From GWAS to signal validation: An approach for estimating SNP genetic effects while preserving genomic context.

Scott Wolf[1,2*], Varada Abhyankar[1,2*], Diogo Melo[1,2], Julien F. Ayroles[1,2,†], and Luisa F. Pallares[1,2,3,†]

1 - Lewis-Sigler Institute for Integrative Genomics, Princeton University;
2 - Department of Ecology and Evolutionary Biology, Princeton University;
3 - Current address: Friedrich Miescher Laboratory, Max Planck Society;
*These authors contributed equally to this work.

† Correspondence: Julien Ayroles *jayroles@princeton.edu* and Luisa F. Pallares *luisa.pallares@tuebingen.mpg.de*

Abstract

Validating associations between genotypic and phenotypic variation remains a challenge, despite advancements in association studies. Common approaches for signal validation rely on gene-level perturbations, such as loss-of-function mutations or RNAi, which test the effect of genetic modifications usually not observed in nature. CRISPR-based methods can validate associations at the SNP level, but have significant drawbacks, including resulting off-target effects and being both time-consuming and expensive. Both approaches usually modify the genome of a single genetic background, limiting the generalizability of experiments. To address these challenges, we present a simple, low-cost experimental scheme for validating genetic associations at the SNP level in outbred populations. The approach involves genotyping live outbred individuals at a focal SNP, crossing homozygous individuals with the same genotype at that locus, and contrasting phenotypes across resulting synthetic outbred populations. We tested this method in *Drosophila melanogaster*, measuring the longevity effects of a polymorphism at a naturally-segregating cis-eQTL for the *midway* gene. Our results demonstrate the utility of this method in SNP-level validation of naturally occurring genetic variation regulating complex traits. This method provides a bridge between the statistical discovery of genotype-phenotype associations and their validation in the natural context of heterogeneous genomic contexts.

Keywords: GWAS validation, midway, genetic effect size

Introduction

Understanding how genetic variation regulates phenotypic differences between individuals is one of the main challenges of modern biology. Advances in genetic mapping methods, such as GWAS, have identified thousands of genetic variants associated with variation in complex traits across a wide range of organisms (Visscher *et al.* 2017; Saul *et al.* 2019; Alsheikh *et al.* 2022; Pallares *et al.* 2023). Although mapping studies have contributed substantially to elucidating the genetic architecture of complex traits, the validation of candidate variants has dramatically lagged behind (Gallagher & Chen-Plotkin 2018) and remains particularly difficult.

Methods for validating genetic association signals fall under two broad categories, methods that operate at the gene level and those that operate at the polymorphism level. Currently, the most common approaches to validate candidate genotype-phenotype associations rely on gene level experimental perturbation using loss-of-function mutations or RNAi constructs, comparing organisms with and without a functional copy of a gene of interest and assessing the phenotypic consequences (Housden *et al.* 2017; Zimmer *et al.* 2019; Bellen, Wangler & Yamamoto 2019). Despite their immense usefulness in defining gene function, these approaches have important limitations when validating genotype-phenotype associations given how disconnected the validation context is from the association study that generated the signal of interest. In addition, the genetic effects assessed with such approaches usually represent genetic variation (e.g., null KO alleles) that does not segregate in natural populations. Recent developments in CRISPR technology provide an alternative that has revolutionized the field. They have allowed for more realistic experiments in which specific single nucleotides can be targeted and replaced to assess the phenotypic effects of alternative variants (Ramaekers *et al.* 2019; Hoedjes *et al.* 2023). These CRISPR-based validation methods still present major drawbacks. First, off-target effects can be substantial and difficult to assess (Schaefer *et al.* 2017; Lessard *et al.* 2017). Secondly, CRISPR assays can be costly, time-consuming, and not readily available depending on the target organism. Critically, both loss-of-function and CRISPR-based studies usually modify the genome of a single genetic background, often using inbred lines (Mokashi *et al.* 2021). We argue that, given that genetic effects are often background-dependent (Chandler, Chari & Dworkin 2013), working on a single genetic background limits the inference and generalizability of the validation results.

