Molecular Evolution and Population Genetics of Duplicated Accessory Gland Protein Genes in *Drosophila*

Alisha K. Holloway* and David J. Begun†

*Section of Integrative Biology, University of Texas-Austin and †Section of Evolution and Ecology, University of California-Davis

To investigate the potential importance of gene duplication in *D. melanogaster* accessory gland protein (*Acp*) gene evolution we carried out a computational analysis comparing annotated *D. melanogaster Acp* genes to the entire *D. melanogaster* genome. We found that two known *Acp* genes are actually members of small multigene families. Polymorphism and divergence data from these duplicated genes suggest that in at least four cases, protein divergence between *D. melanogaster* and *D. simulans* is a result of directional selection. One putative *Acp* revealed by our computational analysis shows evidence of a recent selective sweep in a non-African population (but not in an African population). These data support the idea that selection on reproduction-related genes may drive divergence of populations within species, and strengthen the conclusion that *Acps* may often be under directional selection in *Drosophila*.

At least three classes of models have been proposed to explain the evolutionary processes for the retention and subsequent divergence of gene duplicates. Lynch and Force (2000) suggest that ancestral genes with multiple functions in different tissues or developmental stages may have high rates of retention of duplicates under mutation-selection balance. In this model, degenerative mutations result in subfunctionalization, which favors retention and subsequent evolution of duplicates. A second class of models invokes fixation of duplications by genetic drift (e.g., Lynch and Conery 2003; Walsh 2003). Finally, a third class of models relies on new, beneficial mutations driving adaptive divergence (and thus retention) of duplicates (Hughes 1994). One would expect new duplicates from classes of proteins under chronic directional selection to have unusually high fixation probabilities because a higher proportion of new mutations may be beneficial in such genes. For example, if reproduction-related proteins experience directional selection more frequently than other proteins (Civetta and Singh 1998; Nurminsky et al. 1998; Swanson and Vacquier 2002; Ranz et al. 2003), then perhaps a large number of duplicate reproduction-related genes spread through populations and diverge under directional selection.

We investigated duplication and divergence in reproduction-related accessory gland proteins genes (*Acps*) in Drosophila. *Acps* are male-specific seminal fluid proteins that affect multiple aspects of female physiology and behavior (for review see Wolfner 1997). We carried out Blast comparisons of the 13 annotated *Acps* (see *Methods*) to the *D. melanogaster* reference sequence (Flybase Consortium 2003). These Blast analyses suggested that two genes, *Acp29AB* and *Acp53Ea*, are members of small multigene families.

E-values returned from the tBlastN search (default parameters) with *Acp29AB* as the query sequence were 1.5×10^{-47} and 2.6×10^{-35} for *Lectin29Ca* and *Lectin30A*, respectively. Intraspecific paralogous protein divergence was, on average, 31% between *Acp29AB* and *Lectin29Ca*,

35.5% between *Acp29AB* and *Lectin30A*, and 38% between *Lectin29Ca* and *Lectin30A*. *Lectin29Ca* is 356 bases distal to *Acp29AB* and *Lectin30A* is approximately 1 Mbase distal to these tandem duplicates. *Acp29AB* is 234 amino acids, while *Lectin29Ca* and *Lectin30A* are 236 and 223 amino acids long, respectively. Each gene is composed of a single exon. Our analysis of *Lectin30A* and comparison to its paralogs suggested that the 5′ end was incorrectly annotated. We confirmed this hypothesis by RACE, and we used our annotation in all analyses. The three members of the *Acp29AB* family are predicted to be lectin galactose binding proteins (Theopold et al. 1999) and to have a signal sequence (SignalP v2.0, Nielsen et al. 1997). The tBlastN search returned several other more distantly related putative *Acp29AB* paralogs, primarily lectins (*Lectin21Cb, Lectin24Db, Lectin22C, Lectin 21Ca, Lectin24A, Lectin28C*, and CG15818). However, we will not present data from these genes in this report.

