doi:10.1017/S0016672311000346

Functional genome annotation of *Drosophila* seminal fluid proteins using transcriptional genetic networks

JULIEN F. AYROLES[1,2†], BROOKE A. LAFLAMME[3†], ERIC A. STONE[1,2], MARIANA F. WOLFNER[3] AND TRUDY F. C. MACKAY[1,2*]

[1] *Department of Genetics, North Carolina State University, Raleigh, NC 27695, USA*
[2] *W. M. Keck Center for Behavioral Biology, North Carolina State University, Raleigh, NC 27695, USA*
[3] *Departments of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853-2703, USA*

(*Received 11 May 2011; revised 10 October 2011; accepted 12 October 2011*)

Summary

Predicting functional gene annotations remains a significant challenge, even in well-annotated genomes such as yeast and *Drosophila*. One promising, high-throughput method for gene annotation is to use correlated gene expression patterns to annotate target genes based on the known function of focal genes. The *Drosophila melanogaster* transcriptome varies genetically among wild-derived inbred lines, with strong genetic correlations among the transcripts. Here, we leveraged the genetic correlations in gene expression among known seminal fluid protein (SFP) genes and the rest of the genetically varying transcriptome to identify 176 novel candidate SFPs (cSFPs). We independently validated the correlation in gene expression between seven of the cSFPs and a known SFP gene, as well as expression in male reproductive tissues. We argue that this method can be extended to other systems for which information on genetic variation in gene expression is available.

1. Introduction

The diminishing cost of high-throughput technologies such as whole genome transcript profiling, high-density genotyping and whole genome re-sequencing has shifted the focus of genomic sciences from data production to data interpretation. Foremost among the challenges in interpretation is functional gene annotation, through experimental validation or computational prediction. Even for the best-annotated genomes, a significant proportion of genes are yet to be functionally characterized (Peña-Castillo & Hughes, 2007; Costello *et al*., 2009); less than half in *Drosophila* (Costello *et al*., 2009).

Most knowledge regarding gene function in eukaryotes comes from mutagenesis, single-gene knock-outs and RNA interference (RNAi) knock-down experiments performed in yeasts, *Drosophila*, *Caenorhabditis elegans*, mouse and *Arabidopsis* (Winzeler *et al*., 1999; Alonso *et al*., 2003; Kamath & Ahringer, 2003; Bellen *et al*., 2004; Dietzl *et al*., 2007; Ni *et al*., 2009; Guan *et al*., 2010; Spirek *et al*., 2010). These approaches have provided functions for a large number of genes in many organisms and the basis for making gene function predictions based on gene sequence similarities. However, screening large mutant collections for quantitative phenotypes is highly laborious. Furthermore, unique mutations in the same gene, or the same mutation in multiple genetic backgrounds can give different phenotypes, further complicating the interpretation of such screens (Flint & Mackay, 2009; Mackay *et al*., 2009; Dowell *et al*., 2010).

Computational methods for gene annotation complement experimental approaches. Computational methods rely on the detection of particular sequence motifs (e.g. a binding domain) (Hrmova & Fincher, 2009); strong orthology with a gene of known function in a closely related species; or 'guilt-by-association' (Bréhélin *et al*., 2010). The last approach is based on correlative evidence, such as the co-regulation of gene expression or the existence of known protein–protein interactions. In all cases, the functional annotation of a known gene is transferred to its interacting or correlated partner, providing a hypothesis that can be verified experimentally.

* Corresponding author: North Carolina State University, Raleigh, NC, USA.

† These authors contributed equally to this work.

Traditionally, guilt-by-association annotation has been used in the context of environmental perturbations (Walker *et al*., 1999; Reverter *et al*., 2008; Vandepoele *et al*., 2009; Klie *et al*., 2010). A complementary approach is to utilize natural variation in genetically correlated transcriptional networks to identify co-regulated transcripts. Previously, we used genome wide transcript profiles from 40 lines from the *Drosophila* Genetic Reference Panel (DGRP; Ayroles *et al*., 2009), a set of inbred lines recently derived from the wild, as a source of genetic variation in gene expression. The genetic variation among these inbred lines greatly exceeds that which can be obtained by mutagenesis screens or standard genetic crosses, while sampling multiple genetically identical individuals from each line reduces environmental variance. The genetically variable transcripts are highly correlated among the lines, forming 241 transcriptional co-expression modules (Ayroles *et al*., 2009). These co-expression modules were enriched for common Gene Ontology (GO) categories, expression in the same tissues, common transcriptional factor binding sites and associations of gene expression with the same quantitative traits. These observations suggest that genetic correlation of gene expression with a co-expression module may be due to co-regulation and that transcripts genetically correlated with a target gene of known function are plausibly involved in the same biological process or molecular function as the target gene (Luo *et al*., 2007; Ayroles *et al*., 2009). Here, we test this hypothesis using seminal fluid proteins (SFPs) as the focal genes.

We chose SFPs as focal genes for two reasons. First, many of the gene products of the secretory tissues of the male reproductive tract that produce the SFPs are well understood in *Drosophila melanogaster* (Wolfner, 2009). This is especially true for the male accessory glands (AGs), which produce proteins collectively known as ACcessory gland Proteins (ACPs). ACPs are transferred to females in the seminal fluid and affect a number of post-mating processes (Wolfner, 2009), including sperm storage and maintenance (Neubaum & Wolfner, 1999; Tram & Wolfner, 1999; Ravi Ram & Wolfner, 2007, 2009), egg production and mating receptivity (Heifetz *et al*., 2000; Chapman *et al*., 2003; Liu and Kubli, 2003), female feeding behaviour (Carvalho *et al*., 2006) and sleep patterns (Isaac *et al*., 2010). Proteomic (Findlay *et al*., 2008, 2009) and gene expression (Swanson *et al*., 2001) studies have identified 187 SFPs, most of which are ACPs. Second, we observed strong genetic correlations in expression among the known ACPs (Ayroles *et al*., 2009), suggesting that new SFPs, and potentially genes important for the production or function of these proteins, could be found by analysing the correlation structure between genetically variable transcripts.

