Evolutionary Dynamics of Male Reproductive Genes in the Drosophila virilis Subgroup

Yasir H. Ahmed-Braimah,* Robert L. Unckless,‡ and Andrew G. Clark†
*Department of Biology, University of Rochester, New York 14627, †Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, and ‡Department of Molecular Biosciences, University of Kansas, Lawrence, Kansas 66045

ORCID IDs: 0000-0002-5805-5894 (Y.H.A.-B.); 0000-0001-8586-7137 (R.L.U.)

ABSTRACT Postcopulatory sexual selection (PCSS) is a potent evolutionary force that can drive rapid changes of reproductive genes within species, and thus has the potential to generate reproductive incompatibilities between species. Male seminal fluid proteins (SFPs) are major players in postmating interactions, and are important targets of PCSS in males. The virilis subgroup of Drosophila exhibits strong interspecific gametic incompatibilities, and can serve as a model to study the genetic basis of PCSS and gametic isolation. However, reproductive genes in this group have not been characterized. Here we utilize short-read RNA sequencing of male reproductive organs to examine the evolutionary dynamics of reproductive genes in members of the virilis subgroup: D. americana, D. lummei, D. novamexicana, and D. virilis.

We find that the majority of male reproductive transcripts are testes-biased, accounting for ~15% of all annotated genes. Ejaculatory bulb (EB)-biased transcripts largely code for lipid metabolic enzymes, and contain orthologs of the D. melanogaster EB protein, Peb-me, which is involved in mating-plug formation. In addition, we identify 71 candidate SFPs, and show that this gene set has the highest rate of nonsynonymous codon substitution relative to testes- and EB-biased genes. Furthermore, we identify orthologs of 35 D. melanogaster SFPs that have conserved accessory gland expression in the virilis group. Finally, we show that several of the SFPs that have the highest rate of nonsynonymous codon substitution reside on chromosomal regions, which contributes to paternal gametic incompatibility between species. Our results show that SFPs rapidly diversify in the virilis group, and suggest that they likely play a role in PCSS and/or gametic isolation.

KEYWORDS accessory glands seminal fluid proteins ejaculatory bulb tests RNA-seq

In sexually reproducing organisms, the ability to secure mates is a central component of fitness. Male mating success is largely determined by behavioral traits that are often under sexual selection. Furthermore, in species where females store sperm for extended periods and mate with multiple males, sperm can compete for fertilization and females can bias fertilization to certain males (Birkhead and Pizzari 2002). This additional layer of sexual selection—known as postcopulatory sexual selection (PCSS)—can drive rapid evolution of genes involved in postcopulatory interactions. Indeed, reproductive genes evolve rapidly in many animal taxa, often by positive selection (Swanson and Vacquier 2002). Importantly, rapid evolution of reproductive genes can have direct consequences for speciation by establishing barriers to fertilization between divergent populations (Markow 1997). The pattern of rapid evolution of reproductive genes and the potential involvement of PCSS in this divergence is widely recognized. However, the molecular genetic basis of PCSS is still not well understood (Wilkinson et al. 2015).

Among internally fertilizing organisms, a complex interaction takes place between the female reproductive tract and the male ejaculate, ultimately leading to the union of female and male gametes (Wolffner 2009). Morphological features of these gametes—such as sperm length or female reproductive tract morphology—can play an important role in this interaction, as these features can rapidly diverge between closely
related species (Pitnick et al. 1999). Furthermore, extensive studies in Drosophila melanogaster reveal that these interactions are mediated in part by seminal fluid proteins (SFPs), which are secreted from the paired male accessory glands (AGs) and the ejaculatory bulb (EB) (Ravi Ram and Wolfner 2007). Some of these proteins induce physiological effects in the mated female, such as increased oviposition (Herndon and Wolfner 1995), facilitation of sperm storage (Neubaum and Wolfner 1999), reduction of the female’s propensity to remate (Tram and Wolfner 1998), and reduction of her lifespan (Chapman 2001). To date ~200 SFPs have been identified in D. melanogaster, and several of these have been shown to evolve rapidly between species of the melanogaster group (Begun and Lindfors 2005; Swanson et al. 2001). Because of this rapid evolution, only a subset of these genes have orthologs in distantly related species (Haerty et al. 2007). This pattern suggests that SFPs may differ in content across different taxa (Kelleher et al. 2009). Thus, a comprehensive understanding of postcopulatory interactions and their consequences for speciation would benefit from genetic studies on reproductive genes from a wide range of taxa (Wilburn and Swanson 2016).

Closely related species that exhibit gametic incompatibilities in interspecific crosses—often called postmating prezygotic (PMPZ) reproductive isolation—provide a unique opportunity to study the genetic basis of PCSS because the genes that underlie these gametic incompatibilities likely diverge through PMPZ mechanisms within species. Genetic mapping is the traditional approach to identify regions of the genome that cause reproductive isolation between species. The repertoire of reproductive proteins that reside in these mapped regions represent strong candidates for gametic isolation between species, as the incompatible components in PMPZ necessarily involve the male ejaculate and the female reproductive tract. Male reproductive proteins, particularly SFPs, have been characterized in several taxa, including mosquito (Sirot et al. 2011), honeybee (Baer et al. 2009), mouse (Dean et al. 2011), and D. melanogaster (Findlay et al. 2008; Wasbrough et al. 2010). Many other taxa that are amenable to detailed genetic study and display interspecific PMPZ phenotypes, however, remain largely uncharacterized (Wilburn and Swanson 2016).

