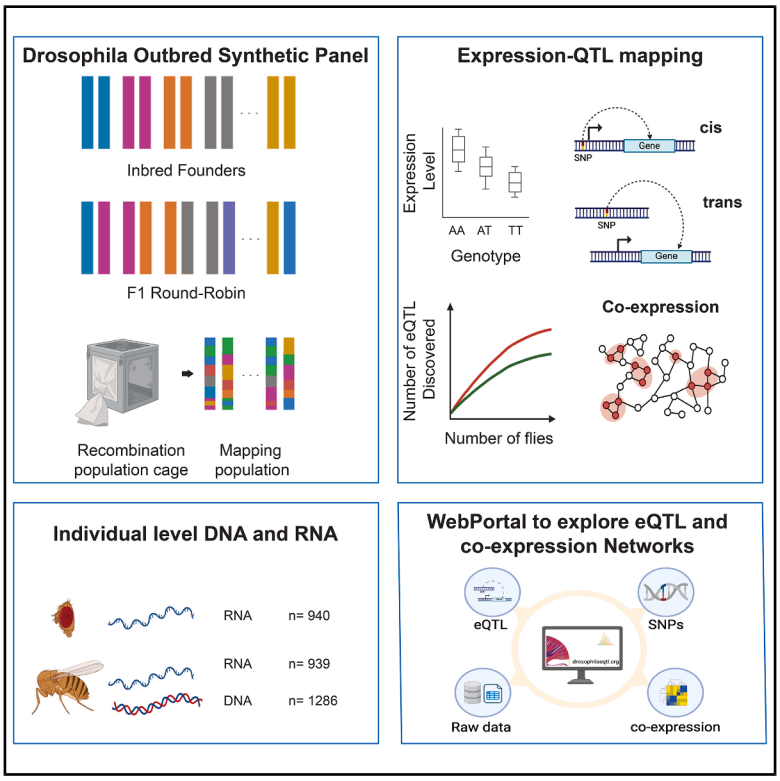
Cell Genomics

Article

Saturating the eQTL map in *Drosophila*: Genome-wide patterns of *cis* and *trans* regulation of transcriptional variation in outbred populations

Graphical abstract



Highlights

- The *Drosophila* Outbred Synthetic Panel as a resource for complex trait mapping
- Tissue-specific *cis* and *trans* regulation for 98% of the genes in *Drosophila*
- *cis* and *trans* eQTLs are subjected to different selection pressures
- Bayesian clustering uncovers non-modular structure of gene co-expression networks

Authors

Luisa F. Pallares, Diogo Melo, Scott Wolf, Evan M. Cofer, Varada Abhyankar, Julie Peng, Julien F. Ayroles

Correspondence

luisa.pallares@tuebingen.mpg.de (L.F.P.), ayroles@berkeley.edu (J.F.A.)

In brief

Pallares et al. collect over a thousand RNA-seq and DNA samples from *Drosophila melanogaster* flies to investigate the genetic basis of gene expression variation. They find eQTLs for 98% of the genes and identify the distinct patterns of tissue specificity, pleiotropy, and selection pressures between *cis* and *trans* regulation.

Pallares et al., 2025, Cell Genomics *5*, 101033
December 10, 2025
https://doi.org/10.1016/j.xgen.2025.101033

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Article

Saturating the eQTL map in *Drosophila*: Genome-wide patterns of *cis* and *trans* regulation of transcriptional variation in outbred populations

Luisa F. Pallares,[1,*] Diogo Melo,[2] Scott Wolf,[3] Evan M. Cofer,[3] Varada Abhyankar,[4] Julie Peng,[5] and Julien F. Ayroles[5,6,*]
[1]Friedrich Miescher Laboratory of the Max Planck Society, Tübingen 72076, Germany
[2]Departamento de Genética e Biologia Evolutiva, Instituto de Biociências, Universidade de São Paulo, São Paulo, SP, Brazil
[3]Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, USA
[4]Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA 90095, USA
[5]Department of Integrative Biology, University of California, Berkeley, Berkeley, CA 94720, USA
[6]Lead contact
*Correspondence: luisa.pallares@tuebingen.mpg.de (L.F.P.), ayroles@berkeley.edu (J.F.A.)
https://doi.org/10.1016/j.xgen.2025.101033

SUMMARY

Most genetic polymorphisms associated with complex traits are found in non-coding regions of the genome. Characterizing their effect presents a formidable challenge, and expression quantitative trait locus (eQTLs) mapping has been a key approach to do so. As comprehensive eQTL maps are available only for a few species, here we developed the *Drosophila* outbred synthetic population (Dros-OSP) and used it to characterize the landscape of transcriptional regulation in *Drosophila melanogaster*. We collected head and body transcriptomes and genomes from 1,286 outbred flies and mapped local and distant eQTLs for 98% of the genes. We characterized the network organization of the transcriptome across tissues and described the properties of local and distal eQTLs in terms of genetic diversity, heritability, connectivity, and pleiotropy. These results provide new insights into the genetic basis of transcriptional regulation in the fruit fly and offer a new mapping resource that will expand the possibilities currently available for the *Drosophila* community.

INTRODUCTION

Two decades of genome-wide association studies and quantitative trait locus (QTLs) mapping experiments in a wide array of species have revealed that most polymorphisms associated with variation in complex traits map predominantly to non-coding regions and are likely to be involved in gene regulation.[1,2] This poses major challenges for the characterization of the functional consequences of sequence variation and ultimately for the understanding of how genetic polymorphisms drive phenotypic variation. To address this problem, considerable effort has been dedicated to mapping genetic variants that regulate variation in gene expression levels genome-wide (i.e., expression quantitative trait loci or eQTLs). While eQTL studies in humans have led the way,[3–5] large-scale projects in model systems such as yeast,[6,7] mice[8] and *Arabidopsis*[9] have, for each of these species, provided detailed annotations of their regulatory genome. These eQTL catalogs have proven invaluable resources that have enabled the identification of general characteristics of the genetic basis of gene expression variation across species. Finally, such catalogs have helped prioritize candidate genes by suggesting mechanisms through which polymorphisms impact gene function in the context of specific traits, diseases, or adaptations.

Gene expression variation is regulated by a large number of genetic variants that can only be detected in experiments with high statistical power.[5,6] As a result, for most organisms, the gene regulatory landscape that explains individual transcriptional variation remains poorly understood. This is also true for *Drosophila melanogaster*, which, despite having one of the best annotated genomes and being an iconic model system, has lagged behind in the identification of eQTLs regulating transcriptional variation at the population level, limiting the reach of this powerful model system. Studies aimed at understanding the genetic basis of transcriptional variation in *D. melanogaster* have, thus far, been limited to a relatively small number of inbred lines[10] or pools of F1 individuals derived from crosses of inbred lines,[11] limiting the statistical power. In addition, the use of inbred genomes complicates the study of gene interactions, such as dominance or epistasis, which are an important part of the genetic architecture of most complex traits.[12]

To address these limitations, we created a synthetic outbred mapping resource that can be used to dissect the genetic basis of any complex trait. We created five mapping populations, one for each geographic location represented in the wild-derived Global Diversity panel[13] with flies collected in Ithaca (USA), the Netherlands, Beijing (China), Zimbabwe, and Tanzania. For each location, 15 to 18 inbred lines were crossed using a round-robin design, and the resulting population was maintained

in large cages and allowed to recombine freely for over ~125 generations. The resulting *Drosophila* outbred synthetic populations (Dros-OSPs) differ from other mapping panels like the Drosophila Synthetic Population Resource (DSPR)[14] and Drosophila Genetic Reference Panel (DGRP)[15] in that individuals are not maintained as inbred lines. The key benefits of this design are as follows: (1) it assures that the allele frequency spectrum is biased toward common alleles, in contrast with wild populations, which improves statistical power to detect associations with alleles that otherwise would be too rare in the wild. (2) More than ~125 generations of random mating and recombination erode long linkage disequilibrium (LD) blocks, which improves mapping resolution. (3) The outbred nature of the Dros-OSP allows for the exploration of complex traits in a natural genomic context, where complex interactions like dominance and epistasis occur. (4) Finally, this mapping panel does not impose any limitations on sample size; experiments can be as large as necessary, which is particularly important for the identification of genotype-phenotype association with small effects sizes, as may be expected for most complex traits.

Here, we characterize in detail the genome and transcriptome of the Dros-OSP from the Netherlands and use it as a case study to show the insight that an outbred mapping population can provide for the *Drosophila* complex traits community. The sample consists of 1,286 individual flies for which we collected whole-genome sequencing data. From the same individuals, we collected 1,879 RNA sequencing (RNA-seq) samples from two tissues, head (n = 940) and body (n = 939). We identified eQTLs for 98% of genes, increasing by thousands the number of genes for which regulatory loci are known in *Drosophila*. Such a comprehensive genome-wide description of the regulatory landscape allowed us to describe the properties of local (*cis*) and distant (*trans*) eQTLs in terms of genetic diversity, heritability, connectivity, and pleiotropy. Additionally, we summarize the results of a companion paper,[16] where we used state-of-the-art clustering algorithms to explore tissue-specific gene co-expression networks offering new insights into the organization of *Drosophila* transcriptional networks. The results of this study are available at http://drosophilaeqtl.org, a website that provides a user-friendly interface to explore the properties of the *Drosophila* transcriptome organization and eQTL map.

RESULTS

A mapping resource: The Dros-OSPs

Our ability to dissect the genetic basis of complex traits depends on the available mapping populations. Traditionally, inbred and recombinant inbred lines panels (RILs) have been a commonly used resource for complex trait mapping. However, research in mice and rats has shown that using outbred populations greatly improves mapping resolution and power to detect genotype-phenotype associations while maintaining a genetic context closer to wild populations.[17–19] Here, we present an outbred mapping resource in *D. melanogaster* (Dros-OSP) for mapping complex trait variation.

The Dros-OSP represents a shift from existing mapping resources in *Drosophila*. Traditionally, the sequence of a given set of inbred lines (or RILs) is provided as a resource. This means that the user only needs to phenotype their trait of interest in the desired lines and can benefit from the available genetic information to perform genetic mapping. This paradigm was a game-changer for the field when DNA sequencing was cost prohibitive. However, the limited number of available inbred lines imposes a cap on the statistical power to detect genotype-phenotype associations. And the fact that such associations tend to have small effect sizes for most complex traits further limits the inferences made with small sample sizes. The availability of outbred mapping populations not only remedies the sample size limitations, improving statistical power, but also allows us to dissect complex traits in a population that is more similar to a natural population in that each individual has a unique heterogeneous genome. The trade-off is that the user will need to sequence every Dros-OSP individual used for mapping. However, novel methods and reduction of sequencing costs have made possible the high-throughput and low-cost collection of genomic and transcriptomic data in thousands of individuals (e.g., Pallares et al.[20]).

Here, we characterize in detail the genome of one Dros-OSP population from the Netherlands and show that this mapping population has ideal properties for genetic mapping: The allele frequency spectrum is biased toward common alleles (Figure S1A); it harbors ample genetic diversity (1,128,092 segregating SNPs, minor-allele frequency [MAF] > 0.01; Figures S1B and S1D–S1I), has rapid LD decay (~200 bp; Figure S1C) and shows virtually no genetic structure (the first SNP-based principal component [PC] explains less than 3% of genetic variation; Figure S2). And finally, the Dros-OSP populations represent genetic and phenotypic diversity of five locations around the world: Tasmania, Beijing, the Netherlands, Zimbabwe, and Ithaca (USA). All outbred populations are currently maintained by the Ayroles and the Pallares labs. Sharing the population cages is very straightforward; shipping *Drosophila* is an everyday procedure for most fly labs, and a shipment of ~30 fly food bottles with eggs is enough to capture the allele frequency spectrum of the populations we currently keep in the lab.

In the present study, we sequence the genome and transcriptome of the Netherlands population and illustrate one application of this mapping resource by characterizing the genetic architecture of transcriptional variation. We have quantified genetic variation in 1,286 outbred *D. melanogaster* flies and transcriptional variation in 1,879 samples including two body parts, the head and the body, referred to as "tissues" hereafter.

