A novel computational approach for predicting complex phenotypes in *Drosophila* (starvation-sensitive and sterile) by deriving their gene expression signatures from public data

Dobril K. Ivanov[1,2]*, Gerrit Bostelmann[1], Benoit Lan-Leung[2], Julie Williams[2], Linda Partridge[3,4,¶], Valentina Escott-Price[2,¶], Janet M. Thornton[1,¶]

[1]European Molecular Biology Laboratory, The European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK

[2]UK Dementia Research Institute at Cardiff (UKDRI), Cardiff University, College of Biomedical and Life Sciences, Hadyn Ellis Building, Cardiff, UK

[3]Max Planck Institute for Biology of Ageing, Cologne, Germany

[4]Institute of Healthy Ageing, and Department of Genetics, Evolution and Environment, UCL, London, UK

* Corresponding author

E-mail: IvanovD1@cardiff.ac.uk (DKI)

[¶]These authors are joint senior authors of this work (LP, VEP and JMT are Joint Senior Authors)

Abstract

Many research teams perform numerous genetic, transcriptomic, proteomic and other types of omic experiments to understand molecular, cellular and physiological mechanisms of disease and health. Often (but not always), the results of these experiments are deposited in publicly available repository databases. These data records often include phenotypic characteristics following genetic and environmental perturbations, with the aim of discovering underlying molecular mechanisms leading to the phenotypic responses. A constrained set of phenotypic characteristics is usually recorded and these are mostly hypothesis driven of possible to record within financial or practical constraints.

We present a novel proof-of-principal computational approach for combining publicly available gene-expression data from control/mutant animal experiments that exhibit a particular phenotype, and we use this approach to predict unobserved phenotypic characteristics in new experiments (data derived from EBI's ArrayExpress and ExpressionAtlas respectively).

We utilised available microarray gene-expression data for two phenotypes (starvation-sensitive and sterile) in *Drosophila*. The data were combined using a linear-mixed effects model with the inclusion of consecutive principal components to account for variability between experiments in conjunction with Gene Ontology enrichment analysis. We present how available data can be ranked in accordance to a phenotypic likelihood of exhibiting these two phenotypes using random forest.

The results from our study show that it is possible to integrate seemingly different gene-expression microarray data and predict a potential phenotypic manifestation with a relatively high degree of confidence (>80% AUC). This provides

thus far unexplored opportunities for inferring unknown and unbiased phenotypic characteristics from already performed experiments, in order to identify studies for future analyses. Molecular mechanisms associated with gene and environment perturbations are intrinsically linked and give rise to a variety of phenotypic manifestations. Therefore, unravelling the phenotypic spectrum can help to gain insights into disease mechanisms associated with gene and environmental perturbations. Our approach uses public data that are set to increase in volume, thus providing value for money.

Introduction

Despite the flood of molecular omics data, with a few notable exceptions, such as the Genotype-Tissue Expression (GTEx) project [1], most datasets are rarely re-used, mainly due to challenges with combining the data from different sources. However, in most experimental studies, additional measures are made of biochemical, and physiological changes and of changes in the phenotypic characteristics that they bring about. Phenotypes can include, for instance, morphology, behaviour and pathology. Usually, a limited number of phenotypes are recorded, due to various study constraints. An intermediate phenotype, or sub-phenotype, is one that underlies the study phenotype, but crucially is influenced by fewer genes[2]. For instance, sub-phenotypes of Parkinson's Disease (PD) can include olfactory impairment, gut function disturbance, motor impairments and cognitive decline, each of which may be mediated by subsets of the genes that together result in PD pathology. Quantifying a wide variety of sub-phenotypes associated with animal models of a disease could therefore help to identify causal mechanisms.

The aim of the present study was to develop an *in-silico* approach for inferring unobserved phenotypic characteristics from published gene-expression data resulting from genetic or environmental perturbations. To do this, we generated molecular signatures for two target phenotypes in the fruit fly *Drosophila*, starvation stress response defective (starvation-sensitive) and sterile, using available gene-expression data. Using machine learning, we were able to show that these molecular signatures are able to reliably predict the starvation-sensitive and sterile phenotypic traits solely using expression datasets from studies where these phenotypes were not originally measured, thus adding value to already deposited data.

Materials and Methods

A schematic overview of the generation of a gene-expression molecular signature for a specific phenotype of interest is presented in Fig 1.

Fig 1. Flow diagram of the overall generation of molecular signatures for a phenotype of interest. a) Building the molecular signature and selecting model parameters for a particular phenotype. b) Predicting phenotypic manifestation in unknown experiments utilising the molecular signature.

Data collection

Linking phenotypes to perturbed genes in *Drosophila*

In order to identify perturbed genes that lead to a particular phenotype, we downloaded several datasets from FlyBase (http://flybase.org/). These comprised: allele phenotypic data, synonyms, annotation identifiers, control vocabulary and alleles to gene identifiers. Using in-house custom programs, we parsed and linked all

these identifiers with the phenotypic data. That is, for each FlyBase phenotype, we obtained a list of identifiers (e.g. FlyBase gene numbers, allele symbols, synonyms).

Obtaining expression data from EBI's ArrayExpress

To maximise the number of experiments for each phenotype chosen for this study, we used the Affymetrix GeneChip Drosophila Genome 2.0 Array (EBI's ArrayExpress identifier A-AFFY-35). At the time of conducting the analysis, the largest number of experiments had been performed using the Affymetrix Genome 2.0 microarray platform (number of experiments: 330).

Using the above-mentioned FlyBase identifiers (linking phenotypes to perturbed genes) we searched EBI's ArrayExpress for any potential match using the textual representation of EBI's web resource, i.e. REST-style queries. The identifiers were used as keywords to form a URL and the XML result was parsed using a custom-made Perl program. The nature of the allele constructs for experiments deposited in EBI's ArrayExpress does not follow a specific nomenclature and the authors/depositors are allowed relative freedom in describing the gene constructs. For example, EBI's ArrayExpress identifier E-GEOD-18576 lists a genotype description as a *DHR96* mutant. We did not assume that different allele constructs for the same gene will exhibit the same phenotype. Therefore, for each of the experiments that matched any of the FlyBase identifiers for the two target phenotypes, we manually curated the data first by reading all the accompanying manuscripts and subsequently retained experiments where the same allele construct was used. Furthermore, only experiments with raw gene-expression data (data with available raw cel files) were retained.

Normalised gene-expression values

Raw gene-expression data (cel files) were downloaded from the EBI's ArrayExpress (https://www.ebi.ac.uk/arrayexpress/). An 'experiment' throughout this manuscript was considered to be a set of control/mutant gene-expression microarray assays, submitted to EBI's ArrayExpress under the same identifier and exhibiting the phenotype of interest, unless otherwise specified (see Fig 2). Separately, for each experiment, the raw data were summarised and normalised by using the rma (bioconductor's package *affy* [3]). Log2-normalised expression data for all experiments that exhibited a particular phenotype were combined in a single dataset.

Fig 2. Definition of an experiment exhibiting a phenotype of interest. EBIs ArrayExpress identifier: E-GEOD-24978

Removal of batch effects within an experiment

Individual experiments for the two target phenotypes were examined for the presence of batch effects. For each ArrayExpress accession number, all individual microarray cel files were downloaded, including any microarray assays that did not exhibit the phenotypes in question but were submitted under the same ArrayExpress identifier. For each experiment, we performed principal component analysis (PCA) of the log2-normalised microarray expression data. Where significant batch effects were detected, we used bioconductor's *ber* package [4] to correct for them. For example, if an experiment that exhibited the phenotype of interest had sets of controls/mutants derived from different tissues, and that therefore exhibited significant heterogeneity in pattern of gene expression, the tissue effect was used as a factor in the batch effect correction.

Generation of the molecular signatures (Linear-mixed effects model)

A random intercept linear mixed-effects model (LMEM) was used to generate normalised residuals for each gene within the Affymetrix Genome 2.0 microarray, accounting for a number of consecutive principal components. Fixed and random effects comprised the principal components and the different experiments, respectively, with gene-expression as the dependent variable. The residuals were then used to perform a logistic regression to assess the statistical significance of each gene. For the LMEM, the *lmer* function in R was used. The number and nature of the underlying biological and technical factors that differ between the different experiments are largely unknown. In order to determine how many principal components to use, the molecular signatures for the two target phenotypes were generated using LMEM, including a number of consecutive principal components to account for these biological/technical effects, e.g. sex, tissue. The consecutive principal components used started with using LMEM with no principal components progressing up to a LMEM with the first 7 consecutive principal components included (8 different models).

Gene Ontology (GO) enrichment analysis

The Wilcoxon rank sum test, as implemented in Catmap [5], was used to perform functional analysis to test for significant enrichment of Gene Ontology categories. Ranks of genes were based on the p-value derived from the logistic

regression, irrespective of beta-coefficients. To account for multiple hypotheses testing the Benjamini-Hochberg false discovery rate was used (FDR). To assess if there was a significant enrichment of GO terms associated with the two target phenotypes of interest in the derived molecular signatures, we selected GO terms that we considered representative of the two phenotypes (S1 and S2 Figs).

Leave-one-out cross-validation

To assess how well the molecular signatures could be used to predict the target phenotype in other experiments that exhibit a phenotype of interest, we used *randomForest* package in R (default parameters with 1,000 trees). We used a leave-one-out cross-validation (LOOCV) in order to calculate an area under the curve (AUC). Iteratively for all experiments we left one experiment out and derived the molecular signature using the rest of the experiments that exhibited the target phenotype. For example, one iteration comprised removing the controls/mutants, part of the *crol* experiment (starvation-sensitive) and generating the molecular signature using the rest of the experiments (*dhr96*, *mir14*, *p53* and *rbf*). Crucially, we derived the residuals from the random intercept LMEM, along with consecutive principal components, for all experiments that exhibited the target phenotype, and then left one experiment out. This ensured that the model was corrected for underlying technical factors before performing the LOOCV. The AUC was calculated using the class (control/mutant) probabilities derived from the *randomForest* package, using the top 200 genes from the molecular signature (based on the p-values from the logistic regression). We also tested a different number of top genes (from 50 to 3,000 genes, Figs S6 and S7 for the starvation-sensitive and sterile phenotypes respectively). In

addition, we also formally tested if the mean of the class probabilities was different from 0.5 using a t-test, separately for controls and mutants, for the left-one-out experiment. The probability of 0.5 is the null hypothesis and it is equivalent to a random assignment of the controls/mutants.

