binding grooves of importin alpha. *J Biol Chem.* 284(1):478–485.
Kursel LE, Malik HS. 2017. Recurrent gene duplication leads to diverse repertoires of centromeric histones in Drosophila species. *Mol Biol Evol*. 34(6):1445–1462.
Larracuente AM. 2014. The organization and evolution of the Responder satellite in species of the *Drosophila melanogaster* group: dynamic evolution of a target of meiotic drive. *BMC Evol Biol.* 14:233.
Le HD, Donaldson KM, Cook KR, Karpen GH. 2004. A high proportion of genes involved in position effect variegation also affect chromosome inheritance. *Chromosoma* 112(6):269–276.
Lee YCG, Leek C, Levine MT. 2017. Recurrent innovation at genes required for telomere integrity in Drosophila. *Mol Biol Evol*. 34(2):467–482.
Lerat E, Burlet N, Biémont C, Vieira C. 2011. Comparative analysis of transposable elements in the melanogaster subgroup sequenced genomes. *Gene* 473(2):100–109.
Levine MT, McCoy C, Vermaak D, Lee YCG, Hiatt MA, Matsen FA, Malik HS. 2012. Phylogenomic analysis reveals dynamic evolutionary history of the Drosophila heterochromatin protein 1 (HP1) gene family. *PLoS Genet.* 8(6):e1002729.
Levine MT, Vander Wende HM, Hsieh E, Baker EP, Malik HS. 2016. Recurrent gene duplication diversifies genome defense repertoire in Drosophila. *Mol Biol Evol*. 33(7):1641–1653.
Levine MT, Vander Wende HM, Malik HS. 2015. Mitotic fidelity requires transgenerational action of a testis-restricted HP1. *Elife* 4:e07378.
Lewis SH, Salmela H, Obbard DJ. 2016. Duplication and diversification of Dipteran Argonaute genes, and the evolutionary divergence of Piwi and Aubergine. *Genome Biol Evol*. 8(3):507–518.
Lewis SH, Webster CL, Salmela H, Obbard DJ. 2016. Repeated duplication of Argonaute2 is associated with strong selection and testis specialization in Drosophila. *Genetics* 204(2):757–769.
Li Y, Jing XA, Aldrich JC, Clifford C, Chen J, Akbari OS, Ferree PM. 2017. Unique sequence organization and small RNA expression of a "selfish" B chromosome. Chromosoma. 126(6):753–768.
Liao Y, Smyth GK, Shi W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* 30(7):923–930.
Lomberk G, Wallrath L, Urrutia R. 2006. The heterochromatin protein 1 family. *Genome Biol*. 7(7):228.
Lopez-Maestre H, Carnelossi EAG, Lacroix V, Burlet N, Mugat B, Chambeyron S, Carareto CMA, Vieira C. 2017. Identification of misexpressed genetic elements in hybrids between Drosophila-related species. *Sci Rep*. 7:40618.
Love MI, Huber W, Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol*. 15(12):550.
Mahajan S, Bachtrog D. 2017. Convergent evolution of Y chromosome gene content in flies. *Nat Commun.* 8(1):785.
Maheshwari S, Ishii T, Brown CT, Houben A, Comai L. 2017. Centromere location in Arabidopsis is unaltered by extreme divergence in CENH3 protein sequence. *Genome Res*. 27(3):471–478.
Marchler-Bauer A, Derbyshire MK, Gonzales NR, Lu S, Chitsaz F, Geer LY, Geer RC, He J, Gwadz M, Hurwitz DI, et al. 2015. CDD: nCBI's conserved domain database. *Nucleic Acids Res.* 43(D1):D222–D226.
Meehan RR, Kao C-F, Pennings S. 2003. HP1 binding to native chromatin in vitro is determined by the hinge region and not by the chromodomain. *EMBO J*. 22(12):3164–3174.
Milner A. 1997. Book review: the chironomidae: biology and ecology of non-biting midges. London: Chapman and Hall. p. 572.
Mishima Y, Watanabe M, Kawakami T, Jayasinghe CD, Otani J, Kikugawa Y, Shirakawa M, Kimura H, Nishimura O, Aimoto S, et al. 2013. Hinge and chromoshadow of HP1alpha participate in recognition of K9 methylated histone H3 in nucleosomes. *J Mol Biol*. 425(1):54–70.
Misof B, Liu S, Meusemann K, Peters RS, Donath A, Mayer C, Frandsen PB, Ware J, Flouri T, Beutel RG, et al. 2014. Phylogenomics resolves the timing and pattern of insect evolution. *Science* 346(6210):763–767.
Mohn F, Sienski G, Handler D, Brennecke J. 2014. The rhino-deadlock-cutoff complex licenses noncanonical transcription of dual-strand piRNA clusters in Drosophila. *Cell* 157(6):1364–1379.
Montchamp-Moreau C, Ogereau D, Chaminade N, Colard A, Aulard S. 2006. Organization of the sex-ratio meiotic drive region in *Drosophila simulans*. *Genetics* 174(3):1365–1371.
Morales-Hojas R, Reis M, Vieira CP, Vieira J. 2011. Resolving the phylogenetic relationships and evolutionary history of the *Drosophila virilis* group using multilocus data. *Mol Phylogenet Evol*. 60(2):249–258.
Muchardt C, Guilleme M, Seeler JS, Trouche D, Dejean A, Yaniv M. 2002. Coordinated methyl and RNA binding is required for heterochromatin localization of mammalian HP1. *EMBO Rep*. 3(10):975–981.
Neafsey DE, Waterhouse RM, Abai MR, Aganezov SS, Alekseyev MA, Allen JE, Amon J, Arcà B, Arensburger P, Artemov G, et al. 2015. Mosquito genomics. Highly evolvable malaria vectors: the genomes of 16 Anopheles mosquitoes. *Science* 347(6217):1258522.
Nielsen PR, Nietlispach D, Mott HR, Callaghan J, Bannister A, Kouzarides T, Murzin AG, Murzina NV, Laue ED. 2002. Structure of the HP1 chromodomain bound to histone H3 methylated at lysine 9. *Nature* 416(6876):103–107.
Nowick K, Hamilton AT, Zhang H, Stubbs L. 2010. Rapid sequence and expression divergence suggest selection for novel function in primate-specific KRAB-ZNF genes. *Mol Biol Evol*. 27(11):2606–2617.
Obbard DJ, Jiggins FM, Halligan DL, Little TJ. 2006. Natural selection drives extremely rapid evolution in antiviral RNAi genes. *Curr Biol.* 16(6):580–585.