Using the DGRP expression data (Ayroles *et al*., 2009), we identified transcripts whose expression patterns correlated with known SFPs. These correlated transcripts are candidates for both previously unknown SFPs and genes that are required for regulation of SFP production. Very little is known about how SFP genes are regulated in the male; this method provides a means to identify candidate regulatory genes for further study. As a proof of principle, the only known transcription factor required for the expression of specific SFP genes (Xue & Noll, 2002) was among the candidate genes we identified. Although proteins encoded by regulatory genes would not necessarily be transferred to females during mating, and are therefore not SFPs per se, we refer to our set of candidate SFPs as cSFPs.

We identified 176 cSFP genes. For validation, we selected seven candidates with varying levels of correlation to known SFP genes and used quantitative real-time PCR (qRT-PCR) to validate the correlation patterns. We also used RT-PCR to test the tissue of expression for these seven genes. We propose that this method can be widely applied to similar datasets, beyond the example of the SFP functional annotation we present.

2. Methods

(i) *Gene expression data*

The gene expression data are from Ayroles *et al*. (2009). Whole genome expression was quantified using Affymetrix *Drosophila* 2.0 arrays for two replicate pools of 3–5-day-old mated males and females for each of 40 DGRP lines. We median-centred the perfect match (PM) data and removed probes that were identified as likely single feature polymorphisms. We used the median $\log_2$ signal intensity of the remaining PM probes in each probe set as the measure of expression. A total of 14 840 (78·9 %) of the 18 767 transcripts on the array were expressed. Because we focus here on highly male-biased transcripts, we only used the male gene expression data to identify genetically variable transcripts. We fitted the following model to the expression data: $Y = L + e$, where Y is the median $\log_2$ signal intensity, L is the line effect and e is the residual. We identified 7151 transcripts as genetically variable at a False Discovery Rate (FDR) $< 0·01$.

The raw microarray data are deposited in the ArrayExpress database (http://www.ebi.ac.uk/arrayexpress) under accession number E-MEXP-1594. The DGRP stocks are available from the Bloomington *Drosophila* Stock Center (Bloomington, Indiana).

(ii) *cSFPs*

Of the 187 known SFP genes, 107 had genetically variable expression levels in the DGRP lines.

https://doi.org/10.1017/S0016672311000346 Published online by Cambridge University Press

We computed pairwise Pearson correlations between the 107 genetically variable SFPs and all 7151 genetically variable transcripts, 1. We then calculated an 'SFP score' for each of the 7151 transcripts by tallying the number of significant correlations ($P<0{\cdot}01$) with known SFPs, divided by 107. For a given transcript, a score of 100 indicates that it is correlated with all 107 known SFPs, and a score of 0·93 ($1/107 \times 100$) indicates the absence of significant correlation between the focal gene and any of the known SFP genes (i.e. only showing correlation to itself). The thresholds used to compute this score are arbitrary, but this method is both simple and intuitive, and gives similar results to more sophisticated statistics such as the identification of eigengenes (Langfelder & Horvath, 2007) following the construction of co-expression gene networks and using the Principal Component Analysis (PCA) loadings to identify correlated transcripts.

In addition to the correlation structure, we used several criteria to identify transcripts as putative SFPs (proteins that are predominately or exclusively expressed in the male reproductive tract and likely to be transferred to females), or as potential regulatory genes (those that produce proteins unlikely to be transferred to females) but whose expression is also predominately limited to male reproductive tissues. We used FlyAtlas (Chintapalli *et al*., 2007), a database of tissue-specific expression for *D. melanogaster*, to examine the tissues of expression for each gene with an SFP score of greater than 8. In addition, because SFPs are secreted proteins, we used SignalP software (http://www.cbs.dtu.dk/services/SignalP/) to identify the presence of predicted signal sequences. The program calculates the probability that the input amino acid sequence contains an N-terminal secretion signal. Here, we used the signal peptide probability score given from the SignalP-HMM prediction method. Signal peptides are usually 15–30 amino acids long and contain a stereotypical pattern of charged, hydrophobic and uncharged residues, although the amino acid sequence itself is not conserved (Emanuelsson *et al*., 2007). However, not all secreted proteins contain predicted signal sequences (Findlay *et al*., 2008), and not all proteins with secretion signals are secreted (Emanuelsson *et al*., 2007). Therefore, we do not exclude genes as being SFPs or ACP candidates based solely on a low SignalP score.

(iii) *Experimental validation of cSFPs*

We chose seven genes identified as cSFPs for validation of the guilt-by-association results as well as further characterization. These genes have a range of SFP scores and a few have predicted biochemical functions, though none were predicted to be involved with SFP function. In addition to the seven candidates, we also included a known ACP gene (*CG9997*; Swanson *et al*., 2001; Ravi Ram & Wolfner, 2007), and a known ejaculatory duct (ED) protein gene (*Dup99B*; Saudan *et al*., 2002), both of whose products are transferred to females, as positive controls. We expect cSFP genes, including those expressed in the ED or bulb, to correlate in expression with the known SFP, *CG9997*. We included *CG34422* as a negative control, given its low SFP score and wide expression pattern across tissues, including the male AGs, brain, eye and hindgut. This gene should not show a significant correlation to *CG9997* in the qRT-PCR experiment, in contrast to the seven cSFPs.

We independently validated the tissue-biased expression results from FlyAtlas (Chintapalli *et al*., 2007) for these 10 genes. We reared Canton-S males on standard yeast-glucose medium under uncrowded conditions at ~24 °C. We dissected 50–60 testes (T), AGs, EDs, ejaculatory bulbs (EB) and male carcasses (C; no reproductive tract). Dissected tissues were placed directly into TRIzol Reagent (Invitrogen) on ice. We collected two biological replicates for each RNA extraction.

