The Evolutionary Implications of Adaptations to Stressful Environments on Time Scales Spanning Days to Millennia

Cory Andrew Weller
White Bear Lake, Minnesota

Bachelor of Arts, Gustavus Adolphus College, 2010

A Dissertation presented to the Graduate Faculty of the University of Virginia
in Candidacy for the Degree of Doctor of Philosophy

Department of Biology

University of Virginia
August, 2019

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

While the number of people who had a hand in shaping me during my dissertation work is innumerable, I'd like to express my thanks to some particular groups.

I am thankful for my dissertation committee and their valuable input on the direction of my dissertation. In particular, I am thankful for Alan having accepted me as his student.

I am thankful for friends and colleagues I met within the UVA Biology department: Megan and Amanda for welcoming me so warmly upon my move to Charlottesville; Alyssa, and Phoebe, always lending an ear when I needed to talk something out; Mike, for being a great roommate for three years; and too many to name, and the entirety of the EEBio community for providing feedback and conversation over the last six years.

I am thankful for friends I met in Charlottesville during my stay, particularly those who became part of our core trivia team: Ana, Chris, TK, Katie, and Marc. There are too many stories and adventures we've shared over the last few years, and am looking forward to keeping in touch as we continue our careers.

I am thankful for my family and their patience with me working so far from home, and thankful for the Press family welcoming me into theirs.

Lastly, I am thankful and looking forward to my time with my fiancé, Eric, whose love and support has pushed me through the last three years of my dissertation. Thank you for making me feel normal for frenetically checking of simulation jobs at odd hours or bringing stacks of papers to read on road trips and holidays. Those things about me aren't changing any time soon!

INTRODUCTION

Life is a perpetual onslaught of less-than-optimal conditions. In order for life to continue despite the ubiquity of otherwise-lethal stressors, organisms develop physiological and behavioral adaptations to escape, tolerate or avoid the stresses they face. Broadly, my dissertation covers the genetic signatures of evolutionary responses to stressful conditions. In the first portion of my dissertation, I investigate the evolutionary consequences of an extreme stress tolerance strategy in Firmicutes bacteria. For the remaining parts of my dissertation, I turn to the humble fruit fly, *Drosophila melanogaster*. Using this model organism, I investigate genetic mechanisms related to survival in starvation and desiccation conditions, such as those encountered at the onset of winter when food becomes scarce. Using both systems, I address how strategies of adapting to stress influences each organism's genome. Using various inferential techniques, I make use of present-day observations to recapitulate the evolutionary trajectories of organisms—trajectories that vary in length from thousands of years to a single day.

The basal unit of molecular evolution is the substation of one nucleotide for another, and rates of evolution are inferred from rate of nucleotide substitutions. In 1969, Laird, McConaughy, and McCarthy noticed that hemoglobin proteins evolved—i.e. fixed nucleotide substitutions—more rapidly in rodents than in ungulates, despite sharing a single point of divergence. The disparity in evolutionary rates was resolved when considering substitutions per generation instead of per year, and the generation time hypothesis is a current explanation for differences in evolutionary rates.

Generation time is predicted to correlate with rates of molecular evolution due to the inheritance pattern of gametes. Nucleotide substitutions occur as a result of mutagenic environmental factors, e.g. UV radiation, or as the result of proofreading errors during DNA replication (Kunkel 2011). While the former should occur at a relatively constant rate in a static environment, the latter will be directly proportional

to the number of times a cell lineage undergoes DNA replication. Thus, a cell line that undergoes fewer rounds of division will exhibit fewer total substitutions as a result of fewer replication-dependent substitutions. If an organism sequesters gametes (or gamete-producing cells) instead of continually dividing, then replication-dependent mutations will accrue at a rate proportional to sexual generations.

There is now a preponderance of evidence for the generation-time effect across the tree of life, including various mammals (Ohta 1973; Li and Tanimura 1987; Nabholz et al. 2008), reptiles (Bromham 2002), birds (Mooers and Harvey 1994), invertebrates (Thomas et al. 2010, but see Thomas et al. 2006) and plants (Gaut et al. 1996; Laroche et al. 1997; Ainouche and Bayer 1999; Laroche and Bousquet 1999; Andreasen and Baldwin 2001; Smith and Donoghue 2008).

While bacteria do not produce gametes, whole genomes are inherited during cell division, thus every mitotic division in bacteria is analogous to a single generation. Subsequently, one might expect rates of molecular evolution in bacteria to correlate with the average time between mitotic divisions—with bacterial "generation time." Yet, evidence for an effect of generation time on the rate of molecular evolution in bacteria was lacking. This is partially due to the intrinsic difficulty of investigating the historic effects of generation time in bacteria. Whereas most plants and animals exhibit relatively constant generation times, bacteria generation times vary across orders of magnitude depending on environmental conditions. As a result, it is difficult to identify appropriate contrasts for testing the generation-time effect for most bacteria species. One alternative could be to infer generation time by proxy—using a measurable phenotype that can be assumed to vary with generation time.

One such proxy of generation time is the ability to form bacterial endospores. A subset of bacteria in the phylum Firmicutes—and *only* the phylum Firmicutes—are capable of entering this vegetative, nonreproductive endospore state. Endospores have

a highly durable water-proof cell wall, do not require nutrition, and can withstand otherwise lethal levels of heat, desiccation, radiation, or lack of oxygen (Nicholson et al. 2000; Onyenwoke et al. 2004), with endospores of *Bacillus subtilis* having been shown to survive conditions akin to asteroid impact. It is believed that a majority of a spore-forming bacteria's life is spent in endospore state (Priest and Grigorova 1990). Endospores can remain dormant for extended durations until a proper environment triggers reanimation, with reports of spores being revived after 40 million years of dormancy inside of fossilized amber (Cano and Borucki 1995). As a consequence, it is a reasonable assumption that endospore-forming Firmicutes exhibit relatively longer "generation times" as measured by duration between mitotic cell divisions compared to close relatives that do not form endospores.

A study by Maughan (2007) contrasted evolutionary rates between spore-forming Firmicutes and non-spore-forming Firmicutes, finding no difference in evolutionary rate. Assuming endospore-formation is an appropriate proxy for time between cell divisions (and is properly assigned to various taxa being contrasted), a lack of evidence of a generation time in bacteria is a curious result in the light of widespread evidence of such effect in plant and animal species. One explanation for the negative result of Maughan (2007), should a generation-time effect exist, is low statistical power from limited taxonomic sampling.

In chapter one, I revisit the generation-time hypothesis using the Firmicutes bacteria system, increasing bacteria representation from 23 taxa in (Maughan 2007) to 197 taxa. With this improved taxonomic sampling, I reevaluated the relationship between generation time and evolutionary rate within the Firmicutes phylum. My comparative phylogenomic analysis of the Firmicutes phylum was capable of estimating putative endospore-forming ability from the history of gene content across time. I show strong evidence that the rate of molecular evolution is negatively correlated with generation time in bacteria, in line with observations throughout the

rest of the tree of life. In order to understand the evolutionary history of endospore formation in ancestral nodes of the Firmicutes tree, I conducted ancestral state reconstruction of spore-forming ability. My results indicate that endospore formation arose once near the root of the Firmicutes tree, and the trait was subsequently lost in multiple lineages—never to be regained.

Thus far, I have introduced endospore formation in Firmicutes as one example of extreme stress avoidance. As a consequence of this adaptive state of dormancy, the genomes of spore-forming Firmicutes contain detectable signatures a reduced rate of nucleotide substitution acting over millions of years. Whereas environmental stresses experienced by Firmicutes bacteria have resulted in slowed evolution, some organism exemplify rapid evolutionary responses to environmental stresses. For the remainder of my dissertation, I address the methods by which we study the genetic basis of local adaptation in natural populations over extremely short time scales—again, in the context of surviving in stressful environments.

The ability to tolerate (and thus survive in) stressful conditions is one outcome of local adaptation. While the ability to tolerate stress may reduce mortality and improve lifespan, such adaptations may not be without cost. Consider the Y-model of resource allocation (van Noordwijk and de Jong 1986), where two traits draw from a single pool of resources, and allocation to one trait (i.e. reproduction) precludes resources from being allocated to another trait (i.e. survival). Drosopholids make for a useful model to study the genetic mechanisms of tradeoffs between survival and reproduction, in part due to the ease at which these traits evolve during artificial selection. Maynard Smith (1985) described the negative correlation between survival and reproduction in *D. melanogaster*, where sterile mutants experienced longer lifespans, and battery of experimental evolution experiments have recapitulated this result. Artificial selection for delayed reproduction results in an improved lifespan (Rose 1984) and improved tolerance to desiccation and starvation (Rose et al. 1985). Selection for

resistance to starvation results in decreased early-adult-life fecundity (Wayne et al. 2006), and selection for decreased early-adult-life fecundity in turn results in greater desiccation tolerance (Nghiem et al. 2000). Desiccation tolerance evolves rapidly during artificial selection experiments (Hoffmann and Parsons 1989) with female survival doubling from 15 to 30 days in 26 generations of selection (Telonis-Scott 2006). Together, results show that for *D. melanogaster*, stress tolerance traits are highly correlated with each other, and in turn negatively correlated with reproduction.

Patterns of genotypic and phenotypic variation in *D. melanogaster* have been shaped by environmental stresses that vary over space and time. As a result, measurements of growth and reproduction in the wild shows a high degree of clinality. Flies from low, warm latitudes develop more quickly (James and Partridge 1995) and have low incidence of diapause (Schmidt et al. 2005), a state of ovarian arrest with low metabolic rate and improved stress tolerance (Tauber et al. 1986). Diapausing flies do not produce eggs and do not senesce as quickly—but treatment with juvenile hormone reactivates both senescence and egg production (Tatar et al. 2017). The same pattern is borne out of expression data, where genes differentially expressed across latitudinal clines are involved in metabolism and growth (Chen et al. 2012). The reproduction-survival tradeoff in response is not unique to Drosopholids: desiccation tolerance is improved during diapause in the mosquito *Culex pipiens*, where diapausing animals are larger and lose water more slowly, at the cost of delayed reproduction (Benoit and Denlinger 2007). In both cases of mosquitos and flies, larger body size will confer greater desiccation resistance due to a lower surface area to volume ratio.

Geographically distinct regions that are independently colonized by Drosophilids provide natural replicates of adaptation to spatially variable environments. There is concordance of genetic differentiation when comparing flies from latitudinal clines in North America to those of Australia (Turner et al. 2008), showing congruence in genetic responses to the same environmental stresses.

Drosophila melanogaster flies from distinct arid environments are consistently more desiccation tolerant than those from tropical environments (Hercus and Hoffmann 1999) and a comparison of cactus specialists reveals shared patterns of expression and positive selection for the same genes (Rane et al. 2019).

