Functional and Evolutionary Analyses of Germline Stem Cell Regulating Genes across Select *Drosophila* and Outgroup Species

Luke Arnce
Aquadro Lab
Field of Genetics, Genomics, and Development
Cornell University

6/6/2022

Abstract

Germline stem cell (GSC) regulating genes are developmentally critical for reproduction in *Drosophila.* Recent comparative analysis identified two GSC genes determined to be essential in *D. melanogaster*, *bag of marbles* (*bam*) and *Female sterile (1) Yb* (*Yb*), are functionally flexible across closely related *Drosophila* species. It is unknown whether this observed flexibility is exceptional in essential GSC genes or a common phenomenon. The growing density of high quality full genome sequences within the *Drosophila* genus makes comparative analysis of GSC genes across *Drosophila* (and closely related outgroups) possible. This project aims to characterize functional flexibility at GSC genes in *Drosophila* by further defining the functional evolutionary history of *bam,* broadly evaluating the extent and characteristics of GSC gene network flexibility, and identifying potential points of functional divergence at GSC genes across species. To achieve these goals, I will generate *bam* nulls, generate a GSC gene ortholog dataset, and conduct analyses of selection for GSC regulating genes.

Specific Aims

Often in evolutionary analyses of diverse species, orthologs are assumed to have conserved function despite only having experimental functional validation in a model species. This comes from the "ortholog conjecture" or standard model of phylogenomics which assumes orthologs generally retain their ancestral function. This assumption, at a minimum, is not universally true considering *bam* and *Yb,* but, historically, comparative analysis of ortholog functional conservation has been difficult to conduct. Currently, the number and quality of publicly available fully sequenced genomes makes comparative analysis of ortholog functional conservation feasible.

I will create an ortholog dataset including 39 *Drosophila* species and two closely related outgroup species selected for having high quality full genome sequences and covering the evolutionary scope of the genus. I will include 366 GSC genes determined essential for fertility in *D. melanogaster*. By using genes with a known essential function in *D. melanogaster*, I can use absences of orthologs for these genes in other species to identify functional divergence. Additional relevant functional data and analyses of selection for GSC genes will add relevant details to further parse our findings, but I will first evaluate the broad null hypothesis: GSC gene orthologs are almost completely conserved across all included species. I will also further define the evolutionary history of *bam* specifically. Generated *bam* nulls have demonstrated that *bam*'s essential "switch" function for GSC differentiation is not universally conserved within the *D. melanogaster* species group. *Bam* is present across the genus *Drosophila* so I can not identify functional divergence by gene absence, but I can generate nulls in *Drosophila* species beyond the *D. melanogaster* group to further understand *bam*'s functional evolutionary history. This will specifically evaluate the narrower hypothesis: *bam*'s critical GSC "switch" function is specific to the *D. melanogaster* species group. Together, these strategies provide a powerful combination of a broad strategy identifying functional flexibility with lower sensitivity and a specific strategy that can precisely determine gene function. To directly test these hypotheses, I propose three related, but independent aims:

Aim 1: Assess the role of *bam* in GSC differentiation in *D. pseudoobscura* and *D. americana*

I will generate *bam* nulls in *D. pseudoobscura* and *D. americana* and conduct cytolological and fertility assays with the generated nulls. These two species represent two major divergent lineages beyond the *D. melanogaster* species group, and nulls in these species will help define the direction of the evolution of bam's essential GSC differentiation function.

Aim 2: Generate a GSC gene ortholog dataset and merge with existing relevant datasets

I will catalog the orthology of the 366 GSC genes identified by an RNAi screen in *D. melanogaster* for 39 *Drosophila* and 2 outgroup species. I will also integrate existing datasets on genetic interactors, physical interactors, and function for included GSC genes to assess the extent and characteristics of GSC gene functional flexibility.

Aim 3: Analysis of selection

I will conduct McDonald-Kreitman tests at *bam* in *Drosophila* species with available polymorphism data to test for signals of recurrent positive selection. I will also conduct divergence analysis of GSC genes using phylogenetic analysis by maximum likelihood (PAML) to test for signals long-term positive selection. These analyses will help further define GSC regulatory flexibility and identify GSC genes of interest for future functional manipulation.

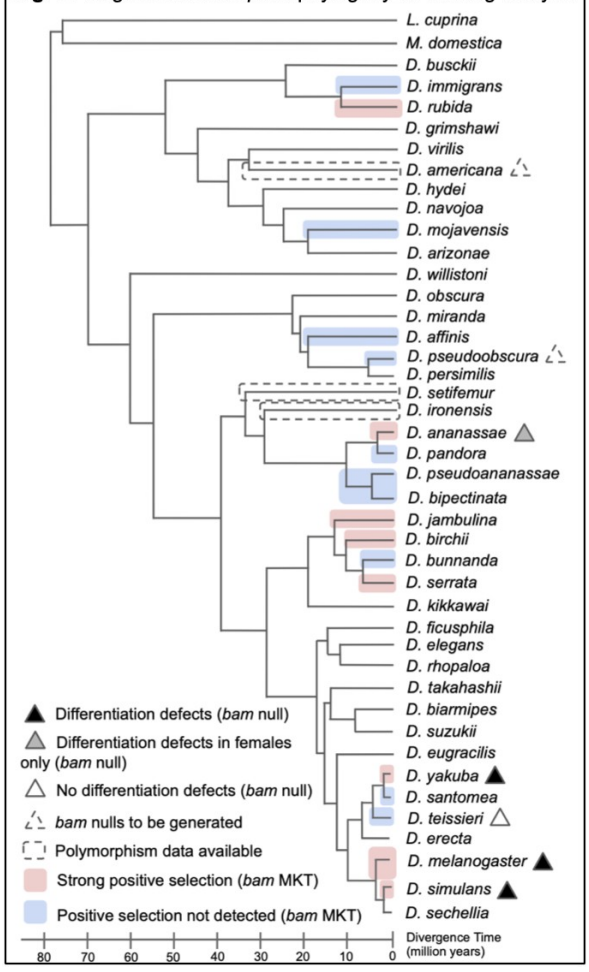
Background and Significance

GSC regulating genes are critical for reproduction in *Drosophila* and have historically been assumed to be functionally conserved given that slight alterations could lead to disruption of critical functions causing sterility. However, a range of evidence including nulls, tests of selection, and ortholog absence have highlighted functional divergence at key GSC genes (*Yb* and *bam*) within the *Drosophila* genus. *Bam* in particular has been directly functionally assessed in several species. *Bam* acts as the switch gene for GSC differentiation for *D. melanogaster* females and is necessary for terminal differentiation of spermatogonia in males (1). In females, *bam* is repressed in GSCs and *bam* expression causes differentiation by binding to several protein partners including *benign gonial cell neoplasm* (*bgcn*) in order to repress the production of self-renewal factors *Nanos* and *eIF4a*. The resultant differentiating cystoblast undergoes several mitotic divisions (2-5). Simultaneously, *bam*

concentrates at the fusome, which connects the cysts, and *bam* and *bgcn* function together to regulate the timing of mitotic divisions between cells. In males, *bam* is expressed in GSCs, and as differentiation continues, expression increases (1,6-8). Once *bam* expression reaches a threshold in the early spermatogonia, it binds to b*gcn* and *tumorous testis* (*tut*) (9,10,11). Binding represses *mei-P26* and ends proliferation, triggering terminal differentiation and beginning meiosis (2-5). Loss of *bam* function prevents differentiation in both sexes and causes over-proliferation of GSCs in females and spermatogonia in males, leading to tumors and sterility in *D. melanogaster* (12,13).

