nature genetics

Systems genetics of complex traits in *Drosophila melanogaster*

Julien F Ayroles[1,2,6], Mary Anna Carbone[1,2,6], Eric A Stone[3,6], Katherine W Jordan[1,2], Richard F Lyman[1,2], Michael M Magwire[1,2,5], Stephanie M Rollmann[1,2,5], Laura H Duncan[1,2], Faye Lawrence[1,2], Robert R H Anholt[1,2,4] & Trudy F C Mackay[1,2]

Determining the genetic architecture of complex traits is challenging because phenotypic variation arises from interactions between multiple, environmentally sensitive alleles. We quantified genome-wide transcript abundance and phenotypes for six ecologically relevant traits in *D. melanogaster* wild-derived inbred lines. We observed 10,096 genetically variable transcripts and high heritabilities for all organismal phenotypes. The transcriptome is highly genetically intercorrelated, forming 241 transcriptional modules. Modules are enriched for transcripts in common pathways, gene ontology categories, tissue-specific expression and transcription factor binding sites. The high degree of transcriptional connectivity allows us to infer genetic networks and the function of predicted genes from annotations of other genes in the network. Regressions of organismal phenotypes on transcript abundance implicate several hundred candidate genes that form modules of biologically meaningful correlated transcripts affecting each phenotype. Overlapping transcripts in modules associated with different traits provide insight into the molecular basis of pleiotropy between complex traits.

Natural populations harbor a wide range of phenotypic variation for all aspects of morphology, physiology, behaviors and disease susceptibility. Knowledge of the genetic basis of this variation is important for understanding adaptive evolution, deriving elite domestic crop and animal strains and improving human health. However, determining the genetic architecture of natural phenotypic variation is challenging because most phenotypic variation is attributable to segregating alleles at many interacting genes with environmentally sensitive effects[1,2].

Recent genome-wide association studies in human populations have reported that diseases and quantitative traits are associated with loci with small effects, which together account for a few percent of the phenotypic variance for that trait[3]. These findings suggest that the bulk of genetic variation for complex traits may be due to alleles with small and possibly context-dependent effects. Further, the reported associated variants are often found in genes with no a priori expected relationship to the trait, in computationally predicted genes, or in gene deserts. In *Drosophila*, quantitative analyses of new mutations have revealed large numbers of loci affecting quantitative traits[4], as have high-resolution maps of segregating quantitative trait loci (QTLs) in *Drosophila*[4] and mice[5]. Single alleles often have pleiotropic effects on multiple traits, epistatic interactions among loci affecting the same trait are common, and allelic effects can be conditional on sex and the environment[4].

Our understanding of the genetic architecture of quantitative traits in model systems, and ultimately humans, will benefit from interrogating a single resource population for variation in DNA sequence, transcript abundance, proteins and metabolites; for multiple organismal phenotypes; and in multiple environments. This 'systems genetics' approach will yield a detailed map of genetic variants associated with each organismal phenotype in each environment, provide a functional context for interpreting the phenotypes, elucidate the genetic underpinnings that govern the interdependence of multiple phenotypes, and address the long-standing question of the genetic basis of genotype by environment interaction[6–8].

Here we report the first step of a systems-genetics analysis of the genetic basis of complex traits in *Drosophila*. We demonstrate ubiquitous variation in transcript abundance among inbred *Drosophila* strains recently derived from the wild, and show that the variable transcripts can be grouped into coherent modules of intercorrelated genes. These lines harbor substantial genetic variation for ecologically relevant complex traits, and variation for hundreds of transcripts is associated with variation for each of the organismal traits. Transcripts associated with each trait form correlated transcriptional modules from which we can construct networks of interacting genes with plausible biological relationships to each other and the traits. These genetic networks provide additional functional annotation of the *Drosophila* genome, and an integrated

[1]Department of Genetics, [2]W. M. Keck Center for Behavioral Biology, [3]Department of Statistics and [4]Department of Biology, North Carolina State University, Raleigh, North Carolina 27695, USA. [5]Present addresses: Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, West Mains Road, Edinburgh EH9 3JT, UK (M.M.M.) and Department of Biological Sciences, University of Cincinnati, PO Box 210006, 614 Rieveschl Hall, Cincinnati, Ohio 45221, USA (S.M.R.). [6]These authors contributed equally to this work. Correspondence should be addressed to T.F.C.M. (trudy_mackay@ncsu.edu).

Received 27 August 2008; accepted 12 January 2009; published online 22 February 2009; doi:10.1038/ng.332

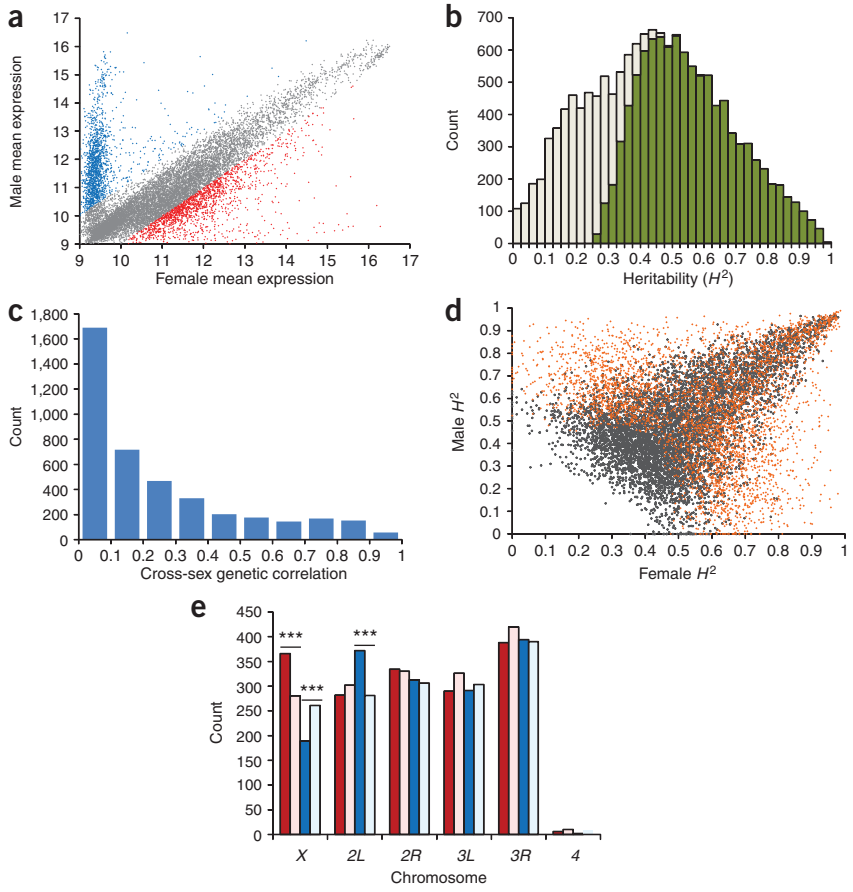


Figure 1 Variation in transcript abundance among 40 wild-derived inbred lines. (**a**) Sex bias for gene expression. Blue and red dots represent genes showing a twofold difference in gene expression between males and females, respectively. (**b**) Distribution of broad-sense heritabilities (H^2). Dark green denotes significant H^2 estimates (line FDR < 0.001) and grey indicates nonsignificant H^2 estimates. (**c**) Distribution of cross-sex genetic correlations for transcripts showing significant variation in sexual dimorphism (significant sex × line interaction variance at FDR < 0.001). (**d**) Bivariate plot of H^2 estimates in males and females. Orange dots indicate significant line-by-sex interaction variance. (**e**) Chromosomal distribution of sex-biased gene expression. The dark blue and red bars are observed male and female counts, respectively, and the light blue and red bars are the expected numbers of male and female transcripts, respectively. Asterisks indicate significant deviation of observed from expected values ($P < 0.001$).

context in which to frame predictions of the behavior of the system following genetic or environmental perturbations.

