Journal of Genetics (2024)103:8
https://doi.org/10.1007/s12041-023-01460-8

RESEARCH ARTICLE

Genomewide architecture of adaptation in experimentally evolved *Drosophila* characterized by widespread pleiotropy

ZACHARY S. GREENSPAN[1], THOMAS T. BARTER[1], MARK A. PHILLIPS[1,2], JOSÉ M. RANZ[1], MICHAEL R. ROSE[1] and LAURENCE D. MUELLER[1]*

[1]*Department of Ecology and Evolutionary Biology, University of California, Irvine, CA 92697, USA*
[2]*Department of Integrative Biology, Oregon State University, Corvallis, OR 97331, USA*
*For correspondence. E-mail: ldmuelle@uci.edu.

Received 16 September 2023; accepted 8 November 2023

Abstract. Dissecting the molecular basis of adaptation remains elusive despite our ability to sequence genomes and transcriptomes. At present, most genomic research on selection focusses on signatures of selective sweeps in patterns of heterozygosity. Other research has studied changes in patterns of gene expression in evolving populations but has not usually identified the genetic changes causing these shifts in expression. Here we attempt to go beyond these approaches by using machine learning tools to explore interactions between the genome, transcriptome, and life-history phenotypes in two groups of 10 experimentally evolved *Drosophila* populations subjected to selection for opposing life history patterns. Our findings indicate that genomic and transcriptomic data have comparable power for predicting phenotypic characters. Looking at the relationships between the genome and the transcriptome, we find that the expression of individual transcripts is influenced by many sites across the genome that are differentiated between the two types of populations. We find that single-nucleotide polymorphisms (SNPs), transposable elements, and indels are powerful predictors of gene expression. Collectively, our results suggest that the genomic architecture of adaptation is highly polygenic with extensive pleiotropy.

Keywords. *Drosophila*; genomics; transcriptomics; pleiotropy; machine learning; experimental evolution.

Introduction

Experimental evolution is a powerful approach for studying the phenotypic and molecular changes caused by selection. Experimental evolution establishes selection on replicated populations in well-defined environments (Bennett and Lenski 1999; Garland and Rose 2009). Genomewide sequencing can provide extensive catalogues of changes in the frequency spectrum of genetic variants in lab evolved populations, as well as a portrait of their expression levels (Remolina *et al.* 2012; Schlotterer *et al.* 2015; Mallard *et al.* 2018). Yet despite recent attempts (Burke *et al*. 2016; Turner *et al.* 2011; Schlotterer *et al.* 2015; Graves *et al.* 2017; Hsu *et al.* 2020), the study of the genomewide architecture of adaptation is still in its infancy (Braendle *et al.* 2011; Topa *et al.* 2015; Taus *et al.* 2017; Kelly and Hughes 2018; Vlachos *et al.* 2019). Multiple questions about the molecular basis of such laboratory adaptation remain unanswered.

Sequences that regulate the expression of neighbouring genes, so-called '*cis*-regulatory elements' have been a focus of many studies on the molecular basis of adaptation. Such studies are predicated on the hypothesis that *cis*-regulatory regions are prime locations that selection targets rather than changes in protein sequence, to drive adaptation (Carroll 2000; Carroll *et al.* 2001; Wray *et al.* 2003; Shapiro *et al.* 2004). However, it is not yet clear that such changes are central in adaptation at the molecular level (Hoekstra and Coyne 2007). Changes in protein-coding regions have been identified in processes of adaptation (Brideau *et al.* 2006). It is also conceivable that changes in distant genomic regions (*trans*-regulation), as opposed to local regions involved in

Zachary S. Greenspan and Thomas T. Barter contributed equally to this work.

Supplementary Information: The online version contains supplementary material available at https://doi.org/10.1007/s12041-023-01460-8.

Published online: 17 January 2024

cis-regulation, might influence transcript abundance and ultimately adaptation. Dating back to the pioneering work of Fisher (1930), it has been suggested that natural selection will limit the degree of pleiotropy exhibited by beneficial mutants (Orr 2000; Otto 2004). This study explores the relative rarity of pleiotropic effects.

Another topic that has been discussed with regard to the molecular basis of adaptation is the role of transposable elements (TEs) in phenotypic change. TEs have been studied extensively with regard to their impact on the genome (Chenais *et al*. 2012; Stapley *et al.* 2015; Van't Hof *et al*. 2016). However, it is not clear to what extent TEs underpin adaptation. Some have concluded that, in ever-changing environments, TEs might play a considerable role in a population's ability to adapt (McClintock 1950; Casacuberta and Gonzalez 2013), and there are clear examples that support the hypothesis that TEs underlie phenotypic variation (Van't Hof *et al*. 2016). Nevertheless, there is again some uncertainty about whether TEs drive adaptation and phenotypic change in the context of complex traits.

Selection can lead to a wide array of reproducible changes in the genome (Topa *et al.* 2015; Graves *et al.* 2017; Taus *et al.* 2017; Hsu *et al.* 2020). In addition, there have been a handful of studies that focus on reproducible changes in the transcriptome between populations that have differing selection regimes (Remolina *et al.* 2012; Mallard *et al.* 2018; Barter *et al.* 2019). However, the relative importance of these two kinds of -omic differentiation in the molecular machinery of adaptation is not clear. Only an integrated analysis encompassing genomic, transcriptomic, and phenotypic data can properly address the *relative* predictive power of each type of molecular differentiation. Further, by having all three types of data in one experimental system of sufficient power, we can test the pattern of connectivity between each level of information (genome, transcriptome, phenotypes), and elucidate whether the flow of information from the genome to the phenotype is simplistic (one gene to one transcript to one phenotype), polygenic (many genes affecting one transcript affecting one phenotype), or more like a network (many genes affecting many transcripts affecting multiple phenotypes) (vid Wright 1980).

Machine learning is a widely used tool across numerous fields of study (Sebastiani 2002; Krizhevsky *et al.* 2012). In particular, machine learning tools are well-suited to the task of parsing causal patterns involving large amounts of data (de Los Campos *et al.* 2013; Petersen *et al.* 2016; Mueller *et al.* 2018; Schrider and Kern 2018). As such, they are potentially well suited to parsing the molecular machinery underlying adaptation (Venier *et al.* 2022), especially when that machinery is complex. The hope is that a combination of high-throughput sequencing, machine-learning tools, and well-characterized life-history characters might help us to understand how genetic variation underpins adaptation generally, and thereby provide some resolution for the issues about the molecular biology of functional evolution adduced above.

Here, we study the interplay between genomics, transcriptomics, and life history traits in 20 experimentally evolved *Drosophila melanogaster* populations. Of these populations, 10 have been selected for accelerated development, while the remaining 10 have been selected for postponed reproduction (figure 1). For all these populations, we have genomic data (Graves *et al.* 2017), female transcriptomic data (Barter *et al.* 2019), and measures of several key life history traits (Burke *et al.* 2016), which show how these populations have become differentiated over time (figures 2 and 3). In this study, we use genomic and transcriptomic data to infer which genes and genomic regions might be causally linked to age-specific mortality and fecundity, two key components of biological fitness. We then evaluate which type of data, genomic or transcriptomic, is most useful when developing predictive models with machine learning. Next, we use different genomic features (e.g. SNPs, TEs) to examine which of them are more predictive when linked to changes in gene expression across the transcriptome. In particular, we test whether or not TEs are a central driving force in adaptation. We also test whether *cis*-regulation plays a major role in functionally influencing the transcriptomes. Finally, we evaluate the pleiotropic network model of Sewall Wright (1980). The specific machine learning approach we employ is the fused lasso additive model, or FLAM (Petersen *et al.* 2016), which previous theoretical work suggests is well suited to the task (Mueller *et al.* 2018).

Materials and methods

Experimental populations

The populations used in this study were subjected to two selection regimes which differed with respect to age-at-reproduction (Rose *et al.* 2004; Burke *et al.* 2016; Graves *et al.* 2017). Each selection regime was applied to two sets of five populations, each with known distinct evolutionary histories (figure 1). The ACO and AO populations, collectively called A-type, were selected for accelerated development and have a generation length of 10 days. The CO and nCO populations, collectively called C-type, have a generation length of 28 days.