We used qRT-PCR to validate the correlation structure between the genes that had been inferred from the microarray experiment. We randomly selected 20 of the 40 DGRP lines used in the microarray study (Ayroles *et al*., 2009), and isolated total RNA from two biological replicates, each with 8–12 males of each line (3–7 days post-eclosion). We then estimated the correlation of gene expression with the known SFP, *CG9997*.

(iv) *RNA extractions and cDNA synthesis*

We extracted total RNA by grinding dissected tissues in 150 μl of TRIzol Reagent (Invitrogen), following the manufacturer's recommendations for RNA isolation, except that 0·5 ml of chloroform was used for every 1 ml of TRIzol. Total RNA was treated with DNase1 (Invitrogen) and converted to cDNA with Superscript II Reverse Transcriptase (Invitrogen) and oligo-dT primers as recommended by the manufacturer. We used 500 ng of total RNA per 20 μl reverse transcription reaction. Negative controls without reverse transcriptase were tested once for all genes and all cDNA samples to exclude potential genomic DNA contamination.

(v) *qRT-PCR*

We quantified mRNA levels by qRT-PCR in 25 μl reactions with the SYBR green detection method (iQ SYBR Green Supermix, Bio-Rad) according to the protocol from MyiQ Single-Color Real-Time PCR Detection System (Bio-Rad). Each reaction was performed with 2 pg of total cDNA, using a

https://doi.org/10.1017/S0016672311000346 Published online by Cambridge University Press

BioRadMyiQ Single-Color Real-Time PCR Detection system. We used the *actin5C* gene as an internal standard. We used Primer3 (http://frodo.wi.mit.edu/primer3/) to design transcript-specific primers to amplify 85–148 bp regions of the genes of interest. *CG34422* primers were designed to encompass the common regions of alternative transcripts. The starting template concentration of each transcript was calculated from the standard curve of that primer pair according to the method described by Qiagen (http://www1.qiagen.com/literature/brochures/pcr/qt/1037490_ag_pcr_0206_int_lr.pdf). We used the linear regression model $Y=mX+b$ to quantify transcript abundance, where Y is the critical threshold (Ct) values from the qRT-PCR experiment, m is the slope, b is the intercept of the standard curve and X is the transcript abundance. We standardized this estimate by dividing by the transcript abundance of *actin5C* in the same sample.

(vi) *GO analysis*

We used the GO analysis to assign functional categories to the cSFP genes tested. We computed the genetic correlations between each of the seven new focal genes with the remainder of the genetically variable transcriptome. We then performed a GO enrichment analysis for the genes most strongly correlated to the focal gene ($P<0{\cdot}001$ and $|r|>0{\cdot}5$). The conclusions regarding enrichment were the same if the threshold was increased to $P<0{\cdot}0001$. We performed this analysis using DAVID 6.7 (Huang *et al.*, 2009).

3. Results and Discussion

Of the 187 known SFPs, 107 had genetically variable transcripts among the 40 DGRP lines (Ayroles *et al.*, 2009). The 107 known SFPs were highly genetically correlated (Fig. 1), reinforcing the idea that gene co-expression may be a reflection of shared function. We attempted to cluster this correlation matrix further into modules using various clustering algorithms, including Modulated Modularity Clustering (MMC) (Stone & Ayroles, 2009), but did not find strong community structure in the graph resulting from this correlation matrix. In addition, we did not find evidence supporting the idea that genes sharing a similar GO term were more strongly correlated with each other than they were to the rest of the genes.

We then analysed the correlation matrix between the 107 known SFPs and 7151 transcripts that were genetically variable in males. We assigned an SFP score to each of the genetically variable transcripts based on the number of significant correlations with known SFPs (Supplementary Table 1 available at http://cambridge.journals.org/GRH). We next asked whether this approach would allow us to recover the known SFPs. We ranked the vector of SFP scores from the highest to the lowest and applied the filter that cSFPs should be expressed in male reproductive tissues based on FlyAtlas (Chintapalli *et al.*, 2007) data. We found that 78% of the known SFPs are in the top 500 transcripts.

We identified 176 cSFP genes that have correlated expression patterns to at least 7 of the 107 genetically variable known seminal protein genes and are expressed in male reproductive tissues (Supplementary Table 1). A total of 37 of the 176 candidates have no known or predicted functions or GO terms. An additional 13 transcripts correspond to probe sets on the Affymetrix array but not annotated genes, and could correspond to new genes. Independent confirmation of cSFP identification comes from a proteomic screen aimed at identifying male proteins transferred during mating (Findlay *et al.*, 2008, 2009). Two candidate transcripts were confirmed as *bona fide* SFPs: *CG34002* (with an SFP score of 15) and *Sfp26Ad* (with an SFP score of 41). *Sfp26Ad* was not annotated as a gene at the time we performed this experiment and corresponded to probe set 637742 at on the Affymetrix array.

We chose seven cSFP genes (*CG9720*, *CG11828, CG31413*, *CG31493*, *CG31496*, *CG32985* and *CG34002*), as well as two positive control genes (the ACP gene *CG9997* and the ED protein gene *Dup99B*) and one negative control gene (*CG34422*, with an SFP score of 0·93) for validation of the microarray correlation results using qRT-PCR in 20 of the DGRP lines. The candidate genes have SFP scores ranging from moderately low (8) to very high (42, the highest SFP score found) (Table 1). The RT-PCR results confirmed the correlation between all seven cSFPs and the known ACP gene *CG9997* across the 20 lines (Fig. 2). As predicted, expression of the negative control *CG34422* was not genetically correlated with that of *CG9997*. However, expression of the ED protein gene *Dup99B*, whose gene product is transferred with the seminal fluid to females, was genetically correlated with *CG9997*, demonstrating that non-ACP SFPs canals to be identified with this method.