Recent analysis has determined that *bam* is rapidly evolving due to positive selection and that *bam* function is not conserved even among closely related species (14,15,41,60,67). B*am* was tested for signals of selection in several species within the *D. melanogaster* species group and a few more distant lineages. Results showed heterogeneous, lineage-specific signals of positive selection (Figure 1). Results for *bam* nulls generated in five species within the *D. melanogaster* species group highlighted functional divergence at *bam* within the *D. melanogaster* species group (60). Homozygous *bam* nulls show three distinct resultant phenotypes: Differentiation defects in both sexes (*D. melanogaster*, *D. simulans*, *D. yakuba*), differentiation defects in females only (*D. ananassae*), and no differentiation defects (*D. teissieri*). These distinct phenotypes highlight *bam* as an example of developmental systems drift (DSD) or divergence in genetic systems that underpin a conserved phenotype (64,65). In the lineages where *bam* is not carrying out its function as a key differentiation gene, these species are still fertile. This means that some other gene or combination of genes must be carrying out the function since GSC differentiation is essential for producing gametes and ultimately reproduction. The lineages in which *bam* nulls show differentiation defects also show signals of positive selection at *bam* while in the lineage where *bam* nulls showed no differentiation defects, signals of positive selection were not detected at *bam*. This could be the result of positive selection being driven by functional refinement or germline conflicts related to *bam*'s essential role in GSC differentiation. Further analysis of *bam* more broadly across *Drosophila* will be necessary to develop a clearer understanding of *bam*'s functional evolutionary history.

Fig. 1 Full genome *Drosophila* phylogeny for otholog analysis



When considering the evolutionary history of *bam*, there are two main possibilities. First, *bam* was a novel gene in the common ancestor of the *Drosophila* genus and, over time, has functionally diversified across the genus, with its role as the switch gene for GSC differentiation recently acquired in the lineage leading to the *D. melanogaster* species group. Alternatively, *bam* could have quickly evolved the switch function in *Drosophila*, but was subsequently altered by lineage-specific germline conflicts. To tease out which evolutionary history is more plausible for *bam*, I will generate *bam* nulls for *D. pseudoobscura* and *D. americana*, both species within the *Drosophila* genus but outside the *D. melanogaster* species group, to determine whether *bam* is necessary for gametogenesis in these more distant species.

Beyond *bam,* there has been significant DNA polymorphism and divergence analysis of some GSC genes within the D. *melanogaster* group highlighting differential signals of adaptive evolution across species (67). There is also already evidence to suggest that essential GSC genes may be gained or lost in related species. *Yb,* which is expressed in cap cells in the germarium, plays an essential role in GSC maintenance and transposable element regulation in *D. melanogaster* (61). In *D. melanogaster* and *D. simulans*, *Yb* shows strong signals of amino acid diversification (62), *Yb* null females are sterile in *D. melanogaster*

(63), and unpublished data in our lab shows the *Yb* ortholog is not in *D. pseudoobscura*, *D. persimilis*, or *D. miranda* lineages. This demonstrates functional divergence for *Yb* and some flexibility in GSC regulation across *Drosophila*.

The majority of functional genetic studies focus on a single species, so there is rarely functional confirmation for orthologs. Still, orthologs are largely assumed to be functionally conserved without interrogation of whether this is the case (68). Assessing the degree and mechanics of ortholog functional conservation is fundamental to understanding how gene networks evolve and could significantly shape interpretations of results for comparative analysis. Degree of ortholog functional conservation, for example, complicates the generation of hypotheses related to potential evolutionary drivers and genetic evolutionary history since orthologous genes are unlikely to be experiencing the same evolutionary pressures in each species if function is not conserved.

To evaluate ortholog functional conservation across GSC regulating genes, I will use ortholog absence, as was done for *Yb*. By starting with a baseline of 366 GSC genes known to be essential in *D. melanogaster* (26), an absence of one of those genes in another species demonstrates developmental systems drift between those lineages. I can build a GSC gene ortholog dataset using available high quality sequences, generate predicted absences for orthologs across the 39 *Drosophila* and two outgroup species, and validate ortholog absence using PCR and sequencing. Present orthologous genes can also highlight points of functional flexibility (*bam*). To capture a degree of the potential functional flexibility for present orthologs, I will test GSC genes for signals of positive selection. This strategy will identify points of known functional flexibility using gene absence and identify potential points of functional flexibility in present orthologs that show signs of adaptive evolution. I will also incorporate relevant functional data for all 366 GSC regulating genes including physical interactors, genetic interactors, annotated function, and defect type as characterized in *D. melanogaster* (70). This data will be considered alongside selection results to enable a deeper understanding of ortholog functional flexibility.

Research and Design Methods

Aim 1: Assess the role of *bam* in GSC differentiation in *D. pseudoobscura* and *D. americana*
Rationale: Recently generated *bam* nulls in *D. melanogaster*, *D. simulans*, and *D. yakuba* demonstrate *bam* is necessary for male and female germline stem cell differentiation and fertility in those species. *D. ananassae* male and *D. teissieri* male and female *bam* nulls demonstrate *bam* is unnecessary for GSC differentiation and fertility (Fig. 1). Recent results also indicate bursts of positive selection at *bam* within the *D. melanogaster* species group, including in *D. melanogaster*, *D. simulans*, *D. yakuba*, and *D. santomea*. *D. teissieri* did not show evidence of positive selection. These findings prompt more direct evaluation of the evolutionary origin of *bam's* essential GSC role and whether the evolution of new *bam* functions could contribute to the bursts of positive selection in the aforementioned species.

To interrogate these topics, I will generate *bam* null alleles in *D. pseudoobscura* and *D. americana*. These species represent major, more divergent outgroups to the *D. melanogaster* species group within the *Drosophila* genus and have been successfully edited with CRISPR/Cas9 (20,21). Creating these nulls will add broader evolutionary scope to our current array of generated *bam* nulls. Results from cytological and fertility assays will help evaluate whether *bam*'s essential role in GSC differentiation is basal to all *Drosophila* species and lost in specific lineages or whether *bam*'s critical role is a gained function within the *D. melanogaster* species group. Resultant null phenotypes will also provide a broader context for understanding the relationship between *bam* function and positive selection.

A: Generate *bam* nulls in *D. pseudoobscura* and *D. americana*
Null hypothesis: *D. pseudoobscura* and *D. americana* null mutants will both display wildtype phenotypes
Methodology: I will use CRISPR/Cas9 to introduce a 3xP3-YFP or a 3xP3-DsRed gene cassette into the first exon of *bam* in *D. pseudoobscura* and *D. americana*, thereby disrupting the *bam* coding

sequence and introducing a premature termination codon. I chose this method instead of a full deletion of *bam* due to past success using this strategy (60) and concerns that large deletions might disrupt regulation of adjacent genes. *Bam* does share a 3' UTR with an adjacent gene. This will generate an allele that is trackable by eye color, which is necessary for non-*melanogaster* species since balancer chromosomes to maintain alleles that cause sterility are unavailable in these non-*melanogaster* species. I will cross the 3xP3-YFP line to the 3xP3-DsRed line and select flies with both DsRed and YFP positive eyes. This scheme also allows us to use the same cross to assay the heterozygous and wildtype *bam* siblings.