Genomic data

We used genomewide SNP, TEs, and structural variant data previously published by Graves *et al.* (2017). The structural variant data were limited to regions between 0.15-kb and 10-kb long. Genomic extraction and sequencing details are provided in Graves *et al.* (2017). We also used the same read mapping protocols, but with a more recent version of the *D. melanogaster* reference genome (Dmel v6.14). The new version of the SNP table can be found in the Dryad directory

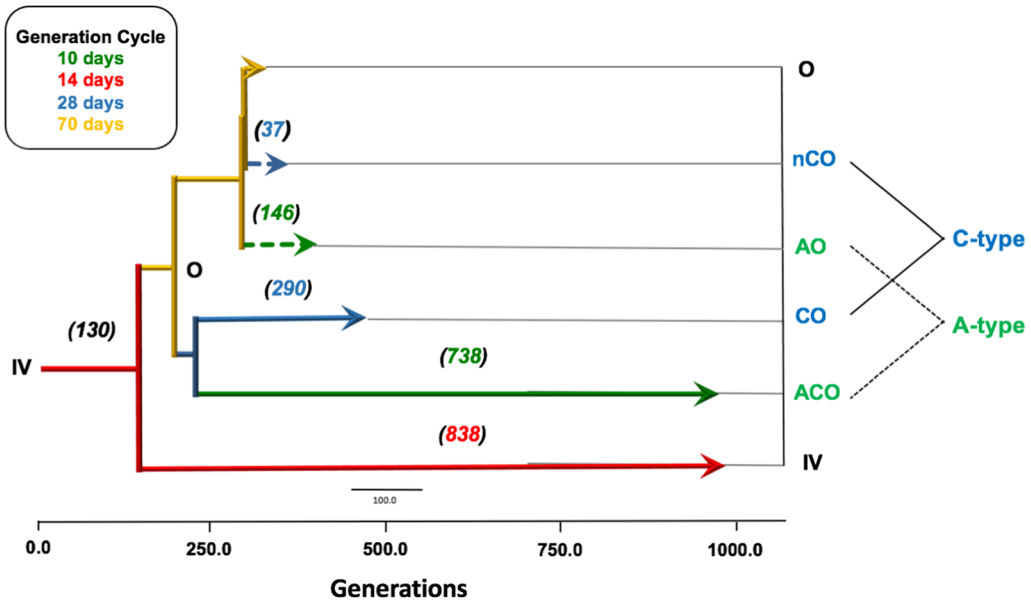

Figure 1. Twenty-population selection history. All treatments in the study share ancestry. Each colour represents a different selection regime and population type (A and C). A-type populations have a 10-day life cycle, while the C-type populations have a 28-day life cycle.

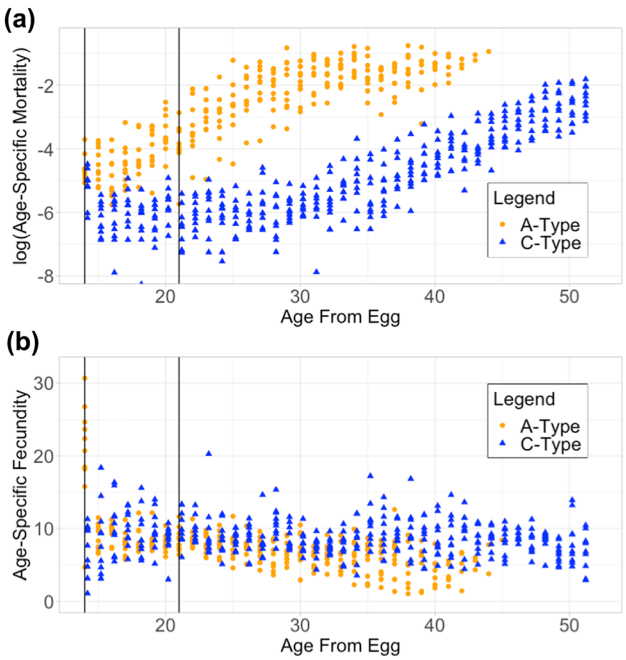

Figure 2. Age-specific phenotypes for 20 large cohorts, one for each of 20 populations. (a) Female age-specific daily mortality (Burke *et al.* 2016). (b) Female age-specific fecundity (eggs per female-day). Orange circles in both graphs represent A-type populations while blue triangles represent C-type populations. The black vertical lines represent the age of the populations when sequenced for the transcriptomics work.

created for this project (data underlying this article are available in the Dryad Digital Repository, at https://doi.org/https://doi.org/10.7280/D1ZT3D). To determine which SNPs would be used in our FLAM analysis, we followed the procedures described by Mueller *et al.* (2018). Briefly, we ran the CMH test as described by Graves *et al.* (2017) along with the same permutation procedure to identify significantly differentiated SNPs. This resulted in a list of 4211 differentiated SNPs spread across all major chromosome arms. Next, we divided each chromosome arm into discrete 50 kb windows starting at position 1 on each chromosome and then discarded any windows that contained less than three significantly differentiated SNPs between population types. Lastly, for each of the remaining windows, we recorded the position within each window with the smallest *p*-value based on the CMH test results. This resulted in a list of 194 positions that serve as representatives of the 50 kb regions that met all of our criteria. These positions and their associated SNP frequencies were then used as inputs in our analyses. In some of these analyses, we also used the differentiated structural variants and TEs described in Graves *et al.* (2017).

Given the magnitude of statistically significant SNPs versus other structural variants, our data sanitization process was only applied to this one genomic feature class. None of the structural variants had thousands of statistically differentiated elements. While there is a small amount of clustering in each remaining feature class, it is much lower than the total original set of differentiated SNPs. For example, of the 71 differentiated TEs, there is only one pair of TEs less than 50 kb apart. Therefore, we used the total list of differentiated structural variants and TEs as presented in Graves *et al.* (2017).

Transcriptomic data

We used previously published stranded PE 75 RNA-seq data corresponding to day 14 and day 21 from the egg (Barter

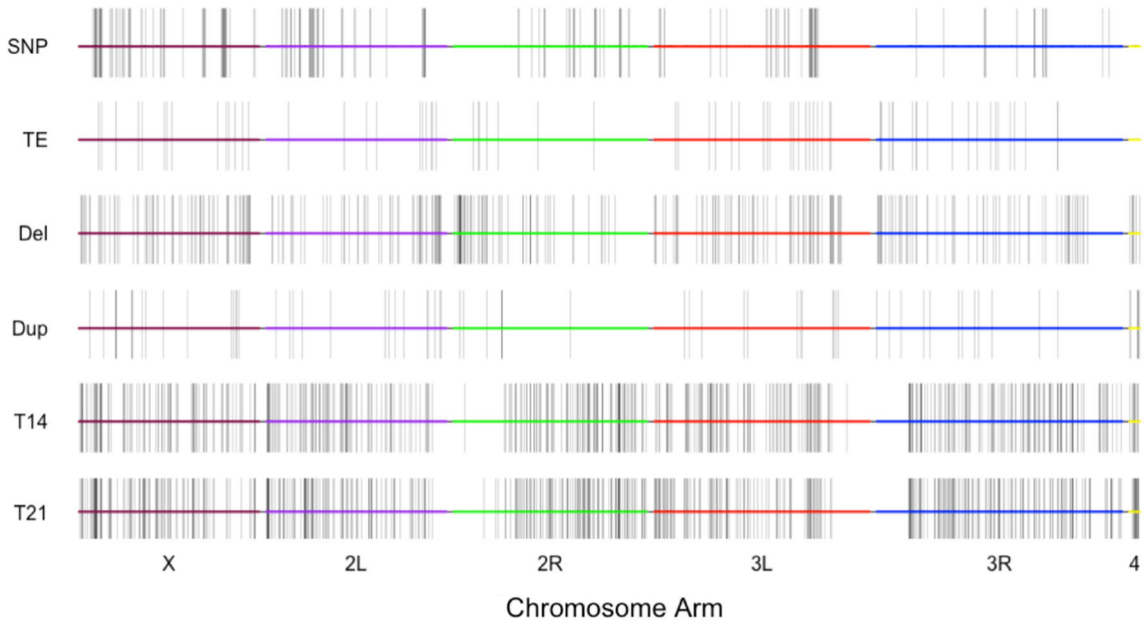

Figure 3. Chromosomal distribution of molecular differences between the two treatments. Each bar shows the position of differentiation across the genome and transcriptome. From top to bottom, each panel shows differentiation in SNPs, TEs, deletions, duplications, expression differences on day 14 (T14), and expression differences on day 21 (T21), respectively.

et al. 2019). Details about extraction, sequencing, and read mapping are given in Barter *et al.* (2019). Briefly, after mapping sequencing reads, alignment post-processing was performed with SAMtools v.0.1.19 (Li *et al.* 2009). Read counts per gene and population were done using HTSeq v0.6.1p1 (Anders *et al.* 2013). For each sample, per gene read counts were normalized using the default DESeq2 settings (Love *et al.* 2014). Genes showing normalized count values greater than 4 in at least 8 out of 10 populations, within at least one of the treatment types, were kept and the rest were discarded. With these normalized gene count values, we used the linear mixed effects model featured in Barter *et al.* (2019) to determine which genes were differentially expressed between our two selection regimes while accounting for any block effects that may be associated with different rounds of extraction and sequencing (R Core Team 2018). Statistical significance for differential expression of any given gene was set at a 5% false discovery rate (FDR) for ∼4000 tests, i.e., the number of expressed genes that passed our filtering criteria (Benjamini and Hochberg 1995). The normalized gene count values for the differentially expressed genes were then used as inputs in our analyses.

Phenotypic data

We relied on age-specific adult mortality and fecundity data for the 10 A and 10 C populations (Burke *et al.* 2016). Mortality and fecundity data were available over the entire adult lifespan of the flies. In our analyses, we focussed on average fecundity and log-transformed mortality measures taken over 3-day intervals. The mortality data were specifically log-transformed due to the fact that a portion of the mortality values were extremely low and FLAM would otherwise treat these low values as 0, instead of fitting the low values.