Table 1 gives SFP scores, secretion signal peptide probability and tissue of expression for these seven genes and for the positive and negative controls. Three genes with high SFP scores were not predicted to have secretion signals. These genes' products may be secreted nevertheless, as has been seen in other cases (Findlay *et al.*, 2008), or they may be non-SFP genes that are important for the regulation of other SFPs.

Among the seven genes, all that were predicted to be expressed in AGs (Chintapalli *et al.*, 2007) were confirmed as expressed in that tissue (Table 1, Fig. 3). To gain insight into the possible biological processes and molecular functions of the candidate genes

https://doi.org/10.1017/S0016672311000346 Published online by Cambridge University Press

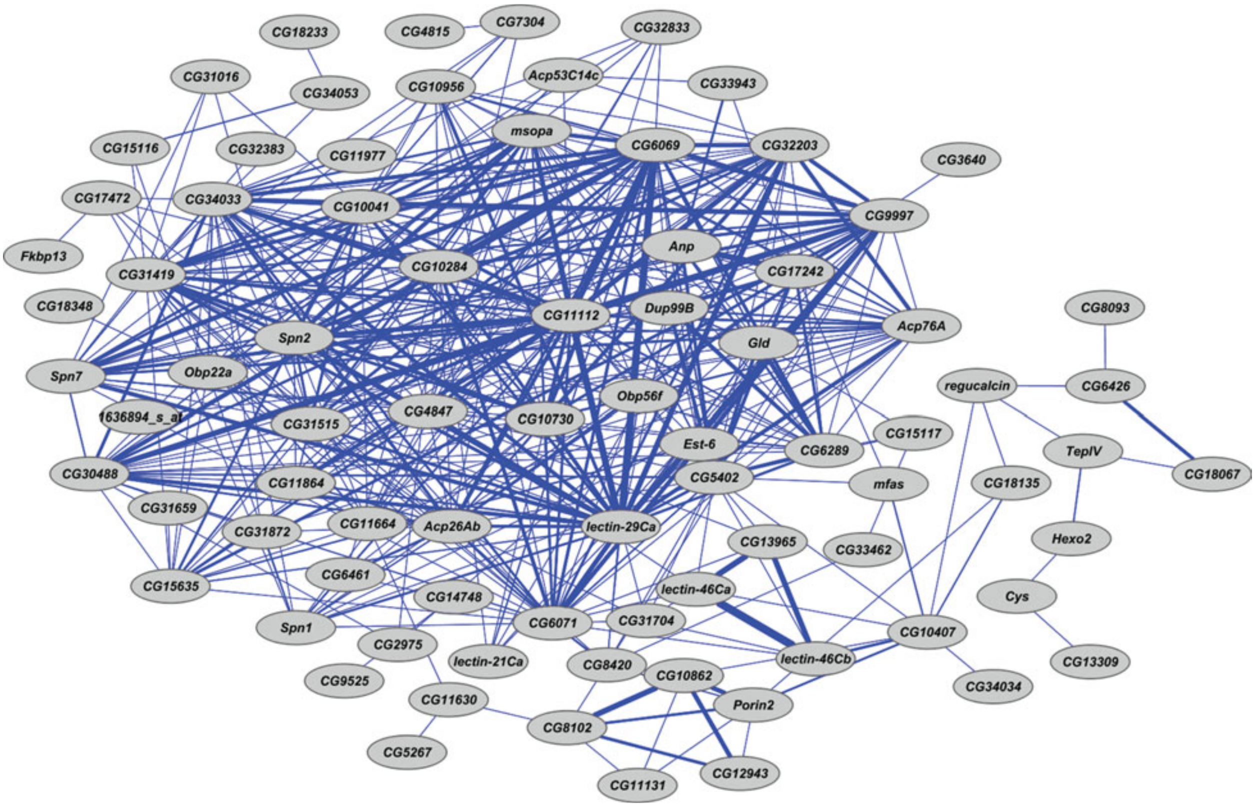

Fig. 1. Graphical representation of the correlation among known SFPs. Each node represents a gene and each edge the correlation between two genes. The thickness of each edge is scaled proportional to the strength of the correlation between two genes. The absolute value of all correlations depicted is greater than 0·5 ($P < 0·001$).

Table 1. *Genes selected for experimental validation. The SFP score is the fraction of known SFPs with which the gene had correlated expression. Sprob is the predicted probability of a secretion signal sequence as given by SignalP. Tissue of expression is given from the FlyAtlas compilation and our RT-PCR data from the male reproductive tract and carcass. ED and bulb are not represented in FlyAtlas. Bold font denotes tissues of predominant expression. AG, accessory glands; ED, ejaculatory duct; EB, ejaculatory bulb; T, testis; LSG, larval salivary glands; HT, heart; HD, head. ED and EB are not represented in FlyAtlas.*

Category	Gene	Affymetrix ID	SFP score	Sprob	Tissue (FlyAtlas)	Tissue (RT-PCR)
Candidate SFA	*CG9720*	1624902_at	35	0·997	AG	**AG**, ED, EB
Candidate SFA	*CG11828*	1633604_at	41	0	AG	**AG**, ED
Candidate SFA	*CG31413*	1635084_at	42	0·987	AG	**AG**, ED
Candidate SFA	*CG31493*	1640609_at	36	0	AG	AG
Candidate SFA	*CG31496*	1628103_at	8	0·721	AG, LSG	**AG**, ED, EB
Candidate SFA	*CG32985*	1632491_at	38	0	AG	**AG**, ED, EB
Candidate SFA	*CG34002*	1625512_s_at	15	0·991	AG	AG
ACP positive control	*CG9997*	1634224_at	39	0·999	AG	AG
ED positive control	*Dup99B*	1639365_at	29	0·98	AG	AG, **ED**, **EB**
ACP negative control	*CG34422*	1641329_at	1	0	All but T, HT, HD	All

chosen for validation, we used a GO enrichment analysis implemented in DAVID (6.7) (Huang *et al.*, 2009). For each candidate gene, we analysed the function of its most correlated transcripts ($P < 0·001$ and $r > 0·5$). Four of the seven candidate genes (*CG11828*, *CG31413*, *CG31493* and *CG34002*) were significantly associated with serine-type endopeptidase inhibitor activity, a predicted function shared by several other SFPs (Wolfner, 2009). However, it is important to note that *CG11828*, *CG31413* and *CG31493* do not contain conserved protease domains but do contain other types of predicted conserved

https://doi.org/10.1017/S0016672311000346 Published online by Cambridge University Press