Several Drosophila species groups provide excellent material for the genetic study of prezygotic interactions (Markow 1997). One such species group is the virilis group (Supplemental Material, Figure S1A in File S1). Members of this group evolve gametic incompatibilities in the mating female, such as increased oviposition (Herndon and Wolfner 1995), facilitation of sperm storage (Neubaum and Wolfner 1999), reduction of the female’s propensity to remate (Tram and Wolfner 1998), and reduction of her lifespan (Chapman 2001). To date ~200 SFPs have been identified in D. melanogaster, and several of these have been shown to evolve rapidly between species of the melanogaster group (Begun and Lindfors 2005; Swanson et al. 2001). Because of this rapid evolution, only a subset of these genes have orthologs in distantly related species (Haerty et al. 2007). This pattern suggests that SFPs may differ in content across different taxa (Kelleher et al. 2009). Thus, a comprehensive understanding of postcopulatory interactions and their consequences for speciation would benefit from genetic studies on reproductive genes from a wide range of taxa (Wilburn and Swanson 2016).

Several Drosophila species groups provide excellent material for the genetic study of prezygotic interactions (Markow 1997). One such species group is the virilis group (Supplemental Material, Figure S1A in File S1). Members of this group evolve gametic incompatibilities rapidly (Sveigart 2010; Jennings et al. 2011, 2014; Sagga and Civetta 2011; Ahmed-Braimah and McAllister 2012). A subset of virilis group species (the D. virilis subgroup: D. americana, D. novamexicana, D. lummei, and D. virilis) show strong gametic incompatibilities in nearly all heterospecific hybridizations (Figure S1B in File S1). In particular, five of the six heterospecific cross combinations between these species produce <2% hatched eggs in at least one direction of the cross (Sveigart 2010; Sagga and Civetta 2011; Ahmed-Braimah and McAllister 2012; Ahmed-Braimah 2016). The only species cross in which both directions show appreciable hatch rates is the cross between D. lummei and D. virilis (~40% fertilization success). The reason for this general reduction in hatch rate was studied in three of these crosses, and appears largely due to defects in sperm storage (Sagga and Civetta 2011; Ahmed-Braimah and McAllister 2012; Ahmed-Braimah 2016), but other postcopulatory defects are likely. Genetic analyses also reveal that the genetic architecture of the incompatibility between D. americana males and D. virilis females is somewhat complex: at least three regions on the centromeric half of chromosome 5 (Muller C) and a large inversion on chromosome 2 (Muller E) carry genes responsible for the paternal component of PMPZ (Sveigart 2010; Ahmed-Braimah 2016). Divergent reproductive genes are likely involved in gametic incompatibilities in the virilis group. The processes disrupted in heterospecific inseminations largely resemble those for which SFPs have been implicated in D. melanogaster (Wolfner 2009). However, almost nothing is known about reproductive genes in the virilis group. Furthermore, given that D. virilis is ~40 MY divergent from D. melanogaster, the virilis group is likely to contain a unique and/or differentiated complement of SFPs.

Here, we use short-read RNA sequence data from male reproductive tissues of the Drosophila virilis subgroup, in addition to whole-genome sequence data, to characterize the repertoire of reproductive genes in this species group. We obtain RNA-seq data from the AG, EB, and testes (Figure 1). These tissues comprise the main sources of ejaculate components, i.e., sperm and seminal fluid. We also obtain RNA libraries from the gonadectomized male carcass to identify reproductive tissue-specific transcripts. Our objectives were to (1) identify candidate SFPs and tissue-biased reproductive genes, (2) identify functional categories that are enriched among reproductive genes and their potential biochemical roles, (3) examine gene expression and sequence divergence among reproductive genes between species, and (4) identify the set of conserved SFPs between the virilis subgroup and D. melanogaster. While we present analyses on all three reproductive tissue types, we focus largely on SFPs as we think they are the main targets of PCSS in males.

We show that several SFPs are rapidly evolving in this species group, as revealed by an excess of nonsynonymous substitutions relative to other reproductive gene classes and the genome average. Interestingly, the most rapidly evolving SFPs reside within an inverted region that has been implicated in PMPZ isolation between species. We also find that several SFPs show confined expression to one or more species, suggesting that expression divergence is driving differentiation in SFP content. Furthermore, we show that enriched functional categories of SFPs are largely similar to those from other known insects, where proteolytic enzymes dominate. Finally, we identify an appreciable number of conserved SFPs between D. melanogaster and the virilis subgroup, suggesting deep conservation of a subset of male ejaculate components.

**MATERIALS AND METHODS**

**Fly strains and hatch rate estimates**

Flies were maintained at a constant temperature (22°C) in a 12-hr day/night cycle on standard cornmeal media. A single strain each of D. virilis (1051.87), D. americana (ML97.3), D. lummei (LM.08) and D. novamexicana (15010-1031.04) were used throughout this study. Conspecific and heterospecific hatch rate estimates for all possible cross combinations were obtained by calculating the percentage of hatched eggs from daily collections (~10 collections) of 100 eggs from population cages containing ~400 males and females. Mean hatch rates represent the average of multi-day collections.

**Tissue dissection, RNA library preparation, and sequencing**

Individual 12- to 14-d-old virgin males and females of each species were paired and allowed to mate overnight. Males were then anesthetized with CO2 and their reproductive tissues removed. The AG, EB, and testes were separated and flash-frozen in liquid nitrogen. Each sample consisted of two (EB) or three (AG, carcass, and testes) replicates, and each replicate contained tissue from ~20 flies. RNA was isolated using the Qiagen RNeasy Mini Kit (cat. no. 74104). Paired-end and single-end libraries were prepared using the Illumina TrueSeq RNA Library Prep Kit v2 (www.illumina.com). Sequencing was performed in two stages. In the first, paired-end libraries (one replicate each of AG, carcass, and testes) were sequenced on an Illumina HiSeq2500 at the Cornell Biotechnology
Resource Center (Cornell University), and single-end libraries (two replicates each of all four tissues) were sequenced on an Illumina HiSeq2500 at the Genomics Resource Center (University of Rochester).