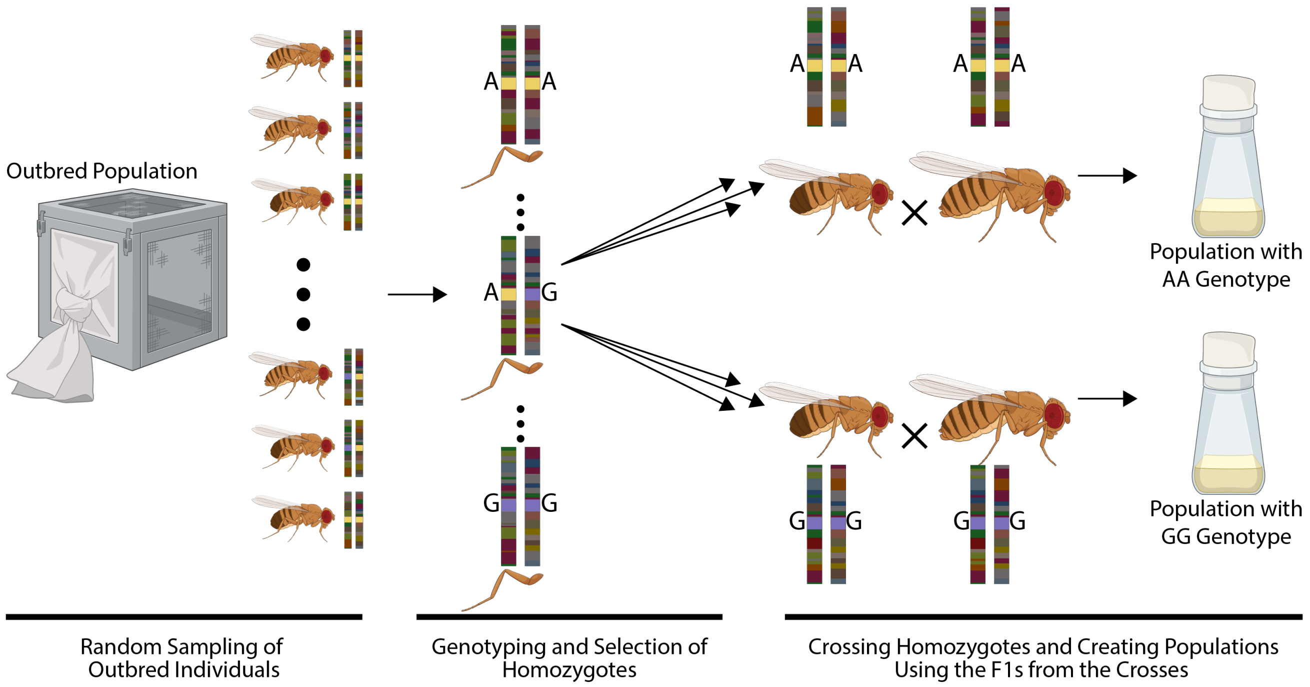

Figure 1: Experimental approach used to create populations with diverse genetic backgrounds fixed at a focal SNP.

Virgin flies from both sexes are collected from an outbred fly population harboring the polymorphisms of interest. While anesthetized with CO2, a single leg is removed from each virgin fly, and individuals are placed into separate vials waiting for the legs to be genotyped. DNA is extracted from each leg and genotyped individually at the focal SNP using PCR and amplicon sequencing. Once males and females homozygous for the same genotype are identified, they are crossed (within genotype) and an equal number of offspring from each cross are transferred to a bottle to ensure that the genomic background of each founder individual is represented in the resulting population. In each bottle, all individuals are fixed for a given genotype at the locus of interest and the populations maintain genetic diversity present in the initial outbred population. The resulting populations are ready to be phenotyped for any trait of interest.