However, correlative studies from sampling natural populations may sometimes yield misleading conclusions. For example, desiccation tolerance in *D. melanogaster* is observed to correlate with melanism in natural populations, where flies residing near the dry or high latitude end of cline are more melanistic (Rajpurohit et al. 2008; Parkash et al. 2009; Telonis-Scott et al. 2011). This observation leads to the Melanism-Desiccation hypothesis: that melanin is an adaptation for desiccation resistance, reducing the rate of water loss. One experiment showed that artificial selection for melanism has improved desiccation tolerance (Ramniwas et al. 2012), potentially affirming the melanism-desiccation hypothesis. However, desiccation tolerance has more recently been shown to evolve wholly independently from melanism (Rajpurohit et al. 2016), showing that melanism cannot, at least by itself, explain size-specific differences in desiccation tolerance.

In order to fully understand the precise mechanisms driving local adaptation in natural populations, we must include all steps from genotype to phenotype. Because it is known that non-coding regulatory genetic variation strongly influences phenotypes (King and Wilson 1975; Andolfatto 2005; Kopf et al. 2015; Nourmohammad et al. 2017), we must connect genotype to expression variation, which is in turn connected to phenotypic variation. The remaining chapters of my dissertation involve the development and use of a novel method of association mapping population to better understand the coding and noncoding genetic variation (e.g., cis-regulatory elements driving allele-specific expression, ASE) that allows evolution of stress tolerance and reproduction life history traits.

Pre-sequenced inbred reference panels are commonly used resources for genetic association studies of life history traits, and such panels exist for mice (Chesler et al. 2008), *Caenorhabditis elegans* (Noble et al. 2019), *Arabidopsis thaliana* (Kover et al. 2009), wheat (Huang et al. 2012), rice (Singh et al. 2013), corn (Krämer et al. 2014), and *D. melanogaster* (King et al. 2012; Mackay et al. 2012; Grenier et al. 2015). While such reference panels provide a replicable source of genetic variation for performing association studies, their genomes are inbred. Because pertinent biological mechanisms such as dominance relationships or allele-specific expression cannot be studied in inbred populations, inbred reference panels are insufficient for holistically informing mechanisms of adaptation in natural populations.

I propose an alternative method of generating a mapping population, which I call the Hybrid Swarm method, where extensive recombination is instead replaced with greater initial haplotype diversity. The Hybrid Swarm method leverages ancestral recombination present in its founders' genomes, requiring fewer (e.g., five) generations of recombination to break up linkage disequilibrium and achieve high mapping resolution to the nucleotide level. This design also provides outbred genomes (allowing the study of dominance and allele-specific expression) without the need for laboriously sampling natural populations.

To date, the Hybrid Swarm method has likely remained unused due to the expensive genotyping requirements of outbred populations, compared to pre-sequenced inbred lines. While genotype imputation (genome reconstructions) can be employed to reduce sequencing coverage requirements (Howie et al. 2009, 2012; Li et al. 2009; Spiliopoulou et al. 2017), the computational requirements of imputation grow at a greater-than-linear rate with the number of founding haplotypes. As a result, a Hybrid Swarm population founded by dozens to hundreds of founding lines would require an intractable amount of computational resources. I proposed to reduce computational requirements by focusing imputation search space on a subset of most-

likely-ancestors for a given chromosome. If a subset of most-likely-ancestors can accurately be determined, the computational requirements of Hybrid Swarm genome reconstructions would become tractable, making the method economically feasible with ultra-low-coverage sequencing data. With the Hybrid Swarm method of association mapping, researchers can quickly and affordably generate outbred mapping populations for the study of complex traits.

In chapter two, I evaluate methods of reconstructing forward-simulated Hybrid Swarm genomes from ultra-low coverage sequencing data. After optimizing parameters of the reconstruction pipeline, I show that it is feasible to generate high-quality genotype estimates while sampling as few as 1/20 to 1/200 variable sites. However, it is not sufficient to solely show that genome reconstructions are accurate—it is further necessary to show that Hybrid Swarm populations are capable of resolving genotype-phenotype relationships. In order to evaluate the effectiveness of the Hybrid Swarm for genetic association mapping, I developed a high-throughput pipeline for simulating GWAS, and then compared my newly developed mapping population to modern alternatives. I show that although inbred populations exhibit greater intrinsic power in mapping additive traits, a Hybrid Swarm population performs similarly to a highly outbred population, e.g. individuals from wild sampling. My results suggest that outbred mapping populations can be quickly generated using the Hybrid Swarm method, as a simpler and cheaper alternative to wild sampling.

In chapter three, I put my low-coverage genome reconstruction pipeline to use to study the mechanisms of rapid adaptation by *D. melanogaster* to desiccation and starvation conditions. After reconstructing nearly 700 individual fly genomes, I conduct a GWAS of survival, with significant associations found for multiple mitochondrial proteins—a class that has been previously implicated in lifespan expansion and tolerance to stress. Using expression data, I categorize genes that are differentially expressed and biological processes differentially represented in fed or starvation

conditions. As expected, fed flies exhibit increased expression of reproductive genes. In the starvation condition, flies exhibit upregulation of multiple metabolic processes, particularly purine synthesis—which has been previously implicated in the tradeoff between fecundity and lifespan. Interestingly, we see that genes primarily expressed in different tissues exhibit variable rates of ASE, with increased representation of ASE in the carcass, salivary glands, heart, and fat body; the only tissue class to show reduced ASE is the ovary. Previous work has shown reduced levels of genetic variation in *D. melanogaster* reproductive genes, as a potential signature of purifying selection. My results suggest that genetic variation in and around reproductive genes themselves do not underly the tradeoff between survival and reproduction.

Literature Cited

Ainouche, A.-K., and R. J. Bayer. 1999. Phylogenetic relationships in Lupinus (Fabaceae: Papilionoideae) based on internal transcribed spacer sequences (ITS) of nuclear ribosomal DNA. Am. J. Bot. 86:590–607.

Andolfatto, P. 2005. Adaptive evolution of non-coding DNA in Drosophila. Nature 437:1149–52.

Andreasen, K., and B. G. Baldwin. 2001. Unequal Evolutionary Rates Between Annual and Perennial Lineages of Checker Mallows (Sidalcea, Malvaceae): Evidence from 18S-26S rDNA Internal and External Transcribed Spacers. Mol. Biol. Evol. 18:936–944.

Benoit, J. B., and D. L. Denlinger. 2007. Suppression of water loss during adult diapause in the northern house mosquito, Culex pipiens. J. Exp. Biol. 210:217–226.

Bromham, L. 2002. Molecular Clocks in Reptiles: Life History Influences Rate of Molecular Evolution. Mol. Biol. Evol. 19:302–309.

Cano, R. J., and M. K. Borucki. 1995. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science 268:1060–4.

Chen, Y., S. F. Lee, E. Blanc, C. Reuter, B. Wertheim, P. Martinez-Diaz, A. A. Hoffmann, and L. Partridge. 2012. Genome-wide transcription analysis of clinal genetic variation in drosophila. PLoS One 7:1–8.

Chesler, E. J., D. R. Miller, L. R. Branstetter, L. D. Galloway, B. L. Jackson, V. M. Philip, B. H. Voy, C. T. Culiat, D. W. Threadgill, R. W. Williams, G. A. Churchill, D. K. Johnson, and K. F. Manly. 2008. The Collaborative Cross at Oak Ridge National Laboratory: Developing a powerful resource for systems genetics. Mamm. Genome 19:382–389.

Gaut, B. S., B. R. Morton, B. C. McCaig, and M. T. Clegg. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc. Natl. Acad. Sci. 93:10274–10279.

Grenier, J. K., J. R. Arguello, M. C. Moreira, S. Gottipati, J. Mohammed, S. R. Hackett,

R. Boughton, A. J. Greenberg, and A. G. Clark. 2015. Global Diversity Lines–A Five-Continent Reference Panel of Sequenced Drosophila melanogaster Strains . G3: Genes|Genomes|Genetics 5:593–603.

Hercus, M. J., and A. A. Hoffmann. 1999. Desiccation resistance in interspecific drosophila crosses: Genetic interactions and trait correlations. Genetics 151:1493–1502.

Hoffmann, A. A., and P. A. Parsons. 1989. An integrated approach to environmental stress tolerance and life-history variation: desiccation tolerance in Drosophila. Biol. J. Linn. Soc. 37:117–136.

Howie, B., C. Fuchsberger, M. Stephens, J. Marchini, and G. R. Abecasis. 2012. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat. Genet. 44:955–959. Nature Publishing Group.

Howie, B. N., P. Donnelly, and J. Marchini. 2009. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet. 5.

Huang, B. E., A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden, M. K. Morell, and C. R. Cavanagh. 2012. A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol. J. 10:826–839.

James, A. C., and L. Partridge. 1995. Thermal evolution of rate of larval development in Drosophila melanogaster in laboratory and field populations. J. Evol. Biol. 8:315–330.

King, E. G., S. J. Macdonald, and A. D. Long. 2012. Properties and power of the Drosophila synthetic population resource for the routine dissection of complex traits. Genetics 191:935–949.

King, M.-C., and A. C. Wilson. 1975. Evolution at Two Levels in Humans and Chimpanzees. Science (80-.). 188:107–116.

Kopf, M., S. Klähn, I. Scholz, W. R. Hess, and B. Voß. 2015. Variations in the non-coding transcriptome as a driver of inter-strain divergence and physiological adaptation in bacteria. Sci. Rep. 5.

Kover, P. X., W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, M. D. Purugganan, C. Durrant, and R. Mott. 2009. A Multiparent Advanced Generation Inter-Cross to

Fine-Map Quantitative Traits in Arabidopsis thaliana. PLoS Genet. 5:e1000551. Public Library of Science.

Krämer, N., N. Ranc, N. Meyer, M. Ouzunova, C. Lehermeier, A. Charcosset, C.-C. Schön, C. Bauland, J. Moreno-González, M. Menz, H. Pausch, H. Walter, A. E. Melchinger, E. Bauer, W. Schipprack, M. Schönleben, P. Flament, C. Camisan, L. Campo, and L. Moreau. 2014. Usefulness of Multiparental Populations of Maize (Zea mays L.) for Genome-Based Prediction . Genetics 198:3–16.

Kunkel, T. A. 2011. Balancing eukaryotic replication asymmetry with replication fidelity. Curr. Opin. Chem. Biol. 15:620–626. Elsevier Ltd.

Laird, C. D., B. L. McConaughy, and B. J. McCarthy. 1969. Rate of Fixation of Nucleotide Substitutions in Evolution. Nature 224:149–154.

Laroche, J., and J. Bousquet. 1999. Evolution of the mitochondrial rps3 intron in perennial and annual angiosperms and homology to nad5 intron 1. Mol. Biol. Evol. 16:441–452.