So far, I have used the NCBI database to obtain nucelotide sequence information for designing constructs in *D. pseudoobscura* (assembly UCI_Dpse_MV25) and *D. americana* (assembly ASM1815291v1). I used geneious for all cloning design. I used the NEB Q5 High Fidelity 2X master mix to generate PCR products. Then I gel extracted and purified PCR products using the NEB Monarch DNA gel extraction kit. IDT primers were used for PCR, sequencing and cloning. I generated donor plasmids for the 3xP3-DsRed and 3xP3-YFP *bam* distruption lines using the NEB HiFi Assembly Cloning kit into the pHD-attP-DsRed vector from flyCRISPR (Gratz et al. 2014) as follows: I amplifed 1.5 kb homology arms from genomic DNA of the appropriate species stock flanking the insertion site for 3xP3-DsRed or 3xP3-YFP. 3xP3-DsRed was amplified from the pHD-attP-DsRed plasmid and YFP was amplified from the *D. simulans* nos-Cas9 line, a gift from David Stern. I then gel extracted the two homology arms and the appropriate 3xP3 marker, purified them, and assembled them into the pHD vector backbone using the manufacturer's protocol. Next, I prepped and purified plasmids for embryo injections with Qiagen plasmid plus midi-prep kit. Plasmid sequences were confirmed with sequencing (Plasmidsaurus). This cloning procedure has been completed for all necessary *D. pseudoobscura* plasmids, and is ongoing for *D. americana* plasmids. gRNAs without off target effects have been chosen and 1-3 gRNAs will be used per injection to increase the chances of a successful CRISPR event. Injections will be carried out by Genetivision and I will then screen resultant *bam* disruption lines for eye color. Mutants will be backcrossed to the stock lines for three generations and all mutants will be maintained as heterozygous stock. I will confirm all CRISPR insertions by Sanger sequencing. I expect the introduced *bam* disruption cassettes will trigger nonsense mediated mRNA decay (NMD) resulting in a loss of function allele (66). Since the only available *bam* antibody is weakly cross-reactive in *D. melanogaster* and *D. simulans*, I cannot directly demonstrate the absence of *bam* protein in the tested species. I will use RT-qPCR to determine whether the 3xP3-DsRed and 3xP3 YFP alleles are expressed in lower levels consistent with NMD in comparison to wildtype *bam* alleles. I will select DsRed and YFP positive flies and cross them to generate homozygous *bam* disruption flies which I will evaluate for GSC differentiation and fertility defects using immunostaining of ovaries and testes and fertility assays as done in previous *bam* null evaluations (67).

Expected outcomes and interpretations: If *D. pseudoobscura* and *D. americana* homozygous null mutants have wildtype *bam* phenotypes, failing to reject the null hypothesis, this would provide evidence that *bam*'s known *D. melanogaster* GSC function is novel to the *D. melanogaster* species group. If the homozygous null mutants exhibit GSC daughter differentiation defects in one or both species, I will conclude that *bam*'s critical GSC function is not novel to the *D. melanogaster* species group, and its evolutionary origin occurred prior to the group's formation.

Caveats and future directions: If I am unsuccessful in generating nulls using the *bam* disruption strategy, I will generate null deletion lines using the methodology outlined in Kanca et. al (24). A potentially interesting future direction would be to generate nulls in more distant species that signal selection at *bam*, like *D. rubida*, and repeat the same assessments of fertility and cytology (67).

Aim 2: Generate a GSC gene ortholog dataset and merge with existing relevant datasets

Rationale: DSD is known to occur in two GSC genes (*bam* and *Yb)*, but the frequency of DSD and where it tends to occur in GSC genes is unknown. I can begin to interrogate these topics by identifying absences of GSC genes essential in *D. melanogaster.* The GSC gene ortholog dataset will be able identify orthologs and absent GSC genes across the 39 *Drosophila* species, spanning the evolutionary scope of the genus and two closely related outgroup species. This will enable us to make informed

predictions regarding the evolutionary history of many GSC genes as well as evaluate the extent and characteristics of GSC gene network flexibility.

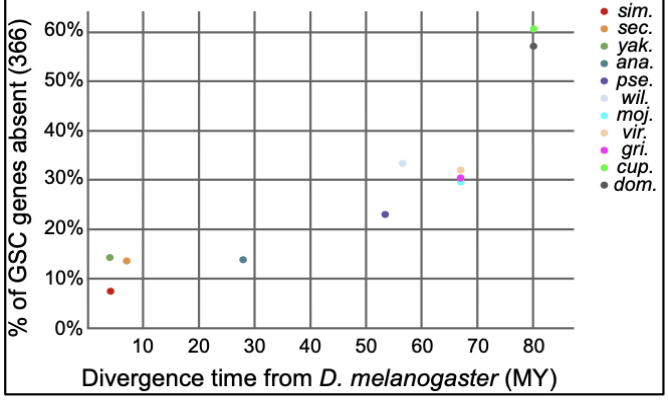
A: Ortholog dataset construction and validation
Null hypothesis: Orthologs of *D. melanogaster* GSC genes are almost all present across the genus *Drosophila* and into included outgroup species
Methodology: I have assessed predicted orthology for GSC genes in 10 *Drosophila* species spanning across the phylogeny and two closely related outgroup species using Ensembl compara (27). This tool catalogs relevant information about each ortholog including sequence alignment, target percent ID (percentage of orthologous sequence matching the *Drosophila melanogaster* sequence), query percent ID (percentage of *Drosophila melanogaster* sequence matching the orthologous sequence), gene order conservation score (evaluating synteny), and high or low ortholog confidence (calculated using results from other categories)(27). To further refine ortholog predictions from Ensembl, coding DNA sequence (CDS) data for all Ensembl species and an additional 29 *Drosophila* species will be downloaded from NCBI. I will use the CDS from the most recent *D. melanogaster* Flybase release to find orthologs among non-*D. melanogaster* species. The CDS of the longest protein sequence will be chosen for each gene. In more distant species, the CDS from a closely related species with the ortholog present will be used to identify orthologs in addition to D. melanogaster CDS. Predicted orthologs will be further validated via reciprocal BLAST (Basic Local Alignment Search Tool)-hit approach of the program INPARANOID (28). Tested predicted orthologs will be divided into orthologs, paralogs, and non-orthologs based on bootstrap values from INPARANOID analysis and synteny. I will realign the ortholog dataset using the phylogeny aware realignment software PRANK. After the alignment, sequences containing large regions of gaps will be removed using maxAlign.

I will then confirm GSC gene ortholog absences via PCR by considering sequence alignments of the species with the ortholog absent and the closest related species with the gene present, designing four primers based on the sequence with the gene present, two flanking the gene and two within the gene body. The genomic DNA of both species will be used as templates and the different primer combinations will be used for PCR reactions. Resulting products will then be visualized using gel electrophoresis and products will be Sanger sequenced. I will use the PCR amplifications of the present gene as positive controls.
Expected outcomes and interpretations: If gel electrophoresis bands are the expected size and Sanger sequencing results reveal there is no significant sequence between the two primers flanking the predicted absence, these results are consistent with ortholog absence. If I see an unexpected large band and sequencing results showing an unexpected region between the flanking primers, these results are consistent with the ortholog not being absent as predicted. The confirmed absences will enable us to further define the extent of observed functional flexibility at the phylogeny, species group, lineage and gene levels. If I find that all GSC genes excluding *Yb* are present across all included species, failing to reject the null hypothesis, I would conclude that GSC regulating genes are highly conserved, at least at the gene level, and that *Yb* is an outlier. If I find that some GSC regulating genes are absent in some included lineages, rejecting the null hypothesis, that confirms that there is some degree of developmental systems drift occurring in GSC genes. I can similarly evaluate lineages and genes for differences in GSC gene functional flexibility. These results will provide a direct test of the assumption that orthologs are generally functionally conserved in GSC regulating genes.