FLAM analyses

Experimental laboratory evolution typically results in the differentiation of many genes and multiple phenotypes (Rose *et al.* 2004; Remolina *et al.* 2012; Burke *et al.* 2016; Graves *et al.* 2017; Phillips *et al.* 2018; Kezos *et al.* 2019; Fabian *et al.* 2021). Documenting these phenotypic and genetic differences is generally straightforward, especially with populations that have been selected for many generations with different culture regimes. Some of the differentiated gene regions (such as SNPs) may affect several phenotypes and thus exhibit pleiotropy. Some genes may only have effects on a single phenotype. Even gene regions which are pleiotropic may not have measurable impact on all the phenotypes which have become differentiated due to adaptation to the novel laboratory environments.

To determine which of the many differentiated regions influence a specific phenotype, we used a statistical learning tool called the FLAM (Petersen *et al.* 2016; Mueller *et al.* 2018). The criteria used to select our causative loci are those that minimize the penalized objective function derived by Petersen *et al.* (2016, eq. 5). At each informative SNP, for example, FLAM builds a step function that describes the relationship between SNP frequency and phenotype. This step function does not have to follow a straight line and can mimic a variety of nonlinear relationships. Statistical

learning tools do not rely on the classic hypothesis testing paradigm in which the distribution of a test statistic is quantified under a null hypothesis. Instead, statistical learning tools make qualitative or quantitative predictions, and the accuracy of these predictions is assessed with test sets of data, e.g., data not used to tune the statistical learning tool. In extensive computer simulations, Mueller *et al.* (2018) showed that, if environmental variation was not too severe, FLAM could reliably distinguish between gene regions that affected a phenotype (causal genes) from those that were differentiated but did not influence the phenotype (noncausal). This ability improved with the number of populations tested. When FLAM is presented with two groups of highly differentiated loci, one of which is causal for a phenotype and the other is not, FLAM can use even subtle differences among replicates to distinguish between these two groups. Standard linear model techniques are unable to do this. Mueller *et al.* (2018) demonstrated that there are two important issues that can interfere with the inference that a SNP is a true biological signal; (i) linkage and (ii) high levels of environmentally induced phenotypic variation. Linkage could make the true biological signal anywhere from 10 to 100 kb away from the identified SNP (Mueller *et al.* 2018). High levels of environmental variation reduce FLAM's ability to separate causal from noncausal differentiated loci.

In principle, a single run of FLAM will be limited to finding no more than N causal SNP's, where N is the total number of independent populations. However, we implemented a permutation procedure that may expand the list of causative loci above N (Mueller *et al.* 2018). In the present study, a total of 100 permutations of the columns of genetic data were done and the final list consisted of genetic variants which occurred at a frequency of at least 50% of the frequency of the most commonly chosen variant, similar to what was tested in Mueller *et al.* (2018).

Results

Predicting age-specific mortality using genomics and transcriptomics

We investigated the relationships between phenotypic, genomic, and transcriptomic data (figures 2 and 3; Material and Methods) by analysing previously collected data from the same populations for the different data types of interest. Both the genome and the transcriptome have the ability to predict some life-history phenotypes, in that they both contain information that ultimately impacts a phenotype, but here we ask whether one of them has more predictive power than the other. Using FLAM, we evaluated whether genomic or transcriptomic data contribute more (i.e., they are chosen more frequently) as predictors in the model, which might suggest which type of -omics is more relevant for understanding the molecular basis of adaptation. To address this, and with the purpose of predicting mortality at days 14 and 21 from egg between the two population types, we performed three different analyses in which we used different types of differentiation data between treatments as predictors: (i) 194 genomic differentiated regions at the SNP level; (ii) normalized expression levels of 539 (day 14) or 625 (day 21) differentially expressed genes; and (iii) both (see Methods for details).

To begin, genomic regions chosen by FLAM are not simply the most differentiated regions with the lowest p-values among the A and C type populations. As previously shown (Mueller *et al.* 2018), FLAM eliminates many highly differentiated SNPs if their pattern of variation within differentiated groups does not follow the pattern of phenotype variation.

The results were consistent across comparisons using both days of transcriptomic data (figure 4; table 1) (figure 1 in electronic supplementary material at http://www.ias.ac.in/jgenet/). Neither day 14, (7/194 vs 7/539 test for equality of proportions, $p = 0.087$), nor day 21, (6/194 vs 8/625; test for equality of proportions, $p = 0.17$) had a sparse set that prefers one type of -omic data over the other. Additionally, when both genomic and transcriptomic sets were combined into one set of input predictors, we once again found that neither genomic nor transcriptomic data were consistently preferred.

We found that both kinds of -omic data contribute to the predictive model (figure 4). The model does not appear to strongly favour one type of -omic data over the other. For both day 14 and day 21 mortality, we found potential predictors from both genomic and transcriptomic data. Therefore, it seems ill-advised to use only one type of -omic data. The reason why one type of -omic data may perform better for some phenotypes but not for others is not apparent at this time.

The predictive power of the transcriptome for later-age phenotypes

Unlike the genome, the transcriptome is subject to change over the course of an organism's life, most notably during developmental transitions. Nevertheless, in *D. melanogaster*, after eclosion, cell division and tissue remodelling are minimal compared to previous stages of its development (Smith *et al.* 1970), which is well reflected in the relative stability of the transcriptome and proteome (Arbeitman *et al.* 2002; Casas-Vila *et al.* 2017).

We used the expression levels of differentially expressed genes from day 14 and day 21 as predictors of mortality and fecundity at days 14–35. For each test of prediction of a phenotype we used a stratified 5-fold cross-validation methodology in order to accurately test said predictive capability. Our data were fractured into two parts: 16 populations were used as a training set and the remaining four populations were used as a testing set. Each testing set

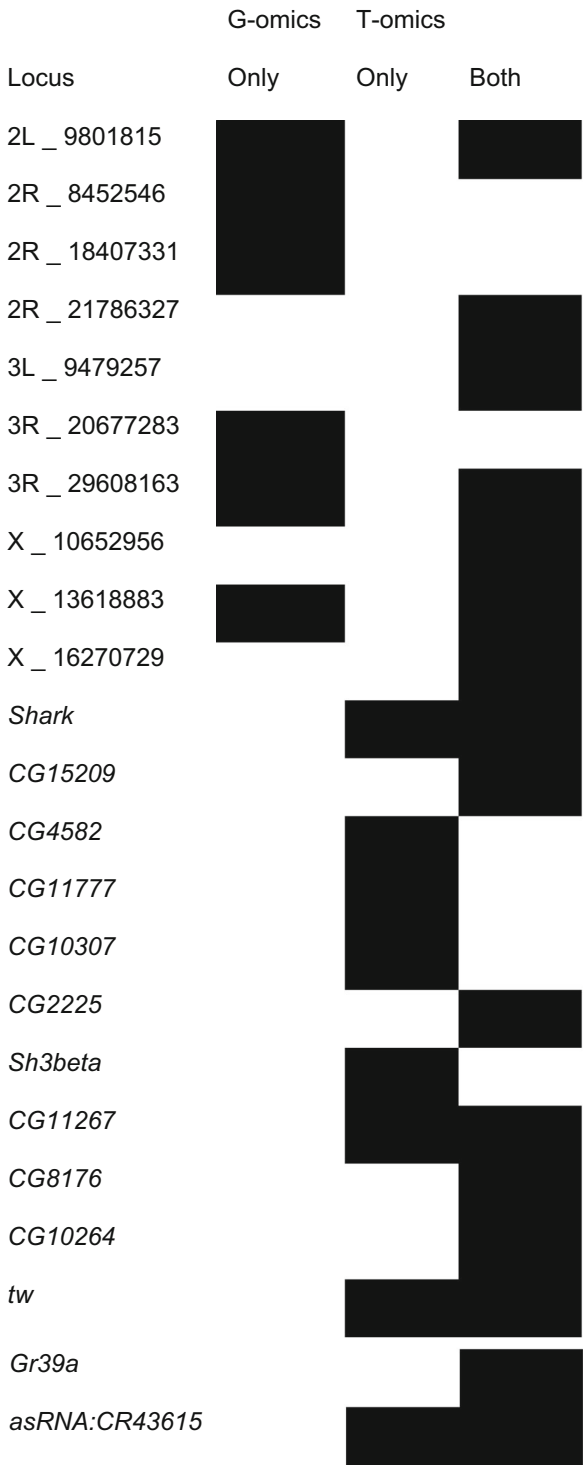

Figure 4. Candidate predictors in predicting age specific mortality at day 21 from genomic SNP data, transcriptomic data, or a combined dataset of both. In all cases, specific elements selected and used in predictive models by the FLAM algorithm are labelled with a black box. The first and second column shows what is selected when statistically differentiated SNPs and transcripts, respectively, are used as the input data. In the third column, both types of data were combined into a single, larger dataset where both types of -omic data can be selected in the same predictive model.

Table 1. Comparative performance among genomic and transcriptomic differences between the two population types in predicting adult trait divergence

	Number of predictors*		
Phenotype	SNP differentiated regions	Differentially expressed genes	Both
Mortality at day 14	10	6	6
Mortality at day 21	17	8	16
Fecundity at day 14	4	4	9
Fecundity at day 21	0	2	0

*According to FLAM.

consisted of two populations from treatment A and two populations from treatment C, and this process was repeated five times to ensure each population was represented in a testing set. We tested both age-specific mortality and fecundity data as outcomes and the expression values from predictors at each age as inputs.