Laroche, J., P. Li, L. Maggia, and J. Bousquet. 1997. Molecular evolution of angiosperm mitochondrial introns and exons. Proc. Natl. Acad. Sci. 94:5722–5727.

Li, W. H., and M. Tanimura. 1987. The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93–6.

Li, Y., C. Willer, S. Sanna, and G. Abecasis. 2009. Genotype Imputation. Annu. Rev. Genomics Hum. Genet. 10:387–406.

Mackay, T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, D. Zhu, S. Casillas, Y. Han, M. M. Magwire, J. M. Cridland, M. F. Richardson, R. R. H. Anholt, M. Barrón, C. Bess, K. P. Blankenburg, M. A. Carbone, D. Castellano, L. Chaboub, L. Duncan, Z. Harris, M. Javaid, J. C. Jayaseelan, S. N. Jhangiani, K. W. Jordan, F. Lara, F. Lawrence, S. L. Lee, P. Librado, R. S. Linheiro, R. F. Lyman, A. J. Mackey, M. Munidasa, D. M. Muzny, L. Nazareth, I. Newsham, L. Perales, L.-L. Pu, C. Qu, M. Ràmia, J. G. Reid, S. M. Rollmann, J. Rozas, N. Saada, L. Turlapati, K. C. Worley, Y.-Q. Wu, A. Yamamoto, Y. Zhu, C. M. Bergman, K. R. Thornton, D. Mittelman, and R. A. Gibbs. 2012. The Drosophila melanogaster Genetic Reference Panel. Nature 482:173–178.

Maughan, H. 2007. Rates of molecular evolution in bacteria are relatively constant

despite spore dormancy. Evolution 61:280–8.

Mooers, A. O., and P. H. Harvey. 1994. Metabolic rate, generation time, and the rate of molecular evolution in birds. Mol. Phylogenet. Evol. 3:344–50.

Nabholz, B., S. Glémin, and N. Galtier. 2008. Strong variations of mitochondrial mutation rate across mammals--the longevity hypothesis. Mol. Biol. Evol. 25:120–30.

Nghiem, D., A. G. Gibbs, M. R. Rose, and T. J. Bradley. 2000. Postponed aging and desiccation resistance in Drosophila melanogaster. Exp. Gerontol. 35:957–969.

Nicholson, W. L., N. Munakata, G. Horneck, H. J. Melosh, and P. Setlow. 2000. Resistance of Bacillus endospores to extreme terrestrial and extraterrestrial environments. Microbiol. Mol. Biol. Rev. 64:548–72.

Noble, L. M., M. V Rockman, and H. Teotónio. 2019. Gene-level quantitative trait mapping in an expanded C . elegans multiparent experimental evolution panel. 1–7.

Nourmohammad, A., J. Rambeau, T. Held, V. Kovacova, J. Berg, and M. Lässig. 2017. Adaptive Evolution of Gene Expression in Drosophila. Cell Rep. 20:1385–1395.

Ohta, T. 1973. Slightly deleterious mutant substitutions in evolution. Nature 246:96–98.

Onyenwoke, R. U., J. A. Brill, K. Farahi, and J. Wiegel. 2004. Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). Arch. Microbiol. 182:182–92.

Parkash, R., V. Sharma, and B. Kalra. 2009. Impact of body melanisation on desiccation resistance in montane populations of D. melanogaster: Analysis of seasonal variation. J. Insect Physiol. 55:898–908.

Priest, F. G., and R. Grigorova. 1990. 18 Methods for Studying the Ecology of Endospore-forming Bacteria. Methods Microbiol. 22:565–591. Elsevier.

Rajpurohit, S., R. Parkash, S. Singh, and S. Ramniwas. 2008. Climate change, boundary increase and elongation of a pre-existing cline: A case study in Drosophila ananassae. Entomol. Res. 38:268–275.

Rajpurohit, S., L. M. Peterson, A. J. Orr, A. J. Marlon, and A. G. Gibbs. 2016. An experimental evolution test of the relationship between melanism and desiccation survival in insects. PLoS One 11:1–17.

Ramniwas, S., B. Kajla, K. Dev, and R. Parkash. 2012. Direct and correlated responses to

laboratory selection for body melanisation in Drosophila melanogaster: support for the melanisation-desiccation resistance hypothesis. J. Exp. Biol. 216:1244–1254.

Rane, R. V., S. L. Pearce, F. Li, C. Coppin, M. Schiffer, J. Shirriffs, C. M. Sgrò, P. C. Griffin, G. Zhang, S. F. Lee, A. A. Hoffmann, and J. G. Oakeshott. 2019. Genomic changes associated with adaptation to arid environments in cactophilic Drosophila species. BMC Genomics 20:1–22. BMC Genomics.

Rose, M. R. 1984. Laboratory Evolution of Postponed Senescence in Drosophila melanogaster. Evolution (N. Y). 38:1004.

Rose, M. R., H. B. Passananti, M. Matos, P. M. SERVICE, E. W. HUTCHINSON, M. D. MACKINLEY, and M. R. ROSE. 1985. Resistance To Environmental Stress in Drosophila Melanogaster Selected for Postponed Senescence. Physiol. Zool. 58:380–389.

Schmidt, P. S., L. Matzkin, M. Ippolito, and W. F. Eanes. 2005. Geographic Variation in Diapause Incidence , Life-History Traits , and Climatic Adaptation in Drosophila melanogaster Author (s): Paul S . Schmidt , Luciano Matzkin , Michael Ippolito and Walter F . Eanes Published by : Society for the Study of Evolutio. Evolution (N. Y). 59:1721–1732.

Singh, R., I. T. Lobina, M. Thomson, S. McCouch, C. Dilla-Ermita, R. Mauleon, E. Redoña, H. Leung, P. Muyco, G. Gregorio, C. Raghavan, C.-W. Tung, N. Bandillo, and M. A. L. Sevilla. 2013. Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6:11.

Smith, S. A., and M. J. Donoghue. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322:86–9.

Spiliopoulou, A., M. Colombo, P. Orchard, F. Agakov, and P. McKeigue. 2017. GeneImp: Fast imputation to large reference panels using genotype likelihoods from ultralow coverage sequencing. Genetics 206:91–104.

Tatar, Chien, and Priest. 2017. Negligible Senescence during Reproductive Dormancy in Drosophila melanogaster. Am. Nat. 158:248.

Tauber, M. J., S. Masaki, C. A. Tauber, S. Masaki, and A. D. Lees. 1986. Seasonal Adaptations of Insects. Oxford University Press.

Telonis-Scott, M. 2006. A new set of laboratory-selected Drosophila melanogaster lines for the analysis of desiccation resistance: response to selection, physiology and correlated responses. J. Exp. Biol. 209:1837–1847.

Telonis-Scott, M., A. A. Hoffmann, and C. M. Sgrò. 2011. The molecular genetics of clinal variation: A case study of ebony and thoracic trident pigmentation in Drosophila melanogaster from eastern Australia. Mol. Ecol. 20:2100–2110.

Thomas, J. A., J. J. Welch, R. Lanfear, and L. Bromham. 2010. A generation time effect on the rate of molecular evolution in invertebrates. Mol. Biol. Evol. 27:1173–80.

Thomas, J. A., J. J. Welch, M. Woolfit, and L. Bromham. 2006. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. Proc. Natl. Acad. Sci. U. S. A. 103:7366–71.

Turner, T. L., M. T. Levine, M. L. Eckert, and D. J. Begun. 2008. Genomic analysis of adaptive differentiation in Drosophila melanogaster. Genetics 179:455–473.

van Noordwijk, A. J., and G. de Jong. 1986. Acquisition and Allocation of Resources: Their Influence on Variation in Life History Tactics. Am. Nat. 128:137–142.

Wayne, M. L., U. Soundararajan, and L. G. Harshman. 2006. Environmental stress and reproduction in Drosophila melanogaster: starvation resistance, ovariole numbers and early age egg production. BMC Evol. Biol. 6:57. BioMed Central.

CHAPTER ONE

A Generation Time Effect on the Rate of Molecular Evolution in Bacteria

As published: Weller C., and M. Wu, 2015 A generation-time effect on the rate of molecular evolution in bacteria. Evolution 69: 643–652.

Abstract

Molecular evolutionary rate varies significantly among species and a strict global molecular clock has been rejected across the tree of life. Generation time is one primary life-history trait that influences the molecular evolutionary rate. Theory predicts that organisms with shorter generation times evolve faster because of the accumulation of more DNA replication errors per unit time. While the generation-time effect has been demonstrated consistently in plants and animals, the evidence of its existence in bacteria is lacking. The bacterial phylum *Firmicutes* offers an excellent system for testing generation-time effect because some of its members can enter a dormant, non-reproductive endospore state in response to harsh environmental conditions. It follows that spore-forming bacteria would—with their longer generation times—evolve more slowly than their non-spore forming relatives. It is therefore surprising that a previous study found no generation-time effect in *Firmicutes*. Using a phylogenetic comparative approach and leveraging on a large number of *Firmicutes* genomes, we found sporulation significantly reduces the genome-wide spontaneous DNA mutation rate and protein evolutionary rate. Contrary to the previous study, our results provide strong evidence that the evolutionary rates of bacteria, like those of plants and animals, are influenced by generation-time.

Introduction

The rate of molecular evolution varies significantly between species and much evidence has rejected the existence of a strict global molecular clock across the tree of life (Bousquet et al. 1992; Thomas et al. 2006; Welch et al. 2008; Kuo and Ochman 2009). It has been demonstrated that organism-level traits such as life history could influence the molecular evolutionary rate in eukaryotes, chief among them is the generation time. The generation-time hypothesis states that organisms with shorter generation time evolve faster, as they copy their genomes more frequently and therefore have more DNA replication errors per unit time.

The generation-time effect has been repeatedly observed in both animals and plants (Laird et al. 1969; Kohne 1970; Li and Tanimura 1987; Ohta 1993; Mooers and Harvey 1994; Bromham et al. 1996; Gaut et al. 1996; Gaut et al. 1997; Laroche et al. 1997; Ainouche and Bayer 1999; Laroche and Bousquet 1999; Andreasen and Baldwin 2001; Bromham 2002; Nabholz et al. 2008; Smith and Donoghue 2008; Welch et al. 2008; Thomas et al. 2010). For example, molecular evolutionary rates in rodents were much faster than those of primates (Laird et al. 1969; Kohne 1970). Perennials, with longer generation times, have been shown to accumulate substitutions more slowly than rapidly-maturing annual plants (Soria-Hernanz et al. 2008; Yue et al. 2010; Buschiazzo et al. 2012; but see Whittle and Johnston, 2003).