To address these challenges, we propose a simple, low-cost experimental scheme for the validation of genetic association at the polymorphism level in outbred populations, thus preserving variation in the genetic background across individual. The approach relies on three steps: 1) genotyping live outbred individuals at a focal SNP, 2) crossing homozygous individuals that share the same genotype at the locus of interest, and maintaining them as outbred populations, and 3) measuring and comparing phenotypes of interest across the resulting synthetic outbred populations that are fixed for the polymorphism of interest but have randomized genetic backgrounds (Figure 1).

We demonstrate the utility of this approach in *Drosophila melanogaster* focusing on the phenotypic effects of a cis-regulatory polymorphism for *midway.* The *midway* gene is known in *Drosophila* to be involved in lipid metabolism (Buszczak *et al.* 2002; Tian *et al.* 2011; Girard *et al.* 2021), immune function (Tschapalda *et al.* 2016), and female fertility (Schüpbach & Wieschaus 1991). And, it was recently identified as a lifespan-regulating gene in a study investigating the genetic basis for variation in lifespan in *D. melanogaster* (Pallares *et al.* 2023)Furthermore, the mammalian ortholog of *midway*, *DGAT1*, has previously been linked with longevity in mice (Streeper *et al.* 2012). While Pallares *et al.* identified that *midway* plays a role in regulating lifespan using a loss-of-function mutation in an inbred line, we do not know if or how naturally-segregating genetic variants linked to *midway* expression are indeed responsible for lifespan variation in this species. Simultaneously, another study aimed at mapping genetic variants that regulate variation in gene expression levels genome-wide (i.e. eQTLs) in an outbred fly population, identified a candidate cis-eQTL upstream of *midway* (Pallares *et al.* unpublished). We used the validation paradigm outlined above to determine whether this regulatory variant identified for *midway* contributes to variation in lifespan.

To accomplish this, we created two fly populations, each homozygous for one of the eQTL focal alleles (i.e. AA vs GG) and, for each, quantified variation in longevity. We were able to validate the role this specific eQTL variant plays in lifespan and indirectly confirm that transcriptional variation in *midway* drives its effect on lifespan. Our experimental approach allowed us to validate a statistical discovery across a randomized set of genetic backgrounds in an outbred population. This study is one of the few validations of naturally occurring genetic variation controlling complex traits at the single nucleotide level.

Methods and Results

Drosophila melanogaster outbred populations

The candidate SNP for the gene *midway* we are investigating was identified in an eQTL mapping study that used an outbred mapping population of *D. melanogaster* derived from the

Netherlands (*Nex* from now on) (Pallares *et al.*, unpublished). This population was generated by crossing 15 inbred Global Diversity Lines (Grenier *et al.* 2015), followed by ~130 generations of recombination. The identification and initial validation of *midway* as a regulator of *Drosophila* lifespan was recently published as part of a GWAS-like study that used an outbred population derived from 600 isofemale lines caught in Princeton, NJ (Pallares *et al.* 2023). To validate the effect of the *midway* eQTL on longevity, we used the same outbred *Nex* population where the eQTL was initially discovered. Flies were maintained at 25°C, 65% relative humidity, and a 12h:12h light:dark cycle, and were fed media with the following composition: 1% agar, 8.3% glucose, 8.3% yeast, 0.41% phosphoric acid (7%), and 0.41% propionic acid (83.6%).

Experimental populations used for SNP validation

To evaluate the effect of the target SNP on longevity, we followed the scheme described in Figure 1 and created two synthetic outbred populations with hundreds of individuals homozygous for each *midway* allele (AA or GG) identified in (Pallares *et al.*, unpublished). The candidate SNP eQTL in *midway* is located at 2L:16,812,901 (3759 bp upstream of the *midway* gene) and has a minor allele frequency of 16% in the *Nex* population (eqtl was identified in pallares unpublished.).