With each specific fit, we used expression levels from the test set of populations to predict their phenotypes and compared those predicted phenotype values to the observed phenotypic values. In order to measure the predictive quality of our models, we used the coefficient of determination, R^2, which represents the proportion of variance in phenotypic outcomes that has been explained by the transcriptomic data in the model. Let y_i (i= 1, …, n) be the observed phenotypes in the test set, $\hat{y}$ is the set of predicted results from our models fit to the training set, and $\bar{y}$ is the mean of the y_i, then R^2 is calculated as,

$$R^2(y,\hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_i - \widehat{y_i})^2}{\sum_{i=1}^{n}(y_i - \overline{y_i})^2}.$$

In more simple terms, R^2 is 1 minus the residual sum of squares divided by the total sum of squares. R^2 was calculated between predictions derived from FLAM and the actual phenotypic data (figure 5). Since the predictions are based on an independent data set it is possible for the predictions, the $\hat{y}$, to fit worse than the mean $\bar{y}$. This possibility can then result in negative values for R^2 which we occasionally see.

We find that the transcriptomic data predicted mortality at all ages quite effectively (average day 14 R^2 = 0.906, day 21 R^2 = 0.898; figure 5). There was no consistent preference for either day of transcriptomic data with respect to predictive accuracy. Models derived from day 14 transcriptomic data had an average R^2 of 0.91 with a 95% confidence interval of 0.86–0.96. Models derived from day 21 transcriptomic data had an average R^2 of 0.90 with a 95% confidence interval of 0.85–0.94. Additionally, we compared predictions made

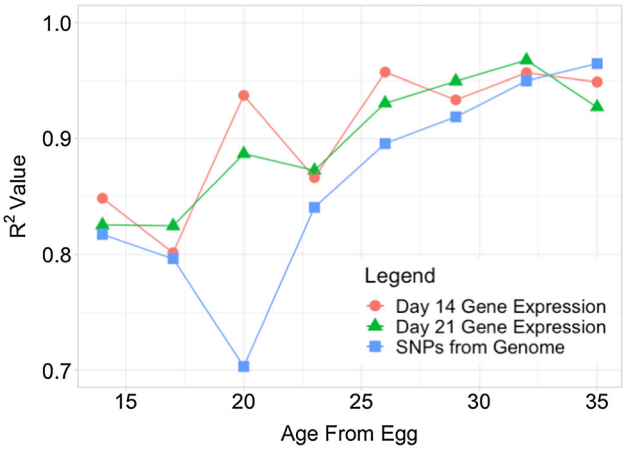

Figure 5. R^2 values between predicted and observed mortality phenotypic values at all age. We compare how well transcriptomic and genetic predictors can predict phenotypes. The values shown correspond to the comparison between the actual phenotypic values and the predicted phenotypic values from our test sets. Statistically differentiated SNPs, and transcriptomics data from day 14 and day 21 were used as predictors while age-specific mortalities were used as outcomes.

from transcriptomic data to predictions made from genotypic data. Accuracy for predictions from genotypic data might appear consistently lower than those from transcriptomic data, but its average R^2 of 0.86 with a 95% confidence interval of 0.78–0.93 for the overall period corroborates the earlier claim from the previous section that both the genome and the transcriptome have comparable predictive power.

The patterns of a strong phenotypic prediction for age-specific mortality do not hold true for age-specific fecundity (figure 2 in electronic supplementary material). FLAM could only create a consistently decent fit from both sets of gene expression data for days 14 and 35. As there is no differentiation between the A and C populations' age-specific fecundity values for days 17–25 (figure 2, Burke *et al.* 2016), it would make sense that FLAM was unable to create accurate predictive models for that time period. Meanwhile, fecundity at days 14–16 and days 35–37 show highly significant differences (p=2.03 × 10^{-5}, and 1.11 × 10^{-5} respectively, Burke *et al.* 2016). Correspondingly, the only days that models generated from both day 14 and day 21 transcriptomic data had any amount of moderate success were day 14, with an R^2 of 0.44 and 0.52 respectively, and day 35, with an R^2 of 0.39 and 0.77 respectively. Even the best days for predicting age-specific fecundity fall short of the precision of age-specific mortality predictions.

However, the difference between the accuracy of the mortality and fecundity results is not simply a function of the levels of differentiation between mortality and fecundity. Phenotypically, both mortality and fecundity have points of extreme differentiation. FLAM can very accurately predict mortality at all ages with both days of transcript data. On the other side, FLAM could not accurately predict fecundity across all the time points of differentiation, and even when it could provide accurate models, the accuracy was far lower than that of the models predicting age-specific mortality. Overall, the fecundity result serves as a negative control, demonstrating that FLAM's use of test data to quantify predictive ability will not result in overfitting.

Unbiased approach to predicting transcriptomic expression using genomic features

Transcription is shaped by both local regulatory sequences (*cis*-) and distantly encoded regulatory factors (*trans*-) (Wray *et al.* 2003). Although we do not focus on detailed mechanisms of gene regulation here, we used different types of genomic features (SNPs, TEs, and structural variants) as predictors and each individual transcript's expression phenotype as outcomes. By doing this, we tried to assess how well those genomic features performed as predictors for the expression of each differentiated transcript. First, we included each type of genomic feature (e.g. SNPs) separately to characterize their effects on the transcriptome. Subsequently, we combined all genomic features to determine which characters were the strongest determinants of mRNA abundance.

In our first analysis, we uncovered numerous interactions between the genomic variants and expression levels (figure 6; figures 3–9 in electronic supplementary material. These interactions often involved predictor loci located on chromosomes different from that in which the differentially expressed transcript resided. The average number of causal candidate predictors varied depending on the time point (day 14 and day 21) and on the type of genetic variant (table 1 in electronic supplementary material).

Our results show evidence of extensive pleiotropy, in which single differentiated genomic regions are reliably contributing to the prediction of expression among numerous genes (day 14: mean (μ) = 12.39, standard deviation (σ) = 8.57; day 21: μ = 14.87, σ = 12.42). Conversely, we see that the differentiated expression of a single gene is predicted by numerous differentiated genomic loci, including many that are located on different chromosomes (day 14: μ = 4.47, σ = 2.68; day 21: μ = 4.56, σ = 2.85). Although we see many interactions between the genome and the transcriptome, all these interactions should not be considered 'regulatory' *per se* due to the fact that the genomic differentiated regions between population types are 50-kb long, and the mechanistic details of how these regions are potentially affecting gene expression remain unknown. In the next section, we also show that gene regions that affect only one phenotype may appear to be pleiotropic and thus our count of pleiotropic regions may include false positives.

We examined the location of each differentially expressed gene in relation to their genomic predictors to determine whether or not these predictors could potentially exert their effects in *cis*. We distinguish between putative *cis*-local (<25 kb from centre of predictor genomic region) and

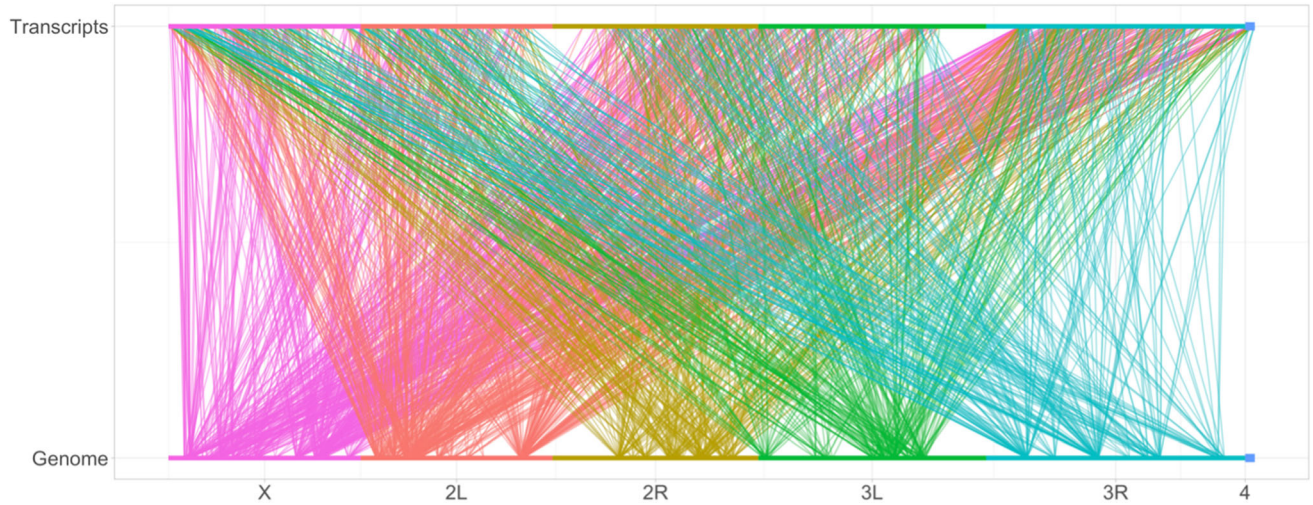

Figure 6. Interactions between predictive SNP regions and differentiated transcripts at day 21 across the *D. melanogaster* chromosomes. Each line maps a predictive SNP region to a transcript for which the SNP region accurately predicts expression. The colour of the line denotes from which chromosome the genomic region originates. The genomic regions are classified as 50 kb windows which contained at least three differentiated SNPs. The transcripts are those classified as significantly differentiated for quantitative expression.

cis-long range (≥25 but <150 kb from centre of predictor genomic region) effects, as the latter are more common than previously thought (Ghavi-Helm *et al.* 2014). Each differentially expressed gene was then checked for any overlap with *cis*-local and *cis*-long range genomic differentiated regions found to have predictive power. The process was repeated for TEs and for duplications. We found a very limited number of both predictor genomic regions that fell within the *cis*-locale of each differentially expressed gene and predictor genomic regions compatible with long range *cis*-effects, although the latter were more prevalent than the former (table 2). This supports the notion that *trans*-effects are more prevalent than *cis*-effects.