E-values returned from the tBlastN search with *Acp53Ea* as the query sequence were 2.1×10^{-5} for CG8626 and 9.4×10^{-4} for CG15616. CG8626 and CG15616 will hereafter be referred to as *Acp53C14a* and *Acp53C14b*, respectively, based on putative function, genomic location, and gene structure. Another more highly diverged putative duplicate identified by B. Wagstaff (personal communication) did not appear in our Blast results. However, this gene (*Acp53C14c*) appears to be another tandem duplicate and shows male-limited expression (B. Wagstaff, personal communication). Intraspecific paralogous protein divergence was 48.5% between *Acp53Ea* and *Acp53C14a*, 42.5% between *Acp53Ea* and *Acp53C14b*, and 45% between *Acp53C14a* and *Acp53C14b*. The divergence of *Acp53C14c* from other putative *Acp53Ea* duplicates was >65%. These genes are tandem duplications, with *Acp53C14a* located 423 bp proximal to *Acp53C14b*, *Acp53Ea* 487 bases distal to *Acp53C14b*, and *Acp53C14c* 519 bp distal to *Acp53Ea*. *Acp53Ea, Acp53C14a, Acp53C14b*, and *Acp53C14c* are predicted to be 120, 121, 132, and 124 amino acids long, respectively. Each is composed of two exons with a 50–60 nt intron roughly 40 bases from the initiation codon. All genes are predicted to be peptide hormones and to have a signal sequence (SignalP v2.0, Nielsen et al. 1997).

The high levels of silent and replacement divergence among putative paralogs suggest that the duplication events

Key words: accessory gland protein, Drosophila, gene duplication, gene expression, molecular evolution, selection.
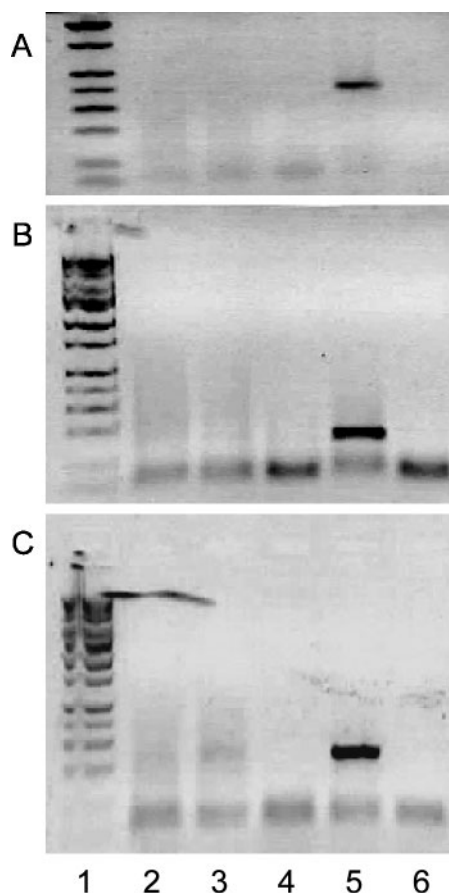
E-mail: aholloway@mail.utexas.edu.

FIG. 1.—RT-PCR analysis of tissue-specific expression of putative duplicates. (*A*) *Lectin29Ca*, (*B*) *Acp53C14a*, and (*C*) *Acp53C14b*. Lane assignments for each gel: (1) 1 kB ladder, (2) whole females, (3) males without reproductive tracts, (4) testes, (5) accessory glands, and (6) negative control.

predate the split of *D. yakuba* from the *D. melanogaster/ D. simulans* lineage. Nevertheless, the conserved gene structures, expression patterns, presence of predicted signal peptides, and, for most cases, tandem organization all indicate that we have correctly identified paralogous genes.

Acp29AB and *Acp53Ea* are expressed only in male accessory glands (Wolfner et al. 1997). Our RT-PCR experiments showed that the only detectable expression of *Lectin29Ca, Acp53C14a*, and *Acp53C14b* is in accessory glands (fig. 1). We were unable to detect an RT-PCR product from *Lectin30A*. However, given that our RACE products were derived from male cDNA, we are certain the gene is expressed in males (perhaps at low levels). The fact that the *Acp* duplicates identified here share accessory-gland enriched expression further supports the inference of paralogy and suggests that subfunctionalization, at least with respect to gene expression (*sensu* Lynch and Force 2000), cannot explain fixation of *Acp* duplicates. Levels of protein polymorphism and divergence for these *Acp* genes and putative duplicates (table 1) were higher than those typically seen in *D. simulans* and *D. melanogaster* genes, as was the case for previous surveys of *Acp* variation (Begun et al. 2000; Swanson et al. 2001). There was, however, a major exception in the *Acp29AB* family.