We found that 9,473 genes are expressed in adult female *D. melanogaster* across both tissues. Most genes are expressed in the two tissues explored here, head and body (n = 7,795), but each tissue has ~10% of genes that are reliably detected only in one tissue (tissue-specific), 1,082 genes (12.6%) and 596 genes (10.4%) in head and body transcriptomes, respectively (Table S1). Overall, the expression level of genes expressed in both tissues (shared) is strongly correlated, although there is notable variation at the gene-specific level (Figure 1A). This is reflected in the fact that 94% of these genes are differentially expressed between tissues (Figure S3A; Table S1). The top differentially expressed genes are enriched in synapsis and sensory-related processes mediating locomotion and mating behavior (Figure S3B). And, as may be expected, tissue-specific genes are enriched for visual perception and neuropeptide

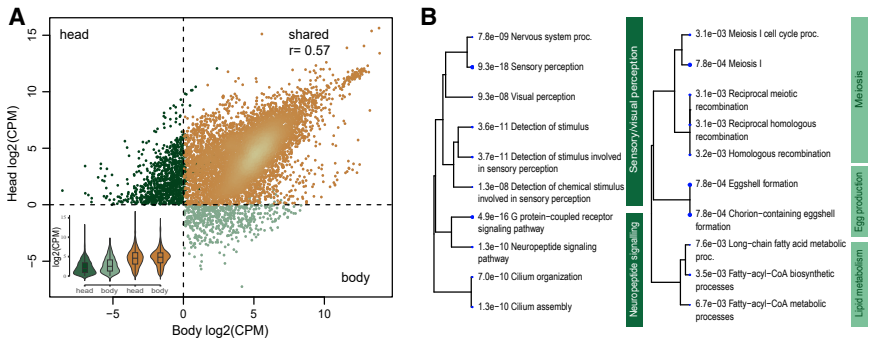


Figure 1. Correlation between head and body gene expression levels in *D. melanogaster* transcriptomes

(A) Correlation between expression levels in head (940 samples) and body (939 samples). Each dot represents a gene; dotted lines indicate the threshold used to determine whether a gene is expressed or not (average CPM > 1 and detected in at least 20% of the samples). Orange, 7,795 genes expressed in both tissues. The Pearson correlation *r* was calculated only including shared genes, *p* value < 2.2e−16. Dark green, 1,082 genes with head-specific expression. Light green, 596 genes with body-specific expression. The inset plot shows the mean expression for tissue-specific (green) and shared genes (orange).

(B) GO Biological Process enrichment for tissue-specific genes (dark green, head-specific; light green, body-specific). The top ten significantly enriched terms at FDR 0.05 are shown grouped by the number of shared genes between terms. Dot size indicates the *p* value of the enrichment, which is also indicated for each term. Analysis done in ShinyGO v0.77.[21]

See also Table S1 and Figure S3.

signaling in the head and egg production in the body (Figure 1B). Notably, tissue-specific genes are expressed at lower levels than shared genes (Figure 1A).

Heritability of gene expression

We estimated SNP heritability per gene using the linear mixed model (LMM) implemented in GEMMA.[22] This estimate corresponds to the amount of gene expression variance in the fly population explained by genetic effects. Non-zero heritability estimates were obtained for 72% (body) and 80% (head) of genes (Figures S4–S6; Table S2). The heritability distribution is skewed toward low values, with an average of 0.07–0.08 (Figure 2A). Genes with higher heritability are expressed at higher levels (Figure 2B) and have higher nucleotide diversity, π (Figure 2C), than low heritability genes. High and low heritability genes are not randomly distributed in the genome; chrX and chr4 are significantly enriched in low heritability genes, while the other main chromosomal arms (chr2L, chr2R, and chr3R) are enriched for high heritability genes. This pattern matches the differences in genetic diversity between chromosomes (Figures S1D–S1I). These groups of genes also differ in their biological function (GO Biological Processes). Low heritability genes are enriched for very general regulatory processes, including protein and RNA metabolism in both tissues and early development and morphogenesis in the body transcriptome (Figure S7). On the other hand, high heritability genes are enriched in more specific processes like carbohydrate and fatty acid metabolism and immune response and detoxification (Figure S8).

Structure of the *Drosophila* transcriptome

Next, we investigate patterns of co-expression to better understand community structure among transcripts. We focus on the average Spearman correlation between gene pairs as a proxy for connectivity. After removing poorly supported edges (Spearman correlation *p* values > false discovery rate [FDR] 1%), the average correlation per gene is 0.13 in both tissues (min = 0.11, max = 0.24; Figure 3A; Table S1). This indicates that the average correlation per gene is rather weak; however, the correlation between specific gene pairs can be very high (max rho = 0.95; Figure 3A). The median number of genes connected to the focal gene (i.e., degree) is 329 and 469 in head and body transcriptomes, respectively (min = 3, max 4,899; Figure 3B; Table S1).

Highly connected genes (correlated with ~20% of genes) are enriched in ATP metabolism, translation in both tissues, and also in behavior and synapsis in the head (Figure S9). Genes with low degree (correlated with <1% of genes) are involved in RNA processing, including ncRNA and tRNA, with specific modules in the head involved in DNA repair, and in carbohydrate and lipid related processes in the body (Figure S10). Expression level (first) and SNP heritability (second) are the gene properties best correlated with connectivity metrics. Highly expressed, highly heritable genes are connected to a large number of genes, and such connections are supported by high correlation values (Figures 3C, 3D, and S11).

To go beyond pairwise correlation and assess tissue-specific transcriptional networks, we applied the weighted nested degree corrected stochastic block model (SBM)[23,24] in our companion study[16] using a subset of the head and body transcriptomes analyzed here. This clustering algorithm, which looks for correlation signatures between genes that go beyond traditional modularity optimization, partitioned body and head transcriptomes into a total of 78 and 82 blocks, respectively, with a median of 32 and 36 genes per block (Figure 4A; Tables S3 and S1). Such fine clustering of genes allows for a clear biological interpretation of most blocks which are, on average, enriched in only six GO Biological Process terms each (Tables S3–S5). In Melo et al.,[16] we also quantified the level of assortativity of each block (Figure S12) and show that despite the similar number of blocks, 94% of the blocks in the head are highly assortative, while only about ~69% of the blocks that constitute the body transcriptome show high assortativity values (Fisher exact test, *p* value 5.88e−5, odds ratio [OR] 6.76, Figure 4B).

Genetic regulation of transcriptional variation

Consistent with the fact that most genes have non-zero heritability estimates, we identified eQTLs (genetic variants regulating

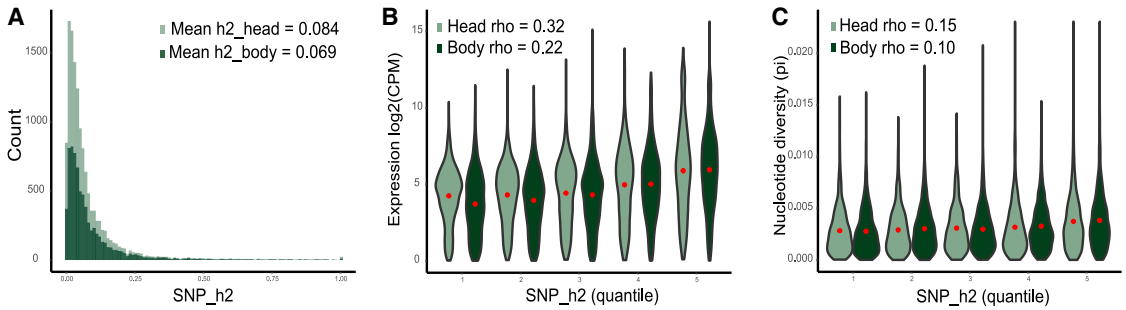


Figure 2. Heritability of gene expression

(A) The percentage of population-level expression variance explained by ~380,000 SNPs (MAF > 0.05), or SNP heritability (SNP_h2), is shown for 6,001 genes expressed in the body and 7,155 genes expressed in the head.

(B) There is a positive correlation between expression levels and SNP heritability. Genes in the first and fifth quantile have a median gene expression of 12.9 CPM and 63.6 CPM, respectively, in the head (23.3 CPM) and in the body (57.5 CPM).

(C) There is a positive correlation between heritability and per-gene nucleotide diversity (π). Average per-gene π in first and fifth SNP heritability quantiles is 0.0027 and 0.0038, respectively, in the head (0.0028) and in the body (0.0037).

The Spearman rank correlation is shown in (B) and (C) for each tissue; *p* value for all tests < 2.2e−16. Rank correlations were calculated using SNP_h2 as a continuous variable and not divided in quantiles. See also Figures S4–S8 and Table S1.

expression levels) for 98% of the genes in the *D. melanogaster* transcriptome, making this one the most comprehensive eQTL map in this species (Figure 5; Table S2). The latest eQTL map in *D. melanogaster* was done with 200 DGRP lines detecting local- and distant-eQTLs for 1,284 and 1,653 genes, respectively.[10] For a comparison, we down-sampled our data to 200 outbred individuals and found genetic regulation for 2,007 genes with local-eQTLs and 5,882 with distant-eQTLs (Figure 5). For genes with eQTLs, 60% have at least one local-eQTL (SNP position within ±10 Kb of the gene body), and 99% at least one distant-eQTL (Figure S13; Table S1). And 58% of genes have significant local and distant eQTLs, highlighting the complexity of gene expression regulation. A single gene has a median number of four local-eQTLs and five distant-eQTLs.

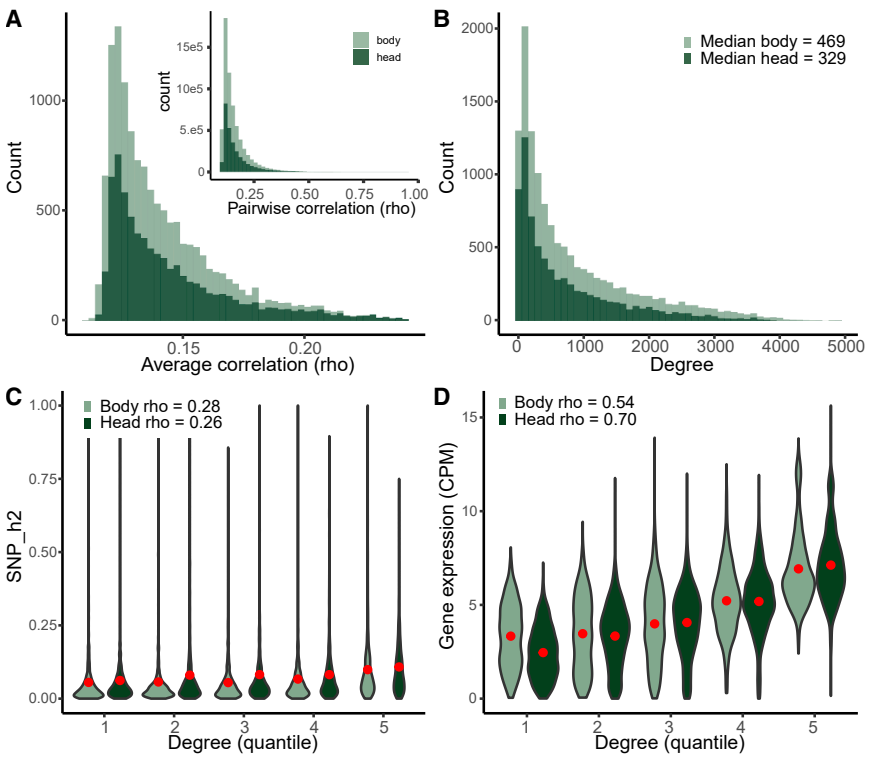


Figure 3. Patterns of gene connectivity based on expression correlation

(A) Two gene-level connectivity measures are shown: First, the average pairwise Spearman correlation between the focal gene and all other genes. Only correlations supported by a *p* value below FDR 1% were used to estimate the average. Median correlation per gene is 0.13 in both tissues (min rho = 0.11, max rho = 0.24). Secondly, the inset shows the distribution of pairwise correlation for all gene pairs with significant correlation *p* values, highlighting the fact that although the distribution is skewed toward low correlations, some gene pairs are highly correlated (max rho = 0.95).

(B) Degree represents the number of genes connected to a focal gene. For each gene, we counted the number of genes with significant Spearman correlation (FDR 1%) (min = 3, max = 4,899).

(C and D) (C) Correlation between gene expression heritability and degree. Median SNP_h2 first vs. fifth degree quantile: 0.034 vs. 0.082 in the head; 0.026 vs. 0.065 in the body. (D) Correlation between gene expression levels (CPM) and gene degree. Median CPM first vs. fifth degree quantile: 4.9 vs. 111.9 in the head; 13.5 vs. 79.5 in the body. Rank correlations in (C) and (D) were calculated using degree as a continuous variable; *p* values for the correlations were both <2.2e−16.

See also Table S1 and Figures S9–S11.