We randomly selected virgin male and female flies from the outbred *Nex* population to identify flies homozygous at the focal locus. While the flies were anesthetized with CO2, we removed one leg from each individual for DNA extraction and genotyping (removing a leg does not alter viability). The flies were kept in separate vials until their genotypes were determined, after which they were paired with flies of the same genotype and mated. DNA was isolated with the QuickExtract™ DNA Extraction (cat no. QE09050) and the region around the focal *midway* polymorphism was amplified using the following primers: Fwd-TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGGAGG AGCCACCAAGTGTTGT and Rev-GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGATC GAACTTCTCTCGCGACT. The sequencing library for these amplicons was generated using Illumina i5 and i7 primers. After bead cleaning, the library was sequenced using a MiSeq Nano flow cell (150bp PE reads) at the Genomics Core Facility of Princeton University. Genotypes were called using bcftools mpileup (Li 2011) with parameters -Ou -B -q 60 -Q 28 -d 1000 -T -b files | bcftools call -Ou -m -o. After identifying the homozygous individuals for the two alleles, five crosses were set up in vials, each with one male and one female of the same genotype. From each cross of the same genotype, 10 male and 10 female offspring were combined in a bottle and allowed to mate freely for two generations before starting the survival assay. This design ensures that the genotype of each founding parent is represented in the synthetic outbred population (see Figure 1).

To assess whether this candidate *midway* variant is associated with variation in lifespan, we performed survival assays in both experimental populations (i.e. one for each alternative genotype). One hundred males and one hundred females (1-2 days old) for each homozygous *midway* cis-eQTL genotype were distributed across 10 vials with 10 individuals of each sex in each vial. Flies were transferred onto fresh media every 3 days, and survival was monitored each time flies were transferred until the last fly died on day 79.

We performed a proportional hazard regression using the empirical survival distribution using the statsmodels Python package (Cox 1972; Seabold & Perktold 2010). Given that male and female *D. melanogaster* differ in average lifespan (Austad & Fischer 2016), we were interested in whether the focal genotype modulated this difference. Therefore, our model was initially specified as:

$$lifespan_i \sim \beta_0 + \beta_1 \delta_i^{Male} + \beta_2 \delta_i^{Genotype\ AA} + \beta_3 \delta_i^{Male} \delta_i^{Genotype\ AA} + \epsilon_i$$

We did not find significant genotype-by-sex effects (95%, CI 0.6635 - 1.4597; p = 0.9365). We then proceeded with the following model:

$$lifespan_i \sim \beta_0 + \beta_1 \delta_i^{Male} + \beta_2 \delta_i^{Genotype\ AA} + \epsilon_i$$

Our results show that after accounting for sex, flies homozygous for the minor GG genotype live longer than flies homozygotes for the major allele AA genotype. Cox's proportional hazard regression using the final model yields a hazard ratio for genotype AA of 0.6767 (95% CI, 0.5544 - 0.8261; p =0.0001).

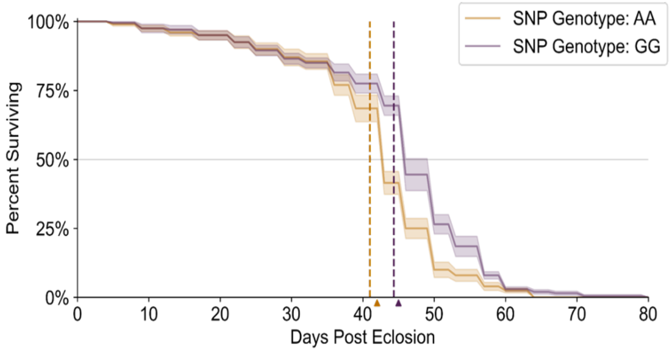

Figure 2: Effect of the midway regulatory polymorphism on D. melanogaster survival.