It is less clear whether there is a generation-time effect on the rate of molecular evolution in bacteria. Testing the generation-time hypothesis in natural bacterial populations can be difficult. Unlike animals and plants that have relatively fixed generation time, bacteria populations of the same species can have highly variable generation times depending on the growth condition. It is therefore difficult to accurately estimate the generation time of bacteria in natural environments, as nutrients and other important environmental factors for growth (e.g., temperature, salinity) vary widely in time and space. Certain members of the bacteria phylum *Firmicutes* are capable of entering an encapsulated, dormant and non-reproductive state known as an endospore. This state allows for bacteria to withstand and survive extreme conditions, such as ultraviolet radiation, desiccation, heat, and lack of nutrients. Spores can stay dormant for extended periods of time. When the environment becomes favorable, spores can exit to the vegetative state. Revival of spores millions of years old has been reported (Cano and Borucki 1995). It is thought that spore-forming bacteria mostly exist as spores in nature (Priest and Grigorova 1990), therefore it is reasonable to assume that spore-forming *Firmicutes* (SFF) have longer generation times than their non-spore-forming relatives (NSFF). As such, *Firmicutes* represents an excellent system to test the generation-time hypothesis in bacteria.

Not replicating while dormant, SFF are expected to have less DNA replication errors per unit time, and thus a lower rate of evolution. It is therefore surprising that a previous study showed no differences in the evolutionary rates when comparing SFF and NSFF (Maughan 2007). However, the use of few *Firmicutes* genomes in that study may have prevented detection of differences between these groups, if such a difference exists. Leveraging on a substantially increased representation of the *Firmicutes* phylum and using a phylogenetic comparative approach, we revisited the relationship between spore formation and evolutionary rates in bacteria. We found strong evidence that rates of molecular evolution are correlated with generation time in bacteria.

Methods

Reconstructing a genome tree of Firmicutes

Using the Phyla-AMPHORA package (Wang and Wu 2013), protein sequences of 168 single-copy, "universal", phylum-level marker genes from 573 completed genomes of *Firmicutes* were identified, aligned, trimmed and concatenated into a single alignment. To reduce the computational cost, a FastTree (Price et al. 2009) built from the concatenated alignment was used to select 200 representatives that maximized the phylogenetic diversity, using a greedy algorithm described in (Steel 2005). A RAxML maximum-likelihood tree was made with 1,000 bootstrap replicates using the LG + gamma model (Stamatakis 2014). Similarly, a bacterial genome tree containing 200 top representatives of major bacteria phyla was reconstructed, using the concatenated protein sequences of 31 "universal markers" (Wu and Scott 2012). The *Firmicutes* genome tree was then rooted using the bacterial genome tree as a guide.

Identifying sporulation genes in *Firmicutes* genomes

199 genes whose annotation contained keywords "spore", "sporulation" or "germination" were downloaded from *Bacillus subtilis* database SubtiList (http://genolist.pasteur.fr/SubtiList/), and combined with 175 *B. subtilis* sporulation

genes used in (Wu et al. 2005). Genes with highly conserved domains (e.g. kinases, phosphatases, ATPases) were removed from the list, resulting in a set of 163 sporulation genes, which were used as query sequences to BLASTP search against 573 complete *Firmicutes* genomes. The mutual-best-hits of each sporulation gene in each *Firmicutes* genome were identified as orthologous genes. The e-value cutoff was 1e-3 in both the forward and reverse BLASTP searches. The same procedure was also used to identify sporulation genes in representatives of non-*Firmicutes* bacteria.

Predicting sporulation potential with phylogenetic profiling

The distribution patterns of the 163 sporulation genes in each of the 200 *Firmicutes* representatives were represented with a binary matrix, where each column represented one species and each row represented one gene. "1" and "0" denoted the presence and absence of genes respectively. Species and genes were then grouped by the gene distribution patterns (phylogenetic profiles) with the CLUSTER program (http://rana.lbl.gov/EisenSoftware.htm), using the absolute correlation (centered) as the similarity metric and complete linkage as the clustering method.

Ancestor state reconstruction

Having predicted sporulation potential by phylogenetic profiling, we reconstructed the ancestral state with the R package 'ape' (Paradis et al. 2004). A binary categorical trait of "spore-forming" and "non-spore-forming" was mapped onto the tips of the *Firmicutes* genome tree. Transition probabilities for trait gain and loss were set to be equal. Results were similar when transition probabilities of trait gain to loss were set to be 1:2, 1:5 and 1:10.

PAML analyses

The CDS (Coding DNA Sequences) of the 168 protein-coding genes were aligned using their protein sequence alignments as a guide. Using the alignment, the likelihood of the corresponding *Firmicutes* genome tree was calculated in PAML (Yang

2007) using three molecular clock models: no molecular clock, in which the rate of every branch varied freely; a global clock, in which every branch shared the same evolutionary rate; and a local clock, where spore-forming and non-spore-forming branches had separate rates. The Akaike Information Criterion (AIC) was calculated for each model to determine which was the best. From the codon alignments, the synonymous substitution rate, *dS*, was calculated in PAML (Yang 2007) using a neutral-site, no-molecular-clock model.

Phylogenetic independent contrast analysis of evolutionary rates

Independent contrasts were performed using the R package 'caper' (Orme 2013). Discrete contrasts were made using the predicted sporulation potential (spore-forming or non-spore-forming). Continuous contrasts were made using the numeric count of identified sporulation genes. Contrasts with studentized residuals having absolute values greater than 3 were excluded from our analyses as potential outliers.

Estimation of codon bias

For each of the 197 *Firmicutes* representatives, CDS of all protein-coding genes in the genome were concatenated into a single nucleotide sequence. The codon bias index (CBI) was calculated for each genome in CodonW (http://codonw.sourceforge.net) based on the usage of a subset of *B. subtilis* optimal codons.

Results

Prediction of a species' sporulation potential

Sporulation is highly complex, but one of the best-studied biological processes. For example, in *Bacillus subtilis*, the model organism for studying sporulation, sporulation involves a cascade of expression of at least 200 genes (Piggot and Losick 2002). For a non-model organism, classifying whether it can sporulate or not based on laboratory observation is unreliable because we do not know exactly what environmental factors might trigger sporulation in the organism. When a species loses

its ability to sporulate (e.g., by losing a key sporulation gene), it is expected to eventually lose most of its sporulation genes since there is no longer selective pressure to maintain them, unless they also perform other biological functions. The tight correlation between the sporulation process and the number of sporulation genes has been used to predict a species' potential to form spores (Onyenwoke et al. 2004; Maughan 2007).

Gene content has been used to group species (Huson and Steel 2004) and in theory could be used to predict phenotypes associated with the genes as well. If we can partition *Firmicutes* into spore-forming and non-spore-forming groups by the distribution patterns of sporulation genes, then we should be able to predict the spore-forming ability of a species based on its membership in the groups. To predict the sporulation potential of *Firmicutes*, we carried out a phylogenetic profiling analysis of sporulation genes in all complete *Firmicutes* genomes. Using a mutual-best-hit approach, we identified orthologs of 163 known *B. subtilis* sporulation genes in other *Firmicutes* species (Supplementary Table S1). The distribution patterns of sporulation genes in 200 representatives of *Firmicutes* are shown in Figure 1. Based on the presence/absence of sporulation genes, *Firmicutes* were partitioned into two major groups at the root of the tree (Figure 1). The first group contained Staphylococcus, Streptococcus and Lactobacillus that are unable to sporulate, while the second group contained well-known spore-forming genera such as Bacillus and Clostridium (Figure 1). We therefore predicted that members of the first and second groups are NSFF and SFF respectively. Consistent with our predictions, not a single species within the first group is known to sporulate. The average number of sporulation genes in the genomes of the first group (95% CI 15.05±0.39 genes per genome) is significantly smaller than that of the second group (95% CI 85.2±4.42, $P < 0.0001$). Accordingly, when plotting the number of sporulation genes against the genome size, SFF form a cluster that is well separated from NSFF and other bacterial phyla (Supplementary Figure S1). However, we note that one butyrate-producing bacterium SM4/1 (NCBI taxonid 245012) in group 2 is

missing 146 of 163 sporulation genes including the master switch gene *spo0A*. We therefore predicted it to be non-spore-forming.

Evolution of sporulation in *Firmicutes*

To test whether SFF and NSFF have different evolutionary rates, it is a prerequisite to obtain a robust species tree of *Firmicutes*. First of all, the tree is required for estimating the overall species evolutionary rates. Secondly, the tree provides a framework of the natural relationships between species, so correlation due to common ancestry can be removed (Felsenstein 1985; Harvey and Purvis 1991) when we test the effect of sporulation on the evolutionary rates. Using the concatenated protein sequence alignment of 168 single-copy genes that are universally present in all *Firmicutes*, we reconstructed a maximum-likelihood tree of 200 *Firmicutes* representatives (Figure 2). The tree was rooted with the last common ancestor (LCA) of *Ammonifex* and *Sulfobacillus* as the deepest lineage, based on the topology of a bacterial species tree containing all major bacterial phyla. The tree is fully resolved and as typical of genome trees, the nodes in the tree are all highly supported with few exceptions.

Consistent with the previous study (Maughan 2007), our ancestral state reconstructions indicate that sporulation arose only once, in the LCA of *Firmicutes*, and subsequently has been lost many times (Figure 2). Interestingly, our ancestral state reconstruction shows that once the ability to form spores was lost, it was never subsequently regained – even with the relaxed assumption that trait gain and loss are equally likely. This is consistent with Dollo's law that complex traits, once lost, may be difficult to re-evolve (Marshall et al. 1994).

Some losses were ancient, for example, in the LCA of *Staphylococcus* and *Lactobacillus*, leading to the diversification of large clades of *Firmicutes* that are all unable to form spores. Other losses were more recent, one striking example being *Clostridiales genomospecies* BVAB3. A previous study (Galperin et al. 2012) and we both predicted

that BVAB3 is a NSFF as it encodes merely 8 sporulation genes. Consistently, endospore formation has not been observed for this species. Our ancestral state reconstructions predicted that BVAB3 lost the ability to sporulate after it split off from its close relatives. These relatives, *Clostridium clariflavum*, *Ruminiclostridium thermocellum*, *Clostridium cellulolyticum* and *Clostridium stercorarium* have all been observed to form spores in the laboratory (Freier et al. 1988; Gehin et al. 1995; Shiratori et al. 2009; Dumitrache et al. 2013). Consistent with our prediction that NSFF evolve more quickly, the branch length of BVAB3 to its last spore-forming ancestor is twice of those of its closest relatives – a clear demonstration of more rapid evolution coinciding with the loss of spore-forming ability.

Similar losses were predicted in species for which sporulation has not been reported: *Bacillus selenitireducens* (Afkar et al., 2003), the LCA of *Finegoldia magna* (formerly *Peptostreptococcus magnus,* Jassem et al., 1996) and *Anaerococcus prevotii* (LaButti et al., 2009), and the LCA of *Acetobacterium woodii* (Balch et al., 1977) and *Eubacterium limosum* (Genthner et al., 1981), among others. In each of these instances, the loss of endospore formation appears to coincide with longer branch lengths, compared to their close relatives.