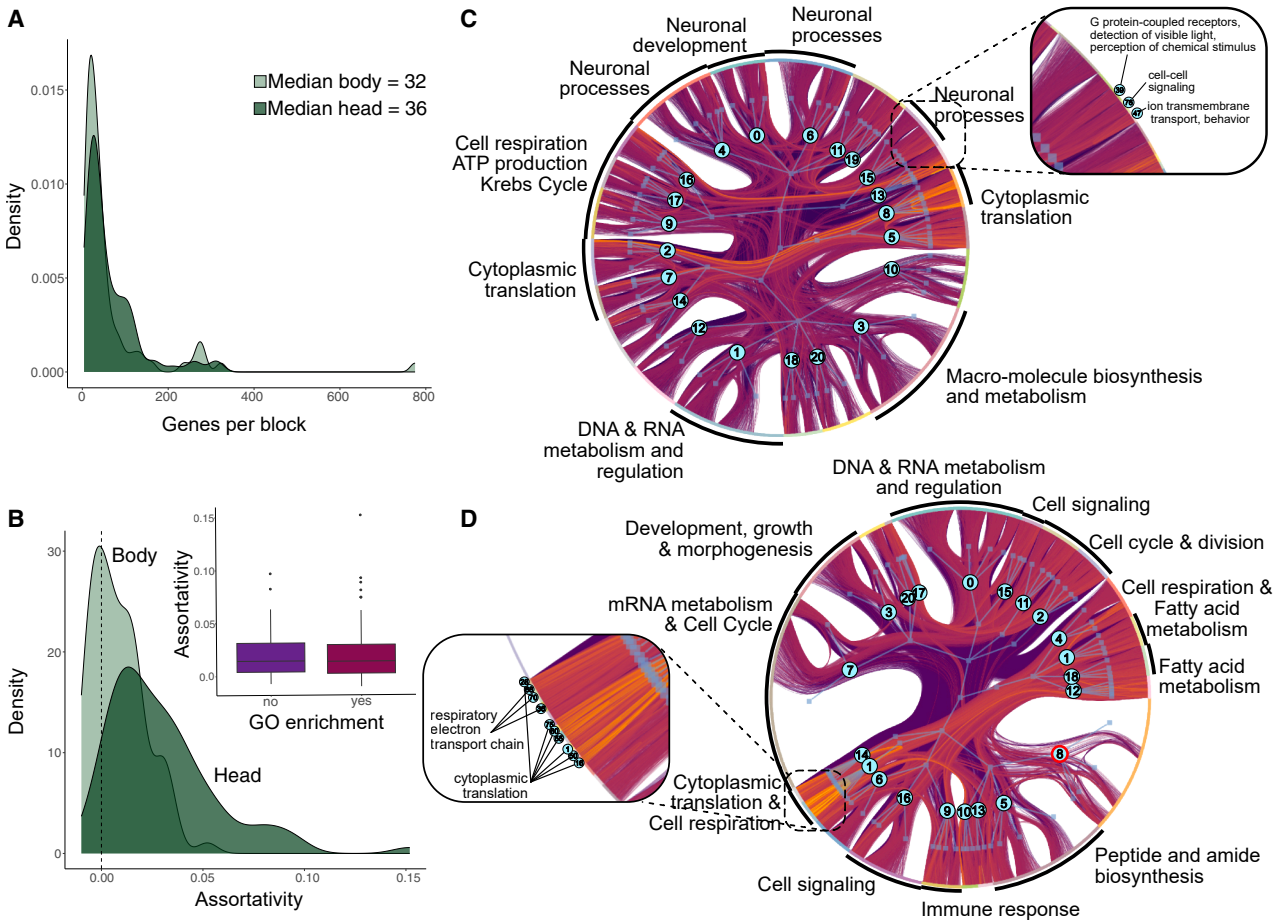


Figure 4. Modular organization of the fly transcriptome

The *Drosophila* transcriptome is highly structured, with 78 gene blocks in the body and 82 in the head.

(A) The distribution of the number of genes per gene block is heavily biased toward small blocks, increasing the resolution of the GO enrichment for each block.

(B) Not all gene clusters follow modularity assumptions. Assortativity values > 0 indicate modular behavior (i.e., genes within a block are more correlated with themselves than with genes in other modules). 94% of head blocks are modular; 67% are modular in the body. The inset plot shows that the level of assortativity is not correlated with GO enrichment; modular and non-modular blocks are associated with functional information (60%–70% of blocks have GO enrichment, Fisher exact test: p value body = 0.59, p value head = 1).

(C and D) Hierarchical representation of the transcriptome structure for head (C) and body (D). Genes are shown on the circle circumference; lines between gene pairs indicate a significant correlation. The line color indicates the strength of the correlation; stronger correlations are more yellow and plotted on top. Light blue lines in the foreground indicate the hierarchical clustering of the transcriptome into blocks (squares). The block in the middle of the circus plot represents the whole transcriptome. The blocks along the edge correspond to the final level of partition (level 1), that is, the smallest gene clusters whose characteristics are plotted in (A) and (B). GO enrichment terms for some level 2 blocks (numbered blue circles) are annotated. The numbered circles in the zoom-in panels indicate block ID in the last level of partition (level 1) and GO enrichment terms for those small blocks are annotated. The zoom-in panels show the fine scale GO partitioning of an assortative block in the head (C) and a non-assortative block in the body (D). (C) and (D) are modified from Melo et al.[16]

See also Tables S1 and S3–S5 and Figure S12.

To further characterize the local-regulatory landscape, we annotated all 258,283 local-SNPs (SNPs located ±10Kb from any gene in the genome) based on the predicted effect that each allele will have on gene expression levels. To that end, we used *DeepArk*, a machine-learning-based approach that uses 1,552 experimentally derived *D. melanogaster* datasets to predict the local regulatory activity of individual SNPs based on sequence information.[25] Compared to non-significant SNPs, the local-eQTLs identified in this study are enriched for non-zero predicted regulatory effects and particularly strongly enriched for large predicted effects (Figure 6A). This relationship between statistical discovery and predicted effect size based on experimental data has been previously shown for local-eQTLs detected in genome-wide-association studies in humans.[26] Our results show that also in *Drosophila*, statistical inference of local gene expression regulation is a good proxy for experimentally derived inference. Finally, local-eQTLs that have regulatory effects in both head and body tend to have larger predicted effect sizes compared to tissue-specific local-eQTLs (Fig. S14A). To facilitate follow-up analysis, for each gene we annotate the best local-eQTL as the one with the largest effect size predicted by *DeepArk* (see https://doi.

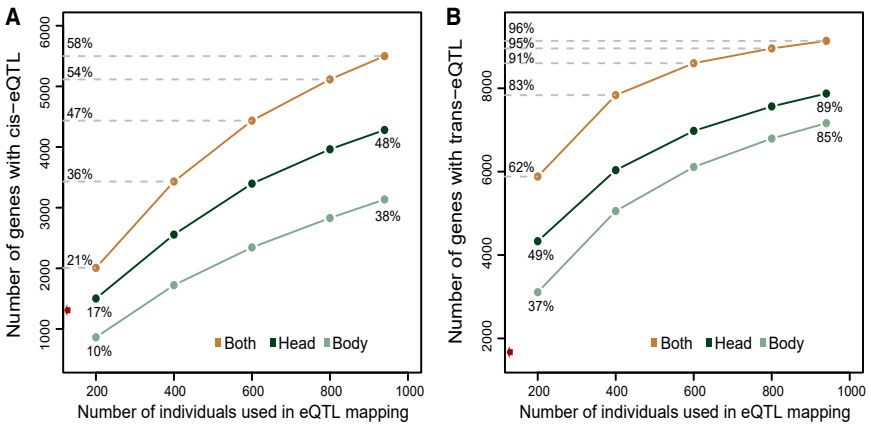


Figure 5. Saturation level of the eQTL map in *D. melanogaster*

(A and B) To evaluate the effect of sample size (i.e., statistical power) in the detection of (A) local-eQTLs and (B) distant-eQTLs, the full dataset of ~1,000 flies per tissue was randomly down-sampled to smaller datasets in increments of ~200 individuals per tissue (*x* axis). To get the overview of the total number of genes with eQTL, the results of both tissues were combined (orange line). We identify local-eQTL for 58% of the genes expressed in female *D. melanogaster* (n = 5,417), and distant-eQTL for 97% of the genes (n = 9,001). The red arrow on the *y* axis represents the number of genes with eQTL from previous eQTL mappings in *D. melanogaster* using 200 lines of the DGRP panel, and RNA-seq collected from whole females (1,284 genes with local-eQTLs and 1,653 with distant-eQTLs[10]). For comparison, with 200 outbred samples, we identify 2,007 local and 5,882 distant-eQTL. See also Tables S1, S2, S8, S10, and S11 and Figures S13 and S15–S17.

org/10.17617/3.8JLETW). With this, we offer an avenue for prioritizing candidate local-eQTLs for 58% (n = 5,499) of the genes expressed in female *D. melanogaster*.

The distant (*trans*) genetic regulation of gene expression tends to be ubiquitous, while local (*cis*) regulation is more tissue specific. To allow this comparison, we focus on genes expressed in both tissues (n = 7,719) and find that while 77% of the genes with distant-eQTLs have distant-eQTLs in both tissues, only 41% of genes with local-eQTLs have local-eQTLs in head and body. This suggests that, at the organismal level, distant-eQTLs might be more pleiotropic. We compared the minor allele frequency between local and distant eQTLs, as has been done previously,[27,28] to indirectly infer whether distant-eQTLs might be subjected to stronger negative selection. The average MAF for distant-eQTLs is indeed significantly lower than local-eQTLs (mean MAF distant = 0.25, mean MAF local = 0.28, *t* test p value < 2.2e−16) and strongly biased toward rare alleles (Figure S15). Interestingly, this pattern looks very different when restricting the analyses to eQTLs shared by both tissues, where local-eQTLs are biased toward intermediate frequency compared to distant-eQTLs (mean MAF local = 0.30, mean MAF distant = 0.272, *t* test p value < 2.2e−16; Figure S15). This suggests that different selection forces are acting on local- and distant-eQTLs, but pleiotropy levels are not the sole driver (see below for *trans*-hotspot detection).

As seen before in humans and other organisms, most eQTLs are located in non-coding regions of the genome (95% of local-eQTLs, 97% of distant-eQTLs in *Drosophila*; Figure S16A; see https://doi.org/10.17617/3.8JLETW). But, in contrast with humans, due to the smaller genome size in *Drosophila,* the majority of eQTLs fall inside genes (73% of local-eQTLs, 67% of distant-eQTLs; Figure S16A). Although this highlights the already well established importance of non-coding genetic variation, we also find an important role for coding variation in regulating gene expression: eQTLs are enriched for SNPs that generate non-synonymous substitutions, compared to SNPs that are not eQTLs (chi-square test, local-eQTLs, p < 2.2 e−16, OR = 2.18; distant-eQTLs, p < 2.2e−16, OR = 1.88). This result suggests two things: first, regulation of gene expression through distant effects can be mediated by protein changes in the focal gene where the eQTL lays (e.g., protein changes in a transcription factor will affect its binding to the promoter of distant genes). Second, and more interesting, genetic variation that results in protein changes in the focal gene also tends to

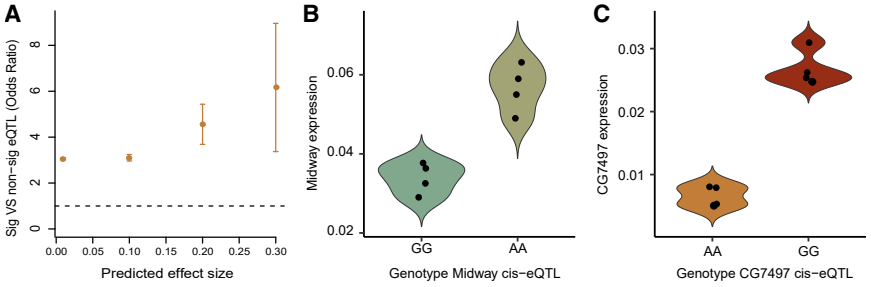


Figure 6. Local-eQTLs are enriched for large predicted regulatory effects and experimentally replicated

(A) The local-eQTLs identified in this study using a statistical approach are enriched in large predicted regulatory effects when compared to non-significant SNPs. Effect size prediction was done using the deep-learning algorithm implemented in DeepArk. Sample size for the comparison: 68,859 candidate local-eQTLs; 258,283 non-significant sites located in *cis* (within 10 Kb from start or end of gene body). For each effect size, groups were compared using Fisher exact tests. All tests have p values < 0.001. 95% confidence intervals for odds ratio are shown.

(B and C) The genotype at candidate local-eQTLs regulates expression levels of the focal gene. Each dot corresponds to a pool of 10 outbred flies fixed for one eQTL allele but with heterogeneous genomic backgrounds. Expression levels of the genes (B) *midway* and (C) *CG7497* were quantified using RT-qPCR (*midway* *t* test p value: 0.0015, fold change 1.68×; *CG7497* *t* test p value 8.088e−5, fold change 4.08×). See also Table S11 and Figure S14.

modify its own expression levels (local-eQTL). The importance of the latter mechanism is further supported by the fact that local-eQTLs located within the gene region (as opposed to upstream or downstream) tend to fall in exons more than distant-eQTLs, which instead are mostly found in introns (Figure S16B).