Survival distribution stratified across *midway* genotype. Shaded regions show standard error, and the dashed vertical lines show mean lifespan by genotype (AA mean lifespan 40.99 days, GG mean lifespan 44.33 days). T_{50}, the last day when 50% of individuals are alive, is denoted by wedges along the x-axis. T_{50} is day 42 for genotype AA and day 45 for genotype GG. Survival data for 200 individuals from each *midway* cis-eQTL genotype is included.

Discussion

The fast-growing power of mapping studies has produced an ever-expanding list of candidate genetic variants associated with complex trait variation. This large number of candidates with small effect sizes presents a formidable challenge for validating these signals, especially if the effects are context-dependent. To overcome some of these challenges, we developed an alternative experimental approach for candidate SNP validation that doesn't rely on genetic engineering (e.g. CRIPRS or RNAi). Our approach simply relies on generating populations of individuals fixed for alternative genotypes of a target polymorphism but randomized across diverse genetic backgrounds. These synthetic outbred populations enable the validation or estimation of additive allelic effects of variants in a natural genetic context. This approach is facilitated by the low cost and high throughput of amplicon sequencing using Illumina's MiSeq sequencer (i.e. hundreds of individuals can be genotyped-by-sequencing in a couple of days and at a low cost).

One of the key benefits of this approach is the gain in statistical power provided by the change in allele frequency between alternative genotypes in the synthetic outbred populations. While in the initial mapping population, potentially low minor allele frequency limits statistical power to detect variation in allelic effects, our approach yields a validation population where jointly, the minor allele frequency is 0.5 (e.g. in this study, the frequency of the minor allele in the mapping population was 0.16 and effectively rose to 0.5 across our validation populations). Thus, this approach maximizes the chance of identifying a phenotypic effect associated with a focal SNP of interest.

An additional benefit of having outbred populations fixed at alternative genotypes is that one can interrogate the pleiotropic effects of a given polymorphism by simply phenotyping each population for any number of traits and without any additional genotyping costs. For example, one could simultaneously and robustly estimate the effect of a SNP on gene expression and its effect on higher-order phenotypes such as behavior. This method can also be directly extended to assess the effect of candidate SNPs across a variety of environmental contexts and treatments, which opens up a wide range of possibilities, such as validating candidate genotype-by-environment interactions.

Here, we have connected two independent discoveries, derived from two genome-wide analyses in different populations of outbred *D. melanogaster*, and show that a cis-regulatory polymorphism associated with expression differences modulates lifespan variation in *Drosophila*. These results show that our experimental design allows for the estimation of the phenotypic effects of a single eQTL. The first report of *midway*'s involvement in *Drosophila* lifespan was recently confirmed using loss-of-function mutants (Pallares *et al.* 2023). The changes in gene expression levels caused by the cis-eQTL that we validated here, contrasts the drastic changes in genome organization caused by complete loss-of-function. While complete loss-of-function mutants for *midway* show a significant decrease in lifespan (T_{50} = 33 for *midway* loss-of-function mutant and T_{50} = 47 for control populations) (Pallares *et al.* 2023), here, we find that the cis-eQTL genotype that reduces *midway* expression actually increases lifespan. The contrast between those results highlights the benefit of validating genetic effects in a relevant genetic context: the biological insight obtained from lab-based experiments will be a more accurate representation of the effect of genetic variation in natural populations.

The broad applicability, simplicity, and low cost of the experimental approach we advance in this study offers a tractable system for validating allelic effects in diverse backgrounds and provides significant gain in statistical power to detect genotype-phenotype associations. While this study focuses on *D. melanogaster*, the general experimental paradigm we outline can be applied broadly to any model (or non-model) systems where focal individuals can be kept alive while being genotyped, and where controlled breeding is an option (e.g. mice can be genotyped from a tail clip, fish from a fin clipping, and fecal sample or venipuncture can be used in birds).

Code and data availability

All code and data for reproducing the analysis presented here can be found at github.com/Wolfffff/midway-code.