Whole-genome short-read sequencing and assembly

The four virilis subgroup species were used to generate whole-genome sequence data. A pool of ~20 males and females from each strain were flash-frozen for DNA library preparation. DNA was isolated using the Qiagen DNeasy Blood & Tissue Kit (cat. no. 69504). Paired-end libraries were prepared using the Illumina TruSeq DNA Library Prep Kit (www.illumina.com). Libraries were sequenced on an Illumina HiSeq2500 at the Genomics Resource Center (University of Rochester).

Paired-end reads from D. americana, D. lummei, and D. novamexicana were mapped to the D. virilis reference genome (r1.06) using Bowtie2 v2.2.2 (Langmead and Salzberg 2012). The reference genome was formatted into chromosome arms (Muller elements) using scaffold placement information (Schaeffer et al. 2008). Reads were mapped with the “–local” bowtie2 setting, and assemblies were used to extract whole-genome FASTA sequences for each chromosome/scaffold using Samtools 1.3.1 (Li et al. 2009), and seqtk (https://github.com/lh3/seqtk).

Transcriptome assembly

We aligned the RNA-seq reads to the D. virilis reference genome (r1.06) using Tophat v2.1.1 (Trapnell et al. 2009) with the following settings: -N 20 –read-gap-length 3 –read-edit-dist 20. Aligned reads for each sample were assembled using Cufflinks v2.2.1 (Trapnell et al. 2010). Assembly annotations from each sample were merged to produce the transcriptome annotation file. We checked whether this annotation contains transcripts not present in the r1.06 annotation, and found none. This file was then used to extract whole-transcript sequence (spliced exons) from the D. virilis reference genome for downstream analysis. We also assembled de novo transcriptomes for each species using Trinity r20140717 (Grabherr et al. 2011). Processed reads from all tissue types within species were pooled and assembled using default parameters.

Differential expression (DE) analysis

To perform DE analyses, we used the genome-based transcriptome to allow comparison across species. DE results using the de novo transcriptomes were virtually identical, but erroneously assembled transcripts complicate the analysis. Thus, we focus on the genome-based analysis. Reads from each sample were mapped to the transcriptome using Bowtie2, and abundance estimates were obtained using xExpress (Roberts and Pachter 2012). Read counts were normalized using the Trimmed Mean of M-values (TMM) method, and a read-count threshold of 200 counts in any given replicate was established as a filter for low-abundance transcripts.

The DE analysis consisted of calculating the fold-change and DE statistics (P-value and FDR) either in comparisons between a given tissue sample against remaining samples within species, or between tissue types across species (edgeR, Robinson et al. 2010). A transcript was considered significantly differentially expressed if abundance differed between samples by more than fourfold with a Bonferroni corrected P-value of <0.001. We defined tissue-biased transcripts as those that are significantly higher in abundance in that particular tissue relative to the remaining tissues.

We also used a tissue specificity index (S) to describe tissue-bias between species. In particular, we calculated specificity scores among the same tissue across the four species. The tissue-specificity index was calculated using the following formula:

\[ S_{g,i} = 1 - D(p_g, q_i) \]

where \( D \) is the Jensen-Shannon distance, \( p_g \) is the expression profile of a given gene \( g \), and \( q_i \) is a unit for complete specificity in a particular condition \( i \) (Cabili et al. 2011).

To examine the distribution of tissue-biased transcripts across D. virilis chromosomes, we calculated the expected number of tissue-biased transcripts on a given chromosome by multiplying the total number of tissue-biased transcripts in the genome by the proportion of all transcripts on that chromosome. The observed and expected number of tissue-biased transcripts were compared using a \( \chi^2 \) test.

Transcriptome annotation and gene ontology (GO) analysis

Genome-based and de novo transcriptomes were annotated using several bioinformatic tools. First, predicted open reading frames (ORFs) for the de novo transcriptomes were obtained using Trans-Decoder (Grabherr et al. 2011) (ORFs for the genome-based transcriptome were downloaded from FlyBase). Second, mRNA and polypeptide sequences were compared to the Swiss-Prot protein database (www.uniprot.org) using BLASTx and BLASTp, respectively (Altschul et al. 1990). Third, conserved protein domains, predicted signal peptides, and predicted transmembrane regions were identified using HMMP v.3.1b1 (Finn et al. 2011), SignalP v.4.1 (Petersen et al. 2011), and tmHMM v2.0 (Krogh et al. 2001), respectively. Finally, GO terms associated with each transcript were extracted from the TrEMBL and SwissProt databases (Trinotate v.2.0). GO term enrichment analyses were performed on sets of candidate tissue-biased genes (GOseq v.1.20.0, Young et al. 2010).

Coding sequence divergence

Coding sequence annotations (CDS) of D. virilis (r1.06) were used to extract CDS from the whole-genome assemblies of D. americana,
D. lummeii, and D. novamexicana described above (Cufflinks v.2.2.1, gffread utility). CDS that contained >20% missing bases (~1% of transcripts) were discarded; >98% of remaining CDS contained <5% missing base calls. CDS and protein sequences from the four species were aligned using ClustalW v.2.1. Alignments were used to estimate pairwise \( K_s/K_a \) between species and the mean \( K_s/K_a \) ratio across the virilis phylogeny (so, PAML v.4.3). In addition, we used PAML’s CODEML program to perform the “branch-site” test on each terminal branch (Yang 2007). In particular, the likelihood of a neutral model in which the \( K_s/K_a \) ratio is fixed at one is compared to the likelihood of a model in which \( K_s/K_a \) is estimated from the data along each branch of the virilis phylogeny (Yang and Nielsen 2002). The test statistic is obtained using the likelihood ratio test (LRT). LRT values were compared to the \( \chi^2 \) distribution with 1 d.f. Multiple test correction was carried out by calculating the false discovery rate (FDR) for each branch class.