Table 1
Silent and Replacement Site Heterozygosity and Divergence for *Acp29AB* and *Acp53Ea* Gene Families in *D. melanogaster* and *D. simulans*

| Gene | No. of Sites | | Sample | θ_{sil} | θ_{repl} | Div.$_{sil}$[a] | Div.$_{repl}$[a] |
	Sil.	Repl.					
Acp29AB	148	554	*mel*	0.0209	0.0028	0.2368	0.0772
			sim	0.0359	0.0052		
Lectin29Ca	143	559	*mel*	0.0000	0.0007	0.1745	0.0807
			Af *mel*[b]	0.0142	0.0116	0.1841	0.0848
			sim	0.0286	0.0073		
Lectin30A	137	532	*mel*	0.0112	0.0029	0.1262	0.0579
			sim	0.0178	0.0015		
Acp53Ea	88	266	*mel*	0.0399	0.0066	0.1574	0.0390
			sim	0.0046	0.0061		
Acp53C14a	85	278	*mel*	0.0174	0.0013	0.1255	0.0068
			sim	0.0289	0.0029		
Acp53C14b	93	303	*mel*	0.0435	0.0024	0.1689	0.0352
			sim	0.0307	0.0027		
Acp53C14c	92	280	*mel*	0.0401	0.0058	0.1387	0.0823
			sim	0.0595	0.0156		
19 other Genes 3R[c]	286 (19)	935 (65)	*sim*	0.0349 (0.0044)	0.0013 (0.0003)	0.1084 (0.0097)	0.0107 (0.0032)

NOTE.—Sil. = silent sites in coding regions; Repl. = replacement.

[a] Divergence (Div.) is between all pairs of *D. melanogaster* and *D. simulans* genes (Jukes-Cantor corrected).

[b] Malawi, Africa.

[c] Data are means with standard error values below in parentheses for 19 unrelated proteins on chromosome arm 3R from Begun and Whitley (2000) and Begun (2002).

Lectin29Ca had no silent polymorphisms and only a single replacement polymorphism in our U.S. *D. melanogaster* sample (table 2). This is highly unusual given the relatively high levels of variation in *D. melanogaster* generally and in *Acp* genes specifically. Low levels of heterozygosity are even more surprising given high levels of silent and replacement divergence at this gene. We used the HKA test (Hudson, Kreitman, and Aguade 1987) to compare polymorphism and divergence data from *Lectin29Ca* and *vermilion* (we chose *vermilion* because of the availability of molecular population genetic data for both U.S. and African samples and because there is no evidence that *vermilion* has been subject to directional or balancing selection [Begun and Aquadro 1995]). *Lectin29Ca* versus the coding region of *vermilion* for the U.S. sample showed a highly significant departure from neutrality ($P = 0.0022$), consistent with a hitchhiking event reducing variation in *Lectin29Ca*. In contrast, the African sample showed considerably higher levels of polymorphism but similar levels of divergence (table 2). The HKA test of African variation at *Lectin29Ca* and *vermilion* showed no deviation from the neutral model ($P = 0.7875$). Given that the North American populations are thought to be recently derived from ancestral African populations (David and Capy 1988), the data support the idea that a selective sweep at *Lectin29Ca* occurred in the very recent past. Note that *Acp29AB* is only 356 bp upstream of *Lectin29Ca*, yet it shows normal levels of heterozygosity in the California sample. This suggests that the window of reduced heterozygosity associated with *Lectin29Ca* may be quite small, though additional population genetic data from the other flanking region

Table 2
Silent and Replacement Variation for *Acp29AB* and *Acp53Ea* Gene Families in *D. melanogaster* and *D. simulans*

Gene		Polymorphic		Fixed		
		Silent	Repl.	Silent	Repl.	Prob.[a]
Acp29AB	*sim*	13	7			
	mel	8	4	24	33	**0.032**
Lectin29Ca	*sim*	10	10			
	mel	0	1	20	38	0.292
	Af *mel*[b]	15	26	21	36	0.979
Lectin30A	*sim*	6	2			
	mel	4	4	12	27	**0.030**
Acp53Ea	*sim*	1	4			
	mel	8	4	10	7	0.730
Acp53C14a	*sim*	6	2			
	mel	4	1	5	1	0.746
Acp53C14b	*sim*	7	2			
	mel	11	2	8	9	**0.022**
Acp53C14c	*sim*	10	8			
	mel	9	4	6	16	**0.025**