Patterning of ortholog presence or absence across species will also help create reasonable hypotheses about evolutionary history of GSC genes. For example, a gene may have orthologs in species throughout *Drosophila* and included outgroup species, which could indicate the gene was present in a common ancestor to all included species. If a gene has orthologs throughout the genus *Drosophila*, but not in any outgroup species, the gene may have been new to the common ancestor of the *Drosophila* genus. If a gene only has orthologs in a subgroup of species or a single species, the gene may have developed in the common ancestor of the species subgroup or even more recently, in the specific species lineage. Similarly, if a gene is absent across a species subgroup, but present elsewhere in the genus, the gene may have been present in the common ancestor of the genus, but subsequently lost in later common ancestors of particular species subgroups.

I have generated a GSC gene ortholog dataset including 10 *Drosophila* species spanning across the phylogeny and two closely related outgroup species with predicted presence and absence of the 366 GSC regulating genes using Ensembl compara, a multi-species database that stores the results of genome-wide species comparisons (18,19,22,23). Ensembl compara generates ortholog predictions based on a combination of factors including percentage of shared sequence identity and gene order conservation (69). I can make preliminary assessments regarding the extent and characteristics of GSC gene network functional flexibility by using predictions from the GSC gene ortholog dataset. For example, in *D. simulans*, a lineage particularly closely related to *D. melanogaster*, predictions show over 7% of GSC genes found to be essential in *D. melanogaster* are without an ortholog in *D. simulans*. This suggests that some GSC regulating genes are gaining or losing essential roles on a rapid evolutionary timescale, quickly following speciation and/or as intraspecies gene loss or gain. Predicted absences of GSC regulating genes generally increases with more divergence time from *D. melanogaster*, peaking at 33% (*D. willistoni*) within *Drosophila* and 61% (*L. cuprina*) when considering outgroups (Fig. 2). If the predictions are true, this represents an unexpected degree of functional flexibility at GSC genes.

Fig. 2 Predicted GSC gene absences and divergence time



The trend of the number of predicted absences increasing with divergence time from *D. melanogaster* is not completely linear. *D. Willistoni* (57 MY divergence time) has a higher degree of predicted gene absences than *D. mojavensis*, *D. virilis*, and *D. grimshawi* (67 MY divergence time). This could be due to stochastic forces or perhaps lineage specific pressures. Though predicted GSC gene absences generally increase with divergence time from *D. melanogaster*, the rate of the percentage of absences to divergence time shows a different relationship. Rates are highest in the species most closely related to *D. melanogaster* (4-7 MY of divergence) and the rates are significantly lower in more divergent species. This suggests that DSD could occur at a faster rate in closely related species. These are predictions that will require downstream validation, but these are some examples of the utility of generating this data.

Caveats and future directions: Assessing functional divergence by gene absence is not an exhaustive analysis of functional divergence. Present genes, like *bam*, can still be involved in developmental systems drift. Gene absence identifies the baseline of functional divergence at GSC genes. Also, while the relationship of orthologs across species can help lay out hypothetical evolutionary histories, parsimonious explanations can be incorrect. Parsimony may suggest that if a gene is lost in a specific lineage while related species in the same subgroup all still have the gene, the gene was present in the common ancestor to the subgroup and lost in the one lineage missing the gene. It could also be the case that the gene was independently gained by the other species in the subgroup and the common ancestor did not have the gene. Future directions could include functional analysis of select GSC genes to better define specific GSC gene functional evolutionary histories.

B: Integration of existing relevant datasets

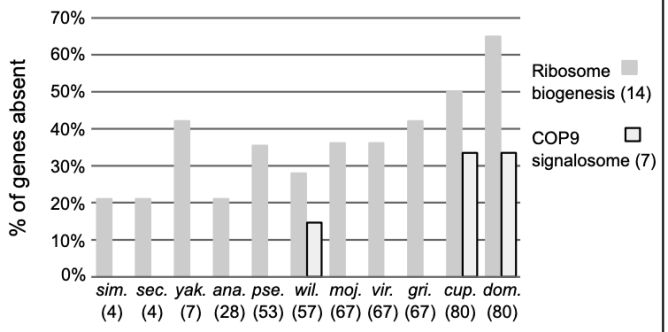
Null hypothesis: Datasets will reveal no common characteristics functionally flexible GSC genes

Methodology: Physical interactors, genetic interactors, and GO annotated molecular function will be incorporated from Flybase datasets using Python (25). GSC gene functional categories, identified by complex-enrichment analysis of the 366 GSC genes, and defect type, defined by the observed phenotypic effect of RNAi knockdown will be incorporated from Yan et. al (2014). I can use these datasets in conjunction with the GSC gene ortholog dataset to parse observed GSC gene and gene interaction network functional flexibility.

Expected outcomes and interpretations: If the datasets fail to reject the null hypothesis, revealing no common characteristics of functionally flexible genes, that will provide evidence that the characteristics incorporated may not have an impact on GSC gene functional flexibility. If the data

rejects the null hypothesis, I will be able to determine which categories are particularly conserved or flexible from the phylogeny to the gene level for GSC genes. For example, results from Ensembl predictions for GSC genes involved in Ribosome biogenesis (14 genes) show a high degree of flexibility across species, generally increasing with divergence time while GSC genes involved in the COP9 signalosome (7 genes) are largely conserved across species, with only one absent gene in a single *Drosophila* species and three genes absent in *L. cuprina* and *M. domestica* (Fig. 3). These different predicted degrees of conservation across functional categories of GSC genes could be related to the function or characteristics of the genes involved in the function. In contrast, predictions for flexibility across defect types show a generally consistent degree of conservation.

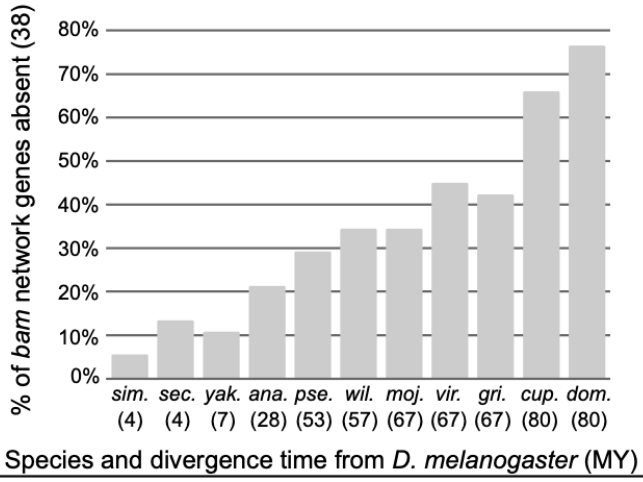
Fig. 3 Predicted absences in select functional categories across species



Physical and genetic interaction networks of GSC genes will also be informative. For example, genes identified as genetic and physical interactors of *bam* do not show predictions of complete conservation across *Drosophila* species considered in Ensembl compara (69). If the predictions are accurate, this indicates that there is at least some degree of functional flexibility within *bam*'s interaction networks across species. If an interacting gene is absent, as predicted, *bam* does not have an interacting relationship at all with the absent gene in that lineage. This does not necessarily indicate that *bam* itself is functionally divergent. Interacting genes could be absent due to a new gene (or genes) interacting with *bam* to carry out the same function. An interacting gene could be lost due to *bam* no longer performing the same function carried out in part by the interaction. The interacting gene could also be absent because it is novel to a distantly related clade and it has not been gained in the lineage I am considering. In lineages where *bam* is not carrying out its essential function as the key switch gene for differentiation (as defined in *D. melanogaster*), another gene or combination of genes must be executing that function.