In our second analysis, we combined 194 differentiated SNP regions, 71 differentiated TE insertions, 323 differentiated small indels, and 69 differentiated duplicated regions into a single set of 657 total potential predictors for differential expression (figure 7). Across 624 genes with differentiated expression, we found 974 candidate SNPs, 1803 candidate indels, 1011 candidate TEs, and 114 candidate duplications (figure 7). Importantly, these numbers include the fact that a singular feature could be incorporated in more than one gene expression prediction model. As an example, one TE insertion was used in 71 different gene expression models.

Table 2. Number of genomic candidate causal predictors for differential gene expression located in *cis* at different distances.

	SNP windows		TEs		Duplications	
Day	Day 14	Day 21	Day 14	Day 21	Day 14	Day 21
Cis-local (<25 kb)	0	0	2	1	2	5
Cis-long range (<150 kb)	11	5	7	5	5	12

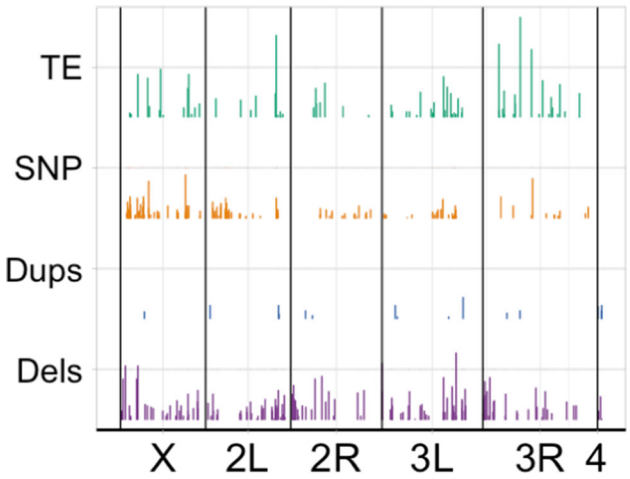

Figure 7. Chromosome distribution of candidate genomic features when all the different individual datasets are combined into one singular dataset. Each bar shows the position of a candidate element, and the height represents how many different predictive models it was involved in. From top to bottom, each panel shows predictive candidate TEs, SNPs, duplications and deletions.

One way to express the relative importance of each genomic feature type would be to compare the proportions of the total number of causal candidates with a null hypothesis ratio of all proportions being equal to their starting input ratio from the full dataset. Small indels (323/657 vs 1803/3902 test for equality of proportions, $p = 0.173$) did not differ from the null hypothesis proportions. SNPs (194/657 vs 974/3902 test for equality of proportions, $p = 0.015$) were slightly underrepresented while duplications (69/657 vs 114/3902 test for equality of proportions,

$p < 2e\text{-}16$) were moderately underrepresented. Consequentially, TE insertions (71/657 vs 1011/3902 test for equality of proportions, $p < 2e\text{-}16$) were overrepresented by this test.

All four classes of data found their way into predictive models for transcriptomic expression. While there is some suggestion of a greater effect size of a random individual transposable element vs a random element of a different data type, the overall contribution of each data type as a whole remains consistent between small indels, SNPs, and TEs. Ultimately, this indicates a greater utility for these three genomic data types as predictors as opposed to the value of duplications.

Comparing the accuracy of predicting transcriptomic expression across four classes of genomic data

From each of the previous analyses, we used each predictor list to determine the accuracy of our FLAM analyses by compiling training and test sets for each set of individual transcriptomic expression prediction from genetic data similar to how the training and testing sets were generated for the predictive power of the transcriptome. We found that all datasets have similar average R^2 values between the actual expression values and the predicted values from the test sets (figure 8; figures 10–16 in electronic supplementary material). The SNP data gave averages of 0.482 and 0.423 R^2 values between the true expression values and the predicted expression for day 14 and day 21, respectively. The transposable element data gave averages of 0.419 and 0.4044 R^2 values between the true expression values and the predicted expression for day 14 and day 21, respectively. The insertion and deletion data gave averages of 0.478 and 0.517 R^2 values for day 14 and day 21, respectively. The duplication data gave averages of 0.343 and 0.336 R^2 values for day 14 and day 21, respectively. Lastly, the combined dataset of the four feature types gave an average of 0.505 R^2 values for day 21.

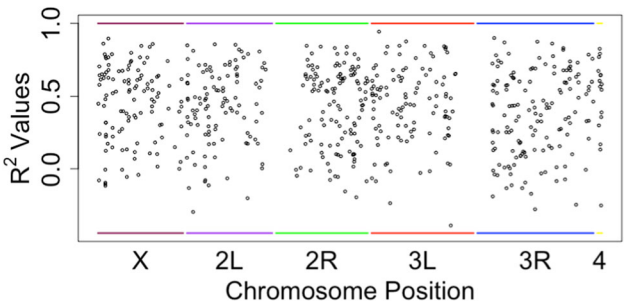

Figure 8. R^2 between predicted value and actual value for each differentially expressed gene. R^2 values were calculated using the predicted expression of each differentiated gene from SNP frequencies and the actual expression values of each differentiated gene at day 21. Each R^2 value was plotted at the location of the differentiated gene above. High positive correlation signifies the FLAM model accurately predicted the actual phenotypic value.

Another way to interpret these results is to look at the numbers of the most accurately predicted transcript expression rather than a simple averaging (figure 9; figure 17 in electronic supplementary material). We have chosen thresholds for R^2 of 0.6 and 0.8 to demonstrate the relative utility of the four genomic classes in our dataset in predicting transcriptomic expression. While somewhat arbitrary, this nonetheless does create a useful segmentation investigating both less stringent and more stringent possibilities. The results are consistent with the above approach. Duplications show very little utility as predictors, having only a couple dozen genes predicted at the lower threshold, and extremely few at the more stringent one. Both SNPs and TEs see a moderate amount of success as predictors, while indels stand out as having the greatest number of well predicted genes. Notably, predictions of gene expression data from day 21 (figure 9) show a greater separation between indels and SNP/TEs than predictions of gene expression data from day 14 (figure 17 in electronic supplementary material) which show those three classes in a tighter grouping.

In contrast to the raw number of TEs selected as useful predictors for gene expression, here the accuracy measurement tells a slightly different story. TEs had some of the lowest accuracy prediction scores, while small insertions and deletions had some of the highest. Even more interestingly, when all genomic features were compared together the average R^2 value was consistent with the R^2 value for indels on the same day. This might also suggest some kind of saturation effect either in the raw number and type of predictors or in the amount of replication in this study.

While different in their nuances, comparing the different accuracy measurements convey the same qualitative result as looking at the raw numbers of predictors chosen; TEs and small insertions/deletions could be of specific interest in future work due to their relatively high inclusion rates in predictive models as well as their accuracy measurements. At the same time, other types of genomic data cannot necessarily be ruled out.

Testing FLAM's ability to detect pleiotropy

Mueller *et al.* (2018) did not examine the ability of FLAM to specifically detect pleiotropy. To evaluate this possibility, we performed a set of simulations, building upon our earlier work but this time incorporating pleiotropic effects of a single gene region on two phenotypes. The purpose of this new set of simulations was to evaluate the frequency with which noncausal gene regions would be included as pleiotropic as well as determining how well FLAM detects pleiotropy.

Simulated datasets were configured using the same approach as that of Mueller *et al.* (2018). Basically, the approach utilized populations that show major genetic and phenotypic differences mimicking populations that had undergone adaptation to different environments. The

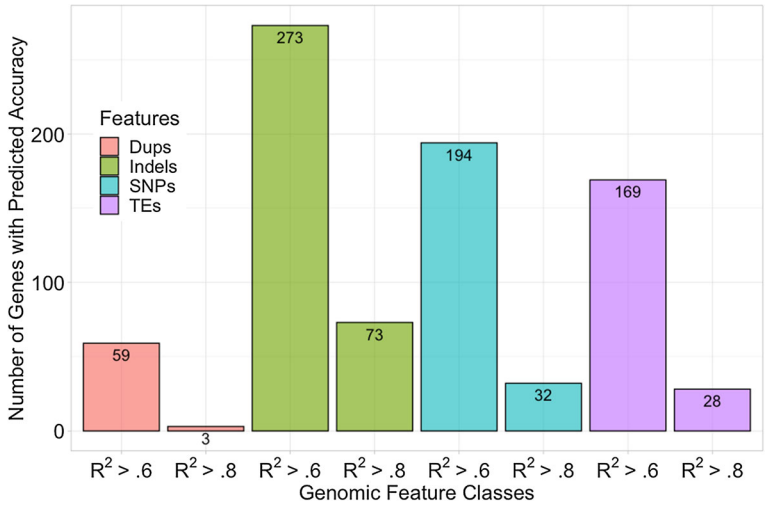

Figure 9. Comparing the number of accurate transcriptomic predictions for each class of genomic data from models constructed for each of the 625 differentiated genes identified on day 21. The height of the bar represents how many different transcripts passed a given accuracy threshold for a single classification.

simulations reported here consisted of 20 populations, with 10 having a low phenotypic value and 10 having a high phenotypic value. Within the replicated high and low populations there were low levels of genetic variation (13 SNPs). Two phenotypes were modelled. Phenotype-1 was determined by SNP-1 and SNP-2 while phenotype-2 was determined by SNP-1 and SNP-3. SNPs 4–13 were differentiated but did not affect either phenotype. Additionally, there were 1987 neutral gene regions which showed only random variation between populations (Mueller *et al.* 2018). If we let a single population's SNP frequencies at SNPs 1–3 be x_1, x_2, and x_3 respectively then phenotype-1 was equal to $(x_1+\ x_2)/2+\varepsilon$, and phenotype-2 was equal to $exp[(x_1+x_3-2)/(x_1+x_3)]+\varepsilon$, where ε is a random number chosen from a normal distribution with mean zero and standard deviation 0.005.