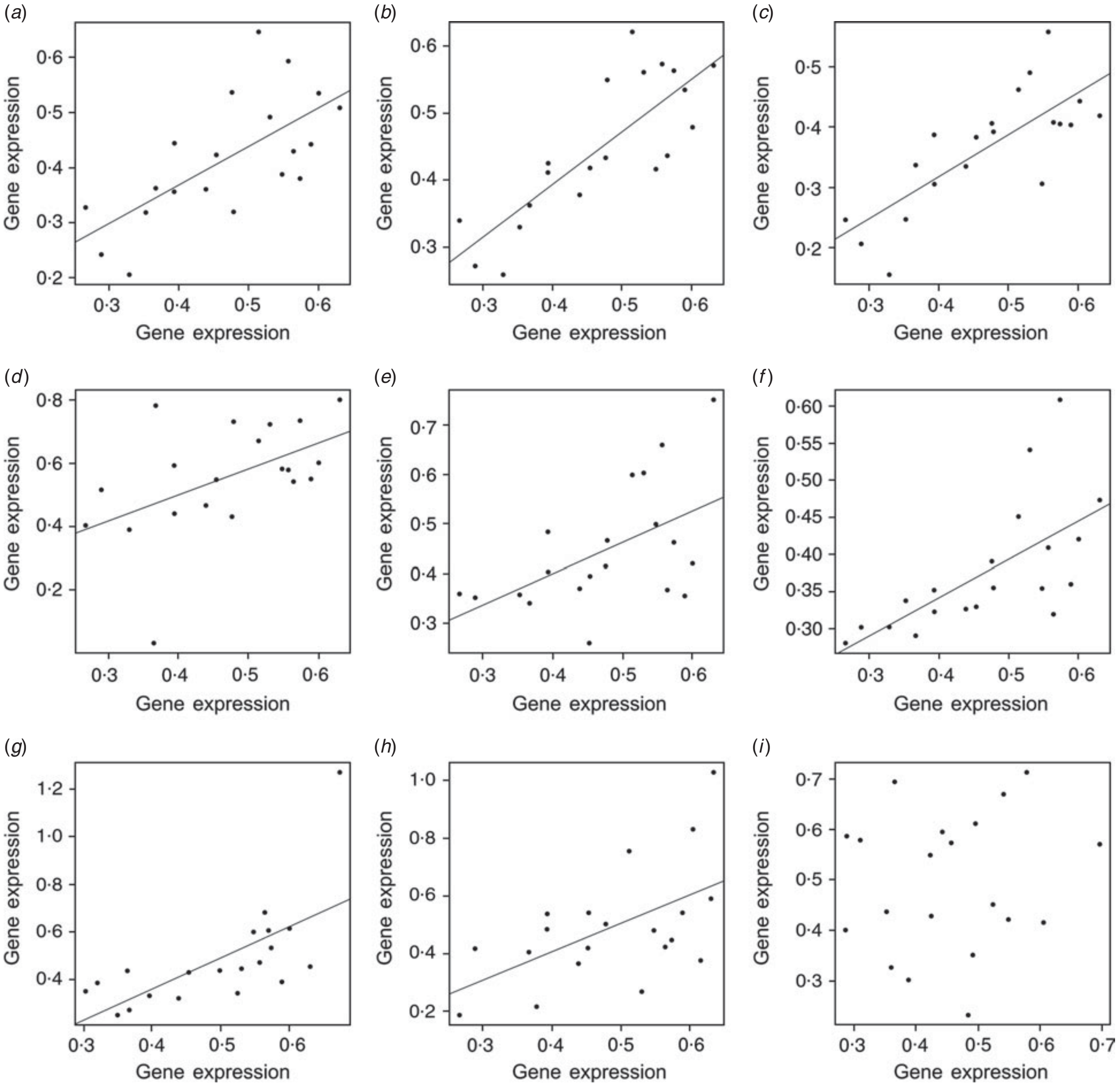

Fig. 2. Correlation of qRT-PCR estimates of gene expression between cSFP genes and positive and negative SFP control genes (y-axis) to a known ACP gene (*CG9997*, x-axis) among males of 20 inbred lines. All estimates of gene expression are normalized to that of *actin5C*. The linear regression line is shown, along with the t-test P-value and the estimate of the correlation coefficient, r. (*a*) *CG11828*, $r=0·68$, $P=0·001$. (*b*) *CG31413*, $r=0·81$, $P=0·000014$. (*c*) *CG31493*, $r=0·77$, $P=0·000063$. (*d*) *CG31496*, $r=0·51$, $P=0·022$. (*e*) *CG32985*, $r=0·53$, $P=0·015$. (*f*) *CG34002*, $r=0·66$, $P=0·0017$. (*g*) *CG9720*, $r=0·66$, $P=0·0016$. (*h*) *Dup99B* (positive control), $r=0·55$, $P=0·012$. (*i*) *CG34422* (negative control), $r=0·12$, $P=0·61$.

domains. No significant GO-class enrichment was observed for *CG9720*, *CG31496* or *CG32985*.