I will also be able to determine whether GSC gene interaction networks show any patterns of conservation. Predicted results for *bam* provide a good example. In species signaling positive selection at *bam* via MK test with *bam* null differentiation defects in both sexes, I have found a few functionally related genes are predicted to be absent, 2 of 38 in *D. simulans* and 4 of 38 in *D. yakuba* (Fig. 4). If these predictions are accurate, there is some degree of DSD occurring quickly in *bam*'s interaction networks across species in which *bam* is essential. *D. ananassae*, which shows a signal of positive selection at *bam* via MK test and *bam* null differentiation defects in females only, has 8 of 38 *bam* interaction network genes predicted to be absent. This is a higher degree of predicted flexibility than in the species where *bam* nulls confirmed *bam*'s essential GSC differentiation function in both sexes. *D. pseudoobscura* Ensembl results predict 11 of 38 *bam* network genes are absent. If the predictions are accurate, this represents a substantial degree of flexibility in *bam*'s interaction network for this species. These absences along with the MK test results detecting no selection at *bam* in the lineage could indicate that *bam* is not performing its essential function in GSC differentiation. If the predicted absences are correct, *bam* is at least carrying out its function with fundamental differences in its interaction network. B*am* could be performing its essential GSC differentiation function with other genes executing the essential roles of the genes predicted to

Fig. 4 *Bam* interaction network predicted absences per species



be absent or *bam* itself could be in some way compensating for these absences. Direct experimental assessment would have to be done to determine *bam*'s specific function.

Another way I can use the gene interaction networks to further define characteristics of developmental systems drift is to compare a gene's interaction networks across species to highlight which interacting genes seem to have particularly high levels of flexibility or conservation. For example, considering *bam*'s interaction networks, *zpg* is weak or absent in all considered species excluding *D. melanogaster* and *Mir-7* is present in all species(70). One possible explanation is that *zpg* has three genetic interactors, all involved in its described GSC function, and no physical interactors. Given its small interaction network and apparent lack of pleiotropic functions, absence of *zpg* is not significantly disruptive. *Mir-7* on the other hand has eighteen physical interactors, five genetic interactors, and is involved in a range of pleiotropic functions including but not limited to regulation of wing growth, development of follicle cells, development of the visual system, and regulation of the Hippo pathway (71-75). Absence of *Mir-7* would lead to effects far beyond GSC differentiation. By considering the functional flexibility in GSC gene interaction networks, I can further characterize which genes seem to be more regularly subject to developmental systems drift and develop hypotheses as to why particular genes are highly conserved or flexible.

Caveats and future directions: The data incorporated from Flybase and Yan et. al (2014) is based on analysis of *D. melanogaster*. If GSC genes are found to be highly functionally flexible, the utility of using *D. melanogaster* interaction networks and functional classifications to make assessments about which specific functional aspects of the interaction networks are flexible will be limited. I will still be able to identify broadly where there is DSD between *D. melanogaster* and other species, but specific functions of interacting genes would require direct interrogation in divergent lineages. RNA-seq expression data for gonads and whole adults without gonads is available for eight species (*D. melanogaster*, *D. yakuba*, *D. ananassae*, *D. pseudoobscura*, D. persimilis, *D. willistoni*, *D. mojavensis*, and *D. virilis*), and I can incorporate expression data for GSC genes in these species to identify potential points of functional divergence by using differences in expression (16).

Aim 3: Analysis of selection

Rationale: Gene absence confirms that there is a functional diversification at the gene across orthologs, considering all included genes are essential in *D. melanogaster*, but gene presence alone does not provide useful insight regarding functional flexibility. To enable more useful evaluation of present genes, I will use the MK test and dN/dS based divergence analysis to detect signals of positive selection. There are lineage-specific MK test results for *bam* in 18 *Drosophila* species, all of which are included in our ortholog dataset (67). Results have shown heterogeneous signals of selection at *bam* across the *Drosophila* phylogeny. There are several possible drivers of positive selection. Some could result in functional divergence while others may not. Positive selection could be driven by a change in function, refinement of a function, life history or environmental factors, conflicts with TEs, endosymbionts, or viruses. Germline conflicts may or may not lead to functional changes at *bam*. For example, germline agents such as *W. pipientis* may be in conflict with *bam* for control of oogenesis leading to rapid evolution at *bam* to evade control (15). This could result in *bam* adaptively evolving while maintaining the same essential function. Germline conflicts might also drive the fixation of novel genes in new roles to evade control. In this case, a different gene or genes could be performing *bam*'s essential differentiation function, potentially relaxing constraint at *bam* and allowing *bam* to lose its now redundant function and/or gain or functionally refine a different function, driving positive selection. There are no reported orthologs of *bam* in non-*Drosophila* species (69), which is consistent with *bam* being novel to the genus *Drosophila*.

For the species with *bam* nulls that show differentiation defects in both sexes and only differentiation defects in females, all show signals of strong positive selection via MK test. Considering *bam* appears to be a gene novel to the *Drosophila* genus, this is consistent with positive selection being driven by functional diversification or refinement of function. Positive selection for refinement of function in particular is common in novel genes, but selection at *bam* could also be driven by germline conflicts. Selection at *bam* will only be driven by W*olbachia* if *bam* retains its essential role, otherwise control of *bam* by W*olbachia* would not give W*olbachia* more control of

reproduction. In D. teissieri, *bam* nulls showed no differentiation defects and no signal of positive selection was detected. This could indicate that *bam* never gained its essential function in this lineage or that *bam* lost its essential function in this lineage. Germline conflict with W*olbachia* is unlikely in this lineage, given *bam* is not essential for GSC differentiation. Functional refinement or change in function for *bam*'s essential role in GSC differentiation also is not occurring. Even with a lack of selection detected, there could still be a relaxation of constraint at *bam* which could indicate recent functional redundancy. The combination of *bam* MK test along with lack of selection detected at *bam* using the PAML site test across 12 *D. melanogaster* group species highlights that the positive selection at *bam* is likely due to lineage specific pressures. MKT results are available for *bam* in *D. pseudoobscura*, but not yet for *D. americana*. Fortunately, polymorphism data for *D. americana* and a few other *Drosophila* species have recently been made available.