Data availability
The raw Illumina reads are available through the Sequence Read Archive under project accession SRP100565. The processed data files and the analysis scripts are available on GitHub (github.com/YazBraimah/VirilisMaleRNAseq).

RESULTS
Transcripts that have specialized reproductive roles are likely to have higher abundance levels in reproductive tissues, or even show tissue-specific expression. We compared transcript abundance levels among tissues within species to identify tissue-biased transcripts. We classified these transcripts using a differential abundance level of more than fourfold with a \( P \)-value of <0.001. We used this higher-than-conventional cutoff to be conservative with respect to what is considered tissue-biased in the dataset. With these criteria, we identified 2493 transcripts that show strong expression bias in the male reproductive tract in all four species (Figure 2). We discuss each tissue-biased gene class below.

AG-biased transcripts and SFPs
The AG are the main source of SFPs, which play an important role in postcopulatory processes (Gillott 2003). SFPs have also been shown to evolve rapidly in postcopulatory processes (Gillott 2003). SFPs have also been shown to be conservative with respect to what is considered tissue-biased in the virilis phylogeny (Yang and Nielsen 2002). The test statistic is obtained using the likelihood ratio test (LRT). LRT values were compared to the \( \chi^2 \) distribution with 1 d.f. Multiple test correction was carried out by calculating the false discovery rate (FDR) for each branch class.

AG-biased genes are expected to be underrepresented on the X chromosome, particularly if those genes exhibit male-specific reproductive functions (Parisi et al. 2003). Indeed, we find that AG-biased transcripts and SFPs are not uniformly distributed across the genome (Figure 3). In particular, we calculated the expected number of tissue-biased genes on each chromosome, and found that the observed number of AG-biased genes and SFPs is significantly lower on the X chromosome than expected (AG-biased: \( \chi^2 = 10.1, P = 0.002 \); SFPs: \( \chi^2 = 8.1, P = 0.005 \)). Furthermore, SFPs were significantly enriched on chromosome 2 (\( \chi^2 = 5.7, P = 0.02 \)), and slightly overrepresented on chromosome 4 (\( \chi^2 = 9.2, P = 0.003 \)). The pattern of reduced representation of AG-biased genes on the X chromosome is consistent with findings in D. melanogaster (Ravi Ram and Wolfer 2007).

In summary, our method identified 71 shared SFPs. These SFPs show functional homology with SFPs in other insect species, and also obey the expected pattern of reduced representation on the X chromosome. Because our approach for identifying SFP candidates is based exclusively on expression abundance and in silico prediction of signal peptide sequence, we are likely underestimating the true number of total SFPs.

EB-biased transcripts
The EB is the source of several seminal fluid components in Drosophila in addition to the waxy substances found in the copulatory plug (Lung et al. 2001). A total of 421 transcripts are classified as EB-biased across the four virilis group species, but only 92 are shared (Figure 2 and Figure S2 in File S1). Of these 92, 20 contain predicted signal sequences, suggesting that these might be components of the ejaculate. Unlike AG-biased genes, EB-biased genes are not significantly underrepresented on the X chromosome (\( \chi^2 = 0.3; P = 0.5 \), Figure 3).

The EB contributes proteins to the seminal fluid; however, the roles of these proteins are poorly understood even in D. melanogaster. Some EB-biased transcripts may have evolutionary fates as SFPs derived from the AG since they are locked in similar coevolutionary dynamics with females. However EB-biased transcripts largely comprise distinct functional classes. In particular, EB-biased transcripts are significantly enriched for proteins that are involved in lipid metabolism, fatty acid biosynthesis, steroid metabolism, and coenzyme binding (Table S1 in File S1), consistent with the likely role of this organ in producing the mating-plug (Lung et al. 2001). For instance, a set of six genes are involved in elongation of very long chain fatty acids (GJ11026, GJ23058, GJ24115, GJ24118, GJ24167, and GJ24664). Another four genes (GJ19303, GJ21360, GJ22269, and GJ26512) are similar to putative fatty acyl-CoA reductases in D. melanogaster.

Finally, we found three EB-biased genes (GJ20447, GJ21330, GJ22262) that are orthologous to the D. melanogaster EB proteins, PEB-me, and PEB-III. These proteins are integral parts of the mating plug that forms at the vaginal canal after mating (Lung and Wolfer 2001). The mating plug is thought to enhance male reproductive success by minimizing sperm loss after copulation and facilitating storage (Bairati 1968). Thus, some of the genetic components of mating plug formation appear conserved among Drosophila species. This data suggests that different species may contain a different number of copies of PEB orthologs/paralogs.

Testis-biased transcripts
The largest set of DE genes are those with testis-biased expression. We identified 3179 genes that are testis-biased in at least one of the four species, but 2211 transcripts show testis-biased expression in all four
species (Figure 2 and Figure S2 in File S1). We also find that testis-biased do not show a pattern of over- or under-representation on the X chromosome ($\chi^2 = 0.83; P = 0.4$, Figure 3), but are significantly over-represented on chromosome 4 ($\chi^2 = 28.2; P = 1 \times 10^{-7}$, Figure 3). Genes with testis-biased expression have previously been shown to have a high turnover rate between species (Ellegren and Parsch 2007), and frequently arise on the X chromosome (Levine et al. 2006).