RESULTS

Natural variation in transcript abundance

We derived 40 highly inbred lines from the natural population of Raleigh, North Carolina, USA. These lines are a living library of common polymorphisms affecting complex traits. We assessed whole-genome variation in transcript abundance for young males and females of each of these lines using Affymetrix *Drosophila* 2.0 arrays. We standardized the raw array data by median centering and used a statistical approach to identify outlier probes in each perfect-match probe set that contained single feature polymorphisms (SFPs) between the wild-derived lines and the reference strain sequence used to design the array. We used the median $\log_2$ signal intensity of the remaining perfect-match probes in each probe set as the measure of expression. Of the 18,800 transcripts on the array, 14,840 (78.9%) were expressed in young adults (**Supplementary Table 1** online). Many genes that have been characterized for their role in early development are also expressed in adults, and may affect adult phenotypes that cannot be predicted from their prior developmental functions[9,10]. We used analysis of variance (ANOVA) to partition variation in expression between sexes, among lines, and the sex × line interaction for each expressed transcript.

The sex term was significant at a conservative false discovery rate (FDR)[11] of 0.001 for 13, 086 (88.2%) of the expressed transcripts (**Fig. 1a** and **Supplementary Table 1**), indicating pervasive sexual dimorphism for gene expression. A total of 3,255 transcripts had twofold or greater differences in expression between the sexes: 1,690 with female-biased expression, and 1,565 with male-biased expression. Previous studies reported largely male-biased expression in *D. melanogaster*[12,13]. Our observation of nearly equal numbers of transcripts with strong male and female expression bias is likely attributable to our greater power to detect smaller sex-biased effects, as the absolute magnitude of the difference in expression between the sexes is less for female-biased than for male-biased transcripts (**Fig. 1a**). Gene ontology analysis[14] showed that both male- and female-biased transcripts were enriched for genes affecting reproduction and gametogenesis. The female-biased transcripts were also enriched for genes affecting basic cellular processes, and the male-biased transcripts for genes involved in reproductive behavior and physiology, mitochondrial energy metabolism and intermediary metabolism (**Supplementary Table 2** online).

The line term was significant at an FDR < 0.001 for 10,096 (68.0%) of the expressed transcripts, indicating considerable genetic variation in gene expression (**Supplementary Table 1**). Estimates of broad-sense heritability (H^2) for significant transcripts ranged from $\approx$0.3–1.0 (**Fig 1b**). Transcripts with high levels of genetic variation ($H^2 > 0.8$) were enriched for genes involved with responses to the environment, whereas transcripts with low levels of genetic variation ($H^2 < 0.2$) were enriched for genes affecting processes essential for survival (**Supplementary Table 2**). The overall correlation (r) between H^2 and mean expression was low, although statistically significant ($r = 0.078$, $P = 3.13 \times 10^{-21}$). The high level of genetic variation in gene expression is partly attributable to the doubling of the additive genetic variation of an outbred population in a population of fully inbred lines, and inflation of broad-sense heritability estimates by any nonadditive genetic variance[1]. Significant heritability of abundance of a particular transcript does not necessarily mean that *cis*-acting genetic polymorphisms cause the variation; it could be due to *trans* regulation by another genetically variable transcript[6–8].

A significant sex × line interaction indicates genetic variation in the magnitude of the sex dimorphism among the lines, or equivalently, a significant departure of the cross-sex genetic correlation (r_{MF}) from unity. The sex × line interaction was significant for 4,108 (40.7%) of

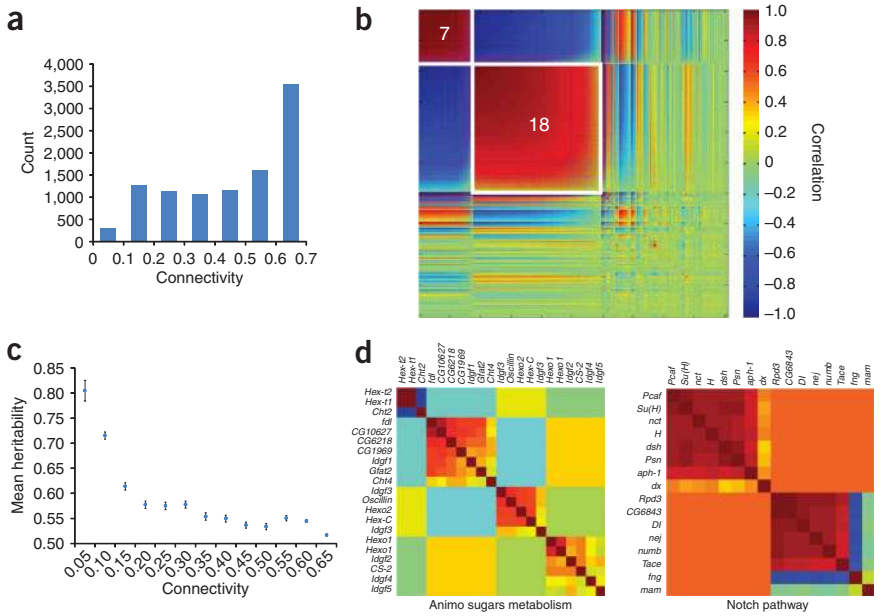
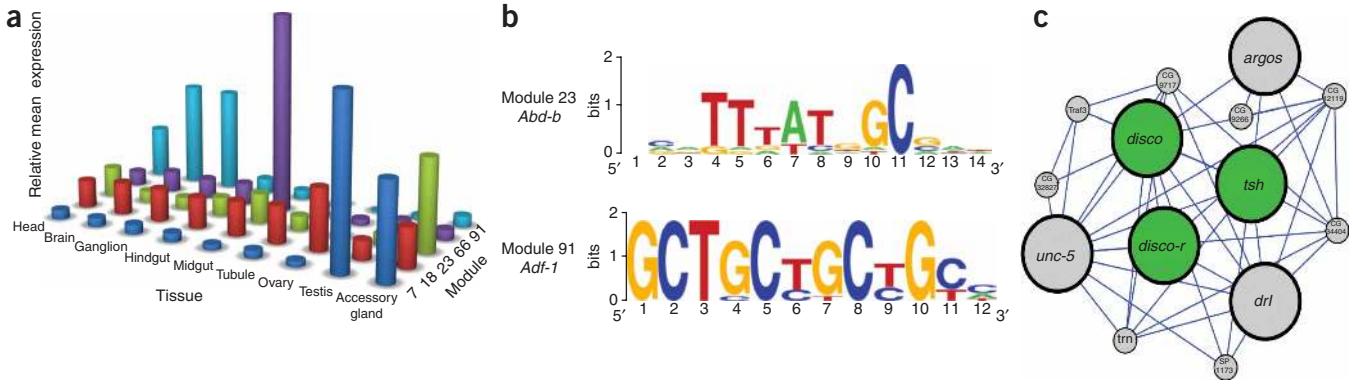


Figure 2 Correlated transcriptional modules. (**a**) Distribution of connectivity (average |*r*|) for the 10,096 genetically variable transcripts (line FDR < 0.001). (**b**) Clustering of the genetically variable transcripts into 241 modules. (**c**) Relationship between transcript H^2 and average connectivity. Error bars, s.e.m. (**d**) Correlated transcriptional modules for genes in the amino sugars metabolism and Notch pathway KEGG ontologies. The colors on the off-diagonal represent the average cross-module correlations.

the expressed transcripts at an FDR < 0.001 (**Supplementary Table 1**). The average cross-sex genetic correlation of the transcripts showing genetic variation in sexual dimorphism was quite low (r_{MF} = 0.234), with a mode at $r_{MF} = 0$ (**Fig. 1c**). Variation in expression of many transcripts among the lines was uncorrelated between males and females, arguing for caution when extrapolating inferences about variation in gene expression from expression profiles drawn from one sex to the other sex. These data also reveal great potential for rapid evolution of sex-biased gene expression, as observed when different *Drosophila* species are compared[13]. Low cross-sex genetic correlations were only partly attributable to transcripts that were expressed and variable in only one sex. Many of the transcripts showing variation in sex dimorphism in expression were actually expressed in both sexes, but had much higher heritabilities in one sex compared to the other (**Fig. 1d**). Transcripts with low cross-sex genetic correlations (r_{MF} < 0.2) were enriched for the same gene ontology categories as sex-biased genes, indicating that sex-biased transcripts are also genetically variable in the sex in which they are highly expressed (**Supplementary Table 2**).