Evolutionary rate varies between lineages

In comparing branch lengths of *Firmicutes* possessing various counts of sporulation genes, it is clear that evolutionary rate is not uniform (Figure 3, Supplementary Table S2). To determine if variation does in fact exist between lineages, various molecular clock models were tested in PAML. Our results show that a global molecular clock is the poorest model compared to both local and no-clock models (Supplementary Table S3), supporting the previous finding that protein evolutionary rates vary between *Firmicutes* lineages (Maughan 2007). In the local clock model, SFF and NSFF were each described by a separate evolutionary rate. The local model fits the data better than the global clock, showing that some variation in evolutionary rate can

be explained by spore-forming ability. Because the no-clock model fits the data best, there exists additional variation in evolutionary rate that cannot be explained by spore-forming ability alone. As the local clock does not take differences due to phylogeny into consideration, this is to be expected.

Spore-forming *Firmicutes* have lower protein evolutionary rates

The presence of many loss of trait events in the species tree of *Firmicutes* provides a powerful framework for us to test the effect of spore formation on evolutionary rate. We first tested whether the amino acid substitution rate variation correlates with spore-forming ability. To remove correlation due to shared evolutionary history, we conducted phylogenetically independent contrast analyses. Independent contrasts performed using 200 *Firmicutes* representatives showed that there is a strong association between the amino acid substitution rate (measured by the tree branch length of the tree in Figure 2) and spore-forming ability ($P < 0.001$, $df = 14$, Supplementary Table S4). One contrast involved three species (*Thermodesulfobium narugense, Coprothermobacter proteolyticus* and *Natranaerobius thermophilus*) that were suspected to be clustered together due to long-branch attraction (LBA). These three species were therefore removed, resulting in a 197-species dataset. Independent contrasts conducted with this reduced dataset also showed a strong association between spore-forming ability and evolutionary rate ($P < 0.0001$, $df = 12$, Table 1).

To ensure that our conclusions are not solely dependent on our predictions of spore-forming ability, as a precautionary measure we removed any SFF not corroborated by Galperin et al. (2012). This resulted in removal of 60 taxa predicted to be spore-forming in our study. We retained species that we classified as NSFF as these classifications are most likely accurate due to the low number of sporulation genes and the lack of master sporulation gene spo0A. Independent contrasts yielded results similar to the 197 and 200 species data set (Supplementary Table S5).

Because there is uncertainty associated with sporulation prediction based on phylogenetic profiles, we also performed phylogenetic contrast analysis using the nominal count of sporulation genes as a continuous variable. These continuous independent contrasts conducted with the 200 and 197 species data sets corroborated the results of discrete contrasts ($P < 0.0001$, $df = 199$ Supplementary Table S4; $P < 0.0001$, $df = 192$, Table 1) without having presumed any spore-forming ability, showing that species with greater counts of sporulation genes evolve more slowly.

To evaluate the effect of the robustness of the tree topology on our results, we removed three contrasts with less than 80% bootstrap support. Specifically, this eliminated 85 species descending from the LCA of *Exiguobacterium* spp and Lactobacillales. Independent contrasts for this 112-species tree yielded results similar to the 197 and 200 species data sets (Supplementary Table S6).

Spore-forming *Firmicutes* have lower spontaneous mutation rates

Because the amino acid substitution rate is the compound result of spontaneous mutation, selection and genetic drift, next we determined whether sporulation affects the spontaneous mutation rate alone. Spontaneous mutation is caused by DNA polymerase errors and other molecular processes that introduce errors during the transmission of genetic information, and therefore directly correlates with the generation time. We expect SFF with longer generation time to have lower spontaneous mutation rates than NSFF.

We measured the spontaneous mutation rate using the synonymous substitution rate (*dS*). Because synonymous mutations in general do not alter the phenotype, they are assumed to be selectively neutral and have been used to estimate the spontaneous mutation rate (Kimura and Ohta 1971, Ohta 1993). *dS* were calculated with codeml of the PAML package using aligned DNA sequences of the 168 genes shared among *Firmicutes* (Supplementary Table S2). Both discrete and continuous *dS*-based

phylogenetic contrasts indicate that spore-forming ability is significantly associated with the spontaneous mutation rates (Table 1).

No evidence of correlation between population size and sporulation

If the population size correlates with spore-forming ability, then genetic drift resulting from small population size could also explain differences in amino acid substitution rates. To test this potential confounding factor, we calculated genome-wide codon bias indices – a proxy for population size – for *Firmicutes* used in our analyses (Supplementary Table S2). Independent contrasts performed on codon bias indices (CBI) revealed no association with either predicted spore-forming ability (df = 13, P >0.05, R^2 = 0.007) or sporulation gene count (df = 192, P > 0.05, R^2 = 0.01154).

Discussion

Using a large dataset of 200 *Firmicutes* species, we found sporulation significantly reduces the rates of molecular evolution in bacteria. This is in sharp contrast to a previous study that found rates of molecular evolution in bacteria are relatively constant despite spore dormancy (Maughan 2007). One possible reason that the study by Maughan (2007) failed to detect significant difference in evolutionary rates between SFF and NSFF is the lack of power in the data: its conclusion was based on only two phylogenetic contrasts. In comparison, we had at least 13 contrasts because of the large number of species included in our study. Perhaps a more important reason is the likely misclassification of sporulation ability of some key *Firmicutes* species in the previous study. One of the two contrasts in Maughan (2007) study involved *Thermoanaerobacter tengcongensis*, which was classified as non-spore forming and was contrasted against three spore-forming *Clostridium* spp. We think the contrast is likely wrong. *T. tengcongensis* encodes 63 sporulation genes (Figure 2) and therefore was predicted to be able to sporulate in this and a previous study (Traag et al. 2013). While *T. tengcongensis* has not been observed to sporulate when it was first identified (Xue et al. 2001), RNA-seq has shown that a cold-shock protein induces expression of sporulation genes in response to low temperatures (Liu et al. 2014). We note that misclassification of SFF or NSFF

will reduce the power to detect generation-time effect. The fact we were able to detect generation-time effect suggests that it is real and might be even stronger.

In this study, we observed a generation-time effect in both synonymous and nonsynonymous substitution rates of protein-coding genes. In theory, synonymous changes are not affected by selection and genetic drift and therefore are expected to reflect the underlying mutation rate. The generation-time effect on evolutionary rate was originally found to be most evident at synonymous sites (Li and Tanimura 1987; Ohta 1993). This is because on top of mutation rate, nonsynonymous substitutions are also affected by selection and genetic drift, which can obscure the generation time effect in amino acid substitutions. For example, mammals with longer generation time tend to have smaller population size, which causes accelerated substitution rates for slightly deleterious mutations. This explains the failure to detect a generation-time effect in nonsynonymous substitution rates in mammals: it is cancelled out by the genetic drift. In bacteria, population size varies between species and is known to have a large impact on nonsynonymous substitution rates. For example, obligate intracellular bacteria usually have much smaller population sizes than free-living bacteria. The increased genetic drift in obligate intracellular bacteria contributes to their highly accelerated amino acid substitution rates (Ochman et al. 1999). The extent to which selection and genetic drift differ between SFF and NSFF is not well known. The fact that we were able to detect a generation-time effect in nonsynonymous substitution rates suggests that neither has obscured the signal in *Firmicutes*. Consistent with this, using the codon usage bias as a proxy for estimating the effective population size, we did not detect significant correlation between population size and sporulation, indicating that genetic drift is not the main cause of rate variation in *Firmicutes*. In fact, in our study we found even a stronger evidence of a generation-time effect in nonsynonymous substitution rate (Table 1). We think this is because the rate differences in synonymous substitution rate were underestimated due to saturated synonymous substitutions in our dataset. Indeed, the average *dS* in our phylogenetic

contrast analysis approached 2.66 substitutions per site, a clear sign of saturated substitutions. In addition, it is possible that endospores experience less selection at nonsynonymous sites because of the dormancy, resulting in fewer substitutions in SFF and contributing to the greater disparity.

Another source of mutations in bacteria is replication-independent mutations such as these caused by unrepaired DNA damages. Because they are independent of DNA replication, they are unaffected by generation time and therefore can obscure the generation-time effect. Spores have tremendous resistance to DNA damage, however, due to a combination of multiple factors including DNA packing assisted by small acid-soluble spore proteins, dehydration, the high concentration of dipicolinic acid (Desnous et al. 2010) and spore photoproduct lyase that repairs UV damage (Van Wang and Rupert 1977). Again, the fact that we observed generation-time effect in *dS* suggests that replication-independent mutations do not play a major role in the evolution of *Firmicutes*.

Our ancestral reconstruction suggests that sporulation evolved only once in the last common ancestor of *Firmicutes* and has since been lost many times in non-spore-forming lineages. Assuming that once the sporulation trait is lost so will most of the sporulation genes, the sporulation gene content can be used not only for predicting the sporulation ability but more importantly the time a species has spent as a spore-former in its evolutionary history. Using sporulation gene number in phylogenetic contrast analysis therefore circumvents the uncertainty associated with the binary trait prediction. Regardless of whether we use a binary (sporulation ability) or continuous variable (number of sporulation genes) in our phylogenetic contrast analyses, we detected significant generation-time effect in *Firmicutes*.

The generation-time effect appears to be genome wide as our results were based on the analysis of 168 genes common among all *Firmicutes*. Our analyses were not based

on laboratory observations, but instead on genomes – the products of bacteria reproducing and evolving over millions of years. As such, they provide strong evidence of generation-time effect in natural bacterial populations throughout an extended evolutionary history.