Another contrast between testis-biased transcripts and AG-biased genes is the low percentage of testis-biased transcripts that contain signal peptide sequences. In particular, only 6.3% of testis-biased transcripts contain predicted signal peptide sequences, whereas 37.3 and 21.7% of AG-biased and EB-biased genes, respectively, contain such sequences. This suggests that testes contribute few, if any, secretory proteins to the seminal fluid. This is further reflected in the overrepresented functional categories among testis-biased transcripts (Table S2 in File S1). Our analysis shows that testes are significantly enriched for intracellular proteins that are involved in gamete production, microtubule synthesis, mitotic processes, ATP binding, and flagellum-mediated motility. These categories highlight the array of genes involved in spermatogenesis and the development of the long sperm tail in these species (Pitnick et al. 1995). In addition, there are many genes (>100) that belong to individual functional categories (e.g., microtubule-based processes, mitotic division, and motor activity).

Production of large numbers of motile sperm is an important component of fitness in sperm competition (Parker 1982). The large number of testis-biased genes and their functional enrichment for sperm production and motility terms may reflect this (Cummins 2009; White-Cooper et al. 2009). Finally, the high rate of gene turnover among reproductive transcripts may represent a process by which novel genes can be recruited and, ultimately, new beneficial functions may arise. We explore these possibilities below.

**Lineage-specific expression patterns:** Lineage-specific genes can arise by changes in gene regulation through tissue-specific cooption of a promoter, or may arise de novo, such that a new gene is derived from previously noncoding DNA (Zhao et al. 2014). Both mechanisms can generate genetic and evolutionary novelty. To identify possible lineage-specific transcripts, we used two approaches. In the first approach, we (1) examined differential expression between genes that we classified as tissue-biased in any of the four species, and (2) calculated the specificity score for each tissue type across all species’ samples. (A high specificity score in this case means that both the tissue and the species show
exclusive expression of that transcript). This approach aims to identify transcripts that are divergent at the regulatory level, show tissue-biased expression in only one tissue type, and are largely confined to one species. In the second approach, we leverage the individual species’ de novo transcriptomes to identify transcripts that are present in one of the species but have no identifiable homologs in the other species.

Using the first approach, we identified a range of lineage-specific expression patterns for reproductive genes, and several cases where genes are expressed exclusively in one species (Figure 4 and Figure S3 in File S1). A significant pairwise difference in abundance (more than fourfold, FDR < 0.001) among tissue-biased genes and a conservative cutoff for the cross-species specific index (0.75) can point to cases where genes are exclusively expressed in one species. Indeed, we identify eight SFP candidates that show exclusive expression in one species (Figure 4). Two genes (GJ10897, GJ23780) are exclusively expressed in D. americana, one gene in D. lummei (GJ23872), two in D. novamexicana (GJ16171 and GJ14396), and three in D. virilis (GJ22755, GJ14247, and GJ13686). Similar lineage-specific patterns of expression are also observed in the other reproductive tissue-biased genes (Figure S3 in File S1). Recruiting new genes may be important in modifying a species’ reproductive capabilities, and thus these results provide opportunities for investigating the functional significance of expression divergence, particularly as it relates to reproduction.

In the second approach, we sought to determine whether any of the species contain de novo transcripts, which we define as being exclusively expressed in one species, and have no ortholog(s) in the other species. We accomplished this by querying tissue-biased transcripts, which we validate as derived from a given species’ genome, against the transcriptomes of the sister species. Under this especially restrictive condition, we find that unique transcripts are rare except in D. americana testes and the D. lummei EB, where 35 and 30 transcripts, respectively, have no homology to transcripts of the sister species (Figure S4 in File S1). The dynamics of losses and gains of reproductive genes among lineages appear complex, and beyond the scope of this current study.

In summary, lineage-specific transcripts that arise by evolutionary changes in gene regulation are common within male reproductive tissue, and, while our results indicate that some tissues may experience higher gene turnover rates within species, additional work is required to fully assess the evolutionary dynamic of gene gain and loss.

**Comparison of D. melanogaster SFPs to D. virilis:** The rapid evolution of SFPs in Drosophila suggests that distantly related species may contain different repertoires of these genes. D. melanogaster is ~40 MY divergent from D. virilis and is undoubtedly the species with the best characterized complement of SFPs. We first examined expression conservation between known D. melanogaster SFPs and their

---

---

Table 1  Significantly enriched GO terms among predicted SFPs (FDR < 0.05)

| Ontology | GO Term                          | # of Genes | P-Value       |
|----------|----------------------------------|------------|---------------|
| BP       | Proteolysis                      | 16         | 1.91E−08      |
|          | Digestion                        | 7          | 7.94E−08      |
|          | Protein metabolic process        | 23         | 3.14E−05      |
| CC       | Extracellular space              | 17         | 1.09E−06      |
|          | Endoplasmic reticulum lumen      | 6          | 0             |
| MF       | Serine-type peptidase activity   | 12         | 3.80E−09      |
|          | Serine hydrolase activity        | 12         | 5.75E−09      |
|          | Peptidase activity               | 16         | 5.97E−09      |
|          | Endopeptidase activity           | 13         | 3.79E−08      |
|          | Serine-type endopeptidase activity| 10        | 2.18E−07      |
|          | Hydrolase activity               | 24         | 4.45E−07      |
|          | Procollagen-proline dioxygenase activity | 4 | 1.10E−06      |
|          | Peptidyl-proline dioxygenase activity | 4 | 1.76E−06      |
|          | Carbohydrate binding             | 7          | 2.82E−06      |
|          | Monosaccharide binding           | 4          | 2.81E−05      |
|          | Oxidoreductase activity          | 4          | 3.93E−05      |
|          | Monosaccharide binding           | 4          | 2.81E−05      |
|          | Oxidoreductase activity          | 4          | 3.93E−05      |

aBP, Biological Process; CC, Cellular Component; MF, Molecular Function.