The patterning of genetic variation within a population depends on effective population size, recombination rate and the selection coefficient of new mutations[1]. These factors vary among chromosomes: *X* chromosomes have smaller effective population sizes than autosomes and are hemizygous in males, and recombination is severely reduced on the *Drosophila* fourth chromosome. Therefore, we asked whether there were differences among chromosomes in mean level and genetic variance of transcript abundance. Consistent with previous studies[12,15], we found that male-biased transcripts were strongly



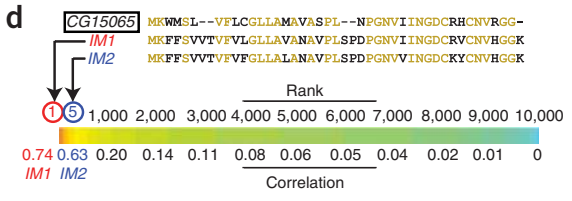


Figure 3 Biology of transcriptional modules. (**a**) Distribution of tissue-specific expression in modules 7, 18, 23, 66, 91. Module 7 is enriched for male-biased transcripts and expression in the testes and accessory glands. Module 18 is enriched for female-biased transcripts and expression in ovaries. Module 23 is enriched for transcripts affecting reproduction and gametogenesis that are highly expressed in ovaries and male accessory glands. Module 66 is enriched for transcripts in the Notch signaling pathway and nervous system development expressed in the midgut. Module 91 is enriched for transcripts affecting the function of the nervous system with high expression in the brain. (**b**) Modules 23 and 91 are, respectively, enriched for the *Abd-b* (P = 0.004) and *Adf-1* (P = 0.001) transcription factor binding motifs. *Abd-b* has been implicated in genital disc development[47] and *Adf-1* in memory and synaptogenesis[48], consistent with the inferred function of genes in these modules. (**c**) Network representation of module 164, emphasizing the genetic correlations between adult transcripts for three transcription factors that interact during embryonic and larval development. (**d**) Putative functional annotation of *CG15065* as a gene encoding an immune-induced molecule. Ranking all genetically variable transcripts according to their correlation to *CG15065* shows that *IM1* is the strongest transcriptional correlate (r = 0.74) and *IM2* is the fifth strongest (r = 0.63). The protein alignments of *CG15065*, *IM1* and *IM2* are highly conserved.

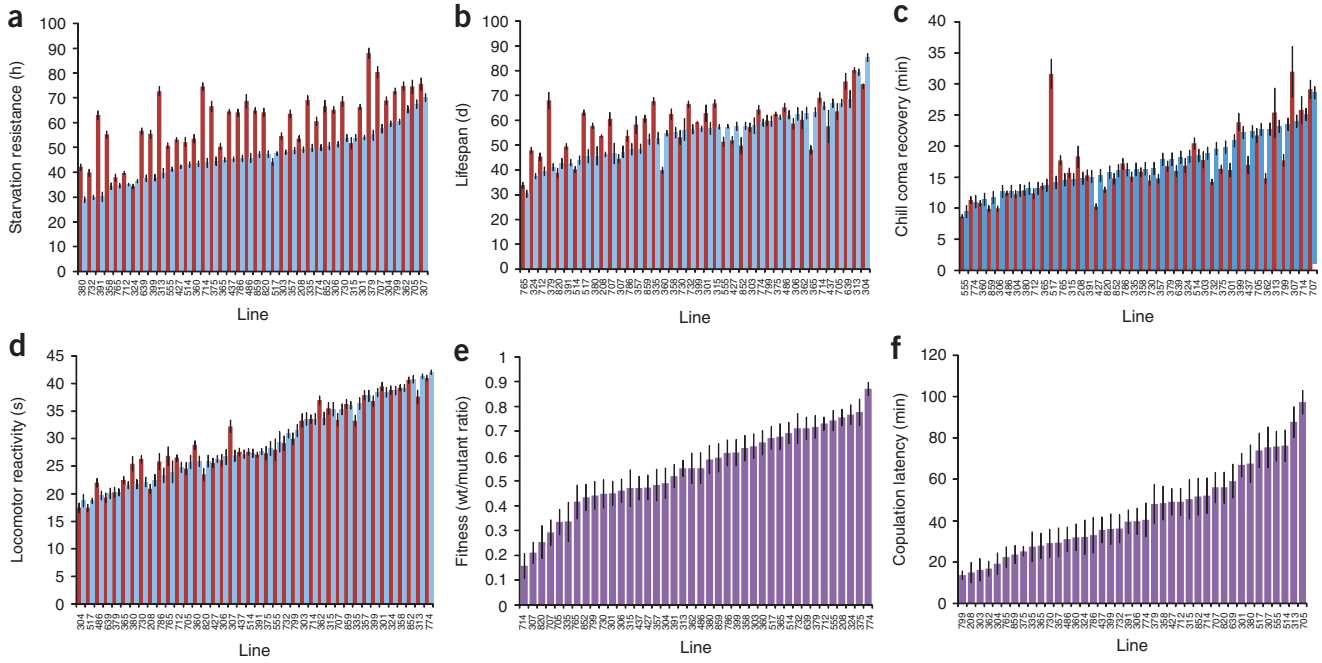


Figure 4 Variation for organismal phenotypes among 40 wild-derived inbred lines. (**a–f**) Distributions of line means among 40 wild-derived inbred lines. The red and blue bars in panels **a**–**d** depict females and males, respectively. Sexes were not measured separately in panels **e**–**f**. Error bars, s.e.m. (**a**) Starvation stress resistance ($H^2 = 0.56$). (**b**) Chill coma recovery ($H^2 = 0.23$). (**c**) Life span ($H^2 = 0.54$). (**d**) Locomotor reactivity ($H^2 = 0.58$). (**e**) Copulation latency ($H^2 = 0.25$). (**f**) Competitive fitness ($H^2 = 0.32$).

underrepresented on the *X* chromosome (and overrepresented on chromosome *2L*). In contrast, female-biased transcripts were strongly overrepresented on the *X* chromosome (**Fig. 1e**). Possibly the *X* chromosome is a hospitable location for female- but not male-biased genes because mutations with *X*-linked deleterious effects on male fitness are quickly purged from populations, but mutations with recessive *X*-linked deleterious effects on female fitness can achieve higher frequencies because they are protected from natural selection when they are rare. We found differences in overall transcript abundance among the chromosomes for both sexes ($P < 0.0001$), with chromosome *4* having the highest mean expression and the *X* chromosome the lowest mean expression (data not shown). We also found differences in H^2 of transcript abundance between the chromosomes ($P < 0.0001$), largely attributable to reduced genetic variation on the *X* and fourth chromosomes (data not shown).