Authors' contributions

V.A., J.F.A., and L.F.P. designed the study. V.A. and L.P. performed the experiments. S.W., V.A., D.M., and L.F.P., performed the analysis. S.W., V.A., and D.M. wrote the original draft. S.W., V.A., D.M., J.F.A., and L.F.P. reviewed and edited the manuscript. All authors approved the manuscript.

Conflict of interest declaration:

The authors have no conflicts of interest to declare.

Funding:

S.W. is supported by the National Science Foundation Graduate Research Fellowship Program (DGE-2039656). D.M. is funded by a fellowship from the Princeton Presidential Postdoctoral Research Fellows Program. J.F.A. is funded by grants from the NIH: National Institute of Environmental Health Sciences (R01-ES029929) and National Institute of General Medical Sciences (R35GM124881). L.F.P. is funded by the Max Planck Society. This study was supported in part by the Lewis-Sigler Institute for Integrative Genomics at Princeton University.

Acknowledgments:

We thank all members of the Ayroles lab for their support.

References

Alsheikh, A.J., Wollenhaupt, S., King, E.A., Reeb, J., Ghosh, S., Stolzenburg, L.R., Tamim, S., Lazar, J., Davis, J.W. & Jacob, H.J. (2022) The landscape of GWAS validation; systematic review identifying 309 validated non-coding variants across 130 human diseases. BMC medical genomics, 15, 74.

Austad, S.N. & Fischer, K.E. (2016) Sex Differences in Lifespan. Cell metabolism, 23, 1022–1033.

Bellen, H.J., Wangler, M.F. & Yamamoto, S. (2019) The fruit fly at the interface of diagnosis and pathogenic mechanisms of rare and common human diseases. Human molecular genetics, 28, R207–R214.

Buszczak, M., Lu, X., Segraves, W.A., Chang, T.Y. & Cooley, L. (2002) Mutations in the midway gene disrupt a Drosophila acyl coenzyme A: diacylglycerol acyltransferase. Genetics, 160, 1511–1518.

Chandler, C.H., Chari, S. & Dworkin, I. (2013) Does your gene need a background check? How genetic background impacts the analysis of mutations, genes, and evolution. Trends in genetics: TIG, 29, 358–366.

Cox, D.R. (1972) Regression models and life-tables. Journal of the Royal Statistical Society, 34, 187–202.

El-Brolosy, M.A. & Stainier, D.Y.R. (2017) Genetic compensation: A phenomenon in search of mechanisms. PLoS genetics, 13, e1006780.

Gallagher, M.D. & Chen-Plotkin, A.S. (2018) The Post-GWAS Era: From Association to Function. American journal of human genetics, 102, 717–730.

Girard, V., Jollivet, F., Knittelfelder, O., Celle, M., Arsac, J.-N., Chatelain, G., Van den Brink, D.M., Baron, T., Shevchenko, A., Kühnlein, R.P., Davoust, N. & Mollereau, B. (2021) Abnormal accumulation of lipid droplets in neurons induces the conversion of alpha-Synuclein to proteolytic resistant forms in a Drosophila model of Parkinson’s disease. PLOS Genetics, 17, e1009921.

Grenier, J.K., Arguello, J.R., Moreira, M.C., Gottipati, S., Mohammed, J., Hackett, S.R., Boughton, R., Greenberg, A.J. & Clark, A.G. (2015) Global diversity lines - a five-continent reference panel of sequenced Drosophila melanogaster strains. G3 , 5, 593–603.

Hoedjes, K.M., Kostic, H., Flatt, T. & Keller, L. (2023) A Single Nucleotide Variant in the PPARγ-homolog Eip75B Affects Fecundity in Drosophila. Molecular biology and evolution, 40.

Housden, B.E., Muhar, M., Gemberling, M., Gersbach, C.A., Stainier, D.Y.R., Seydoux, G., Mohr, S.E., Zuber, J. & Perrimon, N. (2017) Loss-of-function genetic tools for animal models: cross-species and cross-platform differences. Nature reviews. Genetics, 18, 24–40.