It is possible that some of the cSFP genes are important for SFP expression and function but may not encode proteins that are transferred to females as part of the seminal fluid. As proof of principle that such genes can be identified by this method, our analysis detected *paired* (SFP score = 16), which encodes a transcription factor important in AG development and ACP expression (Supplementary Table 2 available online at http://cambridge.journals.org/GRH). This *Pax* gene has a dual function in *Drosophila*: it acts first as a pair-rule gene in early embryo development (Nüsslein-Volhard & Weischaus, 1980; Kilchherr *et al.*, 1986) and later is required for viability and male fertility (Bertuccioli *et al.*, 1996; Xue & Noll, 1996, 2000). AG formation and expression of at least two SFPs expressed in the AG (*ACP26Aa* and *SP*) both require the function of *paired* (Xue & Noll, 2000, 2002).

https://doi.org/10.1017/S0016672311000346 Published online by Cambridge University Press

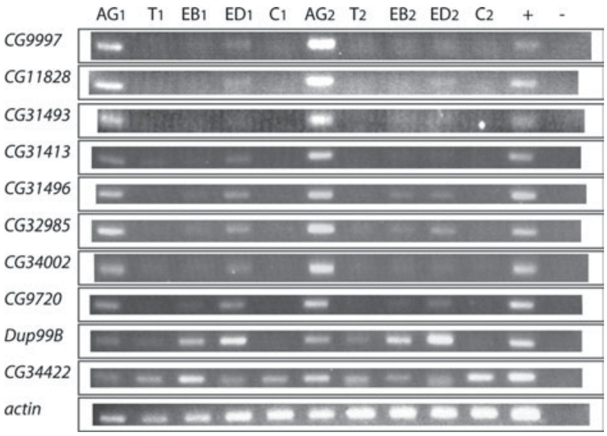

Fig. 3. RT-PCR analysis of gene expression for cSFP genes and positive and negative SFP control genes in five male tissues: AG, testes (T), EB, ED and carcass (C, non-reproductive tissues). The subscripts denote the two biological replicates of each tissue. The number of PCR cycles for each gene was normalized to give non-saturation results. *Actin5C* was used as a control for cDNA synthesis. Whole male cDNA was used as a positive (+) PCR control, and no DNA template was used as a negative (−) PCR control.

Guilt-by-association methods most frequently rely on clustering algorithms to identify the functional membership of a candidate gene or transcript (Aravind, 2000; Miozzi *et al*., 2008; Reverter *et al*., 2008; Klie *et al*., 2010). In its most common use, guilt-by-association is used to assign functions to any or all unannotated genes that respond to a given treatment or are differentially regulated under disease conditions. Here, we have demonstrated the use of guilt-by-association methods in another context: to identify genes in a specific functional class using correlated genetic variation in gene expression among wild-derived inbred lines. This method removes the requirement for relying on arbitrary clustering or reliance on GO terms to assign candidate functions to new genes. Instead, a group of genes that has been annotated and functionally clustered experimentally is used to find correlated transcripts that can then be included in the group. In this case, we used SFPs, a group defined by a biological phenomenon rather than a biochemical function. As in potentially many other cases, for example, identifying genes involved in specific behaviours, GO terms do not define our selected group of genes as belonging to a biologically significant group. The group of genes we identified (cSFPs) have diverse GO functions (ranging from proteases to pro-hormones). A given cSFP gene could not be predicted as an SFP on the basis of GO membership.

SFP genes are well suited for this study since their expression is specific to, or highly biased in, the male reproductive tract, facilitating their confirmation as SFPs; and expression of the known SFPs is genetically variable in the population of lines surveyed. An increasing number of studies are taking advantage of natural genetic variation to better understand the genetic basis of phenotypic variation (Mackay *et al*., 2009). In the future, the availability of sequence information for the *D. melanogaster* population used in this study will allow us to associate co-expression with expression quantitative trait locus (eQTL) analysis (Mackay *et al*., 2009). This additional layer of information will further our understanding of what genetic factors are driving co-expression between SFP genes, and may lead us to rethink what information should be considered when annotating a segment of sequence.

To complement this study, and generalize the simple analysis presented in this manuscript, we have created a web tool (http://dgrp.statgen.ncsu.edu) that allows the user to input the Affymetrix Drosophila 2.0 ID of any focal gene of interest and retrieve a vector of genes, their ranked correlation with the focal gene, as well as the GO of the correlated transcripts. This tool integrates FlyAtlas information (Chintapalli *et al*., 2007), allowing users to restrict the computation of correlations to genes expressed in specific tissue or to genes with strong tissue-biased expression.

Many studies using natural genetic variation to study phenotypic variation also investigate variation in gene expression and gene co-expression (Mackay *et al*., 2009). However, very rarely is this information translated in the form of hypothetical functional annotation for any unannotated genes involved. We advocate that such datasets be used more routinely as patterns should emerge across studies and this information will greatly improve our understanding of genes, their function and regulation. In particular, directed analyses such as the one presented here, in which genes involved in an experimentally defined group are sought, may help to uncover pleiotropy among previously annotated genes and increase our understanding of how various biological systems function together.

This work was funded by National Institutes of Health grants R01 GM45146 (T. F. C. M.) and R0I HD038921 (M. F. W.). B. A. L. was supported by an NSF Predoctoral Fellowship. We thank G. Findlay and E. Kelleher for comments on the manuscript, and J. Mezey and MBG colleagues for use of the qPCR machine. This is a publication of the W. M. Keck Center for Behavioral Biology.