Polymorphism data is not available for most species I am considering in the ortholog dataset. Still, identifying signals selection in these lineages would provide further useful context for understanding the flexibility of *bam* and other GSC regulating genes. I will use dN/dS based divergence analysis to test *bam* for signals of positive selection in species without polymorphism data as well as testing the other 365 GSC regulating genes for selection across the phylogeny. Previous analysis within six *D. melanogaster* group species (*D. melanogaster, D. simulans, D. sechelia, D. yakuba, D. erecta,* and *D. ananassae*)tested GSC genes for signals of positive selection using PAML analysis (33). 76 GSC genes showed signals of positive selection. PAML analysis was not conducted beyond *D. ananassae* due to prohibitive synonymous site saturation preventing incorporation of more distant *Drosophila* species (25,33,34). Specific GSC genes including *stwl*, *ote*, and more extensively *Sxl* have been tested for selection in *Drosophila* species outside of the D. *melanogaster* group (35). Analysis of *Sxl* in *D. pseudoobscura* and many species within the *D. melanogaster* group shows that selective constraint plays a significant role in the molecular evolution of *Sxl* within *Drosophila*, but patterns were also observed suggesting both recent positive selection and episodic bursts of protein evolution at *Sxl* (36). Now, with additional high quality sequences, I can extend PAML divergence analysis of GSC genes to species beyond the *D. melanogaster* group.

A: McDonald-Kreitman test at *bam* in select *Drosophila* species
Null hypothesis: No signals of positive selection will be detected via MK test in newly tested species
Methodology: I will use polymorphism data for *D. americana*, *D. kikkawai*, *D. ironensis*, and *D. setifemur* to test for departures from selective neutrality consistent with positive selection at *bam* using the McDonald Kreitman test. The null hypothesis of the MKT is that the ratio of nonsynonymous to synonymous polymorphism within a species is equal to the ratio of nonsynonymous to synonymous divergence between species. Since *bam* is highly divergent across the *Drosophila* genus and appears to be experiencing episodic signals of adaptive evolution, I will measure lineage-specific divergence.

I will do this by using the codeml package from PAML to generate the predicted common ancestor sequence and align that sequence to the species of interest using PRANK. I will use the MK test webtool at **http://mkt.uab.cat/mkt/mkt.asp** to complete the MK test. Polymorphic sites at <15% frequency will be excluded, since these are likely slightly deleterious alleles that have not yet been purged by purifying selection (76). Resultant values for the contingency table, Chi-Square test, and alpha, the proportion of fixations predicted to be due to positive selection, will be reported.
Expected outcomes and interpretations: If MK test results for *D. americana bam* indicate positive selection and the null results demonstrate *bam* is essential for GSC differentiation, this provides evidence to support that functional refinement for *bam*'s essential function in GSC differentiation in *D. americana* or gametogenic conflict as potential drivers of positive selection at *bam*. This does not rule out other potential drivers of positive selection. If MK test results for *D. americana bam* indicate positive selection and null results demonstrate *bam* is not essential for GSC differentiation, this is consistent with positive selection being driven by functional refinement for a function unrelated to GSC differentiation like *bam*'s role in hematopoietic progenitor maintenance during hematopoiesis (58).
Caveats and future directions: If the MK test does not detect positive selection at a gene, that does not necessarily mean positive selection is not taking place. Using the MK test, when possible, to

evaluate other GSC genes for signals of positive selection could also highlight potential functional divergences.

B: Divergence analysis using PAML
Null hypothesis: PAML analysis reveals positive selection is detected at the same GSC genes and sites across species
Methodology: To identify positive selection, I will estimate dN/dS in all 1-to-1 orthologous GSC regulating genes using the branch-sites model from PAML, since this model enables more effective detection of heterogeneous signals of selection than the site model (25, 33,37,38). The branch-site model of PAML requires an *a priori* phylogenetic tree to test for positive selection in foreground branches. The species designated as the foreground branch will be the branch evaluated for positive selection, and the background branches included in analysis will be those closely related on the species tree (ideally two or more) (34, 38). I may run into the issue of synonymous site saturation for particularly divergent single lineages. I will restrict analysis to groups of lineages that are not saturated, which may result in the exclusion of single long lineages like *D. willistoni*.

I will compare the likelihood of two different models for each orthologous gene: 1. the alternative model that allows for a proportion of the sites to be under positive selection along the foreground branch ($\omega 2 \geq 1$), and the background branches having a proportion of sites being under purifying selection ($\omega 1 < 1$) or neutrally evolving ($\omega 0 = 1$); and 2. the simplistic null model that has the $\omega 2$ fixed at 1 on the foreground branches, and all other branches having $\omega 0 = 1$ and $\omega 1<1$. I will obtain likelihood values after running each transcript under two different models and carrying out likelihood ratio tests between the models to evaluate whether the alternative model outperforms the null model. I will perform a Bonferroni correction to account for multiple comparisons. A significant result from the branch-site model is indicative that a subset of the sites in the coding gene signal positive selection, with the selected sites providing an advantage to the foreground lineage (39).
Expected outcomes and interpretations: If I fail to reject the null hypothesis, this suggests selection acts similarly on GSC genes across the phylogeny. If the null hypothesis is rejected, this suggests selection can act differently on GSC genes across species. Differences in signals of selection will highlight potential points and patterns of functional flexibility in GSC regulating genes. This data in conjunction with the information from the ortholog dataset will give us the ability to evaluate if there are particular functional categories of GSC regulating genes that are particularly common or uncommon targets of positive selection. Results will also help us identify possible patterns of selection in functionally related genes when an interacting gene is lost or gained.

Perhaps when genes are lost in a lineage bursts of positive selection in functionally related genes are observed. This could indicate that functionally related genes may be evolving adaptively to execute the function of the gene that is now absent. Genes that are functionally related to the lost gene could also not show signals of positive selection. If this is the case, perhaps the gene lost had not acquired its essential function in this lineage or it may have been made functionally redundant by another gene using the same functional network. I may also see patterns in functionally related genes when a gene is gained. Genes that appear to be gained in *D. melanogaster* and signal positive selection could indicate refinement of essential function. Interpretation of selection results will be largely context dependent, but results will provide another useful dimension when considering GSC gene functional flexibility.
Caveats and future directions: The selection analysis will be less concrete in highlighting functional flexibility than using presence and absence of genes or generating nulls since function is not directly evaluated. Also, even if selection is detected identically across included species, I can not automatically assume that the genes are not functionally divergent. Using the site model in PAML in addition to the branch site model could help highlight whether I see consistent selection at the same site across species, and incorporating comparisons of evolutionary rates across the phylogeny for GSC genes could highlight rate differences between species that may not have signals of positive selection. This could indicate a relaxation of constraint in the lineage, which could be consistent with the gene no longer being essential in the lineage.