---

**Table 1 Significantly enriched GO terms among predicted SFPs (FDR < 0.05)**

![Figure 3](image-url)  
**Figure 3** Distribution of tissue-biased transcripts across chromosomes: The ratio of observed/expected number of genes on each chromosome averaged between species (dotted line shows random expectation). Significant departures from expectation are depicted by red asterisks (* P < 0.05, ** P < 0.01, *** P < 0.001).
orthologs in the virilis group; here, we relied on orthology assignments described in FlyBase. For the D. melanogaster expression data, we used publicly available RNA-seq data from three D. melanogaster tissues (AG, testes, and head; www.modencode.org).

Of 211 known SFPs in D. melanogaster (Findlay et al. 2008), 93 have clear orthologs in the virilis group (Figure 5). The percent identity among orthologous proteins ranges from 20 to 90%, and several clear orthologs in the virilis group (Figure 5). The percent identity

![Graph](image)

Figure 4 Differential abundance between candidate SFPs across species: cross-species specificity (S) is plotted as a function of fold-change in transcript abundance (log_2 scale) in pairwise comparisons between species (FDR < 0.001). Genes with S > 0.75 are indicated by name. Molecular evolution of male reproductive genes: Reproductive genes can be targets of selection, partly because they play a role in sperm competition, or because male and female reproductive genes coevolve as a consequence of cryptic female choice and/or conflicting reproductive interests. Regardless of the mechanism of PCSS, molecular evolutionary analysis of reproductive genes can reveal the impact of such processes on rates of codon substitution and divergence between species. Here, we use the ratio of nonsynonymous to synonymous (K_s/K_a) codon substitutions to (1) examine the difference in average K_s/K_a across the virilis subgroup phylogeny (ω) for each tissue-biased gene category, (2) examine pairwise K_s/K_a among SFPs, and (3) test for evidence of adaptive codon substitutions among sites within coding sequences and along branches of the phylogeny using PAML’s branch-site test (Yang and Nielsen 2002).

First, we calculated the mean K_s/K_a value (ω) across the phylogeny for each tissue-biased category of genes, and tested for significant differences from the genome average (Figure 6). All tissue-biased categories deviate significantly from the genome average (ω = 0.15), with EB-biased transcripts showing lower ω (ω = 0.11, W = 1.3 × 10^6, P = 0.02; Wilcoxon rank-sum). AG-biased (ω = 0.3) and testes-biased genes (ω = 0.23), on the other hand, show significantly higher mean ω than the genome average (W = 2.9 × 10^6, P < 0.001, and W = 3.4 × 10^7, P < 0.001, respectively; Wilcoxon rank-sum). Among AG-biased genes, those we classify as SFP candidates have the highest mean ω (ω = 0.39, W = 1.3 × 10^6, P < 0.001; Wilcoxon rank-sum). Thus, our results show that AG-derived proteins experience higher selective pressures than other reproductive gene classes.

Nearly all crosses between virilis group members result in strong gametic incompatibilities (PMPZ), suggesting that PCSS processes within species drove significant differentiation of genes involved in postmating interactions. A striking pattern of PMPZ in this group is the strong incompatibility between D. americana males and females from the sister species. Quantitative trait loci (QTL) that contribute to this paternal incompatibility between D. americana and D. virilis females have been mapped in two studies (Sweigart 2010; Ahmed-Braimah 2016), and include a large inversion on chromosome 2 and several adjacent QTL on the centromeric half of chromosome 5. We found that the four SFPs with the highest ω values reside within the chromosome 2 inversion (Figure 7). While other tissue-biased categories do not show this pattern, some of these tissue-biased transcripts with elevated ω coincide with PMPZ QTL (Figure S5 in File S1).

Next, we examined pairwise K_s/K_a for each species pair (six pairwise comparisons) among SFPs, as some lineage-specific patterns might be missed by looking at the average across the phylogeny. Here we find a striking correspondence between SFPs with elevated K_s/K_a and PMPZ QTL on chromosome 2 (Figure S6 in File S1). Similar to the ω pattern described above, several SFPs with K_s/K_a > 1 occur on the
Comparisons involving *D. americana* in particular show 3–4 genes with this pattern, while those comparisons excluding *D. americana* show ≤2 genes. None of the SFPs on the chromosome 5 QTL have $K_{a}/K_{s} > 1$, but a cluster of three elevated genes ($K_{a}/K_{s} > 0.5$) coincide with one of the chromosome 5 QTL. These results suggest that SFPs on chromosome 2 substantially contribute to overall SFP divergence among these species, and particularly along the *D. americana* lineage.

**Figure 5** *D. melanogaster* SFPs and their orthologs in *D. virilis*: expression heatmap of *D. melanogaster* SFPs (right) and their orthologs in *D. virilis* (left). The heat-scale shows expression values in median-centered log$_{2}$ TPM (*virilis* data) or RPKM (*D. melanogaster* data). The left cladogram shows complete-linkage clustering relationships of *virilis* group orthologs, and the color key on the left indicates shared tissue-bias status among *virilis* group species. The cladogram on top of the *virilis* group heatmap depicts species clustering (abbreviations as in Figure 2), and the cladogram on top of the *D. melanogaster* data depicts sample clustering.
Finally, to test for site-specific signatures of positive selection among tissue-biased genes, we performed the branch-site test implemented in PAML (Yang and Nielsen 2002; Yang 2007). This test is conservative, and has higher power to detect positive selection with higher divergence times than the virilis group splits (<5 MYA) and a larger number of species in the alignment (Anisimova et al. 2001; Wong et al. 2004). To perform the test, we calculated the log likelihood of a model where \( \omega = 1 \) (neutral), and compared that to a model in which \( \omega \) is estimated from the sequence alignment (positive selection). This comparison can be performed using the LRT, where the test statistic under the null hypothesis follows a \( \chi^2 \) distribution with 1 d.f. Given that this test is underpowered for our analysis, a signature of positive selection here likely reflects strong adaptive change.