References

Alonso, J. M., Stepanova, A. N., Leisse, T. J., Kim, C. J., Chen, H., Shinn, P., Stevenson, D. K., Zimmerman, J., Barajas, P., Cheuk, R., Gadrinab, C., Heller, C., Jeske, A., Koesema, E., Meyers, C. C., Parker, H., Prednis, L., Ansari, Y., Choy, N., Deen, H., Geralt, M., Hazari, N., Hom, E., Karnes, M., Mulholland, C., Ndubaku, R., Schmidt, I., Guzman, P., Aguilar-Henonin, L., Schmid, M., Weigel, D., Carter, D. E., Marchand, T., Risseeuw, E., Brogden, D., Zeko, A., Crosby, W. L., Berry, C. C. & Ecker, J. R. (2003). Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**, 653–657.

Aravind, L. (2000). Guilt by association: contextual information in genome analysis. *Genome Research* **10**, 1074–1077.

Ayroles, J. F., Carbone, M. A., Stone, E. A., Jordan, K. W., Lyman, R. F., Magwire, M. M., Rollmann, S. M., Duncan, L. H., Lawrence, F., Anholt, R. R. H. & Mackay, T. F. C. (2009). Systems genetics of complex traits in *Drosophila melanogaster*. *Nature Genetics* **41**, 299–307.

Bellen, H. J., Levis, R. W., Liao, G., He, Y., Carlson, J. W., Tsang, G., Evans-Holm, M., Hiesinger, P. R., Schulze, K. L., Rubin, G. M., Hoskins, R. A. & Spradling, A. C. (2004). The BDGP gene disruption project: single transposon insertions associated with 40% of *Drosophila* genes. *Genetics* **67**, 761–781.

Bertuccioli, C., Fasano, L., Jun, S., Wang, S., Sheng, G. & Desplan, C. (1996). *In vivo* requirement for the paired domain and homeodomain of the paired segmentation gene product. *Development* **122**, 2673–2685.

Bréhélin, L., Florent, I., Gascuel, O. & Maréchal, E. (2010). Assessing functional annotation transfers with interspecies conserved coexpression: application to *Plasmodium falciparum*. *BMC Genomics* **11**, 35.

Carvalho, G. B., Kapahi, P., Anderson, D. J. & Benzer, S. (2006). Allocrine modulation of feeding behavior by the sex peptide of *Drosophila*. *Current Biology* **16**, 692–696.

Chapman, T., Bangham, J., Vinti, G., Seifried, B., Lung, O., Wolfner, M. F., Smith, H. K. & Partridge, L. (2003). The sex peptide of *Drosophila melanogaster*: female post-mating responses analyzed by using RNA interference. *Proceedings of the National Academy of Sciences USA* **100**, 9923–9928.

Chintapalli, V. R., Wang, J. & Dow, J. A. (2007). Using FlyAtlas to identify better *Drosophila melanogaster* models of human disease. *Nature Genetics* **39**, 715–720.

Costello, J. C., Dalkilic, M. M., Beason, S. M., Gehlhausen, J. R., Patwardhan, R., Middha, S., Eads, B. D. & Andrews, J. R. (2009). Gene networks in *Drosophila melanogaster*: integrating experimental data to predict gene function. *Genome Biology* **10**, R97.

Dietzl, G., Chen, D., Schnorrer, F., Su, K. C., Barinova, Y., Fellner, M., Gasser, B., Kinsey, K., Oppel, S., Scheiblauer, S., Couto, A., Marra, V., Keleman, K. & Dickson, B. J. (2007). A genome-wide transgenic RNAi library for conditional gene inactivation in *Drosophila*. *Nature* **448**, 151–156.

Dowell, R. D., Ryan, O., Jansen, A., Cheung, D., Agarwala, S., Danford, T., Bernstein, D. A., Rolfe, P. A., Heisler, L. E., Chin, B., Nislow, C., Giaever, G., Phillips, P. C., Fink, G. R., Gifford, D. K. & Boone, C. (2010). Genotype to phenotype: a complex problem. *Science* **328**, 469.

Emanuelsson, O., Brunak, S., von Heijne, G. & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols* **2**, 953–971.

Findlay, G. D., MacCoss, M. J. & Swanson, W. J. (2009). Proteomic discovery of previously unannotated, rapidly evolving seminal fluid genes in *Drosophila*. *Genome Research* **19**, 886–896.

Findlay, G. D., Yi, X., Maccoss, M. J. & Swanson, W. J. (2008). Proteomics reveals novel *Drosophila* seminal fluid proteins transferred at mating. *Public Library of Science Biology* **6**, e178.

Flint, J. & Mackay, T. F. C. (2009). Genetic architecture of quantitative traits in mice, flies, and humans. *Genome Research* **19**, 723–733.

Guan, C., Ye, C., Yang, X. & Gao, J. (2010). A review of current large-scale mouse knockout efforts. *Genesis* **48**, 73–85.

Heifetz, Y., Lung, O., Frongillo, E. A. Jr & Wolfner, M. F. (2000). The *Drosophila* seminal fluid protein Acp26Aa stimulates release of oocytes by the ovary. *Current Biology* **10**, 99–102.

Hrmova, M. & Fincher, G. B. (2009). Functional genomics and structural biology in the definition of gene function. *Methods in Molecular Biology* **513**, 199–227.

Huang, D. W., Sherman, B. T. & Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nature Protocols* **4**, 44–57.

Isaac, R. E., Li, C., Leedale, A. E. & Shirras, A. D. (2010). *Drosophila* male sex peptide inhibits siesta sleep and promotes locomotor activity in the post-mated female. *Proceedings of the Biological Sciences* **277**, 65–70.

Kamath, R. S. & Ahringer, J. (2003). Genome-wide RNAi screening in *Caenorhabditis elegans*. *Methods* **30**, 313–321.

Kilchherr, E., Schumaker, V. N., Phillips, M. L. & Curtiss, L. K. (1986). Activation of the first component of human complement, C1, by monoclonal antibodies directed against different domains of subcomponent C1q. *Journal of Immunology* **137**, 255–262.