Predicting the presence of phenotypic expression in freely available data

Similarly to the LOOCV, we used the molecular signature (top 200 genes based on the p-value from the logistic regression) for the starvation-sensitive and sterile phenotypes to predict the presence of the phenotypes in all available data in ArrayAtlas (Affymetrix GeneChip Drosophila Genome 2.0 Array). Iteratively for each deposited experiment in ArrayAtlas, we first derived residuals from a random intercept LMEM, including consecutive PCs, from the combined log2-normalised data for the experiment and the experiments that were part of the two phenotypes (starvation-sensitive and sterile). This ensured that we accounted for any technical variability between experiments. These residuals were then used to derive the probabilities for class (control/mutant) separation with the *randomForest* package in R. Each individual control/mutant sample within an experiment was assigned a class probability (control or mutant). For each class (control or mutant) the probabilities were averaged across the number of samples, separately for controls and mutants. This mean probability was used to infer quantitatively the target phenotype.

Results

Experiments and expression data

Using the above protocol, we identified five and six experiments, respectively, with specific perturbed genes for which gene-expression data for the starvation stress response defective (FlyBase control vocabulary identifier FBcv:0000708) and the female sterile (FBcv:0000366) target phenotypes were available. These were *dhr96* (E-GEOD-18576), *mir-14* (E-GEOD-20202), *rbf* (E-GEOD-38430), *p53* (E-GEOD-37404) and *crol* (E-GEOD-8775) for the starvation sensitive phenotype and *loj* (E-GEOD-10940), *ovo* (E-GEOD-48145), *pxt* (E-GEOD-29815), *su(HW)* (E-GEOD-36528), *ttk* (E-GEOD-42758) and *vret* (E-GEOD-30360) for the sterile phenotype. Additional information can be found in S1 and S2 Tables. Following normalisation and excluding transcripts that did not match any known or predicted gene, there were 12,630 genes left for analysis. The normalised gene-expression data are available upon request.

GO-terms enrichment analysis

Figs 3 and 4 show the results for the GO-terms associated with the two target phenotypes respectively (full numerical data are shown in S5 and S6 Tables). Enrichment of starvation-related GO terms for the starvation-sensitive phenotype was observed for LMEM with the inclusion of one to four PCs (Fig 3). In contrast, sterile-related GO terms were found to be mostly enriched with LMEM without the inclusion of PCs (Fig 4). This suggests that there is more inter-experiment variability associated

with the starvation-sensitive phenotype as compared to the sterile. All of the individual gene perturbation experiments that exhibited the sterile phenotype comprised female flies and more homogeneous tissue used to derive the expression data (S2 Table), whereas the individual experiments for the starvation-sensitive phenotype were mixed sex and the expression data were derived from a variety of tissues (S1 Table).

We also performed a GO enrichment analysis associated with individual control/mutant experiments exhibiting the two target phenotypes (e.g. *crol* part of E-GEOD-8775). Ranks of genes were derived using the *limma* package in R. Only two experiments showed any statistically significant evidence of GO-terms enrichment associated with the starvation phenotype (*crol* and *p53*; S3 Table), whereas all of the experiments that were identified to exhibit the sterile phenotype showed statistically significant enrichment of reproduction-related GO terms (S4 Table).

Fig 3. Top GO terms for the starvation-sensitive molecular signature. Red vertical line represents FDR p-value 0.05

Fig 4. Top GO terms for the sterile molecular signature. Red vertical line represents FDR p-value 0.05

Removal of batch effects within an experiment

Only one experiment (*loj*), with the sterile phenotype, exhibited a significant batch effect. The controls and mutants comprised two tissues (abdomen and head/thorax). We used the *ber* package to correct for the batch effect using the tissue as a factor. We observed two clusters for the first PC (89.34% variance explained) that separated the *loj* by tissue (S3a Fig). Correcting for the tissue batch effect

eliminated the tissue separation and the *loj* controls/mutants separated by the second PC (S3b Fig).

Determining the number of PCs for unwanted variation

The maximum AUC for the leave-one-out cross-validation for the starvation sensitive phenotype was 97% with six consecutive PCs and 85% with LMEM with no PCs for the sterile phenotype (Figs 5 and 6).

Fig 5. Starvation-sensitive phenotype, leave-one-out cross-validation AUC. AUC- Area Under the Curve; a through h LMEM with 0 to 7 PCs

Fig 6. Sterile phenotype, leave-one-out cross-validation AUC. AUC- Area Under the Curve; a through h LMEM with 0 to 7 PCs

Nevertheless, GO term enrichment analysis showed that the statistical significance of starvation-related GO terms disappeared (FDR p-value >0.05) when the first five or six PCs were included in the LMEM (Fig 3). GO terms enrichment results for the sterile phenotype are shown in Fig 4. Furthermore, PCA of the residuals of the starvation sensitive LMEM with five or six PCs showed near complete separation of the controls and mutants (S4f and S4g Figs). Taken together, these results suggest that the first four PCs account for biological/technical variability, that the overall molecular signature is enriched with starvation-related GO terms, and the 5th and 6th PCs account for the starvation-sensitive phenotype. We hypothesise that when we account for the first 5-6 PCs, the signal that is left is a form of global gene-expression regulation following a gene perturbation. Thus, accounting for the first

five or six PCs results in a prediction of the class separation, rather than the manifestation of the phenotype. A gene perturbation disrupts the global gene-expression equilibrium and results in differential expression of compensatory gene mechanisms. In other words, control/mutant experiments with seemingly different gene perturbations may result in a higher than expected by chance overlap of differentially expressed genes, i.e. genes that are part of the compensatory gene-expression regulatory network. In order to test this hypothesis, we performed 1,000 permutations, whereby we chose five random control/mutant experiments from EBI's ArrayExpress. The number of controls/mutants per experiment was matched to the number of controls/mutants in the five experiments for the starvation-sensitive phenotype. Thus, the number of controls/mutants in a randomly chosen experiment was reduced to match the number of controls/mutants in S1 Table. For each of these experiments we derived normalised gene-expression values using the same procedure as for the starvation-sensitive phenotype. We derived differentially expressed genes using the *limma* package in R. For each of these random sets of experiments, we selected the top 200 genes and calculated the number of genes that overlap within each set of experiments in a pairwise manner. For each of these permutations we calculated the median of the -log10 of the p-value for each pairwise overlap using hypergeometric distribution. We compared these results to the pairwise overlap of random 200 genes as part of 1,000 sets of experiments. The distributions of the results for the random 1,000 sets of experiments and for what is expected by chance are shown in Fig 7.

Fig 7. Distribution of the pairwise overlap of genes in 1,000 random sets of five experiments, derived from ArrayExpress, as compared to expected by chance. y-axis- Median -$\log_{10}$ hypergeometric p-value for significance of pairwise overlap

The results presented in Fig 7 clearly show that a random combination of sets of five experiments exhibit a significantly greater number of differentially expressed genes that overlap between the experiments as compared to purely by chance alone. This observation has been also reported in humans [6]. Thus, for the leave-one-out cross-validation for the starvation-sensitive phenotype we used the first four PCs to account for biological/technical variation. For the sterile phenotype we did not use PCs (LMEM with 0 PCs). PCA graphs for the sterile molecular signature LMEM with 0 to 7 PCs are shown in S5 Fig. For the calculation of the AUC for the LOOCV we tested a range of top genes (from 50 to 3,000). For the starvation-sensitive phenotype there was not a difference in the AUC with different number of top genes, although choosing more genes resulted in a slightly higher AUC (50 genes 87.76% AUC; 3,000 genes 90.31% AUC; Fig S6 with 4PCs). The opposite was noted with the sterile phenotype, fewer number of top genes resulted in higher AUC (50 genes 90.58% AUC; 3,000 genes 73.68% AUC; Fig S7 with 0PCs). These trends could potentially reflect the size of the transcriptional network involved in both phenotype, for example it has been previously reported that the starvation stress resistance involves transcriptional response of ~25% of the genome in *Drosophila* [7].

The mean distribution of the control/mutant class probabilities from the random forest for both the starvation-sensitive and sterile phenotypes were significantly different from 0.5 (Table 1). The results in Table 1, along with the AUC for both phenotypes (Figs 5 and 6), show that we can confidently predict the phenotypic manifestation of a separate experiment that exhibits the phenotype of interest.

Table 1. One sample t-test for class probabilities (controls/mutants) in the two phenotypes following LOOCV

Class	Phenotype	
	starvation-sensitive p-value (μ=0.5)	Sterile p-value (μ=0.5)

Controls	$5.72x10^{-03}$	$3.87x10^{-03}$
Mutants	$5.84x10^{-03}$	$3.10x10^{-03}$

Predicting freely available experiments for the presence of both phenotypes

In order to obtain freely available experiments we utilised EBI's ExpressionAtlas (https://www.ebi.ac.uk/gxa/home) instead of ArrayExpress. We used EBI's ExpressionAtlas due to the availability of normalised gene-expression values for a large number of the already available raw cel gene-expression data in ArrayExpress. This eliminated the need to normalise all of the available raw gene-expression data within ArrayExpress. For all experiments available in EBI's ExpressionAtlas (total number of control/mutant experiments at the time of conducting the study: 211) we used the molecular signatures for the starvation sensitive and sterile phenotypes to derive a mean probability separately for controls and mutants in an experiment. The mean mutant probability was used to suggest a degree of phenotypic manifestation. Ranking of all available experiments is given in S7 and S8 Tables for the starvation-sensitive and sterile phenotypes respectively.

Ranking EBI's ExpressionAtlas experiments for the starvation-sensitive phenotype

The top three ranked experiments were all already used to generate the molecular signature (*dhr96*, *crol* and *rbf*), thus it is not unexpected that we can predict these experiments with the highest accuracy. The *p53* (E-GEOD-37404) and *mir-14* (E-GEOD-20202) experiments are not included in the EBI's ExpressionAtlas datasets.

For the rest of the freely available experiments available in EBI's ExpressionAtlas we found no results from a direct lab-based assay of the starvation sensitivity. Nevertheless, for some of the top-ranked experiments we found additional evidence that can be potentially used to support the results from our prediction. All three gene mutants (*rbf*^{120a}, *rbf*^{120a} *wts*^{latsX1} and *wts*^{latsX1}), part of an experiment (E-GEOD-24978) were ranked with mutant class probabilities of 83%, 74% and 64% respectively. The two genes, *rbf* and *wts* regulate cell proliferation via the p16 and Hippo tumour suppressor pathways. There is only a direct lab-based measurement of the starvation-sensitive phenotype of *rbf*^{120a}, which was used as part of the molecular signature. We speculate that the *wts*^{latsX1} and the double-mutant *rbf*^{120a} *wts*^{latsX1} may also exhibit starvation-sensitive phenotype.

Several of the top-ranked experiments included fly lines from the *Drosophila* Genetic Reference Panel (DGRP) [8]. These included genes (*esg, Pdcd4, mub, Gbs-70E*) that were reported to exhibit a reduced starvation resistance, tested at six weeks.