We find that each eQTL is associated with one or very few genes (median = 1). However, some eQTLs have tens to hundreds of distant effects. To explore the per-eQTL pleiotropy, we asked whether the potential effect on multiple genes might be mediated by direct effects on the expression levels of another gene (i.e., the high-pleiotropy eQTL also regulates a nearby gene). We see that local-eQTLs, which are located close to the target gene, are indeed significantly enriched for having more distant effects than eQTLs that only act distantly (they do not regulate neighboring genes) (Figure S17A; Table S6). Then, we focused on the top most pleiotropic eQTLs or hotspots, which we define as those with more than 20 distant target genes (max eQTL pleiotropy in head tissue = 56, in body tissue = 975; see https://doi.org/10.17617/3.8JLETW). We identify 22 and 234 hotspots in head and body, respectively (Tables S7 and S8). Hotspot number and location is one of the main differences between the eQTL map of head and body parts (Figures S17B and S17C). These highly pleiotropic eQTLs represent the full spectrum of allele frequencies in the fly population (Figure S17D), and although preferentially located in pericentromeric regions, hotspots are found across the genome (Figures S17B and S17C). To better understand the properties of the SNPs that behave as hotspots (associated with >20 genes), we compared them to other significant distant-eQTLs that are not hotspots (associated with <20 genes). Given the few hotspots in the head, we used the body hotspots for the enrichment analysis. We find that they are enriched for non-coding regions of the genome (Fisher exact test, p value = 0.0013, OR = 2.4), in particular for intergenic regions (Fisher exact test, p value = 2.19e−12, OR = 2.5); this contrasts with the location of all eQTLs, local or distant, which is preferentially genic. When the hotspots fall inside genes, those genes are not significantly enriched for predicted transcription factors. However, 59% of the hotspot's target genes in both tissues are significantly enriched for GO terms, indicating the functional cohesiveness of the group of target genes. Four genic hotspots in the body do indeed fall inside transcription factors (TFs) (six4, atf6, scro, and neu2) (Table S9). We used transcription factor enrichment analysis on the target genes to determine whether they had binding sites that correspond to the TF hotspot. In the case of six4 (FBgn0027364), the 36 target genes are indeed significantly enriched for six4 binding sites (TGATA, FDR = 0.0034). Additionally, the target genes for each of the four TF hotspots are enriched for protein-protein interactions (PPIs), GO terms, and KEGG pathways (Table S10), suggesting that, although the current knowledge about TF binding sites does not find an enrichment on predicted binding sites, the target genes are actually functionally connected. This highlights the need to develop and implement new high-throughput experimental methods to survey the unknown parts of the *Drosophila* regulatory genome.[29]

Experimental validation of local-eQTL effects

The experimental validation of regulatory effects on gene expression usually relies on creating transgenic fly lines with the SNP of interest. Given the well-known effects that genomic background has on genetic effects,[30] a reliable validation of SNP effects requires heterogeneous genomic backgrounds. Here, we attempted to validate five candidate local-eQTL-gene pairs using an experimental design that relies on creating pools of individuals that share the same genotype at the focal local-eQTL while retaining heterogeneity in their genomic background.[31] This approach reduces expression variation within a genotype while increasing the statistical power to detect differences in mean expression between alternative genotypes. We chose the candidate local-eQTLs based on their moderate allele frequencies (0.16–0.46), large DeepArk predicted effects (>10%), and non-pleiotropic effects (one SNP is associated with only one nearby gene in a specific tissue) (Table S11) and were able to validate three out of five eQTL-gene pairs at an FDR of 10% (Figures 6B, 6C, and S14B–S14D). These results correspond to one of the very few validations of the regulatory effect of naturally segregating genetic variation using outbred individuals.

DISCUSSION

Outbred populations as windows into the genetic basis of complex traits

To understand how the transcriptome is structured and regulated at the population level, we used the Dros-OSP derived from the Netherlands. Using a total of 1,286 genomic and 1,879 transcriptomic samples derived from two body parts, we identified genomic loci (eQTLs) regulating variation in gene expression for 98% of the genes expressed in female *D. melanogaster*. This comprehensive genome-wide eQTL map enabled us to dissect general properties of *cis* and *trans*-regulation in this species, bringing *D. melanogaster* to a level of understanding previously available only for a few species such as yeast[6] and humans.[4] It should be noted that our analyses, like other mapping studies, only include common genetic variants (MAF > 0.05), and therefore the effect of rare variants remains to be explored.

Focusing on an outbred population, where each individual fly has a unique genomic composition, allowed us to collect a large enough sample size to reach saturation of the eQTL map at the gene level; that is, 98% of the genes have at least one eQTL. Our eQTL catalog offers a comprehensive resource for the complex trait community, where the number of genes with known local-eQTLs increased from 1,284[10] to 5,417 and from 1,653[10] to 9,001 for distant-eQTLs. In addition to the increase in power derived from a large sample size, the use of two tissues allowed us to detect even more eQTLs. This is reflected in the fact that 59% of genes with local-eQTLs and 23% of genes with distant-eQTLs have tissue-specific local- and distant-eQTLs, respectively. This result suggests that to further improve the eQTL map, in particular the local-eQTL map, more tissues, rather than sample size, are necessary.

Recent machine-learning approaches use sequence-based information to predict the impact of individual alleles on expression levels.[25,26] However, they cannot predict which gene will be regulated by a certain SNP. In species with small genomes and high gene density like *D. melanogaster*, single SNPs might act as a local-eQTL for many genes. On the other hand, statistical approaches like the one used here to map eQTLs offer a direct link between a genetic variant and its target gene. To make

use of these two powerful lines of evidence, we have annotated thousands of local-SNPs (located ±10 Kb from any gene) with the statistical probability that they regulate a specific gene as well as with the machine-learning-derived predicted effect size of such regulatory effects. This resource will help to prioritize SNPs for further experiments, as we have done here to validate multiple candidate local-eQTLs (Table S11) using an approach that relies on outbred flies fixed for a candidate eQTL allele while preserving their genomic heterogeneity.[31]

While Dros-OSP allows for substantial increases in statistical power while retaining high mapping resolution, existing inbred line and RILs panels like DSPR[14] and DGRP[15,32] remain keystone resources for the *Drosophila* community. The greatest advantage of using inbred-line-based panels for mapping in contrast with outbred populations is 2-fold, genomes are already available, and measurement error can be minimized by phenotyping multiple individuals from the same line. In particular, the DSPR linkage-based design helps to minimize false positives and supports simultaneous estimation of QTL effect size and frequency. Additionally, the global founder representation in DSPR offers the opportunity to investigate geographically structured traits. The DGRP, in turn, offers unmatched resolution for fine mapping due to its high genetic diversity and extremely rapid LD decay, making it ideal for follow-up studies once candidate regions are known. Both of these inbred panel resources also benefit from a large community of researchers who have measured many phenotypes across these lines. Thus, while Dros-OSP and existing inbred panels rely on distinct mapping paradigms, they are best viewed as complementary tools, each optimized for different experimental goals.

Transcriptional regulation in *Drosophila melanogaster*

Understanding the genetic basis of transcriptional regulation in the fruit fly has been a research focus for more than 20 years. From the first studies tackling between-species differences[33] and the more recent attempts to explain variation within populations,[10,34] we know that there is ample genetic variation underlying gene expression, and that such genetic effects are sex dependent[10,34] and environment dependent,[35] and here we show that they also vary according to tissue.

Our finding that the expression level of most genes is regulated by a median of nine eQTLs, both local (*cis*-eQTLs) and distal (*trans*-eQTLs), suggests that gene-level regulation is more complex than initially thought. We propose that such polygenic architecture might facilitate the evolvability of gene expression by offering different possible genetic routes to adjust expression levels to the requirement of new environments, as it has been shown for other complex traits in *Drosophila*.[36]

We find that local regulation tends to be more tissue-specific, while distal regulation is more ubiquitous. This suggests that distal-eQTLs might be more pleiotropic at the organismal level and therefore be subjected to stronger negative selection compared to local-eQTLs. Indeed, we observe that the minor allele frequency distribution for distal-eQTLs is biased toward rare alleles compared to local-eQTLs, as it has been observed in outcrossing plant species like *Mimulus guttatus* and *Capsella grandiflora*.[27,28] The picture remains the same when focusing only on tissue-specific eQTLs; however, it is strikingly different for ubiquitous eQTLs. This particular set of local-eQTLs shows a bias toward intermediate allele frequencies compared to distal-eQTLs and to the whole genome, suggesting balancing selection.[27] Brown and Kelly[27] observed a similar pattern for *cis*-eQTLs in *M. guttatus* and argued that it could be explained by *cis*-effects compensating for detrimental *trans*-effects that happened to reach higher frequencies. Such compensatory *cis*/*trans* interactions have been shown in *D. melanogaster* and *D. simulans* transcriptomes.[37] Our results show that local- and distal-eQTLs are subjected to different types of selection, but the specific type depends on the tissue specificity of the eQTL. Further analyses are needed to fully understand the evolutionary constraints on the different types of transcriptional regulation. Interestingly, we find that hotspots affecting tens to hundreds of genes are not biased toward low minor allele frequencies as could be expected for highly pleiotropic SNPs. Although hotspots could be attributed to batch effects or population structure,[38] they seem to be a property of the eQTL map of several species, including mice,[8,39] *Drosophila*,[11] *C. grandiflora*,[28] and yeast—where hotspots have been experimentally validated (e.g.,[6,40,41]). Here we show that in *Drosophila*, the genes associated with eQTL hotspots are functionally related to each other in terms of protein-protein interactions, biological function, and pathway enrichment. Although functional analyses are needed to validate these hotspots, our results suggest that hotspots are indeed associated with expression variation in functionally related groups of genes. Our results suggest that the benefits of having groups of genes jointly regulated by hotspots might overcome the pleiotropy costs. Recently, it has been shown that parallel gene expression changes relevant for adaptation to new environments are enriched for *trans*-regulation in *Drosophila*.[37]

Traditionally, our understanding of the organization of transcriptional networks has been driven by the concept of modularity, where a module is defined as a group of genes that are more strongly correlated to each other than to genes outside that module. However, co-expression strength is not the only signature that indicates that genes might be working together. For example, genes might be involved in the same function if they are all, individually, strongly correlated to another group of genes, although weakly connected to each other. We found that this is indeed the predominant pattern in the *Drosophila* body transcriptome. In Melo et al.,[16] we used a subset of the current dataset and implemented the SBM[23,24]), a clustering algorithm that exploits all connectivity information, not only modularity, to show that a considerable proportion of gene blocks in *Drosophila* body transcriptome, in contrast with the head transcriptome, does not show the expected pattern of high connectivity within modules and low connectivity across modules. Importantly, despite violating the modularity expectations that are assumed to correlate with biological function, non-modular blocks are equally enriched in GO terms. This suggests that a full understanding of transcriptional organization needs to go beyond modularity and incorporate other properties of gene connectivity. In Melo et al.,[16] we discuss in detail the advantages of using clustering methods that do not rely on modularity maximization (e.g., SBM) to study the architecture of transcriptional networks and include comparisons to other gene-clustering methods.

This study uncovers the fine scale landscape of genetic variation associated with gene expression variation in *D. melanogaster*, identifies general properties of local- and distant-eQTLs including the selection forces acting on each type of genetic regulation, and describes the global organization of the transcriptome. In doing so, it sets the ground for future studies using gene expression to understand how genetic variation gets translated into phenotypic variation.

Limitations of the study

In this study, we have identified eQTLs in *D. melanogaster* focusing on females reared under standard laboratory conditions. The degree to which genotype-by-environment interactions modify this eQTL map remains to be investigated. Future studies could, for example, use males, different tissues, or different environmental conditions. Here, we used head and body, body parts with heterogeneous tissue composition, to investigate the context dependency of eQTLs. A more detailed understanding of eQTL tissue specificity will require a finer dissection of tissues and cell types. Lastly, despite the high-mapping resolution of the mapping population and the eQTL annotation using machine learning, the residual LD between SNPs needs to be experimentally addressed if causal eQTLs are to be defined. For this, fine-mapping analyses are needed.

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources and reagents should be directed to and will be fulfilled by the lead contact, Julien F. Ayroles (ayroles@berkeley.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- Sequencing data are available on SRA under BioProject: PRJNA1253316. Genotype data (VCF files) and transcriptome data (gene count matrices) have been deposited at Edmond, the open research data repository of the Max Planck Society, and are available at https://doi.org/10.17617/3.8JLETW. In this same repository, we have deposited the output of the eQTL mapping and all files needed to run the eQTL mapping in GEMMA. The results of this study can be explored interactively at http://drosophilaeqtl.org.
- The code can be accessed in https://github.com/Lufpa/eQTLcatalog/(https://doi.org/10.5281/zenodo.15792954) and https://github.com/Lufpa/TM3Seq-Pipeline (https://doi.org/10.5281/zenodo.15793089).
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

ACKNOWLEDGMENTS

We especially thank Paul Durst, Amanda Lea, Luke Henry, Michael Fernandez, Minjia Tang, Annett Schmittfull, and Sudarshan Chari for their help during the 2 weeks of daily fly collection. We thank Annett Schmittfull, Serge Picard, and Elena Filippova for their help in processing the samples. Thanks to Jen Grenier and Andy Clark for sharing the VCF file for GDL. Thanks to Lance Parsons for bioinformatics support and to Andy Dahl for discussions on the statistical approach to map eQTL in this population. During this experiment, our environmental chamber broke down carbonizing all flies at 50°C; thanks to the backup that Andy Clark and Jen Grenier kept at Cornell, we were able to repeat the experiment. J.F.A. was funded by NIH-National Institute of General Medical Sciences R35GM124881 and NIH-National Institute of Environmental Health Sciences R01ES029929. L.F.P. was funded by R35GM124881, a Long-Term Postdoctoral Fellowship from the Human Frontiers Science Program, and by the Max Planck Society.