We performed the PAML test on all coding sequences in the genome and calculated the FDR along each of the four branches. We found 92 genes out of 15,078 that show a signature of adaptive evolution and calculated the FDR along each of the four branches. We found the AG-biased transcripts category. The “all” category represents mean \( \omega \) for all transcripts in the genome. Error bars represent SE.

**DISCUSSION**

We identified male reproductive genes in the virilis subgroup using RNA-seq and differential transcript abundance. Using this approach, we provided an overall description of the evolutionary dynamics of these genes in terms of expression divergence, functional classification, distribution across the genome, and sequence evolution. Because these genes play important roles in gametic interactions, understanding their evolutionary dynamics is critical to gain insights into the genetic basis of PCSS and PMPZ.

Testes-biased genes dominate the repertoire of reproductive genes, with >2000 genes having strong expression bias in that tissue. These genes are highly enriched for GO terms that are linked to gamete development and to basic cell biological processes, often with several hundred genes belonging to individual GO terms. It remains largely a mystery how these genes contribute to variation in sperm competition phenotypes, such as sperm number and length. Virilis group species feature long sperm that is \(~\approx\)2–3 times that of \( D. melanogaster \), suggesting sperm length likely plays an important role in postcopulatory competition within the reproductive tract of females (Pitnick et al. 1999). Additional work is needed to uncover the functional importance of the large number of testes-biased genes in affecting PCSS processes.

EB proteins are also key players in postmating interactions in \( D. sophila \), and some of these proteins play a key role in mating plug formation (Lung and Wolfrer 2001). The mating plug is a nearly ubiquitous component of male ejaculates in many animal taxa, and is thought to facilitate sperm movement, prevent sperm loss, and prevent subsequent matings (Avila et al. 2015). We have identified 92 genes that show strong expression bias in the EB, and the enriched GO categories among them suggest they are involved in lipid biosynthesis, consistent with their presumed functional roles. In \( D. melanogaster \), three proteins (Peb, Peb-II, and Peb-III) play a part in mating plug formation. We identified three genes among virilis species that show homology to two of the \( D. melanogaster \) Peb genes. This suggests that some aspects of mating plug formation are conserved in \( D. sophila \).

The AG are the main source of seminal fluid proteins in \( D. sophila \), and many studies in insects have implicated their importance in postmating interactions (Avila et al. 2011). In \( D. melanogaster \), several of these proteins induce a variety of postmating effects in females (Wolfrer 2007). Several SFPs are also known to evolve rapidly among closely related species (Swanson et al. 2001), which suggests that these genes may be the main targets of PCSS. Because of their rapid evolution, distantly related species may differ in SFP content and/or sequence (Kelleher et al. 2009). It is thus important to independently identify SFPs in species that are distantly related to \( D. melanogaster \) to gain insights into various genetic mechanisms of PCSS.

Similar to other \( D. sophila \) species that produce many proteases in the AG (Findlay et al. 2008; Kelleher et al. 2009), we found that proteins with proteolytic function are enriched among AG-biased genes and SFPs in the virilis group, suggesting that proteases are a conserved functional class among male seminal proteins in the \( D. sophila \) genus. We also found that nearly half of known SFPs in \( D. melanogaster \) have clear orthologs in the virilis group, and nearly a quarter of those show expression conservation. This latter set of highly conserved SFPs across the \( D. sophila \) genus likely
play key roles in postmating interactions. Unfortunately, the majority of these proteins have not yet been characterized in D. melanogaster. Three of them, however, were recently shown to act in the Sex Peptide network in D. melanogaster by facilitating transfer of other SFPs and binding of SP to sperm (Findlay et al. 2014). The majority of D. melanogaster SFP orthologs in D. virilis do not show AG-biased expression among virilis species. These orthologs may have diverged at the expression level and/or have been recruited to other functions. Nonetheless, our general finding of significant similarity between this putatively rapidly evolving class of genes is surprising.

SFPs in the virilis group show the highest rate of nonsynonymous amino acid substitution among the reproductive gene classes. Because SFPs have not been studied in the virilis group, identifying rapidly evolving AG-biased transcripts may reveal interesting candidate genes for further study. The prevalence of gametic incompatibilities as an isolating barrier among members of this group further highlights the utility of investigating rapidly evolving SFPs and their role in PCSS and PMPZ. Indeed, rapidly evolving SFPs in the virilis group are associated with known paternal PMPZ regions. The four SFPs with the highest codon substitution rates reside within a major PMPZ QTL, which is the site of a fixed inversion between D. americana and D. virilis.

The approach that we have used to identify reproductive genes has several strengths. First, RNA-seq provides information on both RNA abundance and sequence, information that is needed to identify transcripts with tissue-biased expression and to measure rates of nucleotide substitution. Second, by sequencing transcripts from each of the three main reproductive organs from the four species we are able to examine rates of loss/gain of reproductive proteins between species, a phenomenon that can play an important role in reproductive divergence between species. Finally, RNA-seq allows considerable improvement in gene annotations and can identify many new transcripts and splice variants.