Ranking EBI's ExpressionAtlas experiments for sterile phenotype

The top four ranked experiments in the EBI's ExpressionAtlas comprise four already used control/mutant experiments for the sterile molecular signature (*ovo* (*ovo* and *ovo/cako*) and *loj* (head and thorax)), thus it is not surprising that we can detect these with high accuracy. The rest of the experiments, part of the molecular signature, were not analysed as part of EBI's ExpressionAtlas (not all experiments from ArrayExpress are analysed in ExpressionAtlas). Similarly to the starvation-sensitive molecular signature we found no direct evidence that the top-ranked experiments will exhibit the sterile phenotype. Nevertheless, there was additional evidence for some of the top-ranked experiments. For example, experiment E-GEOD-55187 comprises

sesb[1] homozygous female mutants that are predicted to exhibit the sterile phenotype with mean probability of 85% across the individual mutants. *Sesb*[1] is listed as female sterile in flybase (http://flybase.org/reports/FBal0015434 - phenotypic_data_sub). Due to lack of information, we could not verify whether the gene-mutant shown as sterile [9] is exactly the same as the gene-mutants with the microarray data in EBI's ArrayExpress [10]. Similarly, in experiment E-MTAB-3546 [11], 3-week reproductive diapause under cold conditions (11C) was predicted to exhibit the sterile phenotype with a mean mutant probability of 91% across the individual mutants. Clearly, the mutant female flies are very likely to exhibit the sterile phenotype as they were induced into a diapause that is associated with a reproductive arrest. The 10 and 40 days aged dietary restricted female flies (E-GEOD-26726) also showed evidence of the sterile phenotype (84% and 79% respectively). There is a well-defined reduction in daily and lifetime fecundity under dietary restriction [12], therefore it is more than likely that the 10 and 40 days old flies will exhibit the sterile phenotype.

Discussion

In this paper we present a novel computational approach for integrating gene-expression data for two specific phenotypes (starvation-sensitive and sterile) in *Drosophila* from the vast and largely unutilised freely available public repositories. This integration is multi-layered with phenotypic information derived from a species-specific database (FlyBase) and gene-expression from the largest repository of publicly available genomic data, the ExpressionAtlas at the European Bioinformatics Institute. Crucially, we present an approach to utilise gene-expression data generated by completely independent groups across the scientific community.

The results of this proof-of-concept study show that it is possible to integrate seemingly different gene-expression microarray data using a combination of linear-mixed effect models and principal components analyses and predict a potential phenotypic manifestation with a relatively high degree of confidence. Nevertheless, the applicability of this methodology to capture a wide range of phenotypes and organisms requires a considerable amount of additional work that is beyond the scope of this article.

The premise of our methodology is based upon the assumption that specific cellular and physiological phenotypes are underlined by or associated with similar gene-expression changes. In addition, the number of such gene-expression changes that are shared between different perturbations and are associated with a specific phenotype, is likely to differ between different phenotypes. Currently, there is no simple way to derive a set number of gene-expression changes that describe a particular phenotype and this number is also likely to depend on the nature of the phenotype. We used an empirically derived number of genes for the two phenotypes that we tested (top 200 genes, based on p-value for differential expression), although this selection can potentially be automated using a different number of genes. Our approach might not be directly applicable if a specific phenotype is underlined by independent biological pathways or caused by mechanisms that do not result in changes in gene-expression. Nevertheless, additional genomic measurements can be incorporated as and when they become available. Furthermore, our methodology relies on freely available gene-expression data, which is only set to increase [13]. Thus, with the increase in repository data, our approach has a great potential to estimate relative degree of independence of biological pathways that influence or give rise to specific phenotypes.

Biological phenotypes are rarely binary features, although they often get binarised for ease of use, for example gravitaxis defective phenotype (movement away from the source of gravity) can be expressed as defective/normal or a more complex measure can be used to account for the continuous nature of the phenotype [14]. Nevertheless, even with considerable efforts to standardise experimental protocols and measurement assays, differences will be exhibited between laboratories across the world. As such, it is difficult to utilise the continuous phenotype response measurements. In this study we only considered control/mutant type experiments. For such experiments the measured phenotypes can be taken as relative with respect to controls, thus minimising the differences in protocols. Nevertheless, for most such experiments in *Drosophila*, there is no unified system/database that collects and archives the outcomes of such measurements and currently these have to be extracted manually from the corresponding manuscripts and assessment made on how similar the protocols are. Our methodology of predicting potential phenotypic manifestation uses a machine learning approach, that is random forest. This could potentially be used to infer the two phenotypes probabilistically, although it is unclear what the relationship is between the similarity in gene-expression and the degree of phenotype manifestation.

Although our study utilises gene-expression microarray data and such type of data is clearly superseded by RNA sequencing [13], we do not foresee any major challenges in adopting our methodology to work with RNA-seq data. For example, raw RNA-seq counts can be relatively easily transformed into transcripts per million (TPM) and log2 of TPM can be used in the linear-mixed effect models.

Our methodology relies on linear-mixed effect models accounting for unwanted biological effects in the form of principal components. In order to estimate

the number of PCs we utilised Gene Ontology enrichment analysis, whereby we chose consecutive number of PCs to maximise GO enrichment. One of the potential limitations is that there might be some degree of circularity when using GO terms to define phenotypic enrichment, since GO categories could have been partially defined using similar data. The other limitation is that the combination of PCs and linear-mixed effect model is likely to be overconservative, such that some variation in the phenotype of interest maybe already included in the PCs. Other approaches, such as probabilistic estimation of expression residuals (PEER) [15] could be used to facilitate estimation of unwanted factors.

The proof-of-concept study presented here is a novel approach of predicting the manifestation of two phenotypes in *Drosophila* from gene-expression data. While, similar attempts have been previously performed [16-19], these studies rely on a single or a few well-defined datasets with few measured phenotypes. Our approach goes beyond single studies and it is not restricted to selective phenotypic measurements in a few datasets. The methodology described here captures the diverse genetic background and gene-perturbations from all the publicly available repository data and links them to phenotypic characteristics, thereby adding value to already deposited and largely unutilised data.

Supporting information

S1 Fig. Representative GO terms associated with the starvation sensitive phenotype. Boxes represent nodes and arrows represent edges. Nodes filled with yellow are the GO terms used to assess if the molecular signature is associated with the starvation sensitive phenotype. Data derived from https://www.ebi.ac.uk/QuickGO

S2 Fig. Representative GO terms associated with the sterile phenotype. Boxes represent nodes and arrows represent edges. Nodes filled with yellow are the GO terms used to assess if the molecular signature is associated with the sterile phenotype. Data derived from https://www.ebi.ac.uk/QuickGO

S3 Fig. PCA before and after batch effect correction for the *loj*. a) *loj* log-2 normalised values without batch effect correction. b) *loj* log-2 normalised values after batch effect correction with *ber*; controls and mutants are labelled with black and red symbols respectively; circles and triangles represent samples derived from abdomen and head/thorax respectively

S4 Fig. PCA plots of LMEM with consecutive PCs of the starvation sensitive phenotype. a) LMEM with 0 PCs; b) LMEM with 1 PCs; c) LMEM with 2 PCs; d) LMEM with 3 PCs; e) LMEM with 4 PCs; f) LMEM with 5 PCs; g) LMEM with 6 PCs; h) LMEM with 7 PCs

S5 Fig. PCA plots of LMEM with consecutive PCs of the sterile phenotype. a) LMEM with 0 PCs; b) LMEM with 1 PCs; c) LMEM with 2 PCs; d) LMEM with 3 PCs; e) LMEM with 4 PCs; f) LMEM with 5 PCs; g) LMEM with 6 PCs; h) LMEM with 7 PCs

S6 Fig. AUC leave-one-out cross-validation using different number of top genes (starvation-sensitive phenotype). AUC- Area Under the Curve; Each bar (from left to right) represents a one leave-one-out cross-validation using 50, 100, 200, 300, 400,

500, 600, 700, 800, 900, 1000, 1500, 2000, 2500 and 3000 genes within each of the PCs (represented by different colours)

S7 Fig. AUC leave-one-out cross-validation using different number of top genes (sterile-sensitive phenotype). AUC- Area Under the Curve; Each bar (from left to right) represents a one leave-one-out cross-validation using 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500 and 3000 genes within each of the PCs (represented by different colours)

S1 Table Data used to derive the molecular signature for the starvation sensitive phenotype (FBcv:0000708). [a]FlyBase gene symbol with gene identifier and gene name in brackets; [b]EBI's ArrayExpress accession identifier with reference in brackets; all perturbed genes are knockouts

S2 Table Data used to derive the molecular signature for the sterile phenotype (FBcv:0000366). [a]FlyBase gene symbol with gene identifier and gene name in brackets; [b]EBI's ArrayExpress accession identifier with reference in brackets; all perturbed genes are knockouts

S3 Table Individual experiment GO enrichment analysis (starvation sensitive phenotype). FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. GO terms directly related to starvation with FDR p-values < 0.05 are labelled in bold

S4 Table Individual experiment GO enrichment analysis (sterile phenotype). FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. FDR p-values < 0.05 are labelled in bold

S5 Table Top GO terms for the starvation sensitive molecular signature. FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. Cov- number of Principal Components included in the linear-mixed effect model. BP- Biological Process; FDR p-values < 0.05 are labelled in bold; AUC- Area Under the Curve

S6 Table Top GO terms for the sterile molecular signature. FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. Cov- number of Principal Components included in the linear-mixed effect model. BP- Biological Process; CC- Cellular Component; FDR p-values < 0.05 are labelled in bold; AUC- Area Under the Curve

S7 Table Ranking EBI's ExpressionAtlas (starvation-sensitive molecular signature top 30 experiments). FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. FDR p-values < 0.05 are labelled in bold; Where there were multiple factor values, these are separated by "|". Factor values comprise genotype, treatment, etc.

S8 Table Ranking EBI's ExpressionAtlas (sterile molecular signature top 30 experiments). FDR p-value is the p-value corrected for multiple hypothesis testing

using False Discovery Rate, accounting for all GO terms tested. FDR p-values < 0.05 are labelled in bold; Where there were multiple factor values, these are separated by "|". Factor values comprise genotype, treatment, etc.

Acknowledgements

We would like to thank the Advanced Research Computing at Cardiff (ARCCA) and EMBL-EBI for providing computational resources.

References

1. Consortium GT. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45(6):580-5. Epub 2013/05/30. doi: 10.1038/ng.2653. PubMed PMID: 23715323; PubMed Central PMCID: PMCPMC4010069.

2. Flint J, Munafo MR. The endophenotype concept in psychiatric genetics. Psychol Med. 2007;37(2):163-80. Epub 2006/09/19. doi: 10.1017/S0033291706008750. PubMed PMID: 16978446; PubMed Central PMCID: PMCPMC2829981.

3. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15. doi: 10.1093/bioinformatics/btg405. PubMed PMID: 14960456.

4. Giordan M. A Two-Stage Procedure for the Removal of Batch Effects in Microarray Studies. Statistics in Biosciences. 2014;6(1):73-84.

5. Breslin T, Eden P, Krogh M. Comparing functional annotation analyses with Catmap. BMC Bioinformatics. 2004;5:193. Epub 2004/12/14. doi: 10.1186/1471-2105-5-193. PubMed PMID: 15588298; PubMed Central PMCID: PMCPMC543458.

6. Crow M, Lim N, Ballouz S, Pavlidis P, Gillis J. Predictability of human differential gene expression. Proc Natl Acad Sci U S A. 2019;116(13):6491-500. Epub 2019/03/09. doi: 10.1073/pnas.1802973116. PubMed PMID: 30846554; PubMed Central PMCID: PMCPMC6442595.

7. Harbison ST, Chang S, Kamdar KP, Mackay TF. Quantitative genomics of starvation stress resistance in Drosophila. Genome Biol. 2005;6(4):R36. Epub 2005/04/19. doi: 10.1186/gb-2005-6-4-r36. PubMed PMID: 15833123; PubMed Central PMCID: PMCPMC1088964.

8. Mackay TF, Richards S, Stone EA, Barbadilla A, Ayroles JF, Zhu D, et al. The Drosophila melanogaster Genetic Reference Panel. Nature. 2012;482(7384):173-8. Epub 2012/02/10. doi: 10.1038/nature10811. PubMed PMID: 22318601; PubMed Central PMCID: PMCPMC3683990.

9. Chen S, Oliveira MT, Sanz A, Kemppainen E, Fukuoh A, Schlicht B, et al. A cytoplasmic suppressor of a nuclear mutation affecting mitochondrial functions in Drosophila. Genetics. 2012;192(2):483-93. Epub 2012/08/02. doi: 10.1534/genetics.112.143719. PubMed PMID: 22851652; PubMed Central PMCID: PMCPMC3454878.

10. Vartiainen S, Chen S, George J, Tuomela T, Luoto KR, O'Dell KM, et al. Phenotypic rescue of a Drosophila model of mitochondrial ANT1 disease. Dis Model Mech. 2014;7(6):635-48. Epub 2014/05/09. doi: 10.1242/dmm.016527. PubMed PMID: 24812436; PubMed Central PMCID: PMCPMC4036471.

11. Kucerova L, Kubrak OI, Bengtsson JM, Strnad H, Nylin S, Theopold U, et al. Slowed aging during reproductive dormancy is reflected in genome-wide transcriptome changes in Drosophila melanogaster. BMC Genomics. 2016;17:50. Epub 2016/01/14. doi: 10.1186/s12864-016-2383-1. PubMed PMID: 26758761; PubMed Central PMCID: PMCPMC4711038.

12. Partridge L, Piper MD, Mair W. Dietary restriction in Drosophila. Mech Ageing Dev. 2005;126(9):938-50. Epub 2005/06/07. doi: 10.1016/j.mad.2005.03.023. PubMed PMID: 15935441.

13. Cook CE, Bergman MT, Cochrane G, Apweiler R, Birney E. The European Bioinformatics Institute in 2017: data coordination and integration. Nucleic Acids Res. 2018;46(D1):D21-D9. Epub 2017/12/01. doi: 10.1093/nar/gkx1154. PubMed PMID: 29186510; PubMed Central PMCID: PMCPMC5753251.

14. Vang LL, Medvedev AV, Adler J. Simple ways to measure behavioral responses of Drosophila to stimuli and use of these methods to characterize a novel mutant. PLoS One. 2012;7(5):e37495. Epub 2012/06/01. doi: 10.1371/journal.pone.0037495. PubMed PMID: 22649531; PubMed Central PMCID: PMCPMC3359294.

15. Stegle O, Parts L, Piipari M, Winn J, Durbin R. Using probabilistic estimation of expression residuals (PEER) to obtain increased power and interpretability of gene expression analyses. Nat Protoc. 2012;7(3):500-7. Epub 2012/02/22. doi: 10.1038/nprot.2011.457. PubMed PMID: 22343431; PubMed Central PMCID: PMCPMC3398141.

16. Takagi Y, Matsuda H, Taniguchi Y, Iwaisaki H. Predicting the phenotypic values of physiological traits using SNP genotype and gene expression data in mice. PLoS One. 2014;9(12):e115532. Epub 2014/12/30. doi: 10.1371/journal.pone.0115532. PubMed PMID: 25541966; PubMed Central PMCID: PMCPMC4277360.

17. Zarringhalam K, Degras D, Brockel C, Ziemek D. Robust phenotype prediction from gene expression data using differential shrinkage of co-regulated genes. Sci Rep. 2018;8(1):1237. Epub 2018/01/21. doi: 10.1038/s41598-018-19635-0. PubMed PMID: 29352257; PubMed Central PMCID: PMCPMC5775343.

18. Li Z, Gao N, Martini JWR, Simianer H. Integrating Gene Expression Data Into Genomic Prediction. Front Genet. 2019;10:126. Epub 2019/03/13. doi: 10.3389/fgene.2019.00126. PubMed PMID: 30858865; PubMed Central PMCID: PMCPMC6397893.

19. Ellis SE, Collado-Torres L, Jaffe A, Leek JT. Improving the value of public RNA-seq expression data by phenotype prediction. Nucleic Acids Res. 2018;46(9):e54. Epub 2018/03/08. doi: 10.1093/nar/gky102. PubMed PMID: 29514223; PubMed Central PMCID: PMCPMC5961118.

Supporting Information for

A novel computational approach for predicting complex phenotypes in *Drosophila* (starvation-sensitive and sterile) by deriving their gene expression signatures from public data

Dobril K. Ivanov[1,2]*, Gerrit Bostelmann[1], Benoit Lan-Leung[2], Julie Williams[2], Linda Partridge[3,4,¶], Valentina Escott-Price[2,¶], Janet M. Thornton[1,¶]

[1]European Molecular Biology Laboratory, The European Bioinformatics Institute (EMBL-EBI), Wellcome Trust Genome Campus, Hinxton, Cambridge, UK
[2]UK Dementia Research Institute at Cardiff (UKDRI), Cardiff University, College of Biomedical and Life Sciences, Hadyn Ellis Building, Cardiff, UK
[3]Max Planck Institute for Biology of Ageing, Cologne, Germany
[4]Institute of Healthy Ageing, and Department of Genetics, Evolution and Environment, UCL, London, UK
* Corresponding author
E-mail: IvanovD1@cardiff.ac.uk (DKI)
[¶]These authors are joint senior authors of this work (LP, VEP and JMT are Joint Senior Authors)

Supplementary Materials and Methods

Normalising gene-expression values

Raw gene-expression data (cel files) were downloaded from the EBI's ArrayExpress (https://www.ebi.ac.uk/arrayexpress/). Separately, for each experiment the raw data were summarised and normalised by using the Robust Multichip Average (*rma* function without background normalisation, part of bioconductor's package *affy* [1]. Summarised probe-sets were mapped to transcripts using bioconductor's package *drosophila2.db*. Transcripts not mapping to any known or predicted genes were excluded from further analysis. Log2-normalised expression data for all experiments that exhibit a particular phenotype were combined in a single dataset.

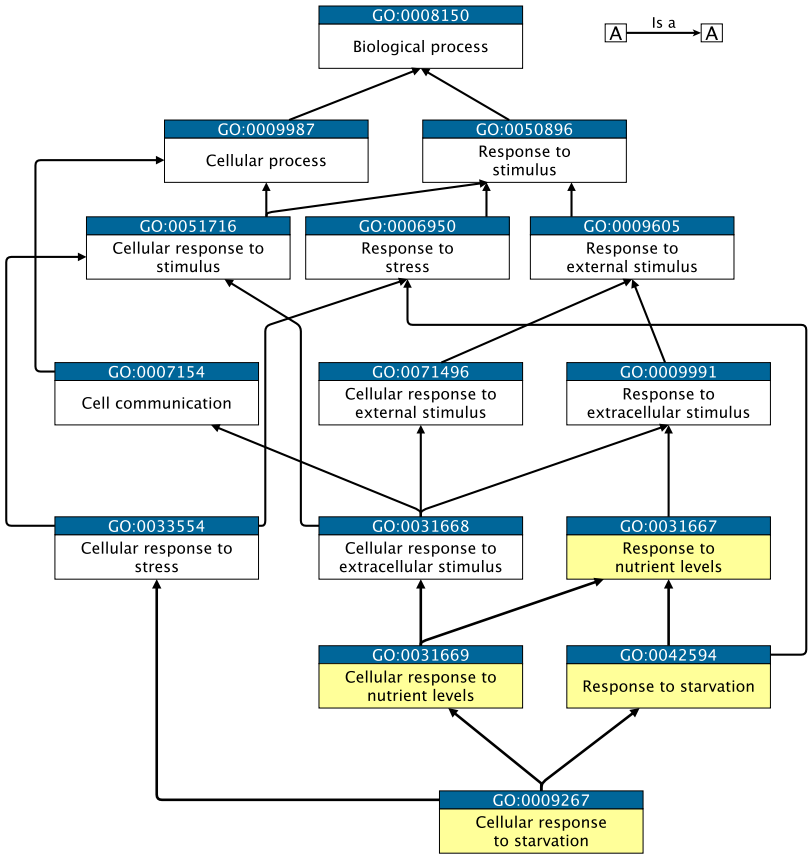
Generation of the Molecular Signatures (Linear-mixed effects model)

To assess the statistical significance of each gene we used logistic regression (*glm* function in R). For the LMEM, we utilised the *lmer* function (bobyqa optimiser) within the *lme4* package, part of bioconductor. For the LMEM, we utilised the *lmer* function within the *lme4* package, part of bioconductor.

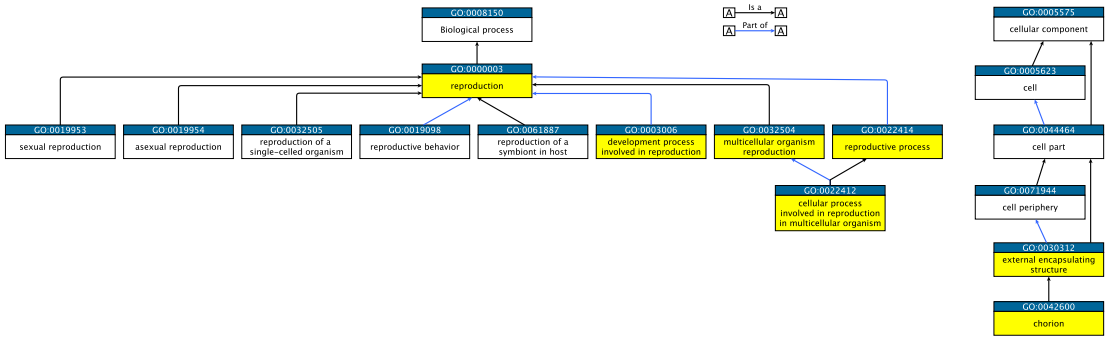
Gene Ontology (GO) enrichment analysis

The Wilcoxon rank sum test, as implemented in Catmap ([2]), was used to perform functional analysis to test for significant enrichment of Gene Ontology categories. FlyBase gene identifiers were mapped to Gene Ontology identifiers (FlyBase version FB2018_02) using custom programs.

