Description of Programs and data furnished with "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

The results in this paper were mostly obtained using analyses in the SAS system. SAS programs have the .sas extension. R scripts have a .R extension.

Data files are in various formats:

Sas7bdat are SAS format data sets. These can be read with R as well as SAS.
.csv are comma delimited
.dat files are tab delimited
.txt files are space delimited

The data and programs that produced the results in this paper are organized into nine folders. To replicate all the results, execute the programs in these folders in the following order. This describes the contents of those folders.

Data

This folder contains data files used in multiple analyses:

1. Multivariate phenotypes – DGRP_Wing_Data_HouleDworkin.csv
2. Eigenvectors for DGRP phenotypes – June. Interlabeigjun.sas7bdat
3. Eigenvectors for the dictionary Nov. Interlabeignov.sas7bdat
4. Genotypes of FDR5% significant SNPs. fdr5sigsnps.sas7bdat
5. MANOVA-significant SNPs and summary statistics – gof2hq3_05.sas7bdat
6. Glmnet and MANOVA results summary. GLMNETwithMan.sas7bdat

SNPdata

Program:

ReadSNPsquality20.sas – This script read the freeze2.vcf file and generates many data sets used by programs in virtually every other folder. The freeze2.vcf file must be obtained from ftp://ftp.hgsc.bcm.edu/DGRP/freeze2_Feb_2013/vcf_files/freeze2.vcf.gz

Data:

1. Inversion typing of DGRP lines from Houle and Marquez, 2015 - Hetinversionscores205.csv
2. Wolbachia status of DGRP lines from Huang et al. 2014 – DGRPWolbachia2014

MANOVA

Program:

MANOVAHighQchunk.sas. Program to do a MANOVA on all SNPs in the DGRP with MAC when inversion genotypes are masked. The script enables calculating these results in 1,000 SNP chunks, for distributing over many processors. Also generates the SAS data set data/interlabsizejun.sas7bdat, which is used by other programs.

LinkageDiseq

Programs:

1. CHCorrHQ184sig.sas calculates the linkage disequilbrium between all pairs of SNPs, and retains a list of those where r2>0.5. Running this program takes a long time.
2. hqcountsigcorrs184.sas summarizes the LD calculated in the previous program.

ClusterAnalysis

Program:

Fastcluster_sigset.sas. Produces clusters used to investigate LD of significant SNPs. The final results of this analysis involved more iterations of this general approach.

GLMnet

Programs:

1. MultregSNP4GLMnet.sas. Produces files of the correlated SNPs with each MANOVA-significant SNPs formatted for analysis in the following R program.
2. GLMnetbycomplexSNP.R. Performs LASSO regressions of phenotypes on the SNPs in LD with each MANOVA-significant SNPs using the files produced by MultregSNP4GLMnet.sas. Also summarizes the results, in particular the regularized effect of each focal SNP.

Data:

The scores on the population structure PCs calculated in smartpca - DGRPallmac5evec.txt

Dictionary

Programs:

1. pipeline_0_ControlsAnalysis.sas. Calculates corrections to knockdown for effects in controls appropriate to each genotype.
2. pipeline_2_MANOVA_RNAi_preculled.sas. Calculates the regressions of phenotype on mifepristone treatment level for each of the RNAi knockdowns.
3. GLMvDict2018.sas. Comparisons of Dictionary vectors with the inferred effects of LASSO-significant SNPs.

Data:

1. Dictionary phenotype data input for each of the pipeline files above. Dictionary_slid_lmdata.dat
2. Precalculated dictionary vectors from the pipeline_2_MANOVA_RNAi_preculled.sas program `pooledregestimates_preculled.sas7bdat`

MENC2

Program:

MENCanalysis.sas. Performs the analyses of the Maine/North Carolina 2 data.

Data:

1. Phenotype data - Maine&NC_output.csv.
2. Genotype data - MNC_genotyping_processed.csv