Modules of correlated transcripts

We assessed the extent to which the 10,096 variable transcripts were genetically correlated among the lines (**Fig. 2**). We computed pairwise genetic correlations among all the variable transcripts, and computed the mean absolute value of the pairwise genetic correlations of each transcript with all other transcripts as a measure of average connectivity ($|r|$, **Supplementary Table 1**). The distribution of $|r|$ was strongly skewed towards high values, with a mode at 0.6 (**Fig. 2a**). Thus, the genome as a whole is highly genetically correlated at the transcriptional level, which imposes constraints on the evolution of transcriptional genetic networks. There is a strong inverse correlation ($r = -0.263$, $P = 1.08 \times 10^{-159}$) between transcript heritability and average connectivity—the most highly heritable transcripts have low mean values of $|r|$ (**Fig. 2c**) and are therefore presumably less evolutionarily constrained. Indeed, there is a significant positive correlation between heritability of transcript abundance and ω, the ratio of nonsynonymous to synonymous substitutions, among single-copy genes with orthologs in six *melanogaster* group species[16] ($r = 0.132$, $P = 7.56 \times 10^{-25}$). Genes encoding transcripts with lower heritabilities experience stronger purifying selection than do genes encoding transcripts with high heritabilities. Although high heritabilities are expected for genes under mutation-drift equilibrium[1], it is not likely that this mechanism accounts for the observed high heritabilities of transcript abundance, as the estimates of ω for these loci are much less than the neutral expectation of unity. In addition, the high heritability transcripts are predominantly for genes affecting responses to the environment, which have been associated with responses to artificial selection for multiple traits[17–20]. Therefore, the high heritabilities for these transcripts could be the result of more complex evolutionary dynamics.

We grouped the genetically variable transcripts into modules using a novel method to identify separable clusters of highly interconnected genes (E.A.S. and J.F.A. unpublished data). The 10,096 transcripts fell into 241 modules (**Fig. 2b** and **Supplementary Table 1**). The two largest modules (7 and 18) consisted of 1,765 and 4,128 transcripts, with average absolute intramodule correlations of 0.89 and 0.77, respectively. Moreover, the expression of genes in these two modules was strongly negatively correlated among the lines (**Fig. 2b**). We carried out gene ontology enrichment analyses[14] and found that module 7 was enriched for the same functional categories as male-biased transcripts, and module 18 was enriched for the same functional categories as female-biased transcripts. We found significant overrepresentations of male-biased genes in module 7 (1,241 of 1,565 male-biased genes, $\chi^2_1 = 4{,}910$; $P \approx 0$) and of female-biased genes in module 18 (1,381 of 1,690 female-biased genes, $\chi^2_1 = 1{,}400$; $P \approx 0$). Thus, the negative correlation between modules 7 and 18 is attributable to higher expression of genes in module 7 in males than females, and higher expression of genes in module 18 in females than males.

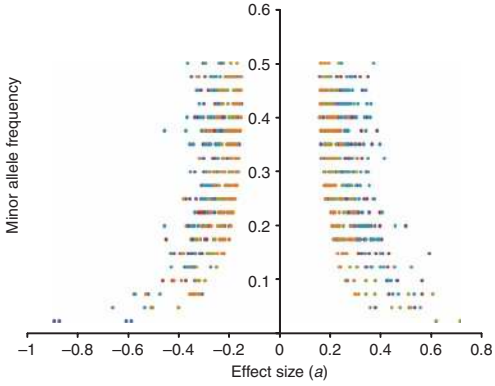


Figure 5 Distribution of SFP effects. The *x* axis is the SFP allele effect, a/σ_G, where *a* is one half the difference in trait mean between the SFP alleles and σ_G is the genetic standard deviations of each trait. The *y* axis is the minor allele frequency. The traits are color-coded: chill coma recovery (dark blue), starvation resistance (red), fitness (green), lifespan (purple), locomotor reactivity (turquoise) and copulation latency (orange).

There are strong correlations among transcripts in the same functional pathways (for example, amino sugars metabolism and the Notch signaling pathway, **Fig. 2d**). Genes within a module often show similar tissue-specific expression patterns[21], which suggests a common biological function, genetic variation in organ size or both (**Fig. 3a**). Many modules are enriched for known transcription factor binding sites (**Fig. 3b** and **Supplementary Table 3** online,). The high degree of connectivity among transcript variation in a natural population allows us to infer potential interacting partners of focal genes in networks for functional studies. For example, *disco*, *disco-r* and *tsh* interact during embryonic and larval development[22], and have genetically intercorrelated transcripts in adults in module 161 (**Supplementary Table 1** and **Fig. 3c**). These genes are expressed in the adult nervous system[21], and are correlated with three other genes in this module (*unc-5*, *drl*, *argos*) that are involved in nervous system development and axon guidance[23] (**Fig. 3c**). Transcriptional correlation also enables functional annotation of computationally predicted genes based on known annotations of other genes in the network. *CG15065* is a putative immune-induced molecule (*IM*) gene, judging from its strong transcriptional correlations with *IM1* ($r = 0.74$), *IM2* ($r = 0.63$) and *IM3* ($r = 0.67$), physical location adjacent to these genes, and notable protein sequence similarity to *IM1* and *IM2* (**Fig. 3d**).

Associations with organismal phenotypes

We quantified variation among the 40 Raleigh inbred lines for several ecologically relevant traits (resistance to starvation stress, time to recover from a chill-induced coma, life span, a startle-induced locomotor response and mating speed) as well as a measure of competitive fitness. We found substantial genetic variation for all traits, with estimates of H^2 between 0.25–0.58 (**Fig. 4a–f** and **Supplementary Table 4** online). The range of variation among these lines is comparable to the difference in mean phenotype between lines subjected to long-term divergent artificial selection for these traits[17,19]. We observed significant sex × line interactions for starvation stress resistance, life span and chill coma recovery time (**Supplementary Table 4**). Estimates of r_{MF} (± s.e.m.) were high for organismal phenotypes (0.72 ± 0.11, 0.73 ± 0.11 and 0.87 ± 0.08 for starvation stress resistance, life span and chill coma recovery, respectively), in contrast to the low cross-sex genetic correlations observed at the level of transcripts.

We asked whether there were significant genetic correlations among these traits, as would occur if segregating alleles have pleiotropic effects on two traits in the same direction[1,2]. Only five genetic correlations showed a significant association (**Supplementary Table 4**). There is a tendency for lines that recover from chill coma quickly to have high competitive fitness and mate rapidly, but at the expense of surviving starvation stress. Lines that are resistant to starvation stress tend to have longer life spans, but reduced competitive fitness. Thus, there is a trade-off between genetic variants affecting recovery from different environmental stresses. We do not observe high positive correlations between all traits with each other and with fitness, as would be the case if variation among the lines was attributable to the fixation of deleterious alleles.

We observed 3,316 probes containing SFPs, and assessed associations of SFPs with organismal phenotypes. We found 119, 118, 141, 217, 245 and 195 SFPs associated with starvation stress resistance, chill-coma recovery time, life span, locomotor behavior, mating speed and fitness, respectively ($P < 0.01$, **Supplementary Table 5** online). Although some SFPs were associated with more than one trait (**Supplementary Table 5**), the number of SFPs associated with multiple traits did not exceed that expected by chance. Because we can estimate the frequency of SFPs with significant phenotype associations, as well as the homozygous effect of the SFPs, we can evaluate the distribution of allelic effects. The homozygous effects follow an exponential distribution for all traits, with larger effects associated with the rarer SFPs, and smaller effects with the common SFPs (**Fig. 5**), as previously predicted[24].