[a] Probability determined by G-test; bold type indicates p < 0.05.
[b] Malawi, Africa.

would be necessary to determine if this is indeed the case. It seems notable that of three solid cases of individual genes with evidence for recent selection in non-African *D. melanogaster* samples (*desat*, Takahashi et al. 2001; *Acp36DE*, Begun et al. 2000; *Lectin29Ca*, this report), two are *Acps*. Such observations support the notion that selection on sexual traits can cause rapid divergence of Drosophila populations (Knowles and Markow 2001; Miller, Starmer, and Pitnick 2003; Pitnick et al. 2003). Additional work will be required to precisely determine the extent of the "swept" region of *Lectin29Ca* in non-African populations and to identify putative mutations that might be targets of selection.

Numbers of polymorphic and fixed, silent, and replacement mutations were compared to predictions of the neutral model (McDonald and Kreitman 1991). Of the six genes in our study, four (*Acp29AB, Lectin30A, Acp53C14b*, and *Acp53C14c*) reject the null hypothesis in a direction consistent with adaptive protein evolution (table 2). Moreover, our population genetic data support the notion that all three members of the *Acp29AB* family have been influenced by recent directional selection. Overall, our results put on firmer ground the conclusion that adaptive protein evolution is a major cause of divergence of *Acp* proteins in *D. melanogaster* and *D. simulans*.

Methods

Amino acid sequences of 13 annotated *Acps* from *D. melanogaster* (*Acp26Aa, Acp26Ab, Acp29AB, Acp32CD, Acp33A, Acp36DE, Acp53Ea, Acp62F, Acp63F, Acp70A, Acp76A, Acp95EF, and Acp 98AB*) were compared to the *D. melanogaster* reference sequence (Genome Release 3.0, Flybase Consortium 2003) by tBlastN searches using default parameters. Sequences were aligned by manual curation and molecular population genetics of putative duplicates with <50% nucleotide divergence from a known *Acp* were investigated. Nucleotide divergence

(or uncorrected pairwise-distance) was calculated as a measure of the number of mismatched nucleotides/total number of nucleotides at 1st and 2nd codon positions. Duplicate genes are and considered putative *Acps* but are referred to as *Acps* for convenience.

Candidate *Acp* duplicates were subjected to RT-PCR to determine whether their expression was restricted to accessory glands. mRNA was extracted from four tissues of *D. melanogaster*: male accessory glands, testes, male carcasses, and whole females. First strand synthesis was carried out using an oligo-dT primer and Superscript Reverse Transcriptase (Invitrogen, San Diego, Calif.). RT-PCR was carried out using gene-specific primers on RNA/DNA heteroduplex isolated from each tissue.

D. simulans population genetic data are from inbred lines established from flies collected at the Wolfskill Orchard in Winters, Calif. (Begun and Whitley 2000). *D. melanogaster* population data are from isochromosomal lines derived from the Wolfskill Orchard or from isofemale lines from Malawi, Africa. *D. yakuba* sequences are from an isofemale line. Some *D. simulans Acp29AB* sequences are from Begun et al. (2000). In most cases DNA sequencing was carried out directly on PCR products. For cases in which inbred lines were not available, PCR products were cloned prior to sequencing.

Sequences were assembled using SeqMan (DNAstar, Inc., Madison, WI) and manually curated in MacClade 4.0 (Maddison and Maddison 2000). Alignments are available upon request from the authors. Summary statistics and tests of the neutral equilibrium model were carried out using DnaSP version 3.53 (Rozas and Rozas 1999). SignalP version 2.0 was used to predict presence/absence of signal peptides characteristic of *Acps* and other secreted proteins (Nielsen et al. 1997). Sequences were submitted to GenBank under accession numbers AY635196-AY635290.

Literature Cited

Begun, D. J. 2002. Protein variation in *Drosophila simulans*, and comparison of genes from centromeric versus noncentromeric regions of chromosome 3. Mol. Biol. Evol. **19**:201–203.

Begun, D. J., and C. F. Aquadro. 1995. Molecular variation at the *vermilion* locus in geographically diverse populations of *Drosophila melanogaster* and *D. simulans*. Genetics **140**:1019–1032.

Begun, D. J., and P. Whitley. 2000. Reduced X-linked nucleotide polymorphism in *Drosophila simulans*. Proc. Natl. Acad. Sci. USA **97**:5960–5965.