Lessard, S., Francioli, L., Alfoldi, J., Tardif, J.-C., Ellinor, P.T., MacArthur, D.G., Lettre, G., Orkin, S.H. & Canver, M.C. (2017) Human genetic variation alters CRISPR-Cas9 on- and off-targeting specificity at therapeutically implicated loci. Proceedings of the National Academy of Sciences of the United States of America, 114, E11257–E11266.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics , 27, 2987–2993.

Mokashi, S.S., Shankar, V., Johnstun, J.A., Huang, W., Mackay, T.F.C. & Anholt, R.R.H. (2021) Systems Genetics of Single Nucleotide Polymorphisms at the Drosophila Obp56h Locus. bioRxiv, 2021.06.28.450219.

Pallares, L.F., Lea, A.J., Han, C., Filippova, E.V., Andolfatto, P. & Ayroles, J.F. (2023) Dietary stress remodels the genetic architecture of lifespan variation in outbred Drosophila. Nature genetics, 55, 123–129.

Ramaekers, A., Claeys, A., Kapun, M., Mouchel-Vielh, E., Potier, D., Weinberger, S., Grillenzoni, N., Dardalhon-Cuménal, D., Yan, J., Wolf, R., Flatt, T., Buchner, E. & Hassan, B.A. (2019) Altering the Temporal Regulation of One Transcription Factor Drives Evolutionary Trade-Offs between Head Sensory Organs. Developmental cell, 50, 780–792.e7.

Saul, M.C., Philip, V.M., Reinholdt, L.G., Center for Systems Neurogenetics of Addiction & Chesler, E.J. (2019) High-Diversity Mouse Populations for Complex Traits. Trends in genetics: TIG, 35, 501–514.

Schaefer, K.A., Wu, W.-H., Colgan, D.F., Tsang, S.H., Bassuk, A.G. & Mahajan, V.B. (2017) Unexpected mutations after CRISPR-Cas9 editing in vivo. Nature methods, 14, 547–548.

Schüpbach, T. & Wieschaus, E. (1991) Female sterile mutations on the second chromosome of Drosophila melanogaster. II. Mutations blocking oogenesis or altering egg morphology. Genetics, 129, 1119–1136.

Seabold, S. & Perktold, J. (2010) Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference SciPy.

Streeper, R.S., Grueter, C.A., Salomonis, N., Cases, S., Levin, M.C., Koliwad, S.K., Zhou, P., Hirschey, M.D., Verdin, E. & Farese, R.V., Jr. (2012) Deficiency of the lipid synthesis enzyme, DGAT1, extends longevity in mice. Aging, 4, 13–

27.

Tian, Y., Bi, J., Shui, G., Liu, Z., Xiang, Y., Liu, Y., Wenk, M.R., Yang, H. & Huang, X. (2011) Tissue-autonomous function of Drosophila seipin in preventing ectopic lipid droplet formation. PLoS genetics, 7, e1001364.

Tschapalda, K., Zhang, Y.-Q., Liu, L., Golovnina, K., Schlemper, T., Eichmann, T.O., Lal-Nag, M., Sreenivasan, U., McLenithan, J., Ziegler, S., Sztalryd, C., Lass, A., Auld, D., Oliver, B., Waldmann, H., Li, Z., Shen, M., Boxer, M.B. & Beller, M. (2016) A Class of Diacylglycerol Acyltransferase 1 Inhibitors Identified by a Combination of Phenotypic High-throughput Screening, Genomics, and Genetics. EBioMedicine, 8, 49–59.

Visscher, P.M., Wray, N.R., Zhang, Q., Sklar, P., McCarthy, M.I., Brown, M.A. & Yang, J. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. American journal of human genetics, 101, 5–22.

Zimmer, A.M., Pan, Y.K., Chandrapalan, T., Kwong, R.W.M. & Perry, S.F. (2019) Loss-of-function approaches in comparative physiology: is there a future for knockdown experiments in the era of genome editing? The Journal of experimental biology, 222.