AUTHOR CONTRIBUTIONS

Conceptualization and design were performed by L.F.P. and J.F.A. DNA and RNA-seq data generation were performed by L.F.P., J.P., and V.A. Data analysis was performed by L.F.P., D.M., and E.M.C. Website design was performed by S.W. Manuscript writing was performed by L.F.P., D.M., and J.F.A., with input from all authors.

DECLARATION OF INTERESTS

The authors declare no competing interests.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS
  - A synthetic global population of Drosophila melanogaster
- METHOD DETAILS
  - Experimental design
  - mRNA extraction and RNAseq library preparation
  - Processing of RNAseq data
  - RNAseq data normalization and covariates
  - DNA extraction and library preparation
  - Processing of DNAseq data and SNP calling
  - SNP annotation
  - Validation of candidate local-eQTL effects in outbred flies
  - RNA extraction and qPCR
- QUANTIFICATION AND STATISTICAL ANALYSIS
  - Patterns of genetic variation in the mapping population
  - eQTL mapping
  - Heritability estimates
  - Saturation curve
  - Differential gene expression analysis
  - Transcriptional structure and modularity
  - Gene connectivity
  - GO terms, chromosomal, and transcription factor enrichment analysis
  - qPCR analysis
- ADDITIONAL RESOURCES

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2025.101033.

Received: October 20, 2024
Revised: February 19, 2025
Accepted: September 9, 2025
Published: October 10, 2025

REFERENCES

1. Albert, F.W., and Kruglyak, L. (2015). The role of regulatory variation in complex traits and disease. Nat. Rev. Genet. *16*, 197–212. https://doi.org/10.1038/nrg3891.
2. Alsheikh, A.J., Wollenhaupt, S., King, E.A., Reeb, J., Ghosh, S., Stolzenburg, L.R., Tamim, S., Lazar, J., Davis, J.W., and Jacob, H.J. (2022). The landscape of GWAS validation; systematic review identifying 309

validated non-coding variants across 130 human diseases. BMC Med. Genomics *15*, 74. https://doi.org/10.1186/s12920-022-01216-w.

3. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D., et al. (2007). Population genomics of human gene expression. Nat. Genet. *39*, 1217–1224. https://doi.org/10.1038/ng2142.

4. GTEx Consortium (2020). The GTEx Consortium atlas of genetic regulatory effects across human tissues. Science *369*, 1318–1330. https://doi.org/10.1126/science.aaz1776.

5. Võsa, U., Claringbould, A., Westra, H.-J., Bonder, M.J., Deelen, P., Zeng, B., Kirsten, H., Saha, A., Kreuzhuber, R., Yazar, S., et al. (2021). Large-scale cis- and trans-eQTL analyses identify thousands of genetic loci and polygenic scores that regulate blood gene expression. Nat. Genet. *53*, 1300–1310. https://doi.org/10.1038/s41588-021-00913-z.

6. Albert, F.W., Bloom, J.S., Siegel, J., Day, L., and Kruglyak, L. (2018). Genetics of trans-regulatory variation in gene expression. eLife *7*, e35471. https://doi.org/10.7554/eLife.35471.

7. Kita, R., Venkataram, S., Zhou, Y., and Fraser, H.B. (2017). High-resolution mapping of cis-regulatory variation in budding yeast. Proc. Natl. Acad. Sci. USA *114*, E10736–E10744. https://doi.org/10.1073/pnas.1717421114.

8. Gonzales, N.M., Seo, J., Hernandez Cordero, A.I., St. Pierre, C.L., Gregory, J.S., Distler, M.G., Abney, M., Canzar, S., Lionikas, A., and Palmer, A.A. (2018). Genome wide association analysis in a mouse advanced intercross line. Nat. Commun. *9*, 5162. https://doi.org/10.1038/s41467-018-07642-8.

9. Lan, Y., Sun, R., Ouyang, J., Ding, W., Kim, M.-J., Wu, J., Li, Y., and Shi, T. (2021). AtMAD: Arabidopsis thaliana multi-omics association database. Nucleic Acids Res. *49*, D1445–D1451. https://doi.org/10.1093/nar/gkaa1042.

10. Everett, L.J., Huang, W., Zhou, S., Carbone, M.A., Lyman, R.F., Arya, G.H., Geisz, M.S., Ma, J., Morgante, F., St Armour, G., et al. (2020). Gene expression networks in the Drosophila Genetic Reference Panel. Genome Res. *30*, 485–496. https://doi.org/10.1101/gr.257592.119.

11. King, E.G., Sanderson, B.J., McNeil, C.L., Long, A.D., and Macdonald, S. J. (2014). Genetic Dissection of the Drosophila melanogaster Female Head Transcriptome Reveals Widespread Allelic Heterogeneity. PLoS Genet. *10*, e1004322. https://doi.org/10.1371/journal.pgen.1004322.

12. Mackay, T.F.C. (2014). Epistasis and Quantitative Traits: Using Model Organisms to Study Gene-Gene Interactions. Nat. Rev. Genet. *15*, 22–33. https://doi.org/10.1038/nrg3627.

13. Grenier, J.K., Arguello, J.R., Moreira, M.C., Gottipati, S., Mohammed, J., Hackett, S.R., Boughton, R., Greenberg, A.J., and Clark, A.G. (2015). Global Diversity Lines–A Five-Continent Reference Panel of Sequenced Drosophila melanogaster Strains. G3 (Bethesda). *5*, 593–603. https://doi.org/10.1534/g3.114.015883.

14. King, E.G., Macdonald, S.J., and Long, A.D. (2012). Properties and power of the Drosophila Synthetic Population Resource for the routine dissection of complex traits. Genetics *191*, 935–949. https://doi.org/10.1534/genetics.112.138537.

15. Mackay, T.F.C., Richards, S., Stone, E.A., Barbadilla, A., Ayroles, J.F., Zhu, D., Casillas, S., Han, Y., Magwire, M.M., Cridland, J.M., et al. (2012). The Drosophila melanogaster Genetic Reference Panel. Nature *482*, 173–178. https://doi.org/10.1038/nature10811.

16. Melo, D., Pallares, L.F., and Ayroles, J.F. (2024). Reassessing the modularity of gene co-expression networks using the Stochastic Block Model. PLoS Comput. Biol. *20*, e1012300. https://doi.org/10.1371/journal.pcbi.1012300.

17. Gileta, A.F., Fitzpatrick, C.J., Chitre, A.S., St Pierre, C.L., Joyce, E.V., Maguire, R.J., McLeod, A.M., Gonzales, N.M., Williams, A.E., Morrow, J.D., et al. (2022). Genetic characterization of outbred Sprague Dawley rats and utility for genome-wide association studies. PLoS Genet. *18*, e1010234. https://doi.org/10.1371/journal.pgen.1010234.

18. Pallares, L.F., Carbonetto, P., Gopalakrishnan, S., Parker, C.C., Ackert-Bicknell, C.L., Palmer, A.A., and Tautz, D. (2015). Mapping of Craniofacial Traits in Outbred Mice Identifies Major Developmental Genes Involved in Shape Determination. PLoS Genet. *11*, e1005607. https://doi.org/10.1371/journal.pgen.1005607.

19. Yalcin, B., Nicod, J., Bhomra, A., Davidson, S., Cleak, J., Farinelli, L., Østerås, M., Whitley, A., Yuan, W., Gan, X., et al. (2010). Commercially Available Outbred Mice for Genome-Wide Association Studies. PLoS Genet. *6*, e1001085. https://doi.org/10.1371/journal.pgen.1001085.

20. Pallares, L.F., Picard, S., and Ayroles, J.F. (2020). TM3'seq: a tagmentation-mediated 3'sequencing approach for improving scalability of RNAseq experiments. G3 (Bethesda). *10*, 143–150.

21. Ge, S.X., Jung, D., and Yao, R. (2020). ShinyGO: a graphical gene-set enrichment tool for animals and plants. Bioinformatics *36*, 2628–2629. https://doi.org/10.1093/bioinformatics/btz931.

22. Zhou, X., and Stephens, M. (2012). Genome-wide efficient mixed-model analysis for association studies. Nat. Genet. *44*, 821–824. https://doi.org/10.1038/ng.2310.

23. Karrer, B., and Newman, M.E.J. (2011). Stochastic blockmodels and community structure in networks. Phys. Rev. E *83*, 016107. https://doi.org/10.1103/PhysRevE.83.016107.

24. Peixoto, T.P. (2017). Nonparametric Bayesian inference of the microcanonical stochastic block model. Phys. Rev. E *95*, 012317. https://doi.org/10.1103/PhysRevE.95.012317.

25. Cofer, E.M., Raimundo, J., Tadych, A., Yamazaki, Y., Wong, A.K., Theesfeld, C.L., Levine, M.S., and Troyanskaya, O.G. (2021). Modeling transcriptional regulation of model species with deep learning. Genome Res. *31*, 1097–1105. https://doi.org/10.1101/gr.266171.120.

26. Zhou, J., Theesfeld, C.L., Yao, K., Chen, K.M., Wong, A.K., and Troyanskaya, O.G. (2018). Deep learning sequence-based ab initio prediction of variant effects on expression and disease risk. Nat. Genet. *50*, 1171–1179. https://doi.org/10.1038/s41588-018-0160-6.

27. Brown, K.E., and Kelly, J.K. (2022). Genome-wide association mapping of transcriptome variation in Mimulus guttatus indicates differing patterns of selection on cis- versus trans-acting mutations. Genetics *220*, iyab189. https://doi.org/10.1093/genetics/iyab189.

28. Josephs, E.B., Lee, Y.W., Wood, C.W., Schoen, D.J., Wright, S.I., and Stinchcombe, J.R. (2020). The Evolutionary Forces Shaping Cis- and Trans-Regulation of Gene Expression within a Population of Outcrossing Plants. Mol. Biol. Evol. *37*, 2386–2393. https://doi.org/10.1093/molbev/msaa102.

29. Li, X.C., Srinivasan, V., Laiker, I., Misunou, N., Frankel, N., Pallares, L.F., and Crocker, J. (2024). TF-High-Evolutionary: In Vivo Mutagenesis of Gene Regulatory Networks for the Study of the Genetics and Evolution of the Drosophila Regulatory Genome. Mol. Biol. Evol. *41*, msae167. https://doi.org/10.1093/molbev/msae167.

30. Chandler, C.H., Chari, S., and Dworkin, I. (2013). Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. Trends Genet. *29*, 358–366. https://doi.org/10.1016/j.tig.2013.01.009.

31. Wolf, S., Abhyankar, V., Melo, D., Ayroles, J.F., and Pallares, L.F. (2023). From GWAS to signal validation: An approach for estimating genetic effects while preserving genomic context. Preprint at bioRxiv. https://doi.org/10.1101/2023.03.09.531909.

32. Huang, W., Massouras, A., Inoue, Y., Peiffer, J., Ràmia, M., Tarone, A.M., Turlapati, L., Zichner, T., Zhu, D., Lyman, R.F., et al. (2014). Natural variation in genome architecture among 205 Drosophila melanogaster Genetic Reference Panel lines. Genome Res. *24*, 1193–1208. https://doi.org/10.1101/gr.171546.113.

33. Wittkopp, P.J., Haerum, B.K., and Clark, A.G. (2008). Regulatory changes underlying expression differences within and between Drosophila species. Nat. Genet. *40*, 346–350. https://doi.org/10.1038/ng.77.

34. Huang, W., Carbone, M.A., Magwire, M.M., Peiffer, J.A., Lyman, R.F., Stone, E.A., Anholt, R.R.H., and Mackay, T.F.C. (2015). Genetic basis of transcriptome diversity in Drosophila melanogaster. Proc. Natl. Acad. Sci. USA *112*, E6010–E6019. https://doi.org/10.1073/pnas.1519159112.

35. Huang, W., Carbone, M.A., Lyman, R.F., Anholt, R.R.H., and Mackay, T.F.C. (2020). Genotype by environment interaction for gene expression in Drosophila melanogaster. Nat. Commun. *11*, 5451. https://doi.org/10.1038/s41467-020-19131-y.

36. Barghi, N., Tobler, R., Nolte, V., Jakšić, A.M., Mallard, F., Otte, K.A., Dolezal, M., Taus, T., Kofler, R., and Schlötterer, C. (2019). Genetic redundancy fuels polygenic adaptation in Drosophila. PLoS Biol. *17*, e3000128. https://doi.org/10.1371/journal.pbio.3000128.

37. Huang, Y., Lack, J.B., Hoppel, G.T., and Pool, J.E. (2021). Parallel and population-specific gene regulatory evolution in cold-adapted fly populations. Genetics *218*, iyab077. https://doi.org/10.1093/genetics/iyab077.

38. Kang, H.M., Ye, C., and Eskin, E. (2008). Accurate Discovery of Expression Quantitative Trait Loci Under Confounding From Spurious and Genuine Regulatory Hotspots. Genetics *180*, 1909–1925. https://doi.org/10.1534/genetics.108.094201.