Ometto L, Cestaro A, Ramasamy S, Grassi A, Revadi S, Siozios S, Moretto M, Fontana P, Varotto C, Pisani D, et al. 2013. Linking genomics and ecology to investigate the complex evolution of an invasive Drosophila pest. *Genome Biol Evol*. 5(4):745–757.

Papa F, Windbichler N, Waterhouse RM, Cagnetti A, D'Amato R, Persampieri T, Lawniczak MKN, Nolan T, Papathanos PA. 2017. Rapid evolution of female-biased genes among four species of Anopheles malaria mosquitoes. *Genome Res*. 27(9):1536–1548.

Paredez AR, Assaf ZJ, Sept D, Timofejeva L, Dawson SC, Wang CJ, Cande WZ. 2011. An actin cytoskeleton with evolutionarily conserved functions in the absence of canonical actin-binding proteins. *Proc Natl Acad Sci U S A*. 108(15):6151–6156.

Parhad SS, Tu S, Weng Z, Theurkauf WE. 2017. Adaptive evolution leads to cross-species incompatibility in the piRNA transposon silencing machinery. *Dev Cell* 43(1):60–70 e65.

Paro R, Hogness DS. 1991. The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of Drosophila. *Proc Natl Acad Sci U S A*. 88(1):263–267.

Pfaffl MW, Tichopad A, Prgomet C, Neuvians TP. 2004. Determination of stable housekeeping genes, differentially regulated target genes and sample integrity: bestKeeper–Excel-based tool using pair-wise correlations. *Biotechnol Lett*. 26(6):509–515.

Phadnis N, Orr HA. 2009. A single gene causes both male sterility and segregation distortion in Drosophila hybrids. *Science* 323(5912):376–379.

Reidenbach KR, Cook S, Bertone MA, Harbach RE, Wiegmann BM, Besansky NJ. 2009. Phylogenetic analysis and temporal diversification of mosquitoes (Diptera: culicidae) based on nuclear genes and morphology. *BMC Evol Biol*. 9:298.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. 2012. MrBayes 3.2: efficient Bayesian phylogenetic inference and model choice across a large model space. *Syst Biol*. 61(3):539–542.

Rosenfeld JA, Foox J, DeSalle R. 2016. Insect genome content phylogeny and functional annotation of core insect genomes. *Mol Phylogenet Evol*. 97:224–232.

Ross BD, Rosin L, Thomae AW, Hiatt MA, Vermaak D, de la Cruz AFA, Imhof A, Mellone BG, Malik HS. 2013. Stepwise evolution of essential centromere function in a *Drosophila neogene*. *Science* 340(6137):1211–1214.

Sanderson MJ. 2003. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* 19(2):301–302.