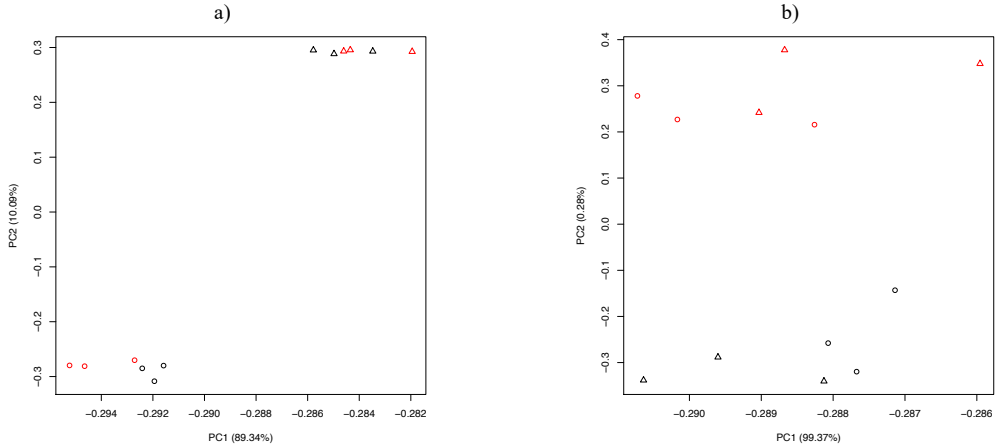
Supplementary Figures and Tables



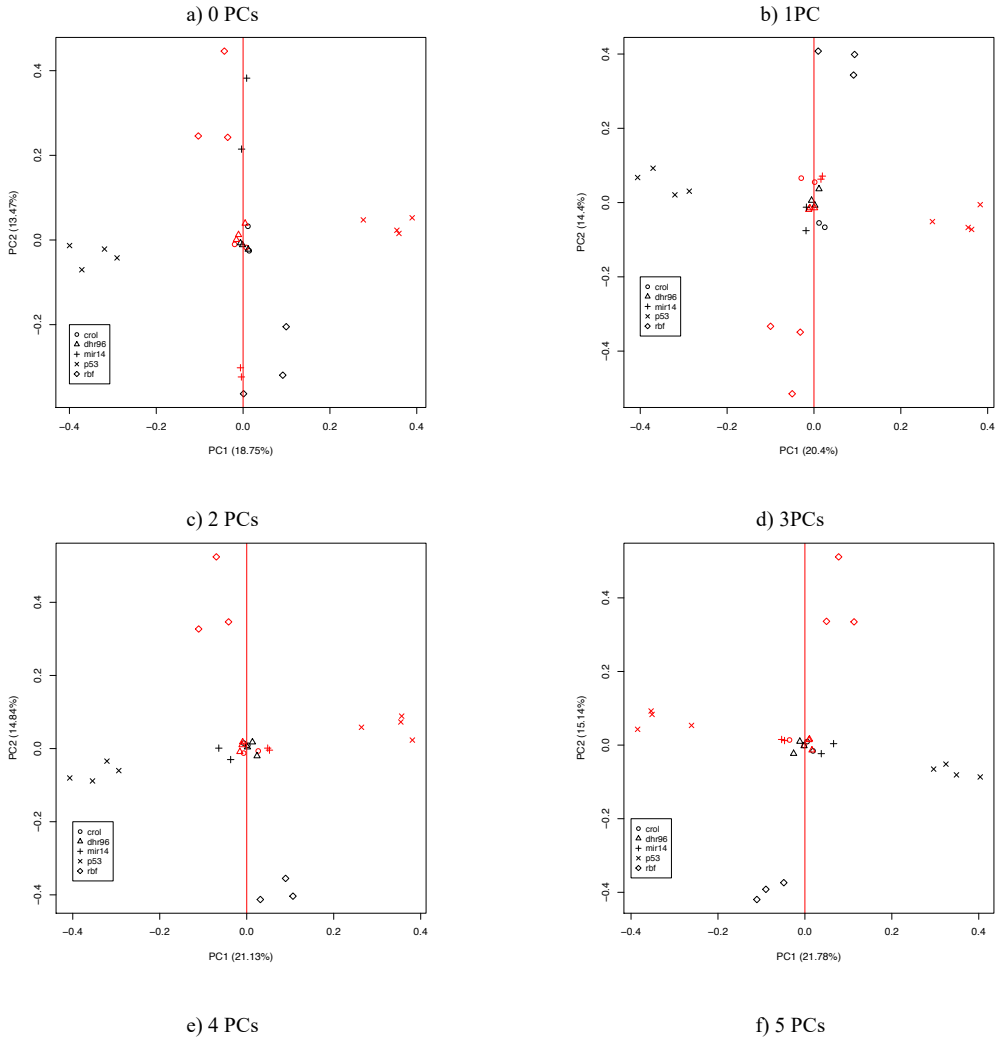
S1 Fig. Representative GO terms associated with the starvation sensitive phenotype. Boxes represent nodes and arrows represent edges. Nodes filled with yellow are the GO terms used to assess if the molecular signature is associated with the starvation sensitive phenotype. Data derived from https://www.ebi.ac.uk/QuickGO

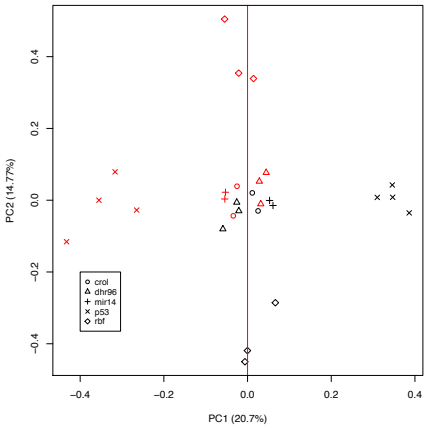


S2 Fig. Representative GO terms associated with the sterile phenotype. Boxes represent nodes and arrows represent edges. Nodes filled with yellow are the GO terms used to assess if the molecular signature is associated with the sterile phenotype. Data derived from https://www.ebi.ac.uk/QuickGO

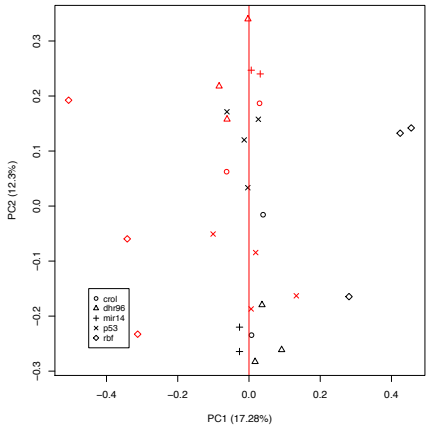


S3 Fig. PCA before and after batch effect correction for the *loj*. a) *loj* log-2 normalised values without batch effect correction b) *loj* log-2 normalised values after batch effect correction with *ber*; controls and mutants are labelled with black and red symbols respectively; circles and triangles represent samples derived from abdomen and head/thorax respectively

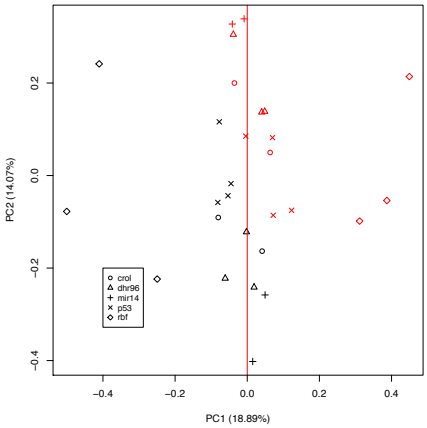


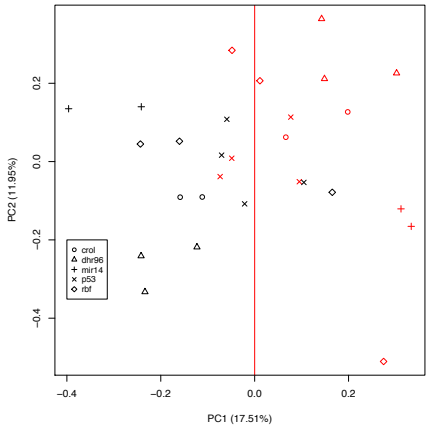


g) 6 PCs



h) 7 PCs

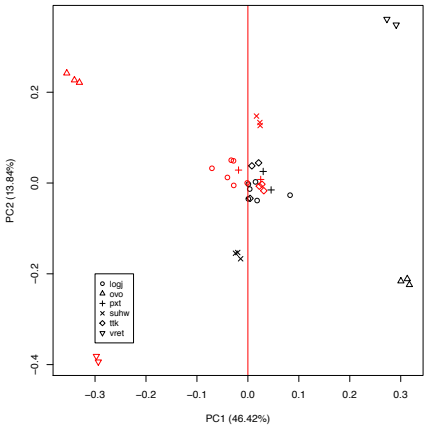




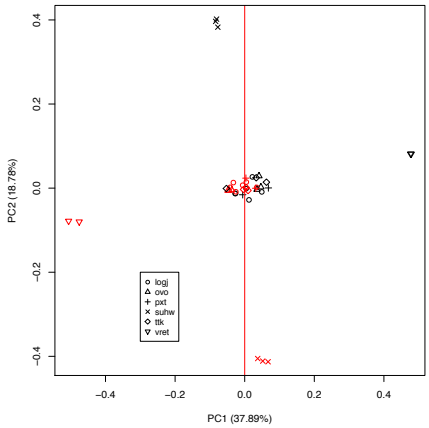
S4 Fig. PCA plots of LMEM with consecutive PCs of the starvation sensitive phenotype. a) LMEM with 0 PCs; b) LMEM with 1 PCs; c) LMEM with 2 PCs; d) LMEM with 3 PCs; e) LMEM with 4 PCs; f) LMEM with 5 PCs; g) LMEM with 6 PCs; h) LMEM with 7 PCs