39. Hasin-Brumshtein, Y., Khan, A.H., Hormozdiari, F., Pan, C., Parks, B.W., Petyuk, V.A., Piehowski, P.D., Brümmer, A., Pellegrini, M., Xiao, X., et al. (2016). Hypothalamic transcriptomes of 99 mouse strains reveal trans eQTL hotspots, splicing QTLs and novel non-coding genes. eLife *5*, e15614. https://doi.org/10.7554/eLife.15614.

40. Smith, E.N., and Kruglyak, L. (2008). Gene–Environment Interaction in Yeast Gene Expression. PLoS Biol. *6*, e83. https://doi.org/10.1371/journal.pbio.0060083.

41. Zhu, J., Zhang, B., Smith, E.N., Drees, B., Brem, R.B., Kruglyak, L., Bumgarner, R.E., and Schadt, E.E. (2008). Integrating large-scale functional genomic data to dissect the complexity of yeast regulatory networks. Nat. Genet. *40*, 854–861. https://doi.org/10.1038/ng.167.

42. Bolger, A.M., Lohse, M., and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics *30*, 2114–2120. https://doi.org/10.1093/bioinformatics/btu170.

43. Dobin, A., Davis, C.A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T.R. (2013). STAR: ultrafast universal RNA-seq aligner. Bioinformatics *29*, 15–21. https://doi.org/10.1093/bioinformatics/bts635.

44. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.

45. Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., and Smyth, G.K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids Res. *43*, e47. https://doi.org/10.1093/nar/gkv007.

46. Robinson, M.D., McCarthy, D.J., and Smyth, G.K. (2010). edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics *26*, 139–140. https://doi.org/10.1093/bioinformatics/btp616.

47. Leek, J.T., and Storey, J.D. (2007). Capturing Heterogeneity in Gene Expression Studies by Surrogate Variable Analysis. PLoS Genet. *3*, 1724–1735. https://doi.org/10.1371/journal.pgen.0030161.

48. McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., and DePristo, M.A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. *20*, 1297–1303. https://doi.org/10.1101/gr.107524.110.

49. Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., and Sham, P.C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575. https://doi.org/10.1086/519795.

50. Cingolani, P., Platts, A., Wang, L.L., Coon, M., Nguyen, T., Wang, L., Land, S.J., Lu, X., and Ruden, D.M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin) *6*, 80–92. https://doi.org/10.4161/fly.19695.

51. Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics *27*, 2987–2993. https://doi.org/10.1093/bioinformatics/btr509.

52. Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T., et al. (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156–2158. https://doi.org/10.1093/bioinformatics/btr330.

53. Wu, T., Hu, E., Xu, S., Chen, M., Guo, P., Dai, Z., Feng, T., Zhou, L., Tang, W., Zhan, L., et al. (2021). clusterProfiler 4.0: A universal enrichment tool for interpreting omics data. Innovation *2*, 100141. https://doi.org/10.1016/j.xinn.2021.100141.

54. Ramakers, C., Ruijter, J.M., Deprez, R.H.L., and Moorman, A.F.M. (2003). Assumption-free analysis of quantitative real-time polymerase chain reaction (PCR) data. Neurosci. Lett. *339*, 62–66. https://doi.org/10.1016/S0304-3940(02)01423-4.

55. Ruijter, J.M., Ramakers, C., Hoogaars, W.M.H., Karlen, Y., Bakker, O., van den Hoff, M.J.B., and Moorman, A.F.M. (2009). Amplification efficiency: linking baseline and bias in the analysis of quantitative PCR data. Nucleic Acids Res. *37*, e45. https://doi.org/10.1093/nar/gkp045.

56. Law, C.W., Chen, Y., Shi, W., and Smyth, G.K. (2014). voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biol. *15*, R29. https://doi.org/10.1186/gb-2014-15-2-r29.

57. Picelli, S., Björklund, A.K., Reinius, B., Sagasser, S., Winberg, G., and Sandberg, R. (2014). Tn5 transposase and tagmentation procedures for massively scaled sequencing projects. Genome Res. *24*, 2033–2040. https://doi.org/10.1101/gr.177881.114.

58. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.

59. Benjelloun, B., Boyer, F., Streeter, I., Zamani, W., Engelen, S., Alberti, A., Alberto, F.J., BenBati, M., Ibnelbachyr, M., Chentouf, M., et al. (2019). An evaluation of sequencing coverage and genotyping strategies to assess neutral and adaptive diversity. Mol. Ecol. Resour. *19*, 1497–1515. https://doi.org/10.1111/1755-0998.13070.

60. Yang, J., Lee, S.H., Goddard, M.E., and Visscher, P.M. (2011). GCTA: A Tool for Genome-wide Complex Trait Analysis. Am. J. Hum. Genet. *88*, 76–82. https://doi.org/10.1016/j.ajhg.2010.11.011.

61. R Core Team (2021). R: A Language and Environment for Statistical Computing (R Foundation for Statistical Computing).

62. Zhang, L., and Peixoto, T.P. (2020). Statistical inference of assortative community structures. Phys. Rev. Research *2*, 043271. https://doi.org/10.1103/PhysRevResearch.2.043271.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Chemicals, peptides, and recombinant proteins		
SuperScript™ III First-Strand Synthesis System	Invitrogen	18080051
OneTaq HS Quick-Load 2x	NEB	M0486L
Critical commercial assays		
Dynabeads™ mRNA DIRECT™ Purification Kit	ThermoFisher	61012
Agencourt AMPure XP beads	Beckman Coulter	NA
Acroprep advance 1mL DNA binding plates	Pall Life Sciences	8132
QuickExtract™ DNA Extraction	Zymo Research	QE09050
Quick-RNA 96 Kit	Zymo Research	R1053
Deposited data		
DNA sequencing files	SRA	BioProject PRJNA1253316
RNA sequencing files	SRA	BioProject PRJNA1253316
Gene count tables	Edmond	https://doi.org/10.17617/3.8JLETW
VCF files	Edmond	https://doi.org/10.17617/3.8JLETW
Output mapping GEMMA	Edmond	https://doi.org/10.17617/3.8JLETW
Files to run GEMMA	Edmond	https://doi.org/10.17617/3.8JLETW
Experimental models: Organisms/strains		
Outbred Drosophila melanogaster	This paper	N/A
Oligonucleotides		
Primers for TM3'seq	Pallares et al.[20]	https://lufpa.github.io/TM3Seq-Pipeline/tm3seq_protocol
Oligos for qPCR	See Table S11	N/A
Software and algorithms		
TM3'seq pipeline	Pallares et al.[20]	https://lufpa.github.io/TM3Seq-Pipeline/pipeline, https://doi.org/10.5281/zenodo.15793089
Code for analyses		https://github.com/Lufpa/eQTLcatalog, https://doi.org/10.5281/zenodo.15792954
Trimmomatic 0.32	Bolger et al.[42]	https://github.com/usadellab/Trimmomatic
STAR	Doblin et al.[43]	https://github.com/alexdobin/STAR
Subread, featurecounts	Liao et al.[44]	https://subread.sourceforge.net/featureCounts.html
limma	Ritchie et al.[45]	https://kasperdanielhansen.github.io/genbioconductor/html/limma.html
edgeR	Robinson et al.[46]	https://bioconductor.org/packages/release/bioc/html/edgeR.html
sva	Leek et al.[47]	https://bioconductor.org/packages/release/bioc/html/sva.html
Picard		http://broadinstitute.github.io/picard
GATK v4.1.3.0	McKenna et al.[48]	https://gatk.broadinstitute.org/hc/en-us
Plink	Purcell et al.[49]	https://github.com/chrchang/plink-ng?tab=readme-ov-file
SNPeff	Cingolani et al.[50]	https://pcingola.github.io/SnpEff/
DeepArk	Cofer et al.[25]	https://github.com/FunctionLab/DeepArk
bcftools	Li et al.[51]	https://samtools.github.io/bcftools/bcftools.html
VCFtools	Danecek et al.[52]	https://vcftools.sourceforge.net/

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
GEMMA	Zhou et al.[22]	https://github.com/genetics-statistics/GEMMA
ShinyGO v0.77	Ge et al.[21]	https://bioinformatics.sdstate.edu/go77/
clusterProfiler v4.2.2	Wu et al.[53]	https://bioconductor.org/packages/release/bioc/html/clusterProfiler.html
LinRegPCR v2020.0	Ramakers et al., Ruijter et al.[54,55]	N/A

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

A synthetic global population of Drosophila melanogaster

Using the Global Diversity Lines[13] as founder lines, we created five Drosophila Outbred Populations that represent genetic diversity from five continents – Tasmania, Beijing, The Netherlands, Zimbabwe, and Ithaca (USA). For each geographical location, 14 to 20 Global Diversity Lines were crossed in a round-robin design. The resulting F1 flies were again crossed in the same way to avoid losing any of the parental chromosomes. The F2 individuals were combined in one cage per geographical location and have been freely recombining for ~200 generations at a population size exceeding ~5000 flies. Each individual fly in these populations has a unique genomic composition, making them suitable for mapping studies in a context analogous to a natural population. In this study, we used the outbred population from The Netherlands, which will be called Nex in the text. At the time of the experiment, flies had been recombining for ~120 generations, with a population size in the order of several thousands. The population was slowly expanded in a large cage for ~10 generations until reaching ~10k individuals before the experiment was conducted. All flies were maintained at 25°C, 65% humidity, and a 12h:12h light:dark cycle, and were fed media with the following composition: 1% agar, 8.3% glucose, 8.3% yeast, 0.41% phosphoric acid (7%), and 0.41% propionic acid (83.6%).

METHOD DETAILS

Experimental design

The experimental flies were collected from the Nex outbred population in three consecutive rounds of egg laying from the same parents. Each egg lay consisted of 15 bottles of fresh media placed inside the cage for ~24 h or less, controlling as much as possible for equal egg density between bottles. Newly eclosed adult flies were transferred to new bottles at the same time every day to assure that all flies were age matched; flies were collected for about a week until most pupae had eclosed. After allowing one or two days for all flies to mate, males and females were sorted and placed in different bottles until they were seven days old. To avoid CO2 exposure, we worked in a 4°C room to cold-shock seven-day-old females and plated them in 96-well plates that were immediately submerged in liquid nitrogen to flash freeze the flies. Individuals that originated from the same egg lay and eclosion batch were always kept in the same plate (allowing us to account for potential batch effect in downstream analysis). Heads and bodies were then separated in the following manner: each plate was sealed with a rubber lid (ThermoFisher, #AB0566) and placed in liquid nitrogen for a couple of minutes, then the plate was bashed against a solid surface causing the head to fall off while the thorax and abdomen remain together; using a custom-made sieve, heads only were transferred to a new pre-cooled 96-well plate while bodies remained in the original plate. Head and body plates were sealed and immediately placed at −80°C for storage. All samples used in this study are female flies.

mRNA extraction and RNAseq library preparation

We used a previously optimized low-cost and high-throughput protocol for mRNA extraction from single fly heads[20]; the protocol is described in detail here: Suppl. File 2 of.[20] Briefly, head and body samples were processed in the original 96-well plates in which they were previously frozen. The tissue was mechanically homogenized in a Talboys High Throughput Homogenizer (#930145) using one 2.8 mm stainless-steel bead (OPS diagnostics, #089- 5000-11) and 100 μl of lysis buffer per well. The lysate was transferred to a new 96-well plate and mRNA extraction was performed using the CyBio FeliX pipetting robot (Analitik Jena) using the Dynabeads mRNA DIRECT Purification Kit (ThermoFisher, #61012). Dynabeads are oligo-(dT) beads that bind the poly-adenylated tails of RNA molecules. For single head mRNA extraction, the final elution was done in 10ul 10mM TRIS-HCl resulting in ~10–20 ng of mRNA per head. For single body extraction, the final elution volume was 30 μl resulting in 90–180 ng of mRNA per body.

RNAseq library preparation was done following the TM3'seq protocol[20]; a detailed description of the protocol, primers, and reagents is available at https://lufpa.github.io/TM3Seq-Pipeline/tm3seq_protocol and in Suppl. File 1 of Pallares et al., 2020.[20] Briefly, during first strand synthesis, adapter B was ligated to the polyA tail of the RNA molecule; this adapter will be used to ligate the sequencing adapter and i7 barcode during the final amplification step. Using home-brew Tn5 transposase, the cDNA was tagmented at the same time the adapter A was ligated to the cDNA fragments. In the final amplification step, Illumina i7 and i5 primers were used to amplify cDNA fragments that contained both, adapter B and adapter A, assuring that only the 3′ end of the molecule is amplified.

Individual libraries processed in the same plate were pooled (2ul per body library, 5ul per head library) and each pool was cleaned and size-selected for a mean size of ~300-400bp using double-sided cleanup approach (0.6x right - 1x left) with Agencourt AMPure XP beads (Beckman Coulter). The quality and concentration of each pool was assessed using Pooled Agilent TapeStation before combining pools for sequencing. Samples were sequenced using 150bp SE reads on the Illumina NovaSeq S4 platform at the New York Genome Center and Illumina NovaSeq S2 at the Genomics Core Facility of the Lewis-Sigler Institute for Integrative Genomics at Princeton University.