Silver N, Best S, Jiang J, Thein SL. 2006. Selection of housekeeping genes for gene expression studies in human reticulocytes using real-time PCR. *BMC Mol Biol*. 7:33.

Smith CD, Shu S, Mungall CJ, Karpen GH. 2007. The Release 5.1 annotation of *Drosophila melanogaster* heterochromatin. *Science* 316TcEyQ(5831):1586–1591.

Smothers JF, Henikoff S. 2000. The HP1 chromo shadow domain binds a consensus peptide pentamer. *Curr Biol*. 10(1):27–30.

Smothers JF, Henikoff S. 2001. The hinge and chromo shadow domain impart distinct targeting of HP1-like proteins. *Mol Cell Biol*. 21(7):2555–2569.

Stuart JJ, Chen MS, Shukle R, Harris MO. 2012. Gall Midges (Hessian flies) as plant pathogens. *Annu Rev Phytopathol*. 50 50:339–357.

Swanson WJ, Vacquier VD. 2002. The rapid evolution of reproductive proteins. *Nat Rev Genet*. 3(2):137–144.

Thomas JH, Schneider S. 2011. Coevolution of retroelements and tandem zinc finger genes. *Genome Res*. 21(11):1800–1812.

Traut W. 1994. Sex determination in the fly *Megaselia scalaris*, a model system for primary steps of sex chromosome evolution. *Genetics* 136(3):1097–1104.

van der Linde K, Houle D, Spicer GS, Steppan SJ. 2010. A supermatrix-based molecular phylogeny of the family Drosophilidae. *Genet Res*. 92(1):25–38.

Vandesompele J, De Preter K, Pattyn F, Poppe B, Van Roy N, De Paepe A, Speleman F. 2002. Accurate normalization of real-time quantitative RT-PCR data by geometric averaging of multiple internal control genes. *Genome Biol*. 3(7):RESEARCH0034.

Vermaak D, Bayes JJ, Malik HS. 2009. A surrogate approach to study the evolution of noncoding DNA elements that organize eukaryotic genomes. *J Hered*. 100(5):624–636.

Vermaak D, Henikoff S, Malik HS. 2005. Positive selection drives the evolution of rhino, a member of the heterochromatin protein 1 family in Drosophila. *PLoS Genet*. 1(1):96–108.

Vermaak D, Malik HS. 2009. Multiple roles for heterochromatin protein 1 genes in Drosophila. *Annu Rev Genet*. 43:467–492.

Vicoso B, Bachtrog D. 2013. Reversal of an ancient sex chromosome to an autosome in Drosophila. *Nature* 499(7458):332.

Vicoso B, Bachtrog D. 2015. Numerous transitions of sex chromosomes in Diptera. *PLoS Biol*. 13(4):e1002078.

Warnes ML, Maudlin I. 1992. An analysis of supernumerary or B-chromosomes of wild and laboratory strains of *Glossina morsitans* morsitans. *Med Vet Entomol*. 6(2):175–176.

Whiteman NK, Gloss AD, Sackton TB, Groen SC, Humphrey PT, Lapoint RT, Sønderby IE, Halkier BA, Kocks C, Ausubel FM, et al. 2012. Genes involved in the evolution of herbivory by a leaf-mining, Drosophilid fly. *Genome Biol Evol*. 4(9):900–916.

Wiegmann BM, Trautwein MD, Winkler IS, Barr NB, Kim J-W, Lambkin C, Bertone MA, Cassel BK, Bayless KM, Heimberg AM, et al. 2011. Episodic radiations in the fly tree of life. *Proc Natl Acad Sci U S A*. 108(14):5690–5695.

Yang Y, Hou Z-C, Qian Y-H, Kang H, Zeng Q-T. 2012. Increasing the data size to accurately reconstruct the phylogenetic relationships between nine subgroups of the *Drosophila melanogaster* species group (Drosophilidae, Diptera). *Mol Phylogenet Evol*. 62(1):214–223.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol*. 24(8):1586–1591.

Zar JH. 2010. Biostatistical analysis. London, UK: Pearson New International Division.

Zhang B, Liu YH, Wu WX, Le Wang Z. 2010. Molecular phylogeny of Bactrocera species (Diptera: tephritidae: dacini) inferred from mitochondrial sequences of 16S rDNA and COI sequences. *Fla Entomol*. 93(3):369–377.

Zhou Q, Bachtrog D. 2012. Sex-specific adaptation drives early sex chromosome evolution in Drosophila. *Science* 337(6092):341–345.

Zhou Q, Bachtrog D. 2015. Ancestral chromatin configuration constrains chromatin evolution on differentiating sex chromosomes in Drosophila. *PLoS Genet*. 11(6):e1005331.