a) 0 PCs

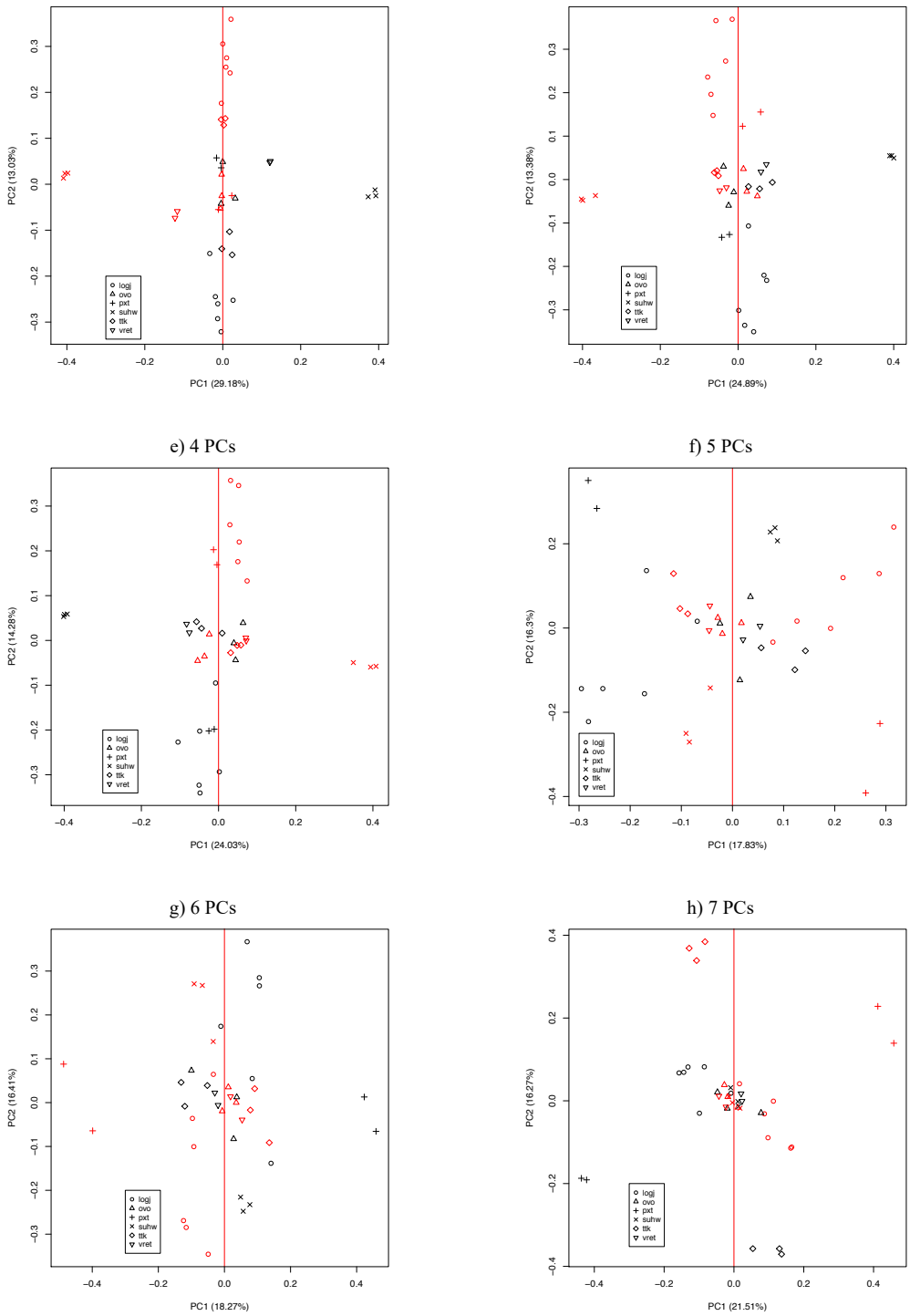
b) 1 PC



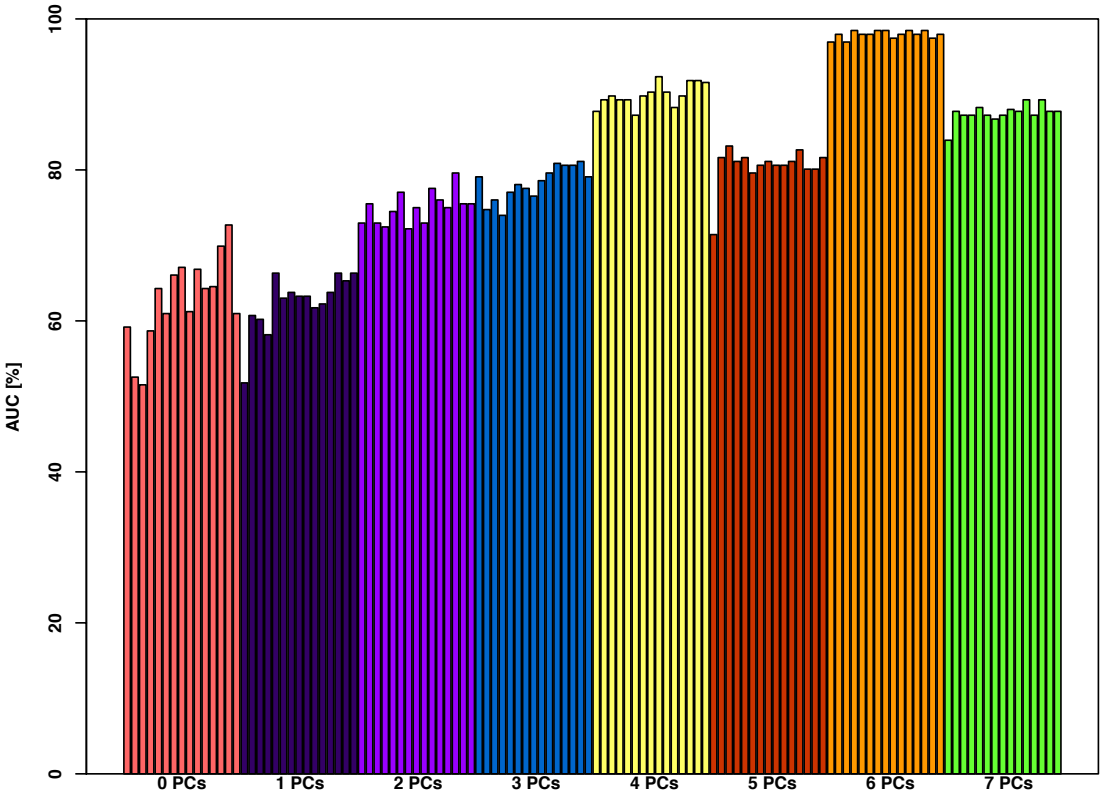
c) 2 PCs



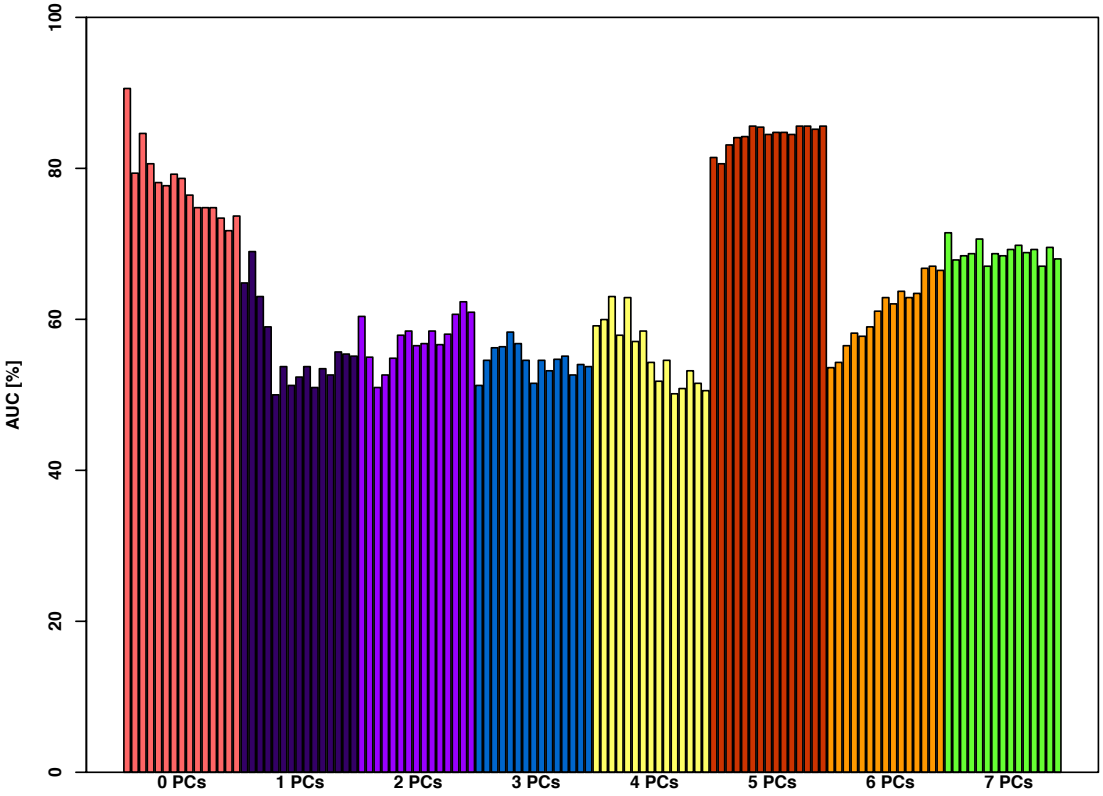
d) 3 PCs



S5 Fig. PCA plots of LMEM with consecutive PCs of the sterile phenotype. a) LMEM with 0 PCs; b) LMEM with 1 PCs; c) LMEM with 2 PCs; d) LMEM with 3 PCs; e) LMEM with 4 PCs; f) LMEM with 5 PCs; g) LMEM with 6 PCs; h) LMEM with 7 PCs



S6 Fig. AUC leave-one-out cross-validation using different number of top genes (starvation-sensitive phenotype)
AUC- Area Under the Curve; Each bar (from left to right) represents a one leave-one-out cross-validation using 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500 and 3000 genes within each of the PCs (represented by different colours)



S7 Fig. Fig. AUC leave-one-out cross-validation using different number of top genes (sterile-sensitive phenotype)
AUC- Area Under the Curve; Each bar (from left to right) represents a one leave-one-out cross-validation using 50, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, 1500, 2000, 2500 and 3000 genes within each of the PCs (represented by different colours)