A database consisted of these two phenotypes for each population and estimated SNP frequencies, $\hat{x}_1, \hat{x}_2, \hat{x}_3$, which reflected sampling error from a standard pooled sequencing experiment (Mueller *et al.* 2018). Our results are based on 100 independent databases. The patterns of phenotype vs SNP frequency change in a single database using the two causal and two randomly chosen noncausal SNPs show extensive overlap (figure 10).

We compared FLAM to linear regression models. We tested all possible linear models with combinations of up to 13 SNPs. The best model was chosen by comparing the mean squared error from test data created from five-fold division of the 20 observations. FLAM correctly identified both causal loci between 71% (phenotype 1) and 88% (phenotype 2) of the time. The best linear model correctly identified both causal loci only half as often (35% and 44%). The application of FLAM following our 50% rule (Mueller *et al.* 2018) shows that 80% of the time the single pleiotropic SNP is correctly identified and 64% of the databases

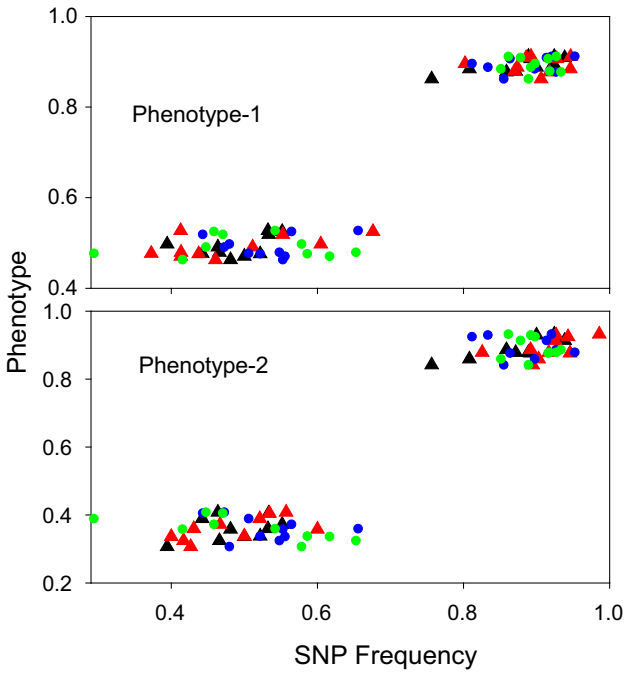

Figure 10. Comparison of simulated differentiated SNPs. A randomly chosen database (out of 100) with a total of 20 populations. The triangles are the causal SNPs. The pleiotropic SNP has black fill. The coloured circles are from two out of 10 noncausal SNPs chosen at random. Each colour identifies a single SNP from each of the 20 simulated populations.

analysed incorrectly included a causal, nonpleiotropic SNP (SNP-2, and SNP-3) as affecting both phenotypes (figure 11). The best linear model correctly identified the pleiotropic locus only 51% of the time. The nonpleiotropic causal SNP was incorrectly identified as pleiotropic in 24% of the databases. A differentiated, noncausal SNP (SNPs 4–13) had only a 20% chance of being identified as a

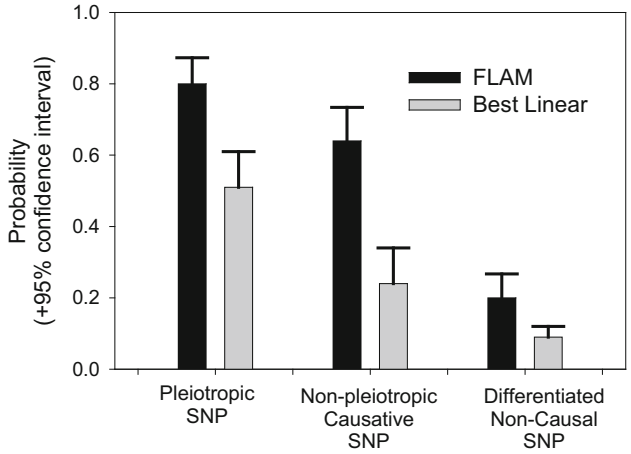

Figure 11. Frequency of identification by FLAM and the best linear model for different SNP categories. Probability of identification of SNPs as pleiotropic based on 100 randomly generated databases, each with 20 populations, one pleiotropic SNP, two nonpleiotropic causative SNPs, and 10 differentiated noncausal SNPs.

pleiotropic SNP (figure 11) with FLAM and a 9% chance with the best linear model. No neutral SNPs were classified as pleiotropic.

These results (figure 11) suggest FLAM is likely to find real pleiotropic genes but may also erroneously include genes which affect only one phenotype and to a much lesser degree may occasionally include genes which have no effect on either phenotype. Thus, the list of pleiotropic gene regions returned can be expected to have false positives.

Comparing FLAM to linear models in phenotypic predictive accuracy

Given the strong phenotypic differentiation (figure 10), it would seem easy to predict phenotypes from genetic data with the standard tools of linear regression. We next evaluated, data from the present study, whether FLAM is superior to standard linear regression methods in its predictive power. We repeated the stratified 5-fold cross validation methodology from the analysis on later-age phenotypes once again to accurately test the predictive capability between models. We generated a predictive model for FLAM and linear models from a subset of the data with 16 populations, called training data, and then predicted phenotypes on the remaining four populations, called test data. This was repeated for nonoverlapping sets of test data. We then calculated the R^2 value between predicted and observed phenotypes.

We used FLAM and linear models to predict mortality at day 14 and day 21. The number of genes selected for the linear models was the same number of genes FLAM had selected on those focal days: seven genes for day 14 mortality and eight for day 21 mortality. We did not use the linear models to choose genes. Instead, we used two different sampling methods, (i) random genes—100 random samples of seven genes out of 539 differentiated genes were chosen to predict day 14 mortality and 100 random samples of eight genes out of N differentiated genes were chosen to predict day 21 mortality, (ii) lowest p-value—only the seven most differentiated genes (as judged by p-value from the CMH test) were used. With these predictor genes we used the R function regsubsets (leaps package) to examine all possible subsets of these genes to find the best model. The best subset incorporated three genes and was chosen to fit the corresponding linear model. The R^2 value for the linear models involving a selection from random genes was averaged from all R^2 values across all 100 random samples.

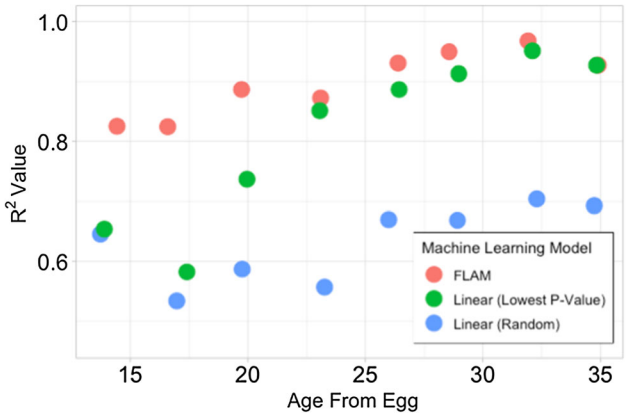

Figure 12. R^2 values between predicted and observed mortality based on models fit to transcriptomic data from age 21.

FLAM had the largest R^2 when predicting all eight ages on day 21 data and the largest R^2 on seven of the eight ages on day 14 data (figure 12; figure 18 in electronic supplementary material). The difference between FLAM and the other linear predictors was greatest at the focal ages of 14 and 21. Also notably, when using the day 21 transcripts with the lowest p-value as predictors of age-specific mortality, the linear model performed well at later ages, but poorly at the earlier ones while FLAM performed consistently across all ages. The strong differentiation of the C and A type populations can explain the high R^2 across all models and days, but FLAM, as a nonlinear model, manages to better capture the patterns of the dataset than even the best linear models.