We used regression models to identify transcripts that were associated with each organismal phenotype. At a *P* value of 0.01, we

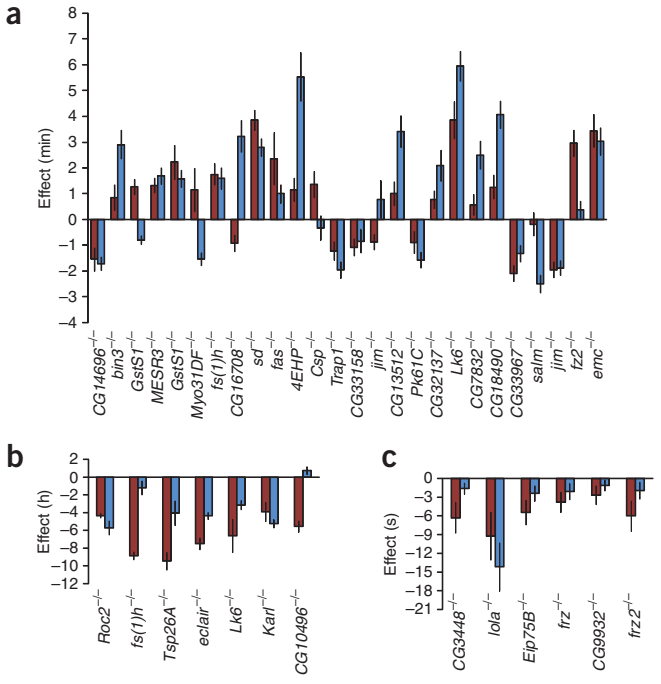


Figure 6 Effects of *P*-element mutations in candidate genes affecting quantitative traits. Mutational effects are given as deviations from the co-isogenic control line. Red and blue bars represent males and females, respectively. Mutations in all genes shown have significant effects in one or both sexes (**Supplementary Table 7**). Error bars, s.e.m. (**a**) Chill coma recovery time. (**b**) Starvation stress resistance. (**c**) Locomotor reactivity (data from ref. 27).

found 355, 1,128, 295, 231, 691 and 414 transcripts associated with starvation stress resistance, chill-coma recovery time, life span, locomotor behavior, mating speed and fitness, respectively (**Supplementary Table 6** online). There was little overlap between associations of variation in transcripts and SFPs for the same phenotypes, further increasing the number of candidate genes potentially associated with each trait.

Transcripts that are significantly associated with organismal phenotypes are candidate genes affecting the phenotype[25]. We compared phenotypes of *P*-element insertional mutations in or near candidate genes with that of their co-isogenic control lines[9,10]. Seven of ten mutations near candidate genes for resistance to starvation stress indeed affected starvation resistance, and 29 of 39 mutations near candidate genes for chill coma recovery time affected this trait (**Fig. 6a**,**b** and **Supplementary Table 7** online). Six of nine mutations in candidate genes affecting locomotor reactivity have been shown previously to affect this trait[26] (**Fig. 6c**).

Transcriptional networks associated with complex traits

Most transcripts associated with phenotypes were either unexpected based on prior mutational analyses of the traits, or from computationally predicted genes. To gain insight about functional relationships among transcripts associated with each trait, we used the residuals of the significant regressions of organismal phenotypes on gene expression to quantify modules of transcripts with coordinated patterns of expression across the 40 lines (**Fig. 7** and

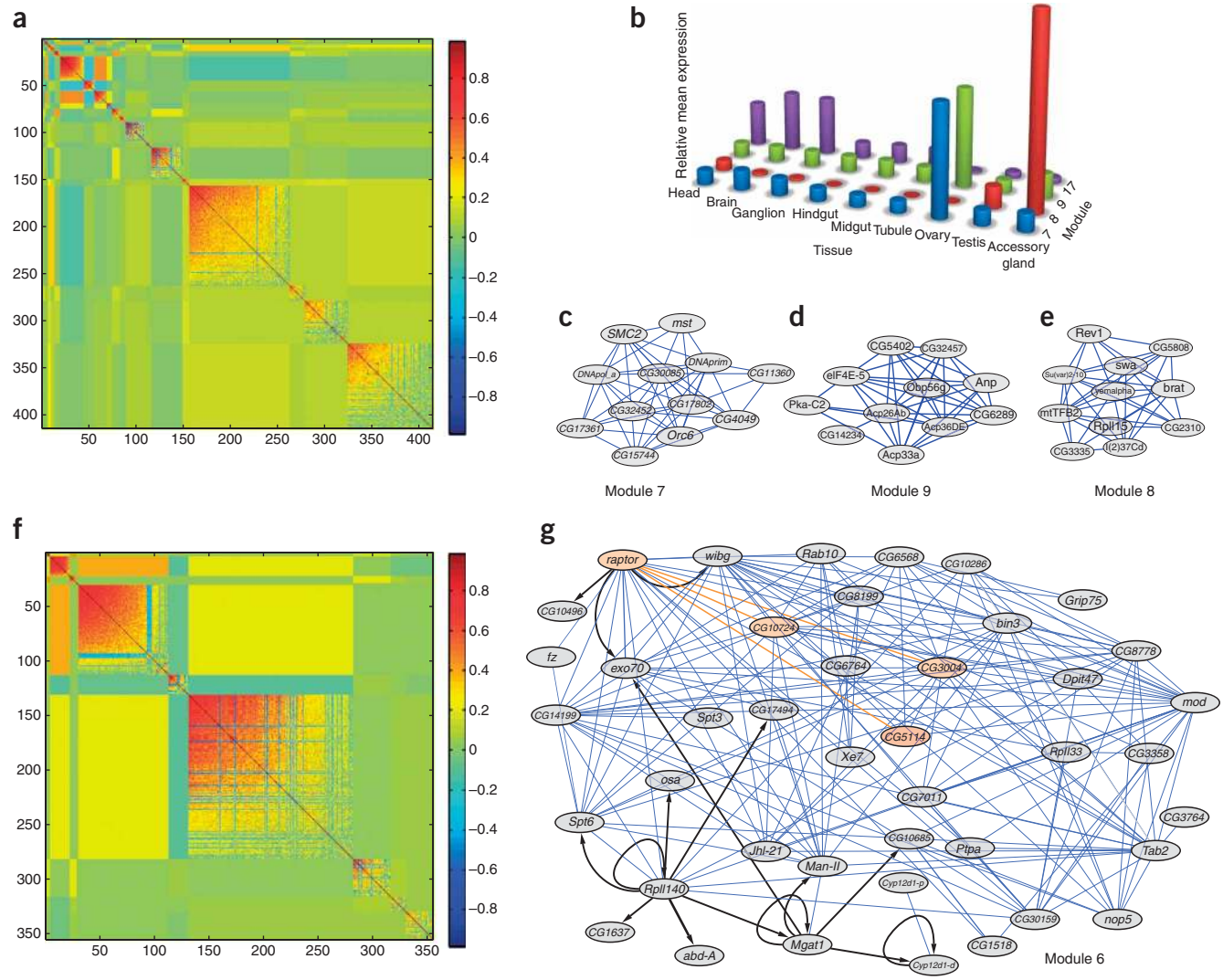


Figure 7 Modules of correlated transcripts associated with organismal phenotypes. (**a**–**e**) Competitive fitness. (**a**) Clustering of the 414 transcripts significantly associated with variation in fitness into 20 modules. (**b**) Tissue-specific expression of transcripts in modules 7 and 9 (ovaries), module 8 (accessory glands and testes) and module 17 (head, brain and thoracicoabdominal ganglion). (**c**) Interaction network for module 7. Each node represents a gene and each edge the correlation between a pair of genes. Module 7 is enriched for female-biased transcripts and transcripts affecting DNA replication. (**d**) Interaction network for module 9. Module 9 is enriched for female-biased transcripts and transcripts affecting oogenesis and transcriptional regulation. (**e**) Interaction network for module 8. Module 8 is dominated by male-biased genes, and is enriched for genes involved in male-induced postmating behaviors, including three genes encoding accessory gland proteins (*Acps*). (**f**–**g**) Starvation stress resistance. (**f**) Clustering of the 355 transcripts significantly associated with variation in starvation resistance into 11 modules. (**g**) Interaction network for module 6. The black arrows indicate SFP variants in a probe set that are associated with variation in expression of the other probes in that probe set (*cis*-acting variants) and with variation in another transcript (*trans*-acting variants). The orange nodes indicate genes with a WD40 protein domain.

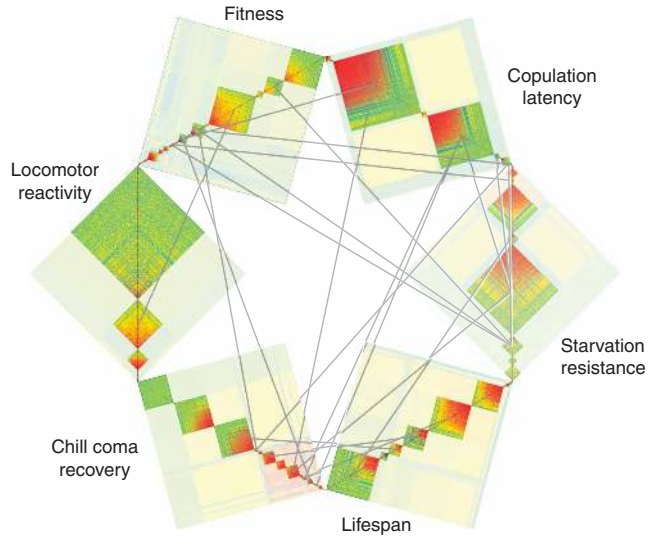


Figure 8 Pleiotropy between phenotypic modules. Grey lines connect modules with a significant overlap of greater than four genes between gene lists, as determined by Fisher's exact tests.

Supplementary Table 6). Transcripts with spurious association to a phenotype are unlikely to correlate with biologically relevant transcripts after removal of the source of the association; conversely, transcripts under coordinated control are likely to show correlated expression patterns even after removal of the effect of their common relationship to a phenotype. Each of the correlated transcript modules associated with a trait can be represented as an interaction network, with edges between transcripts in the network determined by genetic correlations in transcript abundance exceeding a threshold value (**Fig. 7**). We identified 26 modules of correlated transcripts associated with chill coma recovery time, 20 associated with fitness, 11 with starvation stress resistance, 10 with life span and 9 each with locomotor reactivity and copulation latency (**Fig. 7** and **Supplementary Table 6**).

We evaluated the biological significance of these networks by asking whether genes within each module were enriched for gene ontology categories (**Supplementary Table 8** online), expression in particular tissues, known protein–protein interactions or shared domains. As expected, transcripts associated with variation for fitness are enriched for genes that mediate immune response (modules 6 and 11), visual perception and function of the nervous system (module 17) and chemosensation (module 20), and for sex-specific transcripts (modules 7, 8 and 9) (**Fig. 7** and **Supplementary Table 8**). Variation for fitness can be maintained if there are negative genetic correlations between fitness components[1]. Transcripts in modules 7 and 9, which have female-biased expression, are positively genetically correlated with each other but negatively genetically correlated with transcripts in module 8, which have male-biased expression. The genes of module 8 encode proteins that are transferred to females on mating, are thought to benefit male fitness[27,28], and that evolve rapidly[29,30], but which impose a fitness cost on females[31]. The molecular basis of the female response to this sexual conflict is not known, and could plausibly lie in the module 7 and 9 transcripts.

Transcripts associated with variation in starvation resistance are enriched for genes that mediate antimicrobial response (module 4), transcription (module 6) and proteolysis (module 9) (**Supplementary Table 8**). One of the most highly connected genes in module 6, *raptor*, is a member of the TOR (Target Of Rapamycin) pathway, which has a key role in nutrient-sensitive signaling, regulates cell growth and cellular mass[32] and the use of alternative energy resources under starvation conditions[33].

As gene ontology enrichment analysis[14] revealed similarities between modules of correlated transcripts for the six traits (**Supplementary Table 8**), we tested whether there was more overlap of common genes between modules for the different traits than expected by chance, and uncovered substantial modular pleiotropy (**Fig. 8**). For example, genes affecting the mitochondrial ribosome are in common between chill coma recovery module 17 and copulation latency module 3 ($P = 4.74 \times 10^{-4}$), chill coma recovery module 17 and starvation resistance module 8 ($P = 1.17 \times 10^{-3}$), and starvation resistance module 8 and copulation latency module 5 ($P = 7.53 \times 10^{-4}$); whereas genes affecting defense response to bacteria are in common between starvation resistance module 4 and fitness module 11 (4.57×10^{-9}) (**Fig. 8** and **Supplementary Table 6**). These results give insights to the molecular basis of pleiotropy between complex traits.

DISCUSSION

Our quantitative genetic analysis of whole-genome variation in transcript abundance among a wild-derived population of *Drosophila* inbred lines has revealed surprising features of the genetic architecture of transcription. Nearly 80% of the genome is expressed in adult flies, and approximately 90% of the expressed transcripts have sex-biased expression at a stringent false discovery rate. Two-thirds of the expressed transcripts are genetically variable in this sample of lines—a much greater extent of genetic variation than indicated in previous studies[34–43]. Over 40% of the genetically variable transcripts also show genetic variation in sex dimorphism. Notably, the whole transcriptome is highly genetically intercorrelated, with 60% of the variable transcripts belonging to two large modules with high positive genetic correlations within modules, and high negative correlations between modules. One of the large modules is enriched for male-biased transcripts and the other for female-biased transcripts. The genetically correlated transcriptional modules are biologically plausible, with enrichments for transcripts in common pathways, gene ontology categories, tissue-specific expression and transcription factor binding sites. The high transcriptional connectivity at the level of genetic correlation of natural variation in gene expression allows us to infer genetic networks from the transcriptional networks, and the function of computationally predicted genes based on known annotations of other genes in the network.

Several hundred transcripts and SFPs are associated with phenotypic variation in each quantitative trait, and 70% of *P*-element insertional mutations tested in these candidate genes indeed significantly affect the traits. The transcripts associated with each trait group into biologically plausible modules of correlated transcripts, which are in turn correlated between traits, providing insight into the molecular basis of genetic correlations. Variation in transcript abundance in young adults reared under standard culture conditions predicts candidate genes and modules of correlated transcripts associated with variation in stress responses, behaviors and life span. The lines and transcript data are therefore a valuable resource for the *Drosophila* community, enabling similar analyses for any complex phenotype that can be quantified. Future integration of whole-genome DNA sequence variation with variation in transcript abundance and phenotypes will allow us to disentangle causal from consequential associations, and determine the frequency of causal alleles. Further, the lines can be crossed to interrogate heterozygous effects and degrees of dominance

of alleles affecting transcriptional and organismal variation[44]. Knowledge of allele frequencies and homozygous and heterozygous effects will yield unprecedented insight into the nature of evolutionary forces maintaining segregating variation for complex traits in natural populations.

METHODS

Drosophila lines. We derived inbred lines from the Raleigh, USA population by 20 generations of full-sib mating. We used the *C(2L)RM-P1*, b^1; *C(2R)RM-SKIA*, cn^1 bw^1 compound autosome (CA) stock for fitness assays. *P*-element mutations and co-isogenic control lines were a gift of H. Bellen (Howard Hughes Medical Institute, Baylor College of Medicine). We reared flies on cornmeal-molasses-agar medium at 25 °C, 60–75% relative humidity and a 12-h light-dark cycle unless otherwise specified.

Gene expression. We used Affymetrix *Drosophila* 2.0 arrays to assess transcript profiles of 3- to 5-d-old flies from the inbred lines. All samples were frozen between 1 and 3 pm. We extracted RNA from two independent pools (25 flies/sex/line), and hybridized 10 μg fragmented cRNA to each array. We randomized RNA extraction, labeling and array hybridization across all samples, and normalized the raw array data across sexes and lines using a median standardization.

Each transcript is represented by 14 perfect-match 25-bp oligonucleotides. To identify perfect-match probes with SFPs between the wild-derived lines and the strain used to design the array, we quantified the maximal degree to which the variation between lines could be reduced by partitioning the lines into two allelic classes. We computed the sum of squared deviations from each class mean and expressed their sum as a fraction of the total sum of squares. The smallest fraction across all bipartitions was used to score each probe. We identified 3,136 candidate SFPs with scores ≤ 0.1 (a tenfold reduction in the sum of squares). We validated polymorphisms in 20 of 21 of these SFPs by designing primers flanking the SFP and sequencing the PCR products (data not shown).

Our measure of expression for each probe set was the median $\log_2$ signal intensity of perfect-match probes without SFPs. We used negative control probes to estimate the background intensity, and removed probes below this threshold.

Organismal phenotypes. For the starvation stress resistance group, we placed ten same-sex, 2-d-old flies in vials containing 1.5% agar and 5 ml water, and scored survival every eight hours ($N = 5$ vials/sex/line). For the chill coma recovery group, we placed 3- to 7-d-old flies in empty vials on ice for three hours, and recorded the time for each individual to right itself after transfer to room temperature ($N = 20$ flies/sex/line). For longevity, we placed five 1- to 2-d-old same-sex virgin flies into vials containing 5 ml medium, and recorded survival every two days ($N = 5$ vials/sex/line). For locomotor reactivity, we placed single 3- to 7-d-old flies into vials containing 5 ml medium. The following day, between 8 am and 12 pm, we mechanically disturbed each fly[19], and recorded the total activity in the 45 s immediately following the disturbance. We obtained two replicate measurements of 20 flies/sex/replicate/line. For the copulation latency group, we aspirated pairs of 3- to 7-d-old virgin flies into vials containing 5 ml medium between 8 am and 12 pm, and recorded the number of minutes until initiation of copulation, for a maximum of 120 min ($N = 24$ pairs/line). For the reproductive fitness group, we used the competitive index technique[45,46]. We reared all wild-type and CA parents in constant density (10 pairs) vials. We placed six 3- to 4-d-old virgin CA males and females and three 3- to 4-d-old wild-type males and females in a vial containing 10 ml medium, discarding the flies after 7 d. The competitive index was the ratio of the number of wild type to the total number of progeny emerging by day 17 ($N = 20$ replicate vials/line).

Quantitative genetic analyses. We used ANOVA to partition phenotypic variation between sexes (S, fixed), lines (L, random), the $S \times L$ interaction (random) and the error variance (ε). We also carried out reduced ANOVAs by sex. We estimated broad-sense heritabilities (H^2) as $H^2 = (\sigma^2_L + \sigma^2_{SL})/(\sigma^2_L + \sigma^2_{SL} + \sigma^2_E)$, where σ^2_L, σ^2_{SL}, and σ^2_E are the among-line, sex × line and within-line variance components, respectively. For the analyses by sex, $H^2 = \sigma^2_L/(\sigma^2_L + \sigma^2_E)$. We estimated cross-trait (cross-sex) genetic correlations as $r_G = cov_{ij}/\sigma_i\sigma_j$, where cov_{ij} is the covariance of line means between traits i and j (males and females), and σ_i and σ_j are the square roots of the among line variance components for the two traits (males and females).

Transcriptional modules. To identify modules of genetically correlated transcripts, we computed the correlation r_{ij} between all pairs of significant transcripts i and j. The absolute correlationsr $|r_{ij}|$ were transformed to define edge weights $e^{\frac{|r_{ij}|-1}{\sigma^2}}$ in a graph of genes indexed by the free parameter σ. We determined the clustering $P = \{V_1, \ldots, V_k\}$ and value of σ that jointly maximize the modularity function

$$Q(P,\sigma) = \sum_{c=1}^{k} \left[\frac{A_\sigma(V_c, V_c)}{A_\sigma(V,V)} - \left(\frac{A_\sigma(V_c, V)}{A_\sigma(V,V)} \right)^2 \right]$$

where $A_\sigma(X,Y)$ denotes the total edge weight in the graph indexed by σ that connects any vertex in set X to a vertex in set Y. The optimal partition $P = \{V_1, \ldots, V_k\}$ defines k transcriptional modules $V_1, \ldots, V_k$.

Transcript-phenotype associations. We used regression models ($Y = \mu + S + T + S\times T + \varepsilon$, where S denotes sex and T the trait covariate) to identify transcripts significantly ($P < 0.01$) associated with organismal phenotypic variation in both sexes. We used the residuals from regression models ($Y = \mu + E + S + S\times E + \varepsilon$, where E is the covariate median $\log_2$ expression level) to compute genetic correlations between transcripts significantly associated with each phenotype for module construction.

Pleiotropic modules. To identify transcriptional modules common to more than one phenotype, we considered pairs of phenotypes, comparing the lists of significant transcripts for each module from the first phenotype to each module from the second. We used Fisher's exact test to quantify the extent that the overlap between the two modules exceeded the chance expectation.

Transcription factor motifs. We scored 5 UTR sequences of each *D. melanogaster* transcript against position weight matrices for 56 transcription factors; 100 permutations of each sequence were used to generate an empirical distribution of scores for each motif. 'Present' motifs had scores within the top 5% of the permutation distribution. We determined the genome-wide proportion of genes for which each motif was present, and compared the proportion of genes within a gene list or module for which a motif was present to the genome-wide proportion using a one-sided binomial test.

Accession codes. ArrayExpress: microarray data have been deposited under accession number E-MEXP-1594.

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

This work was supported by grants from the National Institutes of Health (R01 GM 45146, R01 GM 076083, R01 AA016560 to T.F.C.M. and R01 GM 59469 to R.R.H.A.). The authors thank S. Heinsohn for technical assistance. This is a publication of the W. M. Keck Center for Behavioral Biology.

AUTHOR CONTRIBUTIONS

T.F.C.M., J.F.A., E.A.S. and R.R.H.A. wrote the paper. R.F.L. constructed the *Drosophila* lines. M.A.C. obtained the gene expression data. K.W.J., M.M.M., S.M.R., L.H.D. and F.L. obtained the organismal phenotype data. J.F.A., E.A.S. and K.W.J. performed the statistical analyses.

Published online at http://www.nature.com/naturegenetics/
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions/

1. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics* (Addison Wesley Longman, Harlow, Essex, UK, 1996).
2. Lynch, M. & Walsh, B. *Genetics and Analysis of Quantitative Traits* (Sinauer Associates, Sunderland, Massachusetts, 1998).
3. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **44**, 66–678 (2007).