Citations

1.McKearin DM, Spradling AC. Bag-of-marbles: A Drosophila gene required to initiate both male and female gametogenesis. Genes Dev. 1990;4: 2242–2251. doi:10.1101/gad.4.12b.2242
2. Shen R, Weng C, Yu J, Xie T. eIF4A controls germline stem cell self-renewal by directly inhibiting BAM function in the Drosophila ovary. Proc Natl Acad Sci U S A. 2009;106: 11623–11628. doi:10.1073/pnas.0903325106
3. Ohlstein B, Lavoie CA, Vef O, Gateff E, Mckearin DM. The Drosophila Cystoblast Differentiation Factor, benign gonial cell neoplasm, Is Related to DExH-box Proteins and Interacts Genetically With bag-of-marbles. 2000. Available: http://www.fruitfly.org
4. Li Y, Zhang Q, Carreira-Rosario A, Maines JZ, Mckearin DM. Mei-P26 Cooperates with Bam, Bgcn and Sxl to Promote Early Germline Development in the Drosophila Ovary. PLoS One. 2013;8: 58301. doi:10.1371/journal.pone.0058301
5. Li Y, Minor NT, Park JK, McKearin DM, Maines JZ. Bam and Bgcn antagonize Nanos-dependent germ-line stem cell maintenance. Proc Natl Acad Sci U S A. 2009;106: 9304–9309. doi:10.1073/pnas.0901452106
6. Sgromo A, Raisch T, Backhaus C, Keskeny C, Alva V, Weichenrieder O, et al. Drosophila Bag-of-marbles directly interacts with the CAF40 subunit of the CCR4–NOT complex to elicit repression of mRNA targets. Rna. 2018;24: 381–395. doi:10.1261/rna.064584.117
7. Pan L, Wang S, Lu T, Weng C, Song X, Park JK, et al. Protein competition switches the function of COP9 from self-renewal to differentiation. Nature. 2014;514: 233–236. doi:10.1038/nature13562
8. Ji S, Li C, Hu L, Liu K, Mei J, Luo Y, et al. Bam-dependent deubiquitinase complex can disrupt germ-line stem cell maintenance by targeting cyclin A. Proc Natl Acad Sci U S A. 2017. doi:10.1073/pnas.1619188114
9. Ting X. Control of germline stem cell self-renewal and differentiation in the Drosophila ovary: concerted actions of niche signals and intrinsic factors. Wiley Interdiscip Rev Dev Biol. 2013;2: 261–273. doi:10.1002/wdev.60
10. Insco ML, Leon A, Tam CH, McKearin DM, Fuller MT. Accumulation of a differentiation regulator specifies transit amplifying division number in an adult stem cell lineage. Proc Natl Acad Sci U S A. 2009. doi:10.1073/pnas.0912454106
11. Insco ML, Bailey AS, Kim J, Olivares GH, Wapinski OL, Tam CH, et al. A selflimiting switch based on translational control regulates the transition from proliferation to differentiation in an adult stem cell lineage. Cell Stem Cell. 2012;11: 689–700. doi:10.1016/j.stem.2012.08.012
12. Shivdasani AA, Ingham PW. Regulation of Stem Cell Maintenance and Transit Amplifying Cell Proliferation by TGF-β Signaling in Drosophila Spermatogenesis. Curr Biol. 2003. doi:10.1016/j.cub.2003.10.063
13. Ohlstein B, McKearin D. Ectopic expression of the Drosophila Bam protein eliminates oogenic germline stem cells. Development. 1997.
14. Dumont VLB, Flores HA, Wright MH, Aquadro CF. Recurrent Positive Selection at Bgcn, a Key Determinant of Germ Line Differentiation, Does Not Appear to be Driven by Simple Coevolution with Its Partner Protein Bam. [cited 14 Apr 2020]. doi:10.1093/molbev/msl141
15. Choi JY, Aquadro CF. The coevolutionary period of Wolbachia pipientis infecting Drosophila ananassae and its impact on the evolution of the host germline stem cell regulating genes. Mol Biol Evol. 2014. doi:10.1093/molbev/msu204
16. Whittle CA, Extavour CG. Selection shapes turnover and magnitude of sex-biased expression in Drosophila gonads. BMC Evol Biol. 2019. 19:60. doi: 10.1186/s12862-019-1377-4.

17. Thode SK, Baekkedal C,Soderberg JJ, Hjerde E, Hansen H, Haegun P. Construction of a *fur* mutant and RNA-sequencing provide deeper global understanding of the *Aliivibrio salmonicida* Fur regulon. PeerJ. 2017. 5: e3461. doi: 10.7717/peerj.3461.
18. Loytynoja Ari Phylogeny-aware alignment with PRANK. Methods in Molecular Biology 2013. Vol.1079 p.155-170.
19. Gouzeia-Oliveira R, Sackett PW, Pedersen AG MaxAlign: maximizing usable data in an alignment. BMC Bioinformatics 2007. 8:312.
20. Vankuren NW, Long M. Gene duplicates resolving sexual conflict rapidly evolved essential gametogenesis functions. Nature Ecology and Evolution 2018. 2, 705-712.
21. Lamb AM, Wang Z, Simmer P, Chung H, Wittkopp P *ebony* Affects Pigmentation Divergence and Cuticular Hydrocarbons in *Drosophila americana* and *D. novamexicana* Front. Ecol. Evol. 2020 https://doi.org/10.3389/fevo.2020.00184
22. Kim BY, Wang JR, Miller DE, Matute DR, Petrov DA, et al. Highly contiguous assemblies of 101 drosophilid genomes Elife. 2021. E66405.
23.Wiegmann BM, Yeates DK, Thorne JL, Kishino H Time Flies, a New Molecular Time-Scale for Brachyceran Fly Evolution Without a Clock Cyst. Biol. 2003. 52(6):745-756.
24. Kanca et al. An efficient CRISPR-based strategy to insert small and large fragments of DNA using short homology arms eLife 2019. 8:e51539.
25. Choi JY, Aquadro CF Molecular Evolution of *Drosophila* Germline Stem Cell and Neural Stem Cell Regulating Genes Genome Biology and Evolution 2015. Vol. 7 Issue 11 p. 3097-3114.
26. Yan et al. A Regulatory Network of *Drosophila* Germline Stem Cell Self-Renewal Developmental Cell 2014 Vol 28 Issue 4 p. 459-473.
27. Kersey PJ et al. Ensembl Genomes: Extending Ensembl across the taxonomic space Nucleic Acids Research 2010. Vol. 38 D563-D569.
28. Ostlund G et al. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis Nucleic Acids Research 2010 Vol. 38 D196-D203.
29. O'Grady PM, DeSalle R Phylogeny of the Genus *Drosophila* Genetics 2018 209(1):1-25.
30. Kearse M et al. Geneious Basic: An integrated and extendable desktop software platform for the organization and analysis of sequence data Bioinformatics 2012 Vol. 28 Issue 12 p. 1647-1649.
31. Hoon MJL, Imoto S, Nolan J, Miyano S Open source clustering software Bioinformatics 2004 20(9):1453-4.
32. Page R TreeView: an application to display phylogenetic trees on personal computers Bioinformatics 1996 DOI:10.1093/bioinformatics/12.4.357.
33. Clark AG et al. Evolution of genes and genomes on the Drosophila phylogeny 2007 8;450(7167):203-18.
34. Gharib WH, Robinson-Rechavi M The Branch-Site Test of Positive Selection is Surprisingly Robust but Lacks Power under Synonymous Substitution Saturation and Variation in GC 2013 3097): 1675-1686.
35. Choi JY, Aquadro CF The Coevolutionary Period of *Wolbachia* pipientis Infecting *Drosophila ananassae* and Its Impact on the Evolution of the Host Germline Stem Cell Regulating Genes Molecular Biology and Evolution Vol. 31 Issue 9 p. 2457-2471.
36. DuMont VLB, White SL, Zinshteyn D, Aquadro CF Molecular population genetics of Sex-lethal (Sxl) in the *Drosophila melanogaster* species group: a locus that genetically interacts with *Wolbachia pipientis* in *Drosophila melanogaster* G3 Vol. 11 Issue 8 jkab197.
37. Bhattacharya T et al. Evidence of Adaptive Evolution in *Wolbachia*-Regulated Gene DNMT2 and Its Role in the Dipteran Immune Response and Pathogen Blocking Viruses 2021 13(8) 1464.
38. Yang Z PAML4: Phylogenetic Analysis by Maximum Likelihood Molecular Biology and Evolution 2007 Vol.24 Issue 8 p. 1586-1591.
39. Yang Z Wong WSW Nielsen R Bayes empirical Bayes inference of amino acid sites under positive selection. *Mol Biol Evol.* 2005 22:1107–1118.
40. Murrell B, et al. FUBAR: a fast, unconstrained Bayesian approximation for inferring selection. *Mol Biol Evol.* 2013 30:1196–1205.