Protocols and subroutines for mRNA extraction, cDNA synthesis, and library preparation were implemented in the CyBio FeliX liquid handling robot and are available upon request.

Processing of RNAseq data

Low-quality bases and adapter sequences were trimmed from the raw RNAseq reads using Trimmomatic 0.32.[42] Reads shorter than 20nt after trimming were discarded. [Trimmomatic parameters: SE ILLUMINACLIP:1:30:7 LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:20]. The trimmed reads were mapped to the Drosophila melanogaster genome r6.14 using STAR,[43] and uniquely mapped reads were assigned to the set of 17727 genes annotated for the r6.14 genome using feautureCounts from the package Subread [parameters: -t exon –g gene_id].[44] At this point, only RNA samples with more than 500,000 reads assigned to genes and matching DNA samples (see below) were kept for further analysis. Finally, only genes present on the autosomes (2L, 2R, 3L, 3R, 4) and chromosome X were kept for downstream analysis. This resulted in 940 samples with RNA derived from head tissue (average gene counts per sample: 3M), and 939 with RNA derived from body tissue (average gene counts: 2.4M) (Figure S18). At the gene level, we only kept genes for which at least 20% of individuals in the population were assayed (with CPM>1 for the particular gene). These filters removed not-expressed and lowly expressed genes which corresponded to roughly half of the 17727 genes present in the Dmel r6.14 annotation file (51% of the head genes, and 53% of body genes). The final dataset used for further analysis consists of RNA samples that have DNA matching samples (see section 'processing of DNAseq data and SNP calling'): 940 RNA samples derived from head tissue expressing 8877 genes, and 939 RNA samples derived from body tissue expressing 8391 genes. 593 samples were derived from head and body tissue from the same individual fly, while the rest (n = 693) correspond to one tissue per fly. The vast majority of genes were detected in more than 90% of the samples (79% and 76% of head- and body-expressed genes). On average, each gene was detected in 873 and 858 samples in head and body tissue, respectively (median = 938, 931 respectively).

RNAseq data normalization and covariates

Raw RNAseq counts per tissue were normalized by library size and composition using the function *voom*[56] implemented in the R package *limma*[45] and the normalization factors estimated with calcNormFactors from the package *edgeR*.[46] The resulting data (log2(CPM)) is normally distributed, conforming to linear modeling assumptions and controlling for outliers typical of RNAseq count data.

All samples were processed in 96-well plates according to their respective round of egg lay and eclosion date (i.e., each plate contains samples from the same egg lay and eclosion date) to library sequencing, and therefore we use plate ID as the main variable accounting for batch effects in our dataset. For RNAseq samples derived from body tissue, accounting for plate effect was enough to remove latent structure in the data, and it was therefore used as the only covariate for downstream analysis. In contrast, after accounting for all known batch effects in RNAseq data derived from head tissue, there was still evident structure left in the dataset. We therefore estimated surrogate variables (SV) using the R package *sva*[47] using a null model that included the main known batch effect, plate, as a predictor. In this way, we estimated SVs that were independent from other known batch effects. Accounting for the first 3SVs in addition to plate batch was sufficient to remove the remaining latent structure. Given that we are interested in mapping the genetic basis of subtle variation between individuals, we were conservative in the number of SVs included in downstream analyses to only remove structure visible in a PCA while not overcorrecting for individual differences.

DNA extraction and library preparation

DNA was extracted from the body or head lysate used to first extract mRNA in the following way: after mRNA bound to the Dynabeads in step 13 of the RNA extraction protocol (see https://lufpa.github.io/TM3Seq-Pipeline/mrna_extraction), the supernatant was transferred to a deep 96-well plate and mixed with 400ul of genomic lysis buffer (Zymo, #D3004-1). The mix was stored at −20C to await further processing. Samples were transferred to an Acroprep advance 1mL DNA binding plates (Pall Life Sciences, #8132) for DNA extraction using the Multi-Well Plate Vacuum Manifold (Pall Life Sciences, #5017) and DNA pre-wash, and gDNA wash buffers from Zymo (#D3004-1, #D3004-5). DNA library preparation was also performed in 96-well plates, following the tagmentation approach outlined in Picelli et al. (2014)[57] and implemented in a CyBio FeliX liquid handling robot (Analitik Jena). In a first step, 10ul (100uM) of forward oligo adapter A (or adapter B) and 10ul (100uM) of reverse oligo Tn5MErev were mixed with 80μL reassociation buffer (10mM Tris pH 8.0, 50mM NaCl, 1mM EDTA) and annealed in thermocycler using the following program: 95°C 10 min, 90°C 1 min, reduce temperature by 1°C/cycle for 60 cycles, hold at 4°C. Second, the pre-annealed adapters A and B were loaded onto the home-brewed Tn5 transposase by mixing 45ul of Tn5 with 9ul each of pre-annealed adapter A and B (10uM) and incubating the mix in a thermocycler for 30 min at 37°C. Finally, the charged Tn5 was diluted 1:1 with reassociation buffer:glycerol (1:1). The DNA tagmentation was performed by incubating for 7 min at 55°C a mix of 5ul DNA sample, 1ul of diluted Tn5, 2μL of 5X TAPS buffer

pH 8.5 (50mM TAPS, 25mM MgCl2, 50% v/v DMF), and 5μL of water. Then, 2.5μL of 0.2% SDS (Promega, #V6551) was added to the mix and incubated in a thermocycler for 7 min at 55°C to dissociate the Tn5 from the DNA. The final step of PCR library amplification was performed by combining 2μL of the tagmentation reaction with 7μL of OneTaq HS Quick-Load 2x (NEB, #M0486L), 1μL of the i5 primer (5uM), 1μL of the i7 primer (5μM), and 4μL of water. The mix was amplified on a thermocycler using the following program: 68°C 3 min, 95°C 30s, [95°C 10s, 55°C 30s, 68°C 30s] for 18 cycles, 68°C 5 min.

For each plate, 2μL per well of the final PCR amplification reactions were pooled, cleaned, and size selected for an average size of 400bp using a double-sided cleanup approach (0.6x-1x) with Agencourt AMPure XP beads (Beckman Coulter). Library size and concentration were checked for each pool using Agilent TapeStation before pooling multiple plates for sequencing. Samples were sequenced using 100bp PE reads on the Illumina NovaSeq S4 platform at the New York Genome Center.

Processing of DNAseq data and SNP calling

DNA reads were mapped to the *D. melanogaster* reference genome r6.14 using BWA[58] while preserving lane identity for samples that were sequenced in multiples lanes or batches. Secondary alignments were removed, and duplicated reads were marked using the function MarkDuplicates in Picard (http://broadinstitute.github.io/picard). Samples with <500 PE reads were removed. The distribution of mean coverage per sample can be seen in Figure S19 (average coverage of samples with head RNAseq data = 6x, and samples with body RNAseq data = 6.6x). To first create a set of high-quality SNPs for this population, we used a set of 210 samples with coverage > 8x (mean coverage 10x). Polymorphic sites were called for individual samples using HaplotypeCaller (-ERC GVCF), and joint variant calling for all samples was done with GenotypeGVCFs, both implemented in GATK v4.1.3.0. The resulting set of SNPs were hard-filtered following the Best Practices recommendations from GATK adjusted to the empirical distribution of our SNPs parameters. Using SelectVariants, we implemented filters at two levels, first to account for overall coverage parameters (remove if: singleton, total coverage <500 or >5000, average coverage <5x or >25x, present in <50% individuals), and second to account for site-specific quality parameters (remove if: QD < 5.0, MQ < 50.0, FS > 10.0, SOR >0.2, MQRankSum < −2.0 or >2.0, -ReadPosRankSum < −2.0 or >2.0, ExcessHet< 10.0 -this corresponds to a probability of violating HW equilibrium due to heterozygote excess <0.1). These filters resulted in a high-quality set of 1,128,092 biallelic SNPs with MAF>1% that we refer to as dbNex. To verify the quality of the SNP calls in the Nex population used here, we compared dbNex with SNP datasets from DGRP2[15,32] and the Global Diversity Lines[13]; 87% and 97.6% of the dbNex sites were validated in these populations, respectively. It is important to note that the DGRP panel of ∼200 lines naturally harbors more genetic diversity (∼1.9M SNPs MAF >0.05) than our panels which started from 14 to 20 lines (∼0.85M SNPs MAF>0.05). And, due to the differences in wild-caught individuals vs. round-robin based panels, LD decays rapidly, but slightly different in DGRP (∼10bp) and our populations (∼200bp).

Using dbNex as ''known dataset'', we performed base quality score recalibration (BQSR) for the 1286 samples that had matching head or body RNAseq data, and then we called SNPs using HaplotypeCaller and GenotypeGVCFs in GATK. We attempted extensively to optimize GATK's VQSR algorithm to filter our final set of variants based on parameters learned from dbNex, DGRP2, and GDL SNP sets, but we did not obtain reliable thresholds to call high-quality SNPs. Therefore, we decided to filter the final set of SNPs using hard thresholds similar to the ones described above following GATK recommendations adapted to the empirical distribution of parameters from our fly samples (filters used in GATK's SelectVariant; keep if: ExcessHet<54.69 – corresponds to a probability to violate HW due to heterozygote excess 3.4e−6, biallelic site, present in at least 500 individuals, mean coverage <22x, MQ > 50, QD > 2.0, FS < 60, SOR<3, −2.0<MQRankSum<2.0, −2.0<ReadPosRankSum<2.0). Additionally, given that many of our individual samples have low coverage (see Figure S19), we set to NA all individual genotypes with Genotype Quality (GQ) < 9 corresponding to a level of confidence of 87.5%. To do this, we first mark these sites using VariantFiltration, and then SelectVariants –set-filtered-gt-to-nocall. 52 Individuals with <20% genotyping rate were removed. Finally, for each tissue, sites with MAF<5%, present in less than 20% of individuals, and with HW p-adj for heterozygote excess <1e−7 and heterozygote deficit <1e−40 were removed. We used two different thresholds for HW equilibrium because heterozygote excess is highly correlated with technical artifacts while heterozygote deficit is a known consequence of low-coverage datasets like ours.[59] These filters resulted in a set of 854,350 SNPs. Using Plink,[49] we removed one member of SNP pairs with r2>0.8 using a sliding window of 1000 SNPs with a 5 SNP offset. This step removed over half of the sites, resulting in 389,896 sites used for mapping. As a final quality check, we removed 4 samples which relatedness was larger than 1 using *–rel-*cutoff 0.99 in Plink. The final VCF file consists of 1286 samples matching 940 head RNA samples and 939 body RNA samples.

Coverage, genotyping rate per individual and number of individuals with genotype data for each site are shown in Figure S19.

SNP annotation

We annotated the SNPs used for eQTL mapping using two approaches. First, we used SNPeff[50] to determine the position in the genome (genic, intergenic, intron, exon, etc.) and the predicted impact on protein coding sequences (synonymous or non-synonymous changes) for all 389,896 SNPs. Secondly, we used the deep learning approach implemented in DeepArk[25] to predict the regulatory impact of local-SNPs (SNPs located within 10Kb of any gene) on the expression of their local-genes (genes identified in our eQTL mapping). DeepArk uses 1552 *D. melanogaster* experimental datasets to predict regulatory impact based solely on the DNA sequence surrounding the SNP of interest. These datasets cover a wide range of tissues and developmental stages, but here we focus on the predictions based on the 245 assays performed in adult Drosophila because this is the developmental stage used in the current study. For each local-eQTL we estimate the impact of an SNP on gene regulation as the predicted effect of reference

allele *minus* predicted effect of alternative allele (ref-alt), and report the largest predicted effect size out of the 245 predictions made by DeepArk. This measure reflects the importance of a site in impacting gene regulation, i.e., if the predicted effect of reference and alternative alleles is very different, then the SNP potential to be involved in regulation of gene expression is high. If the effect, on the contrary, is very similar, then having one or the other allele is not relevant for determining the levels of expression of the focal local-gene.

Validation of candidate local-eQTL effects in outbred flies

To validate the effect of local-eQTL on the expression level of the focal gene, we implemented the approach described in Wolf et al., (2023).[31] We created small outbred populations so that each one was homozygous for one allele at the particular candidate local-eQTL. For this, we randomly sampled hundreds of virgin male and female flies from the same population used for the eQTL mapping (there is a ~100 generations gap between this and the sampling done for the eQTL mapping experiment). While anesthetized with CO2, we removed one leg per fly and used it for DNA extraction and genotyping. In the meantime, each fly was housed in a separate vial and then mated with individuals of the same genotype.

DNA was extracted following the protocol for QuickExtract DNA Extraction (cat no. QE09050). The region around the candidate eQTLs was amplified using custom-made primers (Table S11A). Individually barcoded amplicon libraries were prepared using Illumina i5 and i7 primers. Libraries were pooled and sequenced at the Genomics Core Facility of Princeton University using the Miseq nano 150bp PE reads. Genotypes at each focal eQTL were called using bcftools *mpileup*[51] with parameters -Ou -B -q 60 -Q 28 -d 1000 -T -b files | bcftools call -Ou -m -o.