Discussion

Having two clearly defined sets each of 10 experimentally evolved *D. melanogaster* populations in conjunction with a full suite of genomic, transcriptomic, and phenotypic data for both sets of populations has enabled us to piece together how these three levels of biological machinery interact with one another. Specifically, the 10 A-type populations are clearly differentiated from the 10 C-type populations across the genome (Graves *et al.* 2017), across the transcriptome (Barter *et al.* 2019), and in age-specific mortality and fecundity (Burke *et al.* 2016). Conversely, there is little

differentiation between the populations within a single set of 10 populations for all three sets of data. Lastly, these two sets of populations are closely related, despite their marked differentiation at all three levels, genomic, transcriptomic, and phenotypic.

When examining the predictive power of transcriptomic variation in explaining phenotypic differentiation, we found that the transcriptomic data at day 21 accurately predict mortality for all ages after day 21, which is consistent with a relatively stable transcriptome once the individuals of a given population type have reached sexual maturity. Since the transcriptome data did not result in accurate predictions of fecundity for most ages, we were unable to make any inferences about the relative stability of transcriptome effects on fecundity at later ages. This may have been due to the fact that fecundity is not differentiated between the two treatments at all life stages, or that the environmental variation is too large relative to the genetic differentiation. As it stands, transcriptomic data may only accurately predict phenotypes at ages for which the phenotype is differentiated.

Further, when comparing the genome and the transcriptome in predicting phenotypic outcomes, we find that there is no clear evidence concerning which type of -omic data is preferentially selected when searching for predictors, as both kinds of data are used by the machine learning algorithm to predict phenotypic outcomes as well as outputting similar levels of accuracy in said predictions. This suggests that future studies should consider adding information from differential expression surveys to increase the analytical power of molecular differentiation for predicting phenotypic outcomes.

We investigated the physical patterns among differentiated genomic regions between populations that were good predictors for differential expression and the actual differentially expressed genes. Although these regions may have predicted the expression of a differentiated gene, they should not be considered necessarily specific regulators for the gene. Some of these differentiated genomic regions may contribute to the regulation of gene indirectly through their effects on some of the other differentially expressed genes. It is specifically uncertain whether these regions have evolved solely for their effects on transcript regulation. In addition, the location of each predictive genomic region was not restricted to the *cis*-locale of the differentially expressed genes. In fact, almost all predictive genomic regions were not found in the *cis*-locale of those genes, with a more prevalent role of long-range *cis* effects, which undermines the hypothesis that *cis*-locale evolution is a primary driver of adaptation (Carroll *et al.* 2001; Shapiro *et al.* 2004). Our results instead support the view that transcriptome differentiation during adaptation is affected by many sites across the entirety of the genome, highlighting the importance of *trans*-effects at the intraspecific level, which concurs with previous findings obtained with other approaches such as eQTL mapping and GWAS (Hill *et al.* 2021).

We used different genomic features (SNPs, TEs, small duplications, short insertions/deletions) to determine whether there was a particular genomic feature with enhanced predictive power in relation to differential gene expression. When examined individually, SNPs, TEs, and indels all maintained higher accuracy metrics both on average R^2 and when comparing the number of genes that class highly predicts. On the flipside, duplications performed poorly, having the lowest average R^2 and having only a few strongly predicted genes. When all genomic features were used simultaneously, the machine learning models chose all types of genomic features. However, the candidate frequency of duplications showed a depression in its relative frequency corresponding with an increase in the frequency of TE candidacies. In the case of TEs in particular, our finding does support their importance as a driving force in adaption (Stapley *et al.* 2015; Van't Hof *et al*. 2016), but we would additionally like to stress the general importance of SNPs and indels as useful genomic features in these predictive models.

This lack of predictive power from small duplications is unlikely to be an artifact of the algorithm and models given the low amount of clustering between duplications and other genomics features. Consequentially, this means that the genomic regions that consistently have the highest predictive power across all transcripts do not contain small duplications.

In combining genomic, transcriptomic, and phenotypic data, we were able to investigate whether these three biological levels generally follow a simplistic one-to-one, polygenic many-to-one, or network many-to-many pattern of connectivity (vid Wright 1980). The best predictors for each phenotype, whether genomic or transcriptomic, span the genome. This, at first glance, supports the idea of polygenic functional variation, in that numerous loci affected most phenotypes. Notably, differentiated genomic regions predicted the expression levels of multiple transcripts, while conversely each transcript's expression level was predicted by numerous genomic regions. The complexity of the interactions between the genomic and transcriptomic levels (figures 6 & 8) substantiates the need for machine learning tools to parse the molecular foundation of adaptation. Such complex patterns are not amenable to detection by the otherwise unaided human mind.

Contrary to the predictions of Fisher (1930) and Orr (2000), recent work has suggested that moderate levels of pleiotropy are common (Wagner *et al.* 2008; Frachon *et al.* 2017; Hämälä *et al.* 2020; Rennison and Peichel 2021) and higher levels of pleiotropy are associated with increased per-trait effect size (Wagner *et al.* 2008; Wang *et al.* 2010). Hämälä *et al.* (2020) suggest pleiotropy will be common when a population is far from its adaptive optimum as are the experimental populations in this study. The present paper likewise supports the hypothesis that pleiotropy may be an important component of adaptation even in complex organisms.

Different forms of machine learning have already been used to study how the genome responds to selection. Tools

varying from hidden Markov models (Kern and Haussler 2010), soft/hard inference through classification or S/HIC (Schrider and Kern 2016), and deep learning (Sheehan and Song 2016) have all been used to infer the effects of selection at a genomewide scale. These models primarily focus on discovering hard sweeps and soft sweeps throughout the genome, utilizing genomic data.

Here we have shown that FLAM allows us to determine which differentiated genomic regions or differentially expressed genes between two population types are the best predictors of specific patterns of phenotypic differentiation, whether or not selective sweeps have occurred. FLAM machine learning also allowed us to address long-standing and unanswered questions about the molecular foundations of functional adaptation. In addition, FLAM has enabled us to treat the transcriptome as a phenotype of the genome, and thereby locate some of the regions of the genome that influence the levels of each differentiated transcript. In other words, FLAM can explore the molecular basis of adaptation during replicated experimental evolution, rather than just locating regions of hard or soft sweeps.

Currently, we only have the full suite of genomics, transcriptomics, and phenotypic data for 20 populations. As shown previously (Mueller *et al.* 2018), 20-population analysis is barely sufficient for detecting causal loci, and by no means will detect the full range of causally important sites in genomes undergoing adaptation. Ideally, the number of populations used in analyses of this kind should approach 100. Only at such high levels of replication is it plausible that this experimental strategy will reveal a high proportion of the genomic sites that are involved in the response to selection. Although having the full suite of all three types of data is ideal, just having genomic and phenotypic data in additional populations would result in a drastic increase in our power to detect causal loci. With the addition of more experimentally evolved groups of populations, we can approach the level of 100 populations, at which point thorough penetration of the genomic complexity of adaptation should be achievable.

Acknowledgements

MAP was supported by an NSF Postdoctoral Fellowship (NSF 190624).

Authors' contributions

TTB collected and analysed the genomic data, MRR, TTB and JMR conceived the research, MAP and ZSG helped with the genomic analysis, TTB wrote the first draft and MRR, JMR, MAP, ZSG and LDM edited the manuscript, LDM and ZSG did the FLAM analysis. ZSG and TTB contributed equally to the project.

References

Anders S., McCarthy D. J., Chen Y., Okoniewski M., Smyth G. K., Huber W. and Robinson M. D. 2013 Count-based differential expression analysis of RNA sequencing data using R and Bioconductor. *Nat. Protoc.* **8**, 1765–1786.

Arbeitman M. N., Furlong E. E., Imam F., Johnson E., Null B. H., Baker B. S. *et al*. 2002 Gene expression during the life cycle of *Drosophila melanogaster*. *Science* **297**, 2270–2275.

Barter T. T., Greenspan Z. S., Phillips M. A., Mueller L. D., Rose M. R. and Ranz J. R. 2019 Transcriptomics with and without ageing in *Drosophila*. *Biogerontology* **20**, 699–710.

Benjamini Y. and Hochberg Y. 1995 Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B Methods* **57**, 289–300.

Bennett A. F. and Lenski R. E. 1999 Experimental evolution and its role in evolutionary physiology. *Am. Zool.* **39**, 346–362.

Braendle C., Heyland A. and Flatt T. 2011 Integrating mechanistic and evolutionary analysis of life history variation. In *Mechanisms of life history evolution: the genetics and physiology of life history traits and trade-offs* (ed. T. Flatt and A. Heyland), pp. 1–10. Oxford University Press, New York.

Brideau N. J., Flores H. A., Wang J., Maheshwari S., Wang X. and Barbas D. A. 2006 Two Dobzhanky-Mueller genes interact to cause hybrid lethality in *Drosophila*. *Science* **314**, 1292–1295.

Burke M. K., Barter T. T., Cabral L. G., Kezos J. N., Phillips M. A., Rutledge G. A. *et al*. 2016 Rapid divergence and convergence of life-history in experimentally evolved *Drosophila melanogaster*. *Evolution* **70**, 2085–2098.

Carroll S. B. 2000 Endless forms: the evolution of gene regulation and morphological diversity. *Cell* **101**, 577–580.

Carroll S. B., Grenier J. K., Weatherbee S. D. 2001 *From DNA to diversity: molecular genetics and the evolution of animal design*, Blackwell Publishing, Malden.

Casacuberta E. and Gonzalez J. 2013 The impact of transposable elements in environmental adaptation. *Mol. Ecol.* **22**, 1503–1517.