Tables & Figures

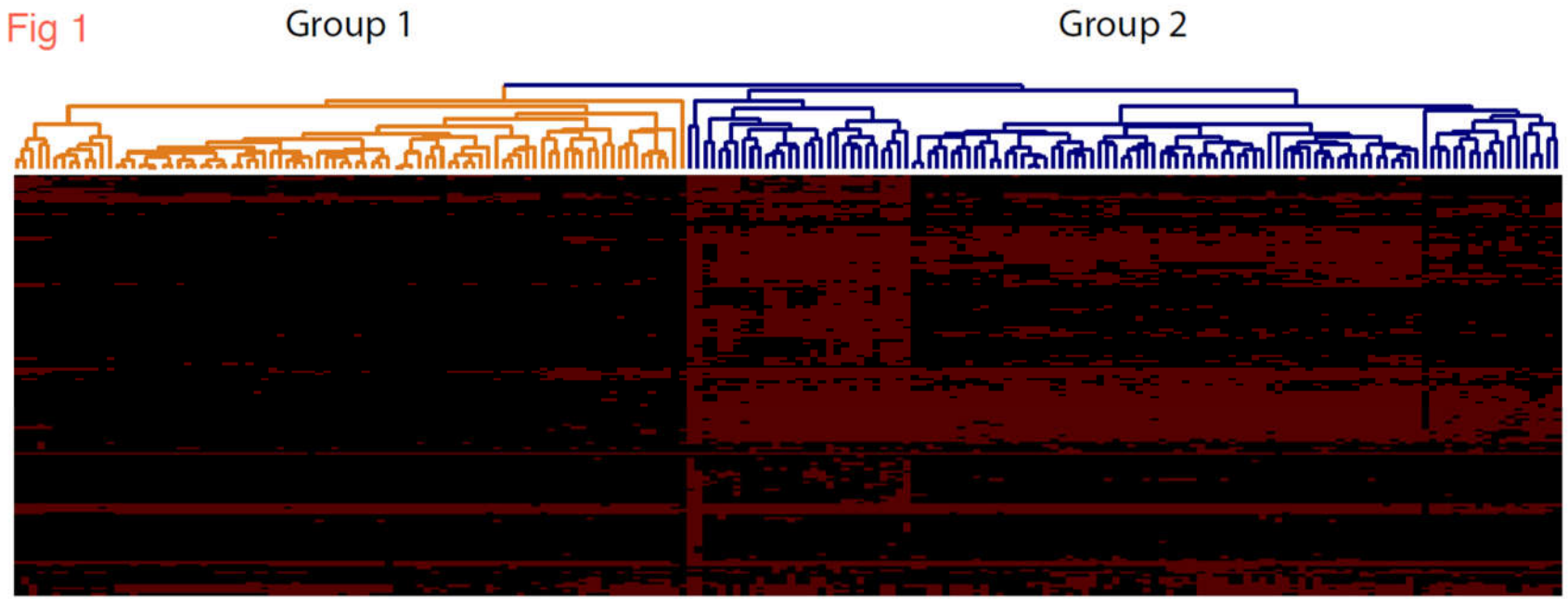


Figure 1. Phylogenetic profile analysis of 163 sporulation genes in genomes of 200 *Firmicutes* representatives.

Each row represents one sporulation gene and each column represents one genome. The presence and absence of genes are indicated in red and black, respectively. *Firmicutes* were clustered into two groups by the similarities of the distribution profiles of sporulation genes in the genomes. The phylogenetic profiles are also described in Supplementary Table S1.

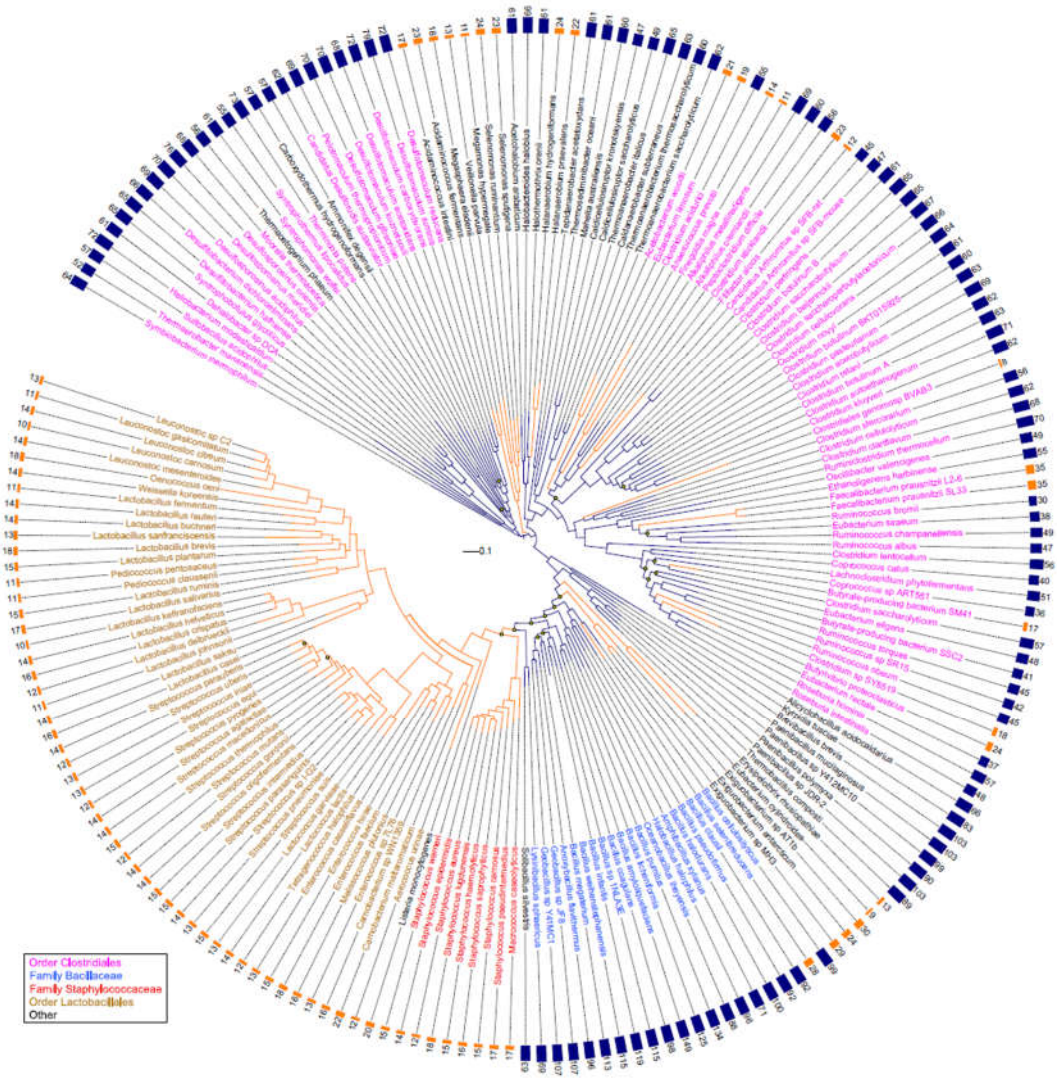


Figure 2. A maximum-likelihood genome tree of 197 *Firmicutes* representatives based on 168 shared marker genes.

The size of the bar on the outer circle represents the number of identified sporulation genes in the genome. The bars were colored by the predicted spore-forming potential of their corresponding genomes (blue: spore-forming; orange: non-spore-forming). The branches were highlighted using the same color scheme based on the maximum-likelihood ancestral state reconstruction of spore-forming ability of the lineages. Bootstrap support for nodes is $\geq$ 80% except for those indicated with a yellow circle. Full-resolution image available as separate file.

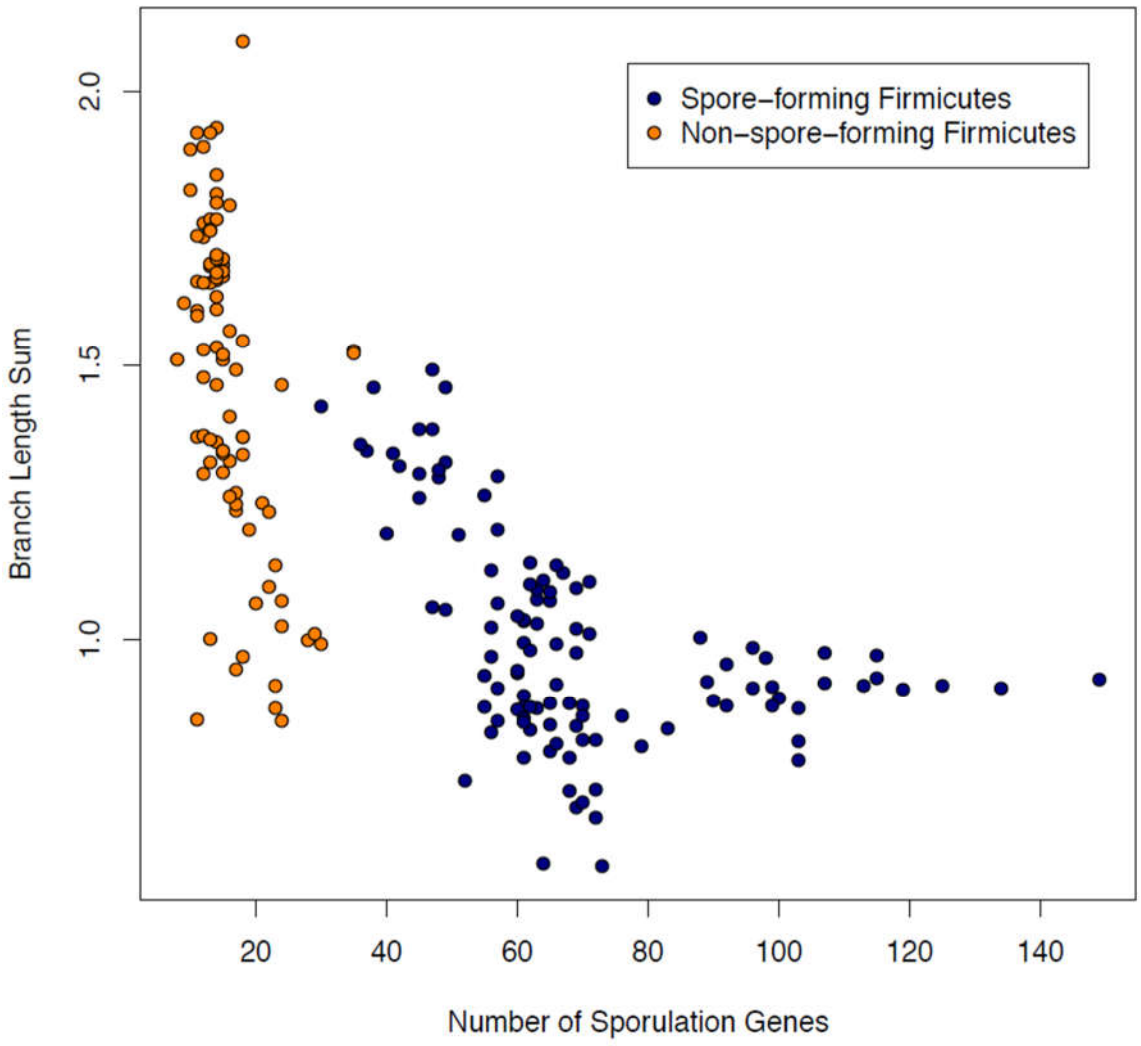
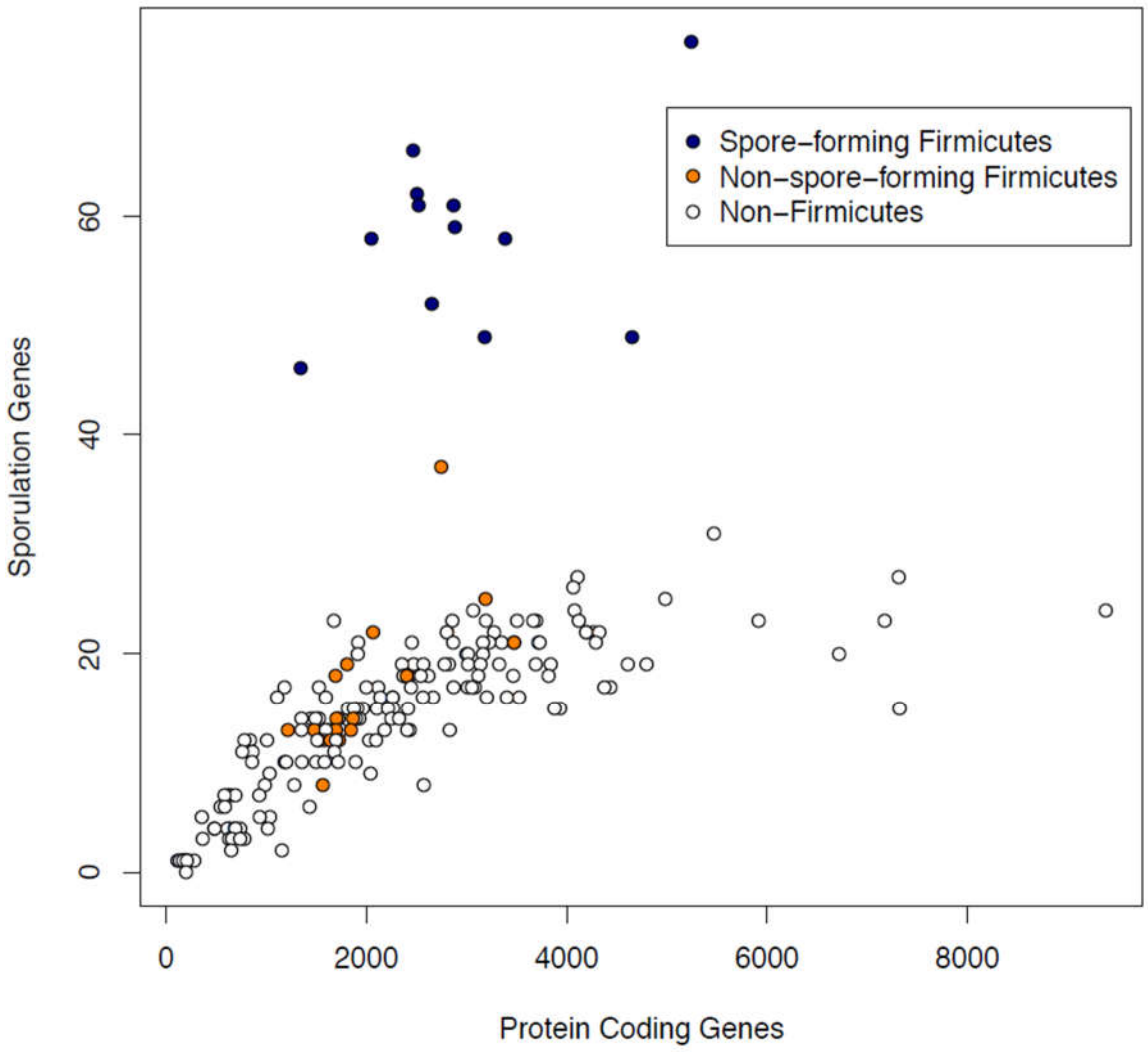