Begun, D. J., P. Whitley, B. L. Todd, H. M. Waldrip-Dail, and A. G. Clark. 2000. Molecular population genetics of male accessory gland proteins in Drosophila. Genetics **156**:1879–1888.

Civetta, A., and R. S. Singh. 1998. Sex-related genes, directional sexual selection, and speciation. Mol. Biol. Evol. **15**:901–909.

David, J., and P. Capy. 1988. Genetic variation of *Drosophila melanogaster* natural populations. Trends Genet. **4**:106–111.

Flybase Consortium. 2003. The FlyBase database of the Drosophila genome projects and community literature. Nucleic Acids Res. **31**:172–175. (www.flybase.org.)

Hudson, R. R., M. Kreitman, and M. Aguade. 1987. A test of neutral molecular evolution based on nucleotide data. Genetics **116**:153–159.

Hughes, A. L. 1994. The evolution of functionally novel proteins after gene duplication. Proc. R. Soc. Lond. B Biol. Sci. **256**:119–124.

Knowles, L. L., and T. A. Markow. 2001. Sexually antagonistic coevolution of a postmating-prezygotic reproductive character in desert Drosophila. Proc. Natl. Acad. Sci. USA **98**:8692–8696.

Lynch, M., and J. S. Conery. 2003. The origins of genome complexity. Science **302**:1401–1404.

Lynch, M., and A. Force. 2000. The probability of duplicate gene preservation by subfunctionalization. Genetics **154**:459–473.

Maddison, D. R., and W. P. Maddison. 2000. MacClade 4: analysis of phylogeny and character evolution. Sinauer Associates, Sunderland, Mass.

McDonald, J. H., and M. Kreitman. 1991. Adaptive protein evolution at the *Adh* locus in Drosophila. Nature **351**:652–654.

Miller, G. T., W. T. Starmer, and S. Pitnick. 2003. Quantitative genetic analysis of among-population variation in sperm and female sperm-storage organ length in *Drosophila mojavensis*. Genet. Res. **81**:213–220.

Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10**:1–6.

Nurminsky, D. I., M. V. Nurminskaya, D. De Aguiar, and D. L. Hartl. 1998. Selective sweep of a newly evolved sperm-specific gene in Drosophila. Nature **396**:572–575.

Pitnick, S., G. T. Miller, K. Schneider, and T. A. Markow. 2003. Ejaculate-female coevolution in *Drosophila mojavensis*. Proc. R. Soc. Lond. B Biol. Sci. **270**:1507–1512.

Ranz, J. M., C. I. Castillo-Davis, C. D. Meiklejohn, and D. L. Hartl. 2003. Sex-dependent gene expression and evolution of the Drosophila transcriptome. Science **300**:1742–1745.

Rozas, J., and R. Rozas. 1999. DnaSP 3: an integrated program for molecular population genetics and molecular evolution analysis. Bioinformatics **15**:174–175.

Swanson, W. J., A. G. Clark, H. M. Waldrip-Dail, M. F. Wolfner, and C. F. Aquadro. 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc. Natl. Acad. Sci. USA **98**:7375–7379.

Swanson, W. J., and V. D. Vacquier. 2002. The rapid evolution of reproductive proteins. Nat. Rev. Genet. **3**:137–144.

Takahashi, A., S.-C. Tsaur, J. A. Coyne, and C.-I. Wu. 2001. The nucleotide changes governing cuticular hydrocarbon variation and their evolution in *Drosophila melanogaster*. Proc. Natl. Acad. Sci. USA **98**:3920–3925.

Theopold, U., M. Rissler, M. Fabbri, O. Schmidt, and S. Natori. 1999. Insect glycobiology: a lectin multigene family in *Drosophila melanogaster*. Biochem. Biophys. Res. Comm. **261**: 923–927.

Walsh, J. B. 2003. Population-genetic models of the fates of duplicate genes. Genetica **118**:279–294.

Wolfner, M. F. 1997. Tokens of love: functions and regulation of Drosophila male accessory gland products. Insect Biochem. Mol. Biol. **27**:179–192.

Wolfner, M. F., H. A. Harada, M. J. Bertram, T. J. Stelick, K. W. Kraus, J. M. Kalb, Y. O. Lung, D. M. Neubaum, M. Park, and U. Tram. 1997. New genes for male accessory gland proteins in *Drosophila melanogaster*. Insect Biochem. Mol. Biol. **27**:825–834.