Casas-Vila N., Bluhm A., Sayols S., Dinges N., Dejung M., Altenhein T. *et al*. 2017 The developmental proteome of *Drosophila melanogaster*. *Genome Res.* **27**, 1273–1285.

Chénais B., Caruso A., Hiard S. and Casse H. 2012 The impact of transposable elements on eukaryotic genomes: From genome size increase to genetic adaptation to stressful environments. *Gene* **509**, 7–15.

de Los Campos G., Hickey J. M., Pong-Wong R., Daetwyler H. D. and Calus M. P. 2013 Whole-genome regression and prediction methods applied to plant and animal breeding. *Genetics* **193**, 327–345.

Fabian D. K., Dönertas H. M., Fuenyealba M., Partridge L. and Thornton J. M. 2021 Transposable element landscape in *Drosophila* populations selected for longevity. *Genome Biol. Evol.*, https://doi.org/10.1093/gbe/evab031.

Fisher R. A. 1930 *The genetical theory of natural selection*, Oxford University Press, Oxford.

Frachon L., Libourel C., Villoutreix R., Carrère S., Glorieux C., Huard-Chauveau C. *et al*. 2017 Intermediate degrees of synergistic pleiotropy drive adaptive evolution in ecological time. *Nat. Ecol. Evol.* **1**, 1551–1561.

Garland T., Rose M. R. 2009 *Experimental evolution*, University of California Press, Berkeley.

Ghavi-Helm Y., Klein F. A., Pakozdi T., Ciglar L., Noordermeer D., Huber W. and Furlong E. E. M. 2014 Enhancer loops appear stable during development and are associated with paused polymerase. *Nature* **512**, 96–100.

Graves J. L., Hertweck K. L., Phillips M. A., Han M. V., Cabral L. G., Barter T. T. *et al*. 2017 Genomics of parallel experimental evolution in *Drosophila*. *Mol. Biol. Evol.* **34**, 831–842.

Hämälä T., Gorton A. J., Moeller D. A. and Tiffin P. 2020 Pleiotropy facilitates local adaptation to distant optima in common ragweed (*Ambrosia artemisiifolia*). *PLOS Genet*, https://doi.org/10.1371/journal.pgen.1008707.

Hill M. S., Vande Zande P. and Wittkopp P. J. 2021 Molecular and evolutionary processes generating variation in gene expression. *Nat. Rev. Genet.* **22**, 203–215.
Hoekstra H. E. and Coyne J. 2007 The locus of evolution: evo devo and the genetics of adaptation. *Evolution* **61**, 995–1016.
Hsu S.-K., Belmouaden C., Nolte V. and Schlotterer C. 2020 Parallel gene expression evolution in natural and laboratory evolved populations. *Mol. Ecol.* **30**, 884–894.
Kelly J. K. and Hughes K. A. 2018 Pervasive linked selection and intermediate-frequency alleles are implicated in an evolve-and-resequencing experiment of *Drosophila simulans*. *Genetics* **211**, 943–961.
Kern A. D. and Haussler D. 2010 A population genetic hidden Markov model for detecting genomic regions under selection. *Mol. Biol. Evol.* **27**, 1673–1685.
Kezos J. N., Phillips M. A., Thomas M. D., Ewunkem A. J., Rutledge G. A., Barter T. T. *et al*. 2019 Genomic and phenotypic effects of selection for starvation resistance in *Drosophila*. *Physiol. Biochem. Zool.* **92**, 591–611.
Krizhevsky A., Sutskever I. and Hinton G. E. 2012 ImageNet classification with deep convolutional neural networks. *Adv. Neural Information Proc. Syst.* **25**, 1097–1105.
Li H., Handsaker B., Wysoker A., Fennel T., Ruan J., Homer N. *et al*. 2009 Genome project data processing S. (2009) The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079.
Love M. I., Huber W. and Anders S. 2014 Moderate estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol.* **15**, 550.
Mallard F., Nolte V., Tobler R., Kapun M. and Schlotterer C. 2018 A simple genetic basis of adaptation to a novel thermal environment results in complex metabolic rewiring in Drosophila. *Genome Biol.* **19**, 119.
McClintock B. 1950 The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. USA* **36**, 344–355.
Mueller L. D., Phillips M. A., Barter T. T., Greenspan Z. S. and Rose M. R. 2018 Genome-wide mapping of gene-phenotype relationships in experimentally evolved populations. *Mol. Biol. Evol.* **35**, 2085–2095.
Orr H. A. 2000 Adaptation and the cost of complexity. *Evolution* **54**, 13–20.
Otto S. P. 2004 Two steps forward, one step back: the pleiotropic effects of favoured alleles. Proceedings: *Biol*. *Sci.* **271**, 705–714.
Petersen A., Witten D. and Simon N. 2016 Fused lasso additive model. *J. Comput. Graph. Stat.* **25**, 1005–1025.
Phillips M. A., Rutledge G. A., Kezos J. N., Greenspan Z. S., Talbott A., Matty S. *et al*. 2018 Effects of evolutionary history on genome wide and phenotypic convergence in *Drosophila* populations. *BMC Genomics* **19**, 743–759.
R Core Team 2018 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (available at https://www.R-project.org/).
Remolina S. C., Chang P. L., Leips J., Nuzhdin S. V. and Hughes K. A. 2012 Genomic basis of aging and life-history evolution in *Drosophila melanogaster*. *Evolution* **66**, 3390–3403.
Rennison D. J. and Peichel C. L. 2021 Pleiotropy facilitates parallel adaptation in sticklebacks. *Mol. Ecol.* **31**, 1476–1486.
Rose M. R., Passananti H. B. and Matos M. (eds.) 2004 *Methuselah flies: a case study in the evolution of aging*, World Scientific, Singapore.
Schlotterer C., Kofler R., Versace E., Tobler R. and Franssen S. U. 2015 Combining experimental evolution with next-generation sequencing: a powerful tool to study adaptation from standing genetic variation. *Heredity* **114**, 431–440.
Schrider D. R. and Kern A. D. 2016 S/HIC: robust identification of soft and hard sweeps using machine learning. *PLoS Genet*, https://doi.org/10.1371/journal.pgen.1005928.
Schrider D. R. and Kern A. D. 2018 Supervised machine learning for population genetics: A new paradigm. *Trends Genet.* **34**, 301–312.
Sebastiani F. 2002 Machine learning in automated texted categorization. *ACM Comput. Surveys* **34**, 1–47.
Shapiro M. D., Marks M. E., Peichel C. L., Blackman B. K., Nereng K. S., Jonsson B. *et al*. 2004 Genetic and developmental basis of evolutionary pelvic reduction in three spine sticklebacks. *Nature* **428**, 717–723.
Sheehan S. and Song Y. S. 2016 Deep learning for population genetic inference. *PLoS Comput. Biol.* **12**, e1004845.
Smith J. M., Bozcuk A. N. and Tebbutt S. 1970 Protein turnover in adult *Drosophila*. *J. Insect Physiol.* **16**, 601–613.
Stapley J., Santure A. W. and Dennis S. R. 2015 Transposable elements as agents of rapid adaptation may explain the genetic paradox of invasive species. *Mol. Ecol.* **24**, 2241–2252.
Taus T., Futschik A. and Schlotterer C. 2017 Quantifying selection with pool-seq time series data. *Mol. Biol. Evol.* **34**, 3023–3034.
Topa H., Jonas A., Kofler R., Kosiol C. and Honkela A. 2015 Gaussian process test for high-throughput sequencing time series: application to experimental evolution. *Bioinformatics* **31**, 1672–1670.
Turner T. L., Stewart A. D., Fields A. T., Rice W. R. and Tarone A. M. 2011 Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet.* **7**, e1001336.
Van't Hof A. E., Campagne P., Rigden D. J., Yung C. J., Lingley J. and Quail M. A. 2016 The industrial melanism mutation in British peppered moths is a transposable element. *Nature* **534**, 102–105.
Veiner M., Morimoto J., Leadbeater E. and Manfredini F. 2022 Machine learning models identify gene predictors of waggle dance behavior in honeybees. *Mol. Ecol. Res.*, https://doi.org/10.1111/1755-0998.13611.
Vlachos C., Burny C., Pelizzola M., Borges R., Futschik A., Kofler R. and Schlotterer C. 2019 Benchmarking software tools for detecting and quantifying selection in evolve and resequencing studies. *Genome Biol.* **20**, 169.
Wagner G. P., Kenney-Hunt J. P., Pavlicev M., Peck J. R., Waxman D. and Cheverud J. M. 2008 Pleiotropic scaling of gene effects and the 'cost of complexity.' *Nature* **452**, 470–472.
Wang Z., Liao B.-Y. and Zhang J. 2010 Genomic patterns of pleiotropy and the evolution of complexity. *Proc. Natl. Acad. Sci. USA* **107**, 18034–18039.
Wray G. A., Hahn M. W., Abouheif E., Balhoff J. P., Pizer M., Rockman M. V. and Romano L. A. 2003 The evolution of transcriptional regulation in eukaryotes. *Mol. Biol. Evol.* **20**, 1377–1419.
Wright S. 1980 Genic and organismic selection. *Evolution* **34**, 825–843.

Corresponding editor: Divya Tej Sowpati