Figure 3. Relationship between sporulation gene count and the branch length (from the tip to the root) of 197 *Firmicutes* representatives.

Each dot represents a genome and is colored by the genome's predicted sporulation potential.



Supplementary Figure S1. The number of sporulation genes plotted against the number of protein coding genes in the genomes.

Each dot represents a genome and is colored as: orange, spore-forming *Firmicutes*; blue, non-spore-forming *Firmicutes*; blank, non-*Firmicutes* bacteria.

Supplementary Table S1. Phylogenetic profiles of sporulation genes in genomes of 200 *Firmicutes* representatives **(Available as separate file).**

Supplementary Table S2. List of molecular evolutionary rates and codon bias index for each of the 197 *Firmicutes* species analyzed in this study **(Available as separate file).**

Supplementary Table S3. Log likelihood and AIC values of different clock models.

Model	-log likelihood	Parameters	AIC
Global	16063256.15	1	32126514.31
Local	16051271.69	2	32102547.38
No clock	15991403.07	391	31983588.14

Supplementary Table S4. Independent contrast analysis of protein evolutionary rates of 200 *Firmicutes* representatives.

Sporulation variable	Evolutionary Rate	Contrasts	P-value (exact)	P-value	R^2
Discrete	amino acid substitution	15	0.0007639	< 0.001	0.5667
Continuous	amino acid substitution	197	2.636E-14	< 0.0001	0.2567

Supplementary Table S5. Independent contrast analysis of evolutionary rates in 137 *Firmicutes* representatives after removing 60 species not positively identified as spore-forming in Galperin et al. (2012).

Sporulation variable	Evolutionary Rate	Contrasts	P-value (exact)	P-value	R^2
Discrete	amino acid substitution	11	0.001255	< 0.01	0.6634
Discrete	*dS*	10	0.001815	< 0.01	0.679
Continuous	amino acid substitution	135	1.12E-10	< 0.0001	0.2669
Continuous	*dS*	135	3.52E-09	< 0.0001	0.2299

Supplementary Table S6. Independent contrast analysis of evolutionary rates in 112 *Firmicutes* representatives after removing three contrast nodes with less than 80% bootstrap support from the 197-species dataset.

Sporulation variable	Evolutionary Rate	Contrasts	P-value (exact)	P-value	R^2
Discrete	amino acid substitution	11	0.004755	< 0.01	0.566
Discrete	*dS*	11	0.04623	< 0.05	0.3409
Continuous	amino acid substitution	110	3.39E-16	< 0.0001	0.4586
Continuous	*dS*	109	2.61E-14	< 0.0001	0.417

Literature Cited

Abecasis, A. B., M. Serrano, R. Alves, L. Quintais, J. B. Pereira-Leal, and A. O. Henriques. 2013. A genomic signature and the identification of new sporulation genes. J. Bacteriol. 195:2101–15.

Ainouche, A.-K., and R. J. Bayer. 1999. Phylogenetic relationships in Lupinus (Fabaceae: Papilionoideae) based on internal transcribed spacer sequences (ITS) of nuclear ribosomal DNA. Am. J. Bot. 86:590–607.

Andreasen, K., and B. G. Baldwin. 2001. Unequal Evolutionary Rates Between Annual and Perennial Lineages of Checker Mallows (Sidalcea, Malvaceae): Evidence from 18S-26S rDNA Internal and External Transcribed Spacers. Mol. Biol. Evol. 18:936–944.

Bousquet, J., S. H. Strauss, A. H. Doerksen, and R. A. Price. 1992. Extensive variation in evolutionary rate of rbcL gene sequences among seed plants. Proc. Natl. Acad. Sci. 89:7844–7848.

Bromham, L. 2002. Molecular Clocks in Reptiles: Life History Influences Rate of Molecular Evolution. Mol. Biol. Evol. 19:302–309.

Bromham, L., A. Rambaut, and P. H. Harvey. 1996. Determinants of rate variation in mammalian DNA sequence evolution. J. Mol. Evol. 43:610–621.

Bueche, M., T. Wunderlin, L. Roussel-Delif, T. Junier, L. Sauvain, N. Jeanneret, and P. Junier. 2013. Quantification of endospore-forming firmicutes by quantitative PCR with the functional gene spo0A. Appl. Environ. Microbiol. 79:5302–12.

Buschiazzo, E., C. Ritland, J. Bohlmann, and K. Ritland. 2012. Slow but not low: genomic comparisons reveal slower evolutionary rate and higher dN/dS in conifers compared to angiosperms. BMC Evol. Biol. 12:8.

Cano, R. J., and M. K. Borucki. 1995. Revival and identification of bacterial spores in 25- to 40-million-year-old Dominican amber. Science 268:1060–4.

Desnous, C., D. Guillaume, and P. Clivio. 2010. Spore photoproduct: a key to bacterial eternal life. Chem. Rev. 110:1213–32.

Dumitrache, A., G. Wolfaardt, G. Allen, S. N. Liss, and L. R. Lynd. 2013. Form and function of Clostridium thermocellum biofilms. Appl. Environ. Microbiol. 79:231–9.

Felsenstein, J. 1985. Phylogenies and the Comparative Method. Am. Nat. 125:1 – 15.

Freier, D., C. P. Mothershed, and J. Wiegel. 1988. Characterization of Clostridium thermocellum JW20. Appl. Environ. Microbiol. 54:204–211.

Galperin, M. Y., S. L. Mekhedov, P. Puigbo, S. Smirnov, Y. I. Wolf, and D. J. Rigden. 2012. Genomic determinants of sporulation in Bacilli and Clostridia: towards the minimal set of sporulation-specific genes. Environ. Microbiol. 14:2870–90.

Gaut, B. S., L. G. Clark, J. F. Wendel, and S. V Muse. 1997. Comparisons of the molecular evolutionary process at rbcL and ndhF in the grass family (Poaceae). Mol. Biol. Evol. 14:769–77.

Gaut, B. S., B. R. Morton, B. C. McCaig, and M. T. Clegg. 1996. Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene Adh parallel rate differences at the plastid gene rbcL. Proc. Natl. Acad. Sci. 93:10274–10279.

Gehin, A., E. Gelhaye, G. Raval, and H. Petitdemange. 1995. Clostridium cellulolyticum Viability and Sporulation under Cellobiose Starvation Conditions. Appl. Environ. Microbiol. 61:868–71.

Harvey, P. H., and A. Purvis. 1991. Comparative methods for explaining adaptations. Nature 351:619–24.

Huson, D. H., and M. Steel. 2004. Phylogenetic trees based on gene content. Bioinformatics 20:2044–9.

Kohne, D. E. 1970. Evolution of higher-organism DNA. Q. Rev. Biophys. 3:327–375. Cambridge University Press.

Kuo, C.-H., and H. Ochman. 2009. Inferring clocks when lacking rocks: the variable rates of molecular evolution in bacteria. Biol. Direct 4:35.

LAIRD, C. D., B. L. McCONAUGHY, and B. J. McCARTHY. 1969. Rate of Fixation of Nucleotide Substitutions in Evolution. Nature 224:149–154.

Laroche, J., and J. Bousquet. 1999. Evolution of the mitochondrial rps3 intron in perennial and annual angiosperms and homology to nad5 intron 1. Mol. Biol. Evol. 16:441–452.

Laroche, J., P. Li, L. Maggia, and J. Bousquet. 1997. Molecular evolution of angiosperm mitochondrial introns and exons. Proc. Natl. Acad. Sci. 94:5722–5727.