Klie, S., Nikoloski, Z. & Selbig, J. (2010). Biological cluster evaluation for gene function prediction. *Journal of Computational Biology* (Epublication ahead of print).

Langfelder, P. & Horvath, S. (2007). Eigengene networks for studying the relationships between co-expression modules. *BMC Systems Biology* **1**, 54.

Liu, H. & Kubli, E. (2003). Sex-peptide is the molecular basis of the sperm effect in *Drosophila melanogaster*. *Proceedings of the National Academy of Sciences USA* **100**, 9929–9933.

Luo, F., Yang, Y., Zhong, J., Gao, H., Khan, L., Thompson, D. K. & Zhou, J. (2007). Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory. *BMC Bioinformatics* **8**, 299.

Mackay, T. F. C., Stone, E. A. & Ayroles, J. F. (2009). The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics* **10**, 565–677.

Miozzi, L., Piro, R. M., Rosa, F., Ala, U., Silengo, L., Di Cunto, F. & Provero, P. (2008). Functional annotation and identification of candidate disease genes by computational analysis of normal tissue gene expression data. *Public Library of Science One* **6**, e2439.

Neubaum, D. M. & Wolfner, M. F. (1999). Wise, winsome, or weird? Mechanisms of sperm storage in female animals. *Current Topics in Developmental Biology* **41**, 67–97.

Ni, J. Q., Liu, L. P., Binari, R., Hardy, R., Shim, H. S., Cavallaro, A., Booker, M., Pfeiffer, B. D., Markstein, M., Wang, H., Villalta, C., Laverty, T. R., Perkins, L. A. & Perrimon, N. (2009). A *Drosophila* resource of transgenic RNAi lines for neurogenetics. *Genetics* **182**, 1089–1100.

Nüsslein-Volhard, C. & Wieschaus, E. (1980). Mutations affecting segment number and polarity in *Drosophila*. *Nature* **287**, 795–801.

Peña-Castillo, L. & Hughes, T. R. (2007). Why are there still over 1000 uncharacterized yeast genes? *Genetics* **176**, 7–14.

Ravi Ram, K. & Wolfner, M. F. (2007). Seminal influences: *Drosophila* ACPs and the molecular interplay between

https://doi.org/10.1017/S0016672311000346 Published online by Cambridge University Press

males and females during reproduction. *Integrative and Comparative Biology* **47**, 427–445.

Ravi Ram, K. & Wolfner, M. F. (2009). A network of interactions among seminal proteins underlies the long-term postmating response in *Drosophila*. *Proceedings of the National Academy of Sciences USA* **106**, 15384–15389.

Reverter, A., Ingham, A. & Dalrymple, B. P. (2008). Mining tissue specificity, gene connectivity and disease association to reveal a set of genes that modify the action of disease causing genes. *BioData Mining* **1**, 8.

Saudan, P., Hauck, K., Soller, M., Choffat, Y., Ottiger, M., Spörri, M., Ding, Z., Hess, D., Gehrig, P. M., Klauser, S., Hunziker, P. & Kubli, E. (2002). Ductus ejaculatorius peptide 99B (DUP99B), a novel *Drosophila melanogaster* sex-peptide pheromone. *European Journal of Biochemistry* **269**, 989–997.

Spirek, M., Benko, Z., Carnecka, M., Rumpf, C., Cipak, L., Batova, M., Marova, I., Nam, M., Kim, D. U., Park, H. O., Hayles, J., Hoe, K. L., Nurse, P. & Gregan, J. (2010). *S. pombe* genome deletion project: an update. *Cell Cycle* **9**, 2399–2402.

Stone, E. A. & Ayroles, J. F. (2009). Modulated modularity clustering as an exploratory tool for functional genomic inference. *Public Library of Science Genetics* **5**, e1000479.

Swanson, W. J., Clark, A. G., Waldrip-Dail, H. M., Wolfner, M. F. & Aquadro, C. F. (2001). Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in *Drosophila*. *Proceedings of the National Academy of Sciences USA* **98**, 7375–7379.

Tram, U. & Wolfner, M. F. (1999). Male seminal fluid proteins are essential for sperm storage in *Drosophila melanogaster*. *Genetics* **153**, 837–844.

Vandepoele, K., Quimbaya, M., Casneuf, T., De Veylder, L. & Van de Peer, Y. (2009). Unraveling transcriptional control in *Arabidopsis* using *cis*-regulatory elements and coexpression networks. *Plant Physiology* **150**, 535–546.

Walker, M. G., Volkmuth, W., Sprinzak, E., Hodgson, D. & Klingler, T. (1999). Prediction of gene function by genome-scale expression analysis: prostate cancer-associated genes. *Genome Research* **9**, 1198–1203.

Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D. J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J. L., Riles, L., Roberts, C. J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R. K., Véronneau, S., Voet, M., Volckaert, G., Ward, T. R., Wysocki, R., Yen, G. S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M. & Davis, R. W. (1999). Functional characterization of the *S. cerevisiae* genome by gene deletion and parallel analysis. *Science* **285**, 901–906.

Wolfner, M. F. (2009). Battle and ballet: molecular interactions between the sexes in *Drosophila*. *Journal of Heredity* **100**, 399–410.

Xue, L. & Noll, M. (1996). The functional conservation of proteins in evolutionary alleles and the dominant role of enhancers in evolution. *European Molecular Biology Organization Journal* **15**, 3722–3731.

Xue, L. & Noll, M. (2000). *Drosophila* female sexual behavior induced by sterile males showing copulation complementation. *Proceedings of the National Academy of Sciences USA* **97**, 3272–3275.

Xue, L. & Noll, M. (2002). Dual role of the Pax gene *paired* in accessory gland development of *Drosophila*. *Development* **129**, 339–346.

https://doi.org/10.1017/S0016672311000346 Published online by Cambridge University Press