Gene[a]	Accession ID[b]	N replicates		Age [days]	Tissue	Sex
		controls	mutants			
dhr96 (FBgn0015240; Hormone receptor-like in 96)	E-GEOD-18576 [3]	3	3	9	whole body	male
mir-14 (FBgn0262447; mir-14 stem loop)	E-GEOD-20202 [4]	2	2	5	head	male
rbf (FBgn0015799; Retinoblastoma-family protein)	E-GEOD-24978 [5]	3	3	0	3rd instar larvae	mixed
p53 (FBgn0039044)	E-GEOD-37404 [6]	4	4	0	3rd instar larvae	mixed
crol (FBgn0020309; crooked legs)	E-GEOD-8775 [7]	2	2	42	whole body	female

S1 Table Data used to derive the molecular signature for the starvation sensitive phenotype (FBcv:0000708). [a]FlyBase gene symbol with gene identifier and gene name in brackets; [b]EBI's ArrayExpress accession identifier with reference in brackets; all perturbed genes are knockouts

Gene[a]	Accession ID[b]	Number replicates		Age [days]	Tissue	Sex
		controls	mutants			
loj (FBgn0061492; logjam)	E-GEOD-10940 [8]	3	3	4	head/thorax	female
ovo (FBgn0003028)	E-GEOD-48145 [9]	3	3	15	whole body	female
pxt (FBgn0261987; Peroxinectin-like)	E-GEOD-29815 [10]	4	3	4	ovarian follicle	female
su(HW) (FBgn0003567; Suppressor of Hairy wing)	E-GEOD-36528 [11]	3	3	0	ovary	female
ttk (FBgn0003870; tramtrack)	E-GEOD-42758 [12]	3	3	1.5	ovary	female
vret (FBgn0263143; vreteno)	E-GEOD-30360 [13]	2	2	7	ovary	female

S2 Table Data used to derive the molecular signature for the sterile phenotype (FBcv:0000366). [a]FlyBase gene symbol with gene identifier and gene name in brackets; [b]EBI's ArrayExpress accession identifier with reference in brackets; all perturbed genes are knockouts

GO ID	GO name	*crol* FDR p-value	*dhr96* FDR p-value	*mir-14* FDR p-value	*p53* FDR p-value	*rbf* FDR p-value
GO:0009267	cellular response to starvation	**2.85E-08**	1.00E+00	6.11E-01	7.41E-02	7.80E-01
GO:0031669	cellular response to nutrient levels	**1.29E-08**	1.00E+00	5.74E-01	7.30E-02	7.93E-01
GO:0042594	response to starvation	**5.73E-05**	1.00E+00	3.46E-01	**3.54E-02**	3.68E-01
GO:0031667	response to nutrient levels	**1.33E-04**	1.00E+00	2.87E-01	**2.72E-02**	4.42E-01
GO:0031668	cellular response to extracellular stimulus	2.96E-08	1.00E+00	5.88E-01	5.05E-02	7.85E-01
GO:0033554	cellular response to stress	9.71E-11	1.00E+00	4.00E-01	1.73E-05	1.12E-07
GO:0009991	response to extracellular stimulus	1.90E-04	1.00E+00	3.01E-01	1.85E-02	4.34E-01
GO:0071496	cellular response to external stimulus	4.10E-07	1.00E+00	4.14E-01	2.70E-02	7.96E-01
GO:0007154	cell communication	3.85E-01	1.00E+00	9.77E-01	6.92E-02	2.67E-01
GO:0009605	response to external stimulus	3.74E-04	1.00E+00	2.07E-02	8.63E-03	1.64E-01
GO:0006950	response to stress	5.47E-11	1.00E+00	2.93E-02	3.69E-08	6.60E-08
GO:0051716	cellular response to stimulus	4.15E-09	1.00E+00	2.72E-01	1.18E-04	1.42E-05
GO:0050896	response to stimulus	4.47E-04	1.00E+00	9.94E-03	6.61E-03	1.85E-03
GO:0009987	cellular process	4.23E-11	1.00E+00	2.27E-04	2.17E-36	1.01E-18
GO:0008150	biological process	2.14E-01	1.00E+00	3.81E-01	1.19E-03	4.00E-02

S3 Table Individual experiment GO enrichment analysis (starvation sensitive phenotype). FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. GO terms directly related to starvation with FDR p-values < 0.05 are labelled in bold

GO ID	GO name	*loj* FDR p-value	*ovo* FDR p-value	*pxt* FDR p-value	*suhw* FDR p-value	*ttk* FDR p-value	*vret* FDR p-value
GO:0003006	developmental process involved in reproduction	**1.62E-03**	**7.25E-14**	**1.70E-29**	**2.49E-05**	**3.70E-08**	**1.39E-21**
GO:0022412	cellular process involved in reproduction in multicellular organism	**1.32E-04**	**6.40E-15**	**4.07E-32**	**1.65E-04**	**4.11E-09**	**3.22E-22**
GO:0022414	reproductive process	**3.38E-03**	**7.65E-12**	**1.13E-25**	**1.65E-04**	**1.01E-08**	**9.45E-19**
GO:0032504	multicellular organism reproduction	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
GO:0000003	reproduction	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00	1.00E+00
GO:0030312	external encapsulating structure	1.00E+00	**1.90E-05**	**9.23E-10**	**1.18E-16**	**1.68E-04**	**9.69E-31**
GO:0042600	chorion	1.00E+00	**3.15E-04**	**1.41E-10**	**1.82E-15**	**4.24E-05**	**1.21E-28**

S4 Table Individual experiment GO enrichment analysis (sterile phenotype). FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. FDR p-values < 0.05 are labelled in bold

GO term	FDR p (0 cov)	FDR p (1 cov)	FDR p (2 cov)	FDR p (3 cov)	FDR p (4 cov)	FDR p (5 cov)	FDR p (6 cov)	FDR p (7 cov)
BP GO:0009267 cellular response to starvation	1.86e-01	**4.13e-03**	**2.90e-03**	**2.37e-03**	**6.18e-03**	7.21e-02	1.65e-01	2.99e-01
BP GO:0031669 cellular response to nutrient levels	1.86e-01	**3.92e-03**	**2.39e-03**	**2.00e-03**	**5.01e-03**	5.97e-02	1.37e-01	2.63e-01
BP GO:0042594 response to starvation	5.59e-01	1.66e-01	1.36e-01	1.23e-01	1.72e-01	5.04e-01	5.29e-01	6.92e-01
BP GO:0031667 response to nutrient levels	6.70e-01	2.64e-01	2.43e-01	2.28e-01	2.65e-01	6.59e-01	6.40e-01	8.31e-01
AUC (200genes)	0.52	0.60	0.73	0.76	0.90	0.83	0.97	0.87

S5 Table Top GO terms for the starvation sensitive molecular signature. FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. Cov- number of Principal Components included in the linear-mixed effect model. BP- Biological Process; FDR p-values < 0.05 are labelled in bold; AUC- Area Under the Curve

GO term	FDR p (0 cov)	FDR p (1 cov)	FDR p (2 cov)	FDR p (3 cov)	FDR p (4 cov)	FDR p (5 cov)	FDR p (6 cov)	FDR p (7 cov)
BP GO:0003006 developmental process involved in reproduction	**2.46e-14**	**1.63e-03**	**9.16e-03**	**6.98e-03**	**3.18e-03**	**1.91e-02**	**1.06e-03**	5.25e-02
BP GO:0022412 cellular process involved in reproduction in multicellular organism	**3.05e-16**	**3.99e-04**	**2.78e-03**	**1.83e-03**	**7.14e-04**	**7.39e-03**	**5.46e-04**	**3.80e-02**
BP GO:0022414 reproductive process	**1.64e-12**	**1.37e-04**	**9.01e-04**	**1.56e-03**	**4.14e-04**	**1.91e-02**	**2.41e-03**	1.70e-01
BP GO:0032504 multicellular organism reproduction	1.00e+00	1.00e+00	1.00e+00	1.00e+00	1.00e+00	9.84e-01	9.97e-01	1.00e+00
BP GO:0000003 reproduction	1.00e+00	1.00e+00	9.99e-01	1.00e+00	1.00e+00	9.72e-01	9.93e-01	1.00e+00
CC GO:0030312 external encapsulating structure	**2.74e-05**	**3.14e-04**	3.06e-01	6.61e-01	9.20e-01	6.57e-01	5.86e-01	5.55e-01
CC GO:0042600 chorion	**1.45e-04**	**3.14e-04**	1.31e-01	4.56e-01	8.72e-01	6.76e-01	6.61e-01	6.18e-01
AUC (200genes)	0.85	0.63	0.51	0.56	0.63	0.83	0.57	0.58

S6 Table Top GO terms for the sterile molecular signature. FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. Cov- number of Principal Components included in the linear-mixed effect model. BP- Biological Process; CC- Cellular Component; FDR p-values < 0.05 are labelled in bold; AUC- Area Under the Curve

EBI Experiment ID	N con	N mut	EBI Control Experiment Factor Value	EBI Mutant Experiment Factor Value	μ Control Prob	μ Mutant Prob
E-GEOD-8775	2	2	female|wild type genotype	female|crol mutant	0.83	0.89
E-GEOD-18576	3	3	wild type genotype	DHR96 mutant	0.85	0.89
E-GEOD-24978	3	3	wild type genotype	rbf120a	0.93	0.87
E-GEOD-8775	2	2	female|wild type genotype	female|BG00817 mutant	0.74	0.78
E-MTAB-849	3	3	control	dL3MBT	0.74	0.76
E-GEOD-31564	3	3	eater-N RNAi|none|30 minute	eater-N RNAi|mixture of Gram-positive and Gram-negative bacteria|90 minute	0.73	0.75
E-MTAB-849	3	3	control	dLint1	0.72	0.68
E-GEOD-8775	2	2	female|wild type genotype	female|esg mutant	0.66	0.68
E-GEOD-24978	3	3	wild type genotype	rbf120a wtsX1Lats	0.77	0.67
E-GEOD-8775	2	2	female|wild type genotype	female|CG10990 mutant	0.66	0.66
E-GEOD-35439	3	3	wild type genotype	key1	0.68	0.66
E-GEOD-37701	3	3	vehicle	protocatechuic aldehyde 0.1	0.63	0.65