Li, W. H., and M. Tanimura. 1987. The molecular clock runs more slowly in man than in apes and monkeys. Nature 326:93–6.

Liu, B., Y. Zhang, and W. Zhang. 2014. RNA-Seq-based analysis of cold shock response in Thermoanaerobacter tengcongensis, a bacterium harboring a single cold shock protein encoding gene. PLoS One 9:e93289.

Marshall, C. R., E. C. Raff, and R. A. Raff. 1994. Dollo's law and the death and resurrection of genes. Proc. Natl. Acad. Sci. U. S. A. 91:12283–7.

Maughan, H. 2007. Rates of molecular evolution in bacteria are relatively constant despite spore dormancy. Evolution 61:280–8.

Mooers, A. O., and P. H. Harvey. 1994. Metabolic rate, generation time, and the rate of molecular evolution in birds. Mol. Phylogenet. Evol. 3:344–50.

Nabholz, B., S. Glémin, and N. Galtier. 2008. Strong variations of mitochondrial mutation rate across mammals--the longevity hypothesis. Mol. Biol. Evol. 25:120–30.

Ochman, H., S. Elwyn, and N. A. Moran. 1999. Calibrating bacterial evolution. Proc. Natl. Acad. Sci. U. S. A. 96:12638–43.

Ohta, T. 1993. An examination of the generation-time effect on molecular evolution. Proc. Natl. Acad. Sci. U. S. A. 90:10676–80.

Onyenwoke, R. U., J. A. Brill, K. Farahi, and J. Wiegel. 2004. Sporulation genes in members of the low G+C Gram-type-positive phylogenetic branch (Firmicutes). Arch. Microbiol. 182:182–92.

Orme, D. 2013. The caper package: comparative analysis of phylogenetics and evolution in R.

Paradis, E., J. Claude, and K. Strimmer. 2004. APE: Analyses of Phylogenetics and

Evolution in R language. Bioinformatics 20:289–90.

Piggot, P. J., and R. Losick. 2002. Sporulation Genes and Intercompartmental Regulation. Pp. 483–518 *in* A. L. Sonenshein, J. A. Hoch, and R. Losick, eds. Bacillus subtilis and Its Closest Relatives. American Society for Micriobiology, Washington, DC.

Price, M. N., P. S. Dehal, and A. P. Arkin. 2009. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol. Biol. Evol. 26:1641–50.

Priest, F. G., and R. Grigorova. 1990. 18 Methods for Studying the Ecology of Endospore-forming Bacteria. Methods Microbiol. 22:565–591. Elsevier.

Shiratori, H., K. Sasaya, H. Ohiwa, H. Ikeno, S. Ayame, N. Kataoka, A. Miya, T. Beppu, and K. Ueda. 2009. Clostridium clariflavum sp. nov. and Clostridium caenicola sp. nov., moderately thermophilic, cellulose-/cellobiose-digesting bacteria isolated from methanogenic sludge. Int. J. Syst. Evol. Microbiol. 59:1764–70.

Smith, S. A., and M. J. Donoghue. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science 322:86–9.

Soria-Hernanz, D. F., O. Fiz-Palacios, J. M. Braverman, and M. B. Hamilton. 2008. Reconsidering the generation time hypothesis based on nuclear ribosomal ITS sequence comparisons in annual and perennial angiosperms. BMC Evol. Biol. 8:344. BIOMED CENTRAL LTD, CURRENT SCIENCE GROUP, MIDDLESEX HOUSE, 34-42 CLEVELAND ST, LONDON W1T 4LB, ENGLAND.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics, doi: 10.1093/bioinformatics/btu033.

Steel, M. 2005. Phylogenetic diversity and the greedy algorithm. Syst. Biol. 54:527–9.

Thomas, J. A., J. J. Welch, R. Lanfear, and L. Bromham. 2010. A generation time effect on the rate of molecular evolution in invertebrates. Mol. Biol. Evol. 27:1173–80.

Thomas, J. A., J. J. Welch, M. Woolfit, and L. Bromham. 2006. There is no universal molecular clock for invertebrates, but rate variation does not scale with body size. Proc. Natl. Acad. Sci. U. S. A. 103:7366–71.

Traag, B. A., A. Pugliese, J. A. Eisen, and R. Losick. 2013. Gene conservation among endospore-forming bacteria reveals additional sporulation genes in Bacillus subtilis. J. Bacteriol. 195:253–60.

Van Wang, T. C., and C. S. Rupert. 1977. Evidence for the monomerization of spore photoproduct to two thymines by the light-independent "spore repair" process in Bacillus subtilis. Photochem. Photobiol. 25:123–7.

Wang, Z., and M. Wu. 2013. A phylum-level bacterial phylogenetic marker database. Mol. Biol. Evol. 30:1258–62.

Welch, J. J., O. R. P. Bininda-Emonds, and L. Bromham. 2008. Correlates of substitution rate variation in mammalian protein-coding sequences. BMC Evol. Biol. 8:53.

Wu, M., Q. Ren, A. S. Durkin, S. C. Daugherty, L. M. Brinkac, R. J. Dodson, R.
Madupu, S. A. Sullivan, J. F. Kolonay, D. H. Haft, W. C. Nelson, L. J. Tallon, K. M. Jones, L. E. Ulrich, J. M. Gonzalez, I. B. Zhulin, F. T. Robb, and J. A. Eisen. 2005. Life in hot carbon monoxide: the complete genome sequence of Carboxydothermus hydrogenoformans Z-2901. PLoS Genet. 1:e65.

Wu, M., and A. J. Scott. 2012. Phylogenomic analysis of bacterial and archaeal sequences with AMPHORA2. Bioinformatics 28:1033–4.

Xue, Y., Y. Xu, Y. Liu, Y. Ma, and P. Zhou. 2001. Thermoanaerobacter tengcongensis sp. nov., a novel anaerobic, saccharolytic, thermophilic bacterium isolated from a hot spring in Tengcong, China. Int. J. Syst. Evol. Microbiol. 51:1335–41.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol. Biol. Evol. 24:1586–91.

Yue, J.-X., J. Li, D. Wang, H. Araki, D. Tian, and S. Yang. 2010. Genome-wide investigation reveals high evolutionary rates in annual model plants. BMC Plant Biol. 10:242.

CHAPTER TWO

Accurate, ultra-low coverage genome reconstruction and association studies in Hybrid Swarm mapping populations

---

A version of this chapter is made available as a preprint on bioRxiv; doi: https://doi.org/10.1101/671925

Abstract

Genetic association mapping studies seek to uncover the link between genotype and phenotype, and often utilize inbred reference panels as a replicable source of genetic variation. However, inbred reference panels can differ substantially from wild populations in their genotypic distribution, and patterns of linkage-disequilibrium and nucleotide diversity. As a result, associations discovered using inbred reference panels may not reflect the genetic basis of phenotypic variation in natural populations. To address this problem, we evaluated a mapping population design where dozens to hundreds of inbred lines are outbred for few (e.g. five) generations, which we call the Hybrid Swarm. The Hybrid Swarm approach has likely remained underutilized relative to pre-sequenced inbred lines due to the costs of genome-wide genotyping. To reduce sequencing costs and make the Hybrid Swarm approach feasible, we developed a computational pipeline that reconstructs accurate whole genomes from ultra-low-coverage (0.05X) sequence data in Hybrid Swarm populations derived from ancestors with phased haplotypes. We compared the power and precision of Genome-Wide Association Studies (GWAS) using the Hybrid Swarm, inbred lines, recombinant inbred lines, and highly outbred populations across a range of allele frequencies and effect sizes, modeling genetic variation from the *Drosophila melanogaster* Genetic Reference Panel as well as variation from neutral simulations. While inbred populations tended to perform best due to the intrinsic power benefits conferred by the lack of heterozygotes, association mapping with the Hybrid Swarm performed comparably to highly outbred (F_{50}) populations and has higher precision than mapping with inbred lines. Taken together, these results demonstrate the feasibility of the Hybrid Swarm as a cost-effective method of fine-scale genetic mapping.

Introduction

Genetic mapping studies seek to describe the link between genotype and phenotype. For experimental crosses, mapping was traditionally conducted by scoring the phenotypes of recombinant offspring descended from a limited number of parental

lines. While such Quantitative Trait Loci (QTL) mapping studies can have high power to detect associations, they offer minimal mapping resolution (Cheng *et al.* 2010), often detecting broad regions of phenotypic association (Bergland *et al.* 2012). If linkage disequilibrium is lowered, spurious associations become rarer (Li *et al.* 2005) and associations can be resolved at the gene or nucleotide level, as in GWAS of large outbred populations (Nikpay *et al.* 2015; Wu *et al.* 2017; Monir and Zhu 2017). However, GWAS suffer from reduced power to detect associations, necessitating a large sample size relative to QTL mapping (Spencer *et al.* 2009).

To generate higher resolution mapping populations than the traditional biparental F2 design, Multiparent Populations (MPPs) are commonly used. By crossing together multiple genetically diverse inbred lines, researchers can leverage maintainable diversity from inbred lines to produce mapping populations without sampling wild individuals. The mapping resolution of MPPs depends on the extent of linkage disequilibrium, and resolution is improved by allowing for more recombination between haplotypes, or by incorporating a greater number of genetically diverse haplotypes (Mott *et al.* 2000; Chia *et al.* 2005). Advanced intercross lines (AILs) are outbred MPPs generated by crossing two inbred lines together for a limited number of generations (ca. 5-20), providing greater mapping resolution than biparental F_2 crosses (Darvasi and Soller 1995). Further increases to mapping resolution can be achieved with Recombinant Inbred Lines (RILs), whereby founding lines are extensively crossed for fifty or so generations, followed by isogenization for long-term maintenance. Such multi-parent populations are commonly used for the dissection of complex traits in model organisms (Chesler *et al.* 2008; Kover *et al.* 2009; King *et al.* 2012b) and agriculturally important crops (Huang *et al.* 2012; Singh *et al.* 2013; Krämer *et al.* 2014).

One alternative approach for generating a high-resolution mapping population is to substitute extensive recombination for increased haplotype diversity. By crossing dozens to hundreds of inbred lines for a limited number of generations, heterozygous

mapping populations can be generated quickly (as with AILs) with sufficiently reduced LD to detect associations with high resolution (as with RILs). Unfortunately, the many-haplotypes few-generations method is not without its drawbacks. First, including many haplotypes decreases the frequency of the rarest alleles, reducing power to detect associations. Second, such an outbred population would require genotyping efforts, unlike pre-sequenced homozygous lines. The net requirement of genotyping a large sample size may explain the continued widespread use of pre-genotyped recombinant inbred lines for genetic association experiments in model systems (Huang *et al.* 2011; King *et al.* 2012b; MacKay *et al.* 2012; Srivastava *et al.* 2017).

Here, we describe computational methods that allow for cost-effