File S1. Supplementary Materials, Methods and Results for Pitchers et al., 2018

A multivariate genome-wide association study of wing shape in *Drosophila melanogaster*

William Pitchers, Jessica Nye, Eladio J. Márquez, Alycia Kowalski, Ian Dworkin, and David Houle

Supplementary Materials and Methods

Rearing, handling of flies and imaging of wings

In both labs, each line was reared in vials for at least one generation in the experimental conditions prior to the start of the experiment, and then reared at low density.

In the Houle lab, flies were reared in a series of 10 temporal blocks over a 14 month period. Twenty-four lines were reared and measured in two or more blocks. Each vial was initiated with four parental males and females per vial, who were allowed to lay eggs for three to four days, until visual inspection suggested that a sufficient number of larvae had been obtained. The parents were discarded, and the experimental progeny were transferred to new vials containing no more than 20 adults to avoid wing damage due to overcrowding. The dorsal surface of the left wings of live flies were imaged using the 'Wingmachine' system (Houle et al. 2003) using Optem macroscopes with an integrated camera. Annotation, scale information, images as grey-scale TIFF files and guide landmarks were recorded using Image-Pro Plus software (Versions 4, 5 and 6). We sought to obtain images from at least 40 flies per line (20 of each sex). After excluding damaged wings and unsplinable images, data was obtained for a total of 7878 wings from 182 lines, for a mean of 43.3 wings per line. We obtained data from fewer than 40 wings in 23 lines, and from less than 30 wings for only four lines.

In the Dworkin lab, flies were reared in an incomplete balanced block design. Blocks consisted of two replicate bottles of each line reared using food made from the same batch. Each block contained lines that had been reared previously for comparison. Media was physically scored and live yeast was added prior to introducing adult flies to promote egg-laying. Flies were reared separately at 24°C, 60% relative humidity at low density (10 pairs of adult flies per bottle) in a Percival incubator. After 3-5 days (depending on egg density) adults were transferred to new bottles. While eggs were not counted, density was controlled for qualitatively, by removing adults once the desired low egg density was approximately achieved. For those lines with low fecundity, adults were left a few days longer (up to 7 days). After 3-5 days in the second bottle, adult flies were discarded. Water, yeast and paper towel were added to bottles as needed to provide an optimal environment for the larvae. After eclosion and hardening of the cuticle, flies were stored in 70% ethanol at room temperature prior to dissection. Bottles were checked daily as needed until a sufficient number of flies was collected. We dissected between 20–24 wings (left wing of each fly) for each replicate/sex/line.

Dworkin lab wings were imaged at 40X magnification using an Olympus DP30BW camera mounted on an Olympus BX51 microscope and controlled with DP controller software V3.1.1. Images were saved in greyscale as TIFF files. We used the program 'tpsDig2' (Rohlf 2011) to record annotation and the guide landmarks. After excluding damaged wings or unsplinable images, data was obtained for a total of 16,272 wings from 165 lines, for a mean number of wings/line of 98.6. We obtained data from fewer than 40 wings in 9 lines, and from less than 30 wings for only four lines.

Handling of morphometric data

Once the data for the 66,890 wings was superimposed as described in Materials and Methods, outliers for the superimposed data were detected in CPR (Márquez 2012-2014), and then re-examined in Wings 3.72 to allow us to determine whether they represented an unusual wing, or mis-splined specimens, which were corrected. Occasionally a very unusual wing was removed from the data set as an outlier. In all cases, these outlier wings were more than 4 S.D.

in Mahalanobis distance from the multivariate mean. The positions of the semi-landmarks were slid along each wing vein (or margin) segment to minimize deviation along the segment. To put numerical results on a more convenient scale we multiplied shape (Procrustes) coordinates by 100.

The 96 superimposed x and y coordinates from the 48 points recorded generate less than 96 dimensional data, for two reasons. First, each semi-landmark is approximately constrained to lie on a 1-dimensional function, so contributes only 1 degree of freedom (df) to the data. Second, Procrustes superimposition uses 3 df for rotation and translation, and transfers size to a new 1 df variable, centroid size. A 58=2 X 48 – (4+34) dimensional space thus captures shape variation. The shape data was projected into a 58-dimensional space using principal components analysis of the combined DGRP and validation data, with no adjustment for the fixed sex and lab effects. Thus, PC1 has a large contribution of variation due to the effects of sex. The scores on the first 58 eigenvectors, plus ln centroid size were used for subsequent analyses.

Univariate residuals for shape were generally heavy-tailed (average kurtosis=2.7, defining the kurtosis of a normal distribution as 0). Residuals for principal components 1 and 2 were slightly right-skewed (skew 0.22 and 0.16 respectively), while the remaining shape variables showed no notable skew. Sex-specific ln(centroid size) was heavy tailed (kurtosis=0.63) and left-skewed (skew=-0.53). Tests for normality of univariate residuals always rejected the normal distribution, which is expected given the large sample size. Association analyses were done on lab, sex and block means, so these departures from normality should have no effect on our results.

Clustering significant SNPs

We quantified LD as the squared gametic correlation between sites

$$r^2_{LD} = \frac{D^2}{p_1 q_1 p_2 q_2}$$

where p_1, p_2, q_1, and q_2 are the major and minor allele frequencies at the two sites, and $D = x_{11} - p_1 q_1$, where x_{11} is the frequency of gametes carrying both the alleles indexed by the frequencies alleles p_1 and q_1 (Weir 1996).

We performed an LD-based cluster analysis on the 2,396 SNPs judged to be significant in our MANOVA-based association tests. To find an initial set of clusters, we used the SAS/FASTCLUS Procedure (SAS 9.3), which uses q vectors of SNP genotypes as seeds to group input SNPs into up to k clusters with a radial spread equal to R, where k and R are user-defined parameters. In a first run, we had FASTCLUS impute missing genotype data, and instructed it to choose a large number of groups k=2000. In a second run, we submitted the previously imputed data to FASTCLUS, and save the output as seeds for subsequent iterations of the same algorithm. We then iterated this step until both the number of clusters and a least squares optimization criterion plateaued. We chose the radius R for our clusters to match the r^2>0.5 cutoff. From the law of cosines, the distance, d, between two SNP vectors is related to their correlation by $d = \sqrt{2(1-r)}$, leading to R = 0.7654.

The above algorithm does not ensure that the clusters identified are discrete. To compensate for this, we carried out a second, refinement phase. This phase consists of three steps: first, we scan each non-singleton cluster to determine whether any of its members do not conform to the clustering criterion (i.e., its squared correlation with every other member of the cluster does not equal or exceed 0.5). SNPs that violate the criterion are marked as singletons for

subsequent processing; second, squared correlations between singleton and all other SNPs are computed to allow for orphan SNPs to join established clusters, or for pairs of singletons to cluster when the $r^2 > 0.5$ criterion is met. If a SNP is correlated with more than one cluster, it is allowed to join the cluster with the most members; finally, the last step merges clusters with highly correlated SNPs. Specifically, two clusters were combined into a single cluster when the minimum of the maximum squared correlations computed between all pairs of members of different clusters exceeds 0.5. All of these steps were iterated until convergence. The result from our algorithm is a series of clusters comprising SNPs each satisfying the correlation criteria $r^2 \geq 0.5$ with at least some other SNPs within the cluster, and $r^2 < 0.5$ with every SNP that does not belong in the same cluster.

For the analyses below for the MANOVAs, they were written in SAS macros and were run at the High Performance Computing facility at North Carolina State University, the Research Computing Center at Florida State University, and a standalone Linux server at the Biological Science Department at Florida State University.

Testing significance in the MANOVAs

To approximate the mixed model tests in the MANOVA analyses, we used the following procedure. We first estimated the sum of squares and cross-products (SSCP) matrices using a least squares method in SAS Proc GLM, designating terms involving line nested in SNP as random with variates weighted by their sample sizes. Because sample sizes over labs and sexes were always unbalanced, the denominators of within-group SSCP matrices, **W,** were assembled as weighted averages of the SSCP matrices obtained in this first analysis. The weights were obtained from the coefficients of the expected mean squares calculated in a univariate analysis of the same SNP in SAS Proc GLM using the Random/Test option. We assessed the statistical significance of model terms using an *F*-distributed statistic based on Wilks' Λ (Rao 1973), computed as $\Lambda = 1/\det(\mathbf{I} + \boldsymbol{W}^{-1}\boldsymbol{B})$, where **B** is the between-group SSCP matrix.

LASSO regressions

For the *f*th focal SNP, we included as predictors the family of *t* SNP variants confounded with the focal SNP due to proximity or LD, plus scores on the 13 significant population structure principal components. Thus, the total number of predictors is $p = t+14$. The median *t* is 65, and the range is from 0 to 5291. The total number of SNPs considered in each model (including the focal SNP) is shown in Column W "N SNPs considered" in File S3. Missing genotype calls were not imputed at focal SNP *f*, but missing calls in all *t* non-focal SNPs were imputed to the allele frequency of the *t*th SNP. Correspondingly, the dependent variable matrix, $\bar{\mathbf{Y}}_h$, includes only the least-squares line means for wing shape and size for lines with non-missing data for SNP *f*.

The LASSO algorithm solves

$$\hat{\boldsymbol{\beta}}_{f\lambda} = \arg\min_{\boldsymbol{\beta},\boldsymbol{\beta}_0} \left[\frac{1}{2n} \sum_{h=1}^{n} \left(\bar{\mathbf{Y}}_h - \boldsymbol{\beta}_{f0} - \boldsymbol{\beta}_f{}^T \mathbf{X}_h \right)^2 + \lambda_f \sum_{j=1}^{p} \left| \boldsymbol{\beta}_{fj} \right| \right] \quad (1)$$

where *n* is the number of DGRP lines with genotype data for the *f*th focal SNP, , $\mathbf{X}_h$ is the vector of *p* predictor variables for DGRP line *h*, $\boldsymbol{\beta}_{f0}$ is the fitted intercept vector, $\boldsymbol{\beta}_f$ is the *p*-predictor

by 59-phenotype matrix of regression coefficients, $\left|\boldsymbol{\beta}_{f\square j}\right|$ is the L1 norm of the vector of coefficients for the *j*th predictor variable, and λ_f is a penalty factor that determines the amount of shrinkage from the least-squares solution that is imposed for SNP model *f*. The L1 norm is the sum of the absolute values of the coefficients.

The first term in brackets in equation (2) is standard least-squares measure of fit, while the second term shrinks the lengths of the inferred vectors to a degree dependent on the magnitude of λ_f. For each SNP model, a value of λ_f was chosen by 5-fold cross-validation. The choice of the L1 norm shrinks the vectors $\boldsymbol{\beta}_{f\square j}$ non-uniformly, such that for sufficiently large values of λ some, or even all, prediction vectors $\boldsymbol{\beta}_{f\square j}$ have 0 length, resulting in variable selection as well as shrinkage.

To check the stability of the LASSO solutions we used an elastic net regularization (Zou and Hastie 2005) with α=0.95 (95% of the weight on the L1 norm and 5% on the L2 norm).

Geneswitch knockdowns

We backcrossed the Tub-5 GS construct into a wild-type Oregon R (OreR+) background before these experiments. The Tub-5 GS driver used in these experiments is strongly inducible by mifepristone, although there is some residual Gal4 activity in the absence of Mifepristone. For each concentration of mifepristone, four replicate vials were set up; a fifth replicate was set up for 2.7 µM due to low survivorship in many experiments. We placed ten virgin females with five males in each vial.

Three different control crosses with their respective reciprocals were also set up: Tub-5GS x the appropriate RNAi background (either *yv* or w^{1118}), UAS-[GOI]RNAi x OreR+, and RNAi background (either *yv* or w^{1118}) x OreR+. Reciprocal and control crosses were set up at the same time on medium from the same batch. After six days, all the parents were moved to fresh vials with the appropriate mifepristone concentration, and then discarded after an additional six days. Offspring were moved to vials with fresh food without mifepristone, sorted by sex, and their wings were imaged at least two days after eclosion. We imaged wings from 20 F_1 females and males from each treatment in each reciprocal cross.

The distribution of within reciprocal, sex and treatment data was frequently heteroscedastic; higher mifepristone RNAi treatments generally had higher variance, often showing outliers along the major axis of RNAi effects. Consequently, we analyzed the within-sex-treatment-reciprocal medians. Further analyses (in prep.) of control and experimental data suggests that mifepristone has background-specific effects on wing shape across UAS-[GOI]RNAi crosses, and data were adjusted for these effects before further analyses. Finally, we calculated the linear effect of mifepristone on the 58 shape dimensions in a linear model with sex and reciprocal as categorical effects and mifepristone as a continuous predictor. In some cases, the reciprocals differed significantly in their effects, and were analyzed separately. These are designated by the sex of the Tub-5 GS parent in File S4. The parameters of the multivariate regression of mifepristone were retained as the effect vector of the manipulated gene of interest.

Vector comparisons

SNP effects and gene knockdowns result in vectors of phenotypic effects in phenotypic space. We used vector correlations to compare the directions of vectors. The correlation of column vectors *x* and *y* is

$$r = \frac{x^T y}{\|x\| \cdot \|y\|}$$

where ^{T} indicates transpose and $\|x\|$ denotes the length (2-norm) of vector x. Like all correlations, $-1 \le r \le 1$. The sign of the correlation is arbitrary, because we could take either the major or the minor allele as the reference, so we report the absolute value of vector correlations. A correlation of 1 means that the vectors point in the same direction, while r=0 means that the two vectors are orthogonal (at 90 degrees).

Supplementary Results

Relatedness among phenotyped lines

Coancestries among the 184 phenotyped DGRP lines were estimated from the genomic data using a principal-component-based approach (Patterson *et al.* 2006). Thirty-one of the nearly 17,000 line pairs (0.2%) had coancestries of 0.2 or more, and probably reflect sampling of close relatives from the original population. Four-hundred thirty-five line pairs (2.6%) have coancestries of 0.02 or more. These are strongly enriched for pairs of lines that are both homozygous for the three common cosmopolitan inversions present in seven or more DGRP lines (In(2L)t, In(2R)NS, In(3R)Mo), and therefore probably represent pairs that share sub chromosome-arm scale haplotypes.

The first 13 genomic principal components had eigenvalues that were significantly greater than the value from the Tracy-Widom distribution expected if lines were unrelated. The contrast between lines carrying the common cosmopolitan inversions and those with the standard karyotype dominates the distribution on three of these PCs (eigenvectors). The remaining PCs with significant variation are dominated by small groups of related lines. Of the 31 pairs of lines with coancestries>0.2, twenty-five are clear outliers in bivariate plots of scores on the significant PCs.

Linkage disequilibrium among phenotyped lines

We enumerated all of the highly correlated ($r^2 \ge 0.5$) SNP pairs for the 184 phenotyped DGRP lines using the approach of Houle and Márquez (2015). The average number of highly correlated SNPs with each MANOVA-significant SNP is very large at low MAF, but still substantial at high MAF, as shown in Figure S2. The probability is greater than 0.5 that at least one other SNP in the genome is highly correlated with each significant SNP at all MAF, as shown in Figure S3. More striking is the fact that SNPs with low MAF have a substantial probability of being correlated with SNPs more than 100kb distant. While there is a difference in the mean number of correlated SNPs between regions inside and outside of inversions, the probability that there is at least one such correlation is affected very little by inversions (Houle and Márquez 2015). Instead, we interpret the bulk of this LD as being due to 'rarity disequilibrium' (Houle and Márquez 2015) due to the large number of low MAF SNPs, and the relatively few combinations of line genotypes that can generate a low MAF as opposed to a high MAF. Twenty-five percent of the SNPs that we analyzed have MAF<0.06, and 50% have MAF<0.137.

Table S3 also includes several variables to help understand whether each SNP is likely to be a causal SNP, and whether the locus that is closest to that SNP is likely to be affected even if the wrong causal SNP was selected by the LASSO. Most important are the number of perfectly correlated SNPs (nperfglm), their identities, and the maximum distance between these SNPs (maxpdist). In most cases perfectly correlated SNPs map very close, so that even if the SNP is not causal, it will be annotated to the same gene. Second, we give the size of the cluster of significant genes in high LD, and the maximum distance among the members of that cluster.

Validating the LASSO

We investigated the behavior of the LASSO as a tool for primary screening of SNPs as candidates by analyzing 2,396 MANOVA-insignificant SNPs chosen at random from >2.5 million SNPs previously analyzed. Eighty-eight percent of random SNPs had non-0 effects in a multiple regression with structure PC scores as covariates, while 48% retained non-zero effects when both the family of highly correlated SNPs and structure PC scores are included as covariates. The effect size was median effect size following LASSO analysis was substantially smaller for random SNPs (median 0.13) than for MANOVA-significant SNPs (median 0.22). This indicates that the LASSO by itself is far more liberal than the MANOVA in implicating SNPs as potentially causal. Consequently, we restrict its use to controlling effect sizes and compensating for population structure and LD.

To check the numerical stability of the LASSO results on the MANOVA-significant SNPs, we compared those results to elastic net (Zou and Hastie 2005) results with a 0.95 weight on L1 and 0.05 on L2 norms. The Spearman correlation of vector lengths for focal SNPs was 0.93, and just 4% of focal predictors had a length>0 in one analysis and zero in the other. This strong similarity indicates that LASSO estimates are numerically stable.

Literature Cited

Houle, D., and E. Márquez, 2015 Linkage disequilibrium and inversion-typing of the *Drosophila melanogaster* Genome Reference Panel. G3: Genes|Genomes|Genetics 5**:** 1695-1701.

Houle, D., J. Mezey, P. Galpern and A. Carter, 2003 Automated measurement of *Drosophila* wings. BMC Evolutionary Biology 3**:** 25.

Márquez, E. J., 2012-2014 CPR: Using Drosophila Wing Shape Data, pp.

Patterson, N., A. L. Price and D. Reich, 2006 Population structure and eigenanalysis. PLOS Genetics.

Rao, C. R., 1973 *Linear Statistical Inference*. Wiley, New York, NY.

Rohlf, F. J., 2011 tpsRegr, pp., State University of New York, Stony Brook.

Weir, B. S., 1996 *Genetic Data Analysis II*. Sinauer, Sunderland, MA.

Zou, H., and T. Hastie, 2005 Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67**:** 301-320.