Each population was started from ten individuals of the same focal genotype (five males, five females) which were used to set up five crosses. 10 female and 10 male offspring from each cross were then combined in one bottle where they mated freely for two generations. The resulting populations are the mosaic of ten founder genomes, and share a particular allele at the candidate local-eQTL. Given the large undertaking that this experiment implies, we chose for validation only five localeQTL-gene pairs (corresponding to three unique SNPs, and to three genes Midway, CG7497, and Ach. Table S11B). Besides being genome-wide significant in the eQTL mapping in both tissues (head and body), the candidate eQTLs have MAF>0.2 to increase the probability that a random sample of the population will contain some flies homozygous for the minor allele, are associated with only one nearby gene (but could be regulating different genes in each tissue), and have large DeepArk predicted effects >10%.

RNA extraction and qPCR

For each population, four pools of seven-day-old female flies (mated) were collected. If the candidate eQTL had an effect in head tissue, each pool consisted of 11 heads, if the effect was to be tested in the body, each pool consisted of three bodies. Gene expression was only quantified in females to match the eQTL mapping experiment. Each pool of heads or bodies was placed in a well of 96-well plates and total RNA was extracted using Quick-RNA 96 Kit by Zymo Research (Catalog no. R1053). 350 ng of RNA were used as input for cDNA synthesis using SuperScript III First-Strand Synthesis System by Invitrogen (Catalog no. 18080051), and rt-qPCR (using PowerTrack SYBR Green Master Mix by Applied Biosystems Catalog no. A46012) was used to quantify gene expression in each pool. The rt-qPCR primers used to amplify the target genes as well as the reference gene rpl32 are shown in Table S11C. The rt-qPCR assay was run in a 384-well plate that included all samples for all candidate eQTL-gene pairs.

QUANTIFICATION AND STATISTICAL ANALYSIS

Patterns of genetic variation in the mapping population

To characterize the genetic structure and variation in the mapping population we analyzed all 1286 DNA samples and 854,350 SNPs passing the quality control filters described above, but that had not been pruned based on LD. Using VCFtools[52] we first estimated r^2 between all pairs of SNPs in 1000bp windows (–geno-r2 –ld-window-bp 1000) excluding the pericentromeric and telomeric low recombination regions by removing the first 5Mb and last 15Mb from chr 2L, 2R, 3L, 3R and X, and the first 100kb and last 900kb of chr 4. The plots shown in Figure S1 represent the smoothed relationship between LD and genomic distance calculated using Loess regression implemented in the loess function in R, using the smoothing parameter span = 0.50. MAF was calculated using –freq in VCFtools; note that the allele frequency spectrum is truncated at MAF 0.05 because the sites used here were previously filtered based on MAF. Inbreeding coefficient (f) was estimated using the –het function, and relatedness using –relatedness. Nucleotide diversity per-site (π) was estimated in a subsample (n = 320) of the best covered DNA samples (8x > coverage <12x). VCF files for these 320 samples were re-generated to include variant as well as invariant sites using GenotypeGVCFs with the flag –include-non-variant-sites in GATK. Then, in VCFtools, while removing indels (–remove-indels) and keeping only mono- and bi-allelic sites (–min-alleles 1, –max-alleles 2), π per site and π in 1Kb windows was estimated using –site-pi and –window-pi 1000, respectively. Windowed π estimates were used to estimate chromosomal level π shown in Figure S1, and using custom scripts π per gene was estimated as the average of per site-pi across gene length (see Figure 2).

To evaluate whether large-scale genetic structure was present in the Nex population, we first generated the Genetic Relatedness Matrix (GRM) for all samples using SNPs pruned for LD > 0.8 (n = 389,896) using the command –make-grm-bin in GCTA,[60] and then extracted the first 10 PCs with the option –pca (Figure S2).

eQTL mapping

To identify genetic variation regulating transcript abundance, we used GEMMA.[22] In contrast with Matrix eQTL (Shabalin, 2012) which has been explicitly designed for eQTL mapping, GEMMA is much slower and requires post-hoc processing of *cis*-eQTL and *trans*-eQTL results. However, given the nature of our experimental design, where each female fly can lay many eggs during each round of egg lay, we expect a wide range of relatedness between individuals (see Figure S2). GEMMA provides the flexibility to fit both fixed and random effect models (Mixed Linear Models - MLM) that account for variation in relatedness between individuals using a genetic relatedness matrix (GRM). Matrix eQTL, on the other hand, implements a linear model, the GRM effect is the same for all genes resulting in an overly conservative correction of the effect of relatedness between individuals. In Figure S20 we show the comparison between GEMMA and Matrix-eQTL results for a subset of 199 randomly selected genes measured in 856 fly bodies. GEMMA recovers 99.25% of *cis*-eQTLs and 86.5% of *trans*-eQTLs detected with Matrix-eQTL while finding many more eQTLs that have been penalized in Matrix-eQTL. For contrast, we also show the results from Matrix-eQTL when relatedness between samples is not incorporated in the model.

Given that not all individual flies were sampled for both, body-RNAseq and head-RNAseq (i.e., head and body tissue taken from the same individual), the individuals used for eQTL mapping in each tissue are different. We therefore estimated in GEMMA two GRM, one for each tissue, using the 389,896 quality-filtered and LD pruned SNPs (MAF>0.05) using the following parameters: -grm -miss 0.5 -gk 2. Voom-transformed gene counts were modeled independently for each tissue: head eQTL mapping, 940 samples, 8877 genes, plate ID and SV1-3 as covariates; body eQTL mapping, 939 samples, 8391 genes, plate ID as covariate. GEMMA parameters: -miss 0.5 -lmm 1.

To distinguish local-eQTL and distant-eQTL, the results from GEMMA were processed in the following manner: First, local SNPs were defined as sites within 10kb from the gene body, and distant SNPs as any other SNPs in the genome. Out of the 8877 genes expressed in the head, 8834 have local SNPs. Out of the 8391 genes expressed in the body, 8357 have local SNPs. This resulted in a total of 675,425 local and 3,153,807,913 distant SNP-gene tests in the head dataset, and 632,149 local and 3,096,239,321 distant tests in the body dataset. Then, because not all *p*-values were recorded due to the very large number of tests, *p*-values for local- and distant-eQTL tests were adjusted for multiple testing using an FDR approach modified for the situation where *p*-values for all tests are not known, but the total number of tests is known. For this, we used the function *p.adjust* implemented in R,[61] setting *n* to the known number of local and distant tests. A threshold of 5% FDR was used to determine significance in each tissue. Shared eQTLs between tissues were defined as eQTLs with FDR <0.05 in both tissues, and tissue-specific eQTLs as FDR <0.05 in one tissue and FDR >0.05 in the other.

Heritability estimates

We estimated the amount of variance in gene expression explained by the set of 389,896 LD-pruned SNPs (MAF>0.05) and the standard error of the estimate using GEMMA.[22] GEMMA provides the percent variance explained (PVE) which corresponds to the 'SNP heritability'. The minimum SNP heritability output by GEMMA is 9.9999e−6 and is usually associated with extremely large error intervals. Given that there is no confidence in such estimates, we exclude those genes from further analyses involving SNP heritability estimates.

Saturation curve

To assess the effect of sample size in eQTL discovery, we mapped eQTLs using subsets of 200, 400, 600 and 800 samples, in addition to the full dataset of 939 body and 940 head samples (Table S2). The datasets were created in the following way for each tissue: the full dataset of raw gene counts was randomly down-sampled to 200 individuals, genes with CPM<1 and detected in less than 20% of the samples were removed. The treatment of the raw RNAseq counts, eQTL mapping, and heritability estimates was the same as described before for the full dataset.

Differential gene expression analysis

To detect the genes that are differentially expressed between head and body, raw gene counts of genes expressed in each tissue were merged, resulting in 7795 genes common to both tissues. The raw gene count dataset containing both tissues was processed through the following steps: first, TMM normalization factors were calculated with the *calcNormFactors* function from the R package *edgeR*.[46] Then, using the *voom* function in the R package *limma*[45,56] raw counts were log-transformed and normalized using the normalization factors estimated in *edgeR* to control for library size and library composition, and precision weights for each gene-sample pair were estimated to control for the mean-variance relationship inherent to RNAseq data. Finally, given the presence of latent factors driving structure in the head transcriptomes (see section "RNAseq data normalization and covariates"), surrogate variables (SV) were estimated for the joint voom-transformed dataset using the R package *sva*.[47] The first four SVs were sufficient to correct for the remaining structure in head transcriptomes and therefore were used as covariates, in addition to plate ID (see section "RNAseq data normalization and covariates").

As described in the section "eQTL mapping", some RNAseq samples were derived from the head and body of the same individual fly, making our dataset partially paired, with 593 individuals where both, head and body RNAseq data was collected. To account for this, we estimated the within-individual correlation for each gene using the function *duplicateCorrelation* from the R package *limma*[45] on the voom-derived object, while specifying the model design and blocking the individual fly ID. Lastly, we modeled differential

expression between tissues by fitting a linear model to the voom-derived object (contains log-transformed normalized counts and precision weights) using the function *lmFit* in the same R package, and including plate ID and SV1-SV4 as covariates. Parameters used in lmFit: block = individual ID, correlation = consensus correlation obtained from duplicateCorrelation as explained above. Differentially expressed genes were called at 5% FDR.

Transcriptional structure and modularity

The details associated with our application of the Stochastic Block Model used to explore the structure of *D. melanogaster* transcriptome are provided in great detail in.[16] Briefly, we analyzed the subset of 248 samples with best coverage per tissue, each one with more than three million RNAseq reads assigned to genes (average gene counts: head = 4.65M, body = 4.58M), filtering for genes with consistent expression across samples (CPM>1). We estimated Spearman correlations between gene pairs and retained only statistically significant edges (FDR 1% for head, 0.1% for body). We then applied the Weighted Nested Degree Corrected Stochastic Block Model[23,24] to cluster genes into hierarchical blocks, using arctanh-transformed correlation values as edge weights. This approach identifies transcriptional structure through Bayesian posterior probability maximization rather than traditional modularity metrics alone. For each gene block in each level of the hierarchy, we estimated its assortativity, i.e., the contribution of each block to the transcriptome-wide level of modularity[62] (Figure 4B (adapted from,[16]; Table S3). Assortativity ranges from −1 to 1, with negative numbers indicating that genes in the respective block are less correlated with each other than to genes in other blocks, and therefore that block does not contribute to the modularity of the transcriptome at that level of the hierarchy. Positive numbers indicate that genes in that block behave like modular genes. To see a detailed description of the SBM method, a comparison with traditional modularity-based clustering approaches like WGCNA, and a discussion of its implications for the understanding of transcriptome structure, see.[16]

Gene connectivity

To determine the level of connectivity of each gene in the transcriptome we used two different metrics estimated using the whole dataset (head samples and genes = 940, 8877; body samples and genes = 939, 8391). First, we estimated the Spearman correlation between all gene pairs, and retained the correlations significant at an FDR 1%. This threshold retained all genes, meaning that all genes retained at least one gene partner after non-significant correlations were removed. Then, we estimated for each gene: a) the number of genes connected to the focal gene (degree), and b) the average correlation between the focal and the connected genes (average correlation).

GO terms, chromosomal, and transcription factor enrichment analysis

ShinyGO v0.77[21] and clusterProfiler v4.2.2[53] were used to calculate the enrichment of gene sets using Biological Process GO terms. The background of genes varies depending on the analysis, e.g., in Figure 1 the background for each tissue corresponds to all genes expressed in that tissue, but for Figure S1 the background includes only genes expressed in both tissues. The number of enriched GO terms shown and the FDR threshold is indicated in each figure. Chromosomal enrichment was also calculated in ShinyGO using the same background as for GO term enrichment, for this we used sliding windows of 6Mb in steps of 3Mb and the hypergeometric test. Significantly enriched regions are defined as FDR 1e−5. For the hotspot analysis, we estimated the enrichment in transcription factors (TF) using the database from the Drosophila transcription factor database – FlyTF which has 753 putative site-specific TFs, and run Fisher exact tests with specific gene backgrounds for head and body data. Odds ratio and *p*-values are presented in the text. GO BP, TF binding sites, and KEGG pathway enrichment analysis for the hotspot data was done in ShinyGO v0.77.[21]

qPCR analysis

Raw qPCR data without baseline correction was imported into LinRegPCR v2020.0[54,55] where the PCR efficiency for each primer pair was estimated. Deviation from the expected PCR efficiency of two is known to cause large biases in the estimation of fold changes in gene expression.[54,55] LinRegPCR uses linear regression of log (fluorescence) to estimate the PCR efficiency per sample, and uses the average efficiency of all samples per amplicon group (i.e., per primer pair) to estimate the amount of starting RNA concentration per sample (N0) in arbitrary fluorescence units. The amount of starting RNA per sample is then estimated as the ratio of N0 for the target gene (gene affected by eQTL) and N0 for the reference gene (rpl32). The difference in gene expression levels between genotypes at a particular eQTL was tested using t.test (see Table S11B for results and sample size).

ADDITIONAL RESOURCES

The results of this study can be explored interactively at drosophilaeqtl.org.