4. Mackay, T.F.C. & Anholt, R.R.H. Of flies and man: Drosophila as a model for human complex traits. *Annu. Rev. Genomics Hum. Genet.* **7**, 339–367 (2006).
5. Valdar, W. *et al.* Genetic and environmental effects on complex traits in mice. *Genetics* **174**, 959–984 (2006).
6. Sieberts, S.K. & Schadt, E.E. Moving toward a system genetics view of disease. *Mamm. Genome* **18**, 389–401 (2007).
7. Emilsson, V. *et al.* Genetics of gene expression and its effect on disease. *Nature* **452**, 423–428 (2008).
8. Chen, Y. *et al.* Variations in DNA elucidate molecular networks that cause disease. *Nature* **452**, 429–435 (2008).
9. Rollmann, S.M. *et al.* Pleiotropic fitness effects of the *Tre1/Gr5a* region in *Drosophila*. *Nat. Genet.* **38**, 824–829 (2006).
10. Sambandan, D., Yamamoto, A., Fanara, J.J., Mackay, T.F.C. & Anholt, R.R. Dynamic genetic interactions determine odor-guided behavior in *Drosophila melanogaster*. *Genetics* **174**, 1349–1363 (2006).
11. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. USA* **100**, 9440–9445 (2003).
12. Ellegren, H. & Parsch, J. The evolution of sex-biased genes and sex-biased gene expression. *Nat. Rev. Genet.* **8**, 689–698 (2007).
13. Zhang, Y., Sturgill, D., Parisi, M., Kumar, S. & Oliver, B. Constraint and turnover in sex-biased gene expression in the genus *Drosophila*. *Nature* **450**, 233–237 (2007).
14. Dennis, G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, 3 (2003).
15. Sturgill, D., Zhang, Y., Parisi, M. & Oliver, B. Demasculinization of *X* chromosomes in the *Drosophila* genus. *Nature* **450**, 238–241 (2007).
16. Drosophila 12 Genomes Consortium. Evolution of genes and genomes on the *Drosophila* phylogeny. *Nature* **45**, 203–218 (2007).
17. Mackay, T.F.C. *et al.* Genetics and genomics of *Drosophila* mating behavior. *Proc. Natl. Acad. Sci. USA* **102**, 6622–6629 (2005).
18. Edwards, A.C., Rollmann, S.M., Morgan, T.J. & Mackay, T.F.C. Quantitative genomics of aggressive behavior in *Drosophila melanogaster*. *PLoS Genet.* **2**, e154 (2006).
19. Jordan, K.W., Carbone, M.A., Yamamoto, A., Morgan, T.J. & Mackay, T.F.C. Quantitative genomics of locomotor behavior in *Drosophila melanogaster*. *Genome Biol.* **8**, R172 (2007).
20. Morozova, T.V., Anholt, R.R.H. & Mackay, T.F.C. Phenotypic and transcriptional response to selection for alcohol sensitivity in *Drosophila melanogaster*. *Genome Biol.* **8**, R231 (2007).
21. Chintapalli, V.R., Wang, J. & Dow, J.A.T. Using FlyAtlas to identify better *Drosophila* models of human disease. *Nat. Genet.* **39**, 715–720 (2007).
22. Robertson, L.K., Bowling, D.B., Mahaffey, J.P., Imiolczyk, B. & Mahaffey, J.W. An interactive network of zinc-finger proteins contributes to regionalization of the *Drosophila* embryo and establishes the domains of HOM-C protein function. *Development* **131**, 2781–2789 (2004).
23. Wilson, R.J., Goodman, J.L. & Strelets, V.B. FlyBase: integration and improvements to query tools. *Nucleic Acids Res.* **36**, D588–D593 (2008).
24. Robertson, A. in *Heritage From Mendel* (ed. Brink, A.) 265–280 (Univ. Wisconsin, Madison, Wisconsin, 1967).
25. Passador-Gurgel, G., Hsieh, W.P., Hunt, P., Deighton, N. & Gibson, G. Quantitative trait transcripts for nicotine resistance in *Drosophila melanogaster*. *Nat. Genet.* **39**, 264–268 (2007).
26. Yamamoto, A. *et al.* Neurogenetic networks for startle-induced locomotion in *Drosophila melanogaster*. *Proc. Natl. Acad. Sci. USA* **105**, 12393–12398 (2008).
27. Lung, O., Kuo, L. & Wolfner, M.F. *Drosophila* males transfer antibacterial proteins from their accessory gland and ejaculatory duct to their mates. *J. Insect Physiol.* **47**, 617–62 (2001).
28. Wolfner, M.F. "S.P.E.R.M." (seminal proteins (are) essential reproductive modulators): the view from *Drosophila*. *Soc. Reprod. Fertil. Suppl.* **65**, 183–199 (2007).
29. Date-Ito, A., Kasahara, K., Sawai, H. & Chigusa, S.I. Rapid evolution of the male-specific antibacterial protein andropin gene in *Drosophila*. *J. Mol. Evol.* **54**, 665–670 (2002).
30. Wong, A., Turchin, M.C., Wolfner, M.F. & Aquadro, C.F. Evidence for positive selection on *Drosophila melanogaster* seminal fluid protease homologs. *Mol. Biol. Evol.* **25**, 497–506 (2008).
31. Wigby, S. & Chapman, T. Sex peptide causes mating costs in female *Drosophila melanogaster*. *Curr. Biol.* **15**, 316–321 (2005).
32. Kim, D.-H. *et al.* mTOR interacts with Raptor to form a nutrient-sensitive complex that signals to the cell growth machinery. *Cell* **110**, 163–175 (2002).
33. Kamada, Y. *et al.* Autophagy in yeast: a TOR-mediated response to nutrient starvation. *Curr. Top. Microbiol. Immunol.* **279**, 73–84 (2004).
34. Jin, W. *et al.* The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.* **29**, 389–395 (2001).
35. Monks, S.A. *et al.* Genetic inheritance of gene expression in human cell lines. *Am. J. Hum. Genet.* **75**, 1094–1105 (2004).
36. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
37. Oleksiak, M.F., Churchill, G.A. & Crawford, D.L. Variation in gene expression within and among natural populations. *Nat. Genet.* **32**, 261–266 (2002).
38. Schadt, E.E. *et al.* Genetics of gene expression surveyed in maize, mouse and man. *Nature* **422**, 297–302 (2003).
39. Stranger, B.E. *et al.* Genome-wide associations of gene expression variation in humans. *PLoS Genet.* **1**, e78 (2005).
40. Brem, R.B. & Kruglyak, L. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc. Natl. Acad. Sci. USA* **102**, 1572–1577 (2005).
41. Chesler, E.J. *et al.* Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function. *Nat. Genet.* **37**, 233–242 (2005).
42. Cheung, V.G. *et al.* Mapping determinants of human gene expression by regional and genome-wide association. *Nature* **437**, 1365–1369 (2005).
43. Hubner, N. *et al.* Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. *Nat. Genet.* **37**, 243–253 (2005).
44. Lemos, B., Araripe, L.O., Fontanillas, P. & Hartl, D.L. Dominance and the evolutionary accumulation of *cis*- and *trans*-effects on gene expression. *Proc. Natl. Acad. Sci. USA* **105**, 14471–14476 (2008).
45. Knight, G.R. & Robertson, A. Fitness as a measureable character in *Drosophila*. *Genetics* **42**, 524–530 (1957).
46. Hartl, D.L. & Jungen, H. Estimation of average fitness of populations of *Drosophila melanogaster* and the evolution of fitness in experimental populations. *Evolution* **33**, 371–380 (1979).
47. Foronda, D. *et al.* Requirement of *Abdominal-A* and *Abdominal-B* in the developing genitalia of *Drosophila* breaks the posterior downregulation rule. *Development* **133**, 117–127 (2005).
48. DeZazzo, J. *et al. nalyot*, a mutation of the *Drosophila myb*-related *Adf1* transcription factor, disrupts synapse formation and olfactory memory. *Neuron* **27**, 145–158 (2000).