				millimolar		
E-GEOD-8775	2	2	female|wild type genotype	female|mub mutant	0.64	0.65
E-GEOD-8775	2	2	female|wild type genotype	female|CG9238 mutant	0.64	0.65
E-GEOD-31564	3	3	eater-N RNAi|none|30 minute	eater-N RNAi|mixture of Gram-positive and Gram-negative bacteria|30 minute	0.63	0.65
E-GEOD-25267	3	3	GMR-Gal4/+	GMR-Gal4/+; UAS-dE2F1,UAS-dDP/+	0.63	0.61
E-GEOD-37148	3	3	45 day|wild type drosophila SOD1 expressed in motoneurons	45 day|G85R expressed in motoneurons	0.63	0.61
E-GEOD-8775	2	2	male|wild type genotype	male|BG00817 mutant	0.63	0.61
E-GEOD-31564	3	3	pBR322 RNAi|none|30 minute	pBR322 RNAi|mixture of Gram-positive and Gram-negative bacteria|90 minute	0.58	0.60
E-GEOD-26717	3	3	w; sensDF2RES/+; sensE1/+	w; sensDF2RES/sensDF2RES; sensE1/sensE1	0.61	0.60
E-GEOD-10940	3	3	control|abdomen	Logjam mutant|abdomen	0.61	0.60
E-GEOD-31564	3	3	pBR322 RNAi|none|30 minute	pBR322 RNAi|mixture of Gram-positive and Gram-negative bacteria|30 minute	0.58	0.60
E-GEOD-8938	3	3	uninfected|2 to 5 hour	Leptopilina boulardi (strain Lb17)|2 to 5 hour	0.61	0.60
E-GEOD-24978	3	3	wild type genotype	wtsX1Lats	0.72	0.60
E-GEOD-26246	3	3	wild type genotype|2 day	Wild type Atro transgene|2 day	0.59	0.59
E-GEOD-31564	3	3	eater-N RNAi|none|30 minute	eater-N RNAi|mixture of Gram-positive and Gram-negative bacteria|180 minute	0.60	0.59
E-GEOD-25267	3	3	control	GMR-Gal4/UAS-miR-11; UAS-dE2F1,UAS-dDP/+	0.55	0.59
E-GEOD-10940	3	3	control|head/thorax	Logjam mutant|head/thorax	0.64	0.59
E-GEOD-14058	3	3	control	delg613 mutant	0.55	0.59
E-MEXP-2082	3	4	0 g gravitation (0g*)|19 degree Celsius|male|22 day	1 g gravitation control|19 degree Celsius|male|22 day	0.53	0.58

S7 Table Ranking EBI's ExpressionAtlas (starvation-sensitive molecular signature top 30 experiments). FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. FDR p-values < 0.05 are labelled in bold; Where there were multiple factor values, these are separated by "|". Factor values comprise genotype, treatment, etc.

EBI Experiment ID	N con	N mut	EBI Control Assay ID	EBI Mutant Assay ID	μ Control Prob	μ Mutant Prob
E-GEOD-48145	3	3	wild type genotype	ovoD mutant	1.00	0.99
E-GEOD-48145	3	3	wild type genotype	CA knockout with ovoD mutant	0.97	0.96
E-GEOD-10940	3	3	control|abdomen	Logjam mutant|abdomen	0.95	0.96
E-GEOD-10940	3	3	control|head/thorax	Logjam mutant|head/thorax	0.92	0.94
E-MTAB-3546	4	4	3 week|normal conditions	3 week|response to cold	0.89	0.91
E-GEOD-8775	2	2	female|wild type genotype	female|mub mutant	0.84	0.87
E-GEOD-8775	2	2	female|wild type genotype	female|esg mutant	0.74	0.85
E-GEOD-55187	3	4	wild type genotype|female	Sesb1 mutation|female	0.90	0.85
E-GEOD-26726	3	3	10 day|normal|wild type genotype|Canton-S	10 day|restricted|wild type genotype|Canton-S	0.88	0.84
E-GEOD-26726	3	3	40 day|normal|wild type genotype|Canton-S	40 day|restricted|wild type genotype|Canton-S	0.80	0.79
E-GEOD-12834	4	4	unmated	double mated	0.77	0.75
E-GEOD-55187	3	4	wild type genotype|male	Sesb1 mutation|male	0.76	0.71
E-MTAB-1066	3	3	wild type genotype	cycC mutant	0.68	0.70
E-GEOD-12834	4	4	unmated	single mated	0.71	0.70
E-GEOD-14531	3	3	normal|EP2449(precise excision)/KG08976(precise excision)	starvation|EP2449(precise excision)/KG08976(precise excision)	0.64	0.70
E-MEXP-2082	3	3	1 g gravitation control|14 degree Celsius|female|26 hour	1 g gravitation (1g*)|14 degree Celsius|female|26 hour	0.72	0.69
E-TABM-297	3	3	wild type genotype	24BGal4/UAS-lbe	0.64	0.69
E-MEXP-1208	3	3	wild type genotype	Ada2a delta 189	0.70	0.67
E-GEOD-30362	3	3	wild type genotype	pex1 homozygous mutant	0.64	0.67
E-GEOD-48997	5	5	wild type genotype	pri -/- mutant	0.66	0.67
E-GEOD-8938	3	3	uninfected|2 to 5 hour	Leptopilina boulardi (strain Lb17)|2 to 5 hour	0.60	0.66
E-MTAB-1066	3	3	wild type genotype	cdk8 mutant	0.62	0.66
E-GEOD-26726	3	3	10 day|normal|sir2 control|yw, w1118	10 day|normal|sir2 overexpression|yw, w1118	0.73	0.66
E-GEOD-8775	2	2	male|wild type genotype	male|mub mutant	0.45	0.65
E-GEOD-31875	4	4	control	elav-GAL4; UAS-DsRed-CAG100	0.57	0.65

				(5x)		
E-GEOD-44090	3	3	tub-Gal4	sage overexpressed tub-Gal4	0.56	0.65
E-MEXP-1208	3	3	wild type genotype	Gcn5[E333st] / Gcn5[E333st]	0.61	0.65
E-GEOD-25988	3	3	wild type genotype	BxJ mutant	0.61	0.65
E-MEXP-2011	3	3	wild type genotype	Nurf301-A/Nurf301-B/Nurf301-C knockout	0.57	0.65
E-GEOD-31564	3	3	eater-N RNAi|none|30 minute	eater-N RNAi|mixture of Gram-positive and Gram-negative bacteria|90 minute	0.65	0.64

S8 Table Ranking EBI's ExpressionAtlas (sterile molecular signature top 30 experiments). FDR p-value is the p-value corrected for multiple hypothesis testing using False Discovery Rate, accounting for all GO terms tested. FDR p-values < 0.05 are labelled in bold; Where there were multiple factor values, these are separated by "|". Factor values comprise genotype, treatment, etc.

References

1. Gautier L, Cope L, Bolstad BM, Irizarry RA. affy--analysis of Affymetrix GeneChip data at the probe level. Bioinformatics. 2004;20(3):307-15. doi: 10.1093/bioinformatics/btg405. PubMed PMID: 14960456.

2. Breslin T, Eden P, Krogh M. Comparing functional annotation analyses with Catmap. BMC Bioinformatics. 2004;5:193. Epub 2004/12/14. doi: 10.1186/1471-2105-5-193. PubMed PMID: 15588298; PubMed Central PMCID: PMCPMC543458.

3. Sieber MH, Thummel CS. The DHR96 nuclear receptor controls triacylglycerol homeostasis in Drosophila. Cell Metab. 2009;10(6):481-90. Epub 2009/12/01. doi: 10.1016/j.cmet.2009.10.010. PubMed PMID: 19945405; PubMed Central PMCID: PMCPMC2803078.

4. Varghese J, Lim SF, Cohen SM. Drosophila miR-14 regulates insulin production and metabolism through its target, sugarbabe. Genes Dev. 2010;24(24):2748-53. Epub 2010/12/17. doi: 10.1101/gad.1995910. PubMed PMID: 21159815; PubMed Central PMCID: PMCPMC3003191.

5. Nicolay BN, Bayarmagnai B, Islam AB, Lopez-Bigas N, Frolov MV. Cooperation between dE2F1 and Yki/Sd defines a distinct transcriptional program necessary to bypass cell cycle exit. Genes Dev. 2011;25(4):323-35. Epub 2011/02/18. doi: 10.1101/gad.1999211. PubMed PMID: 21325133; PubMed Central PMCID: PMCPMC3042156.

6. van Bergeijk P, Heimiller J, Uyetake L, Su TT. Genome-wide expression analysis identifies a modulator of ionizing radiation-induced p53-independent apoptosis in Drosophila melanogaster. PLoS One. 2012;7(5):e36539. Epub 2012/06/06. doi: 10.1371/journal.pone.0036539. PubMed PMID: 22666323; PubMed Central PMCID: PMCPMC3362589.

7. Magwire MM, Yamamoto A, Carbone MA, Roshina NV, Symonenko AV, Pasyukova EG, et al. Quantitative and molecular genetic analyses of mutations increasing Drosophila life span. PLoS Genet. 2010;6(7):e1001037. Epub 2010/08/06. doi: 10.1371/journal.pgen.1001037. PubMed PMID: 20686706; PubMed Central PMCID: PMCPMC2912381.

8. Boltz KA, Carney GE. Loss of p24 function in Drosophila melanogaster causes a stress response and increased levels of NF-kappaB-regulated gene products. BMC Genomics. 2008;9:212. Epub 2008/05/10. doi: 10.1186/1471-2164-9-212. PubMed PMID: 18466616; PubMed Central PMCID: PMCPMC2396179.

9. Yamamoto R, Bai H, Dolezal AG, Amdam G, Tatar M. Juvenile hormone regulation of Drosophila aging. BMC Biol. 2013;11:85. Epub 2013/07/20. doi: 10.1186/1741-7007-11-85. PubMed PMID: 23866071; PubMed Central PMCID: PMCPMC3726347.

10. Tootle TL, Williams D, Hubb A, Frederick R, Spradling A. Drosophila eggshell production: identification of new genes and coordination by Pxt. PLoS One. 2011;6(5):e19943. Epub 2011/06/04. doi: 10.1371/journal.pone.0019943. PubMed PMID: 21637834; PubMed Central PMCID: PMCPMC3102670.

11. Soshnev AA, Baxley RM, Manak JR, Tan K, Geyer PK. The insulator protein Suppressor of Hairy-wing is an essential transcriptional repressor in the Drosophila ovary. Development. 2013;140(17):3613-23. Epub 2013/07/26. doi: 10.1242/dev.094953. PubMed PMID: 23884443; PubMed Central PMCID: PMCPMC3742144.

12. Peters NC, Thayer NH, Kerr SA, Tompa M, Berg CA. Following the 'tracks': Tramtrack69 regulates epithelial tube expansion in the Drosophila ovary through Paxillin, Dynamin, and the homeobox protein Mirror. Dev Biol. 2013;378(2):154-69. Epub 2013/04/03. doi: 10.1016/j.ydbio.2013.03.017. PubMed PMID: 23545328; PubMed Central PMCID: PMCPMC4141043.

13. Zamparini AL, Davis MY, Malone CD, Vieira E, Zavadil J, Sachidanandam R, et al. Vreteno, a gonad-specific protein, is essential for germline development and primary piRNA biogenesis in Drosophila. Development. 2011;138(18):4039-50. Epub 2011/08/13. doi: 10.1242/dev.069187. PubMed PMID: 21831924; PubMed Central PMCID: PMCPMC3160098.