41. Flores HA, et al. Adaptive evolution of genes involved in the regulation of germline stem cells in Drosophila melanogaster and D. simulans *G3* 2015 9;5(4):583-92.
42. Brunette GJ, Jamalruddin MA, Baldock RA, Clark NL, Bernstein KA Evolution-based screening enables genome wide prioritization and discovery of DNA repair genes *PNAS* 2019 116 (39) 19593-19599.
43. Whelan S and Goldman PLN Molecular phylogenetics: state-of-the-art methods for looking into the past Trends Genet. 2001 17(5):262-72.
44.Yang Z Among-site rate variation and its impact of phylogenetic analyses *Trends in Ecology and Evolution* 1996 Vol. 11 Issue 9 p. 367-372.
45. Clark NL, Alani E, Aquadro CF Evolutionary rate covariation reveals shared functionality and coexpression of genes *Genome Res.* 2012 22(4):714-20.
46. Quax TEF, Claassens NJ, Soll D, Oost Jvd Codon Bias as a Means to Fine-Tune Gene Expression *Mol Cell.* 2015 59(2):149-161.
47. Ye Y and Nurmi P Gestimator: Shape and Stroke Similarity Based Gesture Recognition *ICMI* 2015 p. 216-226.
48. Andolfatto P Controlling Type-I Error of the McDonald-Kreitman Test in Genomewide Scans for Selection on Noncoding DNA *Genetics* 2008 Vol. 180 Issue 3 p. 1767-1771.
49. McDonald J and Kreitman M Adaptive protein evolution at the Adh locus in Drosophila *Nature* 1991 351: 652-654.
50. Nielsen R Molecular signatures of natural selection *Annu. Rev. Genet.* 39: 197-218.
51. Smith NCG and Eyre-Walker Adam Adaptive protein evolution in Drosophila *Nature* 2002 28;415(6875):1022-4.
52. Levine MT and Begun DJ Evidence of Spatially Varying Selection Acting on Four Chromatin-Remodeling Loci in *Drosophila melanogaster Genetics* 2008 179(1): 475-485.
53. Lee Y and Langley CH Long-Term and Short-Term Evolutionary Impacts of Transposable Elements on *Drosophila Genetics* 2012 vol. 192 no. 4 1411-1432.
54. Mensch J, Serra F, Lavagnino NJ, Dopazo H, Hasson E. Positive selection in nucleoporins challenges constraints on early expressed genes in *Drosophila* development. *Genome Biol Evol*. 2013 5(11):2231–2241.
55. Tang, S., and D. C. Presgraves. Evolution of the *Drosophila* nuclear pore complex results in multiple hybrid incompatibilities. *Science* 2009 323:779–782.
56. Maheshwari S., Wang J., Barbash D. A. Recurrent positive selection of the Drosophila hybrid incompatibility gene *Hmr*. *Mol. Biol. Evol.* 2008 25: 2421–2430.
57. Maheshwari S and Barbash DA An Indel Polymorphism in the Hybrid Incompatibility Gene *Lethal Hybrid Rescue* of *Drosophila* is Functionally Relevant *Genetics* 2012 vol. 192 no. 2 683-691.
58. Tokusumi T et al. Germ line differentiation factor Bag of Marbles is a regulator of hematopoietic progenitor maintenance during Drosophila hematopoiesis *Development.* 2011 138(18):3879-84.
59. Zhai W, Nielsen R, Slatkin M An Investigation of the Statistical Power of Neutrality Tests Based on Comparative and Population Genetic Data *Molecular Biology and Evolution* 2008 26(2):273-283.
60. Bubnell JE, Ulbing CKS, Fernandez-Begne P, Aquadro CF Functional divergence of the *bag of marbles* gene in the *Drosophila melanogaster* species group *bioRxiv* 2021 doi.org/10.1101/2021.06.25.449946.
61. Szakmary A, Reedy M, Qi H, Lin H The Yb protein defines a novel organelle and regulates male germline stem cell self-renewal in *Drosophila melanogaster J Cell Biol.* 2009 185(4): 613-627.
62. Flores HA, Bubnell JE, Aquadro CF, Barbash DA. The Drosophila bag of marbles Gene Interacts Genetically with Wolbachia and Shows Female-Specific Effects of Divergence *PLOS Genetics* 2015. doi:10.1371/journal.pgen.1005453
63. Swan A, Hijal S, Hilfiker A, Suter B Identification of new X-chromosomal genes required for Drosophila oogenesis and novel roles for fs(1)Yb, brainiac and dunce *Genome Res.* 2001 11(1):67-77.
64. Weiss KM, Fullerton SM. 2000. Phenogenetic drift and the evolution of genotype-phenotype relationships. *Theor Popul Biol*. 57(3):187-195.

65. True JR, Haag ES. 2001. Developmental system drift and flexibility in evolutionary trajectories. *Evol Dev*. 3(2):109-119.
66. Brogna S, Wen J. Nonsense-mediated mRNA decay (NMD) mechanisms. Nature Structural and Molecular Biology. 2009. doi:10.1038/nsmb.1550
67. Bubnell JE, Ulbing CK, Fernandez-Begne P, Aquadro. Functional divergence of the *bag of marbles* gene in the *Drosophila melanogaster* species group bioRxiv. 2021. https://doi.org/10.1101/2021.06.25.449946
68. Tekaia Fredj Inferring orthologs: open questions and perspectives *Genomics Insights* 2016. 9:17-28.
69. Herrero J, et.al. Ensembl comparative genomic resources Database (Oxford) 2016:bav096.
70.Flybase Nucleic Acids Res. 49(D1) D899–D907. N.B.
71. Carre et al. AutomiG, a Biosensor to detect alterations in miRNA Biogenesis in small RNA silencing guided by perfect target complementarity *PLoS ONE* 8(9): e74296. 2013.
72. Pek et al. Drosophila maelstrom ensures proper germline stem cell lineage differentiation by repressing microRNA-7 *Dev. Cell* 17(3):417-424.
73. Da Ros et al. Dampening the signals transduced through hedgehog via MicroRNA miR-7 facilitates notch-induced tumorigenesis *PLoS Biol.* 11(5): e1001554. 2013.
74. Li and Carthew A microRNA mediates EGF receptor signaling and promotes photoreceptor differentiation in the Drosophila eye *Cell* 123(7):1267-1277. 2005.
75. Gilboa et al. Germ line stem cell differentiation in Drosophila requires gap junctions and proceeds via an intermediate state *Development*130(26): 6625-6634. 2003.
76. Charlesworth J, Eyre-Walker A. The McDonald-Kreitman test and slightly deleterious mutations. *Mol Biol Evol.* Doi:10.1093/molbev/msn005. 2008.