Identifying reproductive genes via our approach does have disadvantages. For example, a gene might well play an important role in reproductive processes but not show expression that is specific to reproductive tissues. This limitation can obviously compromise our ability to identify biologically important transcripts. We also used stringent cutoffs in our classification of differentially expressed transcripts to avoid cases of uncertainty due to large variance among replicates, especially with transcripts that have low abundance. Finally, because our inference of secreted ejaculate proteins relies on in silico prediction of signal peptides, we may miss proteins that are present in the male ejaculate (e.g., membrane-bound vesicles) but do not contain a signal peptide. Despite these caveats, our approach succeeded in identifying many candidate reproductive transcripts for each tissue type, allowing preliminary analysis of such genes in this group.

In summary, we have reached several broad conclusions that establish parallels between the virilis and melanogaster species groups in SFP evolution. First, SFPs show an elevated rate of nonsynonymous codon substitution, and that the most rapidly evolving SFPs coincide with known paternal PMPZ QTL. Second, candidate reproductive genes are evolutionarily dynamic such that species may differ in reproductive transcript content, often due to regulatory divergence. Third, AG-biased transcripts are underrepresented on the X chromosome relative to autosomes. Finally, we identify several orthologs of D. melanogaster SFPs, with a subset showing conserved expression patterns suggesting likely functional conservation. Our data and findings provide a powerful platform for further studies of reproductive gene evolution in the virilis group.

ACKNOWLEDGMENTS

We thank Amanda Manfredo for preparing the sequencing libraries, and David Lambert and Daven Presgraves for comments on earlier versions of the manuscript. We especially thank H. Allen Orr for guidance and support throughout this work. This work was supported by National Institutes of Health grants R01-GM051932 to H. Allen Orr and R01-HD059060 to A.G.C.

LITERATURE CITED

Aguadé, M., 1999 Positive selection drives the evolution of the Acp29AB accessory gland protein in Drosophila. Genetics 152: 543–551.
Ahmed-Braimah, Y. H., 2016 Multiple genes cause postmating prezygotic reproductive isolation in the Drosophila virilis group. G3 (Bethesda) 6: 4067–4076.
Ahmed-Braimah, Y. H., and B. F. McAllister, 2012 Rapid evolution of assortative fertilization between recently allopatric species of Drosophila. Int. J. Evol. Biol. 2012: 1–9.
Autschl, S. F., W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, 1990 Basic local alignment search tool. J. Mol. Biol. 215: 403–410.
Anisimova, M., J. P. Bielawski, and Z. Yang, 2001 Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol. Biol. Evol. 18: 1585–1592.
Avila, F. W., L. K. Sirot, B. A. LaFlamme, C. D. Rubinstein, and M. F. Wolfer, 2011 Insect seminal fluid proteins: identification and function. Annu. Rev. Entomol. 56: 21–40.
Avila, F. W., A. Wong, J. L. Sitnik, and M. F. Wolfer, 2015 Don’t pull the plug! the Drosophila mating plug preserves fertility. Fly (Austin) 9: 62–67.
Baer, B., J. L. Heazlewood, N. L. Taylor, H. Eubel, and A. H. Millar, 2009 The seminal fluid proteome of the honeybee Apis mellifera. Proteomics 9: 2085–2097.
Bairati, A., 1968 Structure and ultrastructure of the male reproductive system in Drosophila melanogaster meig. Monitore Zoologico Italiano-Italian Journal of Zoology: 2: 105–182.
Begun, D. J., and H. A. Lindfors, 2005 Rapid evolution of genomic Acp complement in the melanogaster subgroup of Drosophila. Mol. Biol. Evol. 22: 2010–2021.
Birkhead, T. R., and T. Pizzari, 2002 Evolution of sex: postcopulatory sexual selection. Nat. Rev. Genet. 3: 262–273.
Cabili, M. N., C. Trapnell, L. Goff, M. Koziel, B. Tazon-Vega et al., 2011 Integrative annotation of human large intergenic non-coding RNAs reveals global properties and specific subclasses. Genes Dev. 25: 1915–1927.
Chapman, T., 2001 Seminal fluid-mediated fitness traits in Drosophila. Heredity 87: 511–521.
Cummins, J., 2009 Sperm motility and energetics, pp. 185–206 in Sperm Biology. Elsevier, Amsterdam.
Dean, M. D., G. D. Findlay, M. R. Hoopmann, C. C. Wu, M. J. MacCoss et al., 2011 Identification of ejaculated proteins in the house mouse (Mus domesticus) via isotopic labeling. BMC Genomics 12: 306.
Ellegren, H., and J. Parsch, 2007 The evolution of sex-biased genes and sex-biased gene expression. Nat. Rev. Genet. 8: 689–698.
Findlay, G. D., X. Yi, M. J. MacCoss, and W. J. Swanson, 2008 Proteinomics reveals novel Drosophila seminal fluid proteins transferred at mating. PLoS Biol. 6: e178.
Findlay, G. D., J. L. Sitnik, W. Wang, C. F. Aquadro, N. L. Clark et al., 2014 Evolutionary rate covariation identifies new members of a protein network required for Drosophila melanogaster female post-mating responses. PLoS Genet. 10: e1004108.
Finn, R. D., J. Clements, and S. R. Eddy, 2011 HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 39: W29–W37.
Gillcott, C., 2003 Male accessory gland secretions: modulators of female reproductive physiology and behavior. Annu Rev Entomol. 48: 163–184.
Grabherr, M. G., B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson et al., 2011 Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. 29: 644–652.
Haerty, W., S. Jagadeeshan, R. J. Kulathinal, A. Wong, K. Ravi Ram et al., 2007 Evolution in the fast lane: rapidly evolving sex-related genes in Drosophila. Genetics 177: 1321–1335.
Herndon, L. A., and M. F. Wolfer, 1995 A Drosophila seminal fluid protein, Acp26Aa, stimulates egg laying in females for 1 day after mating. Proc. Natl. Acad. Sci. USA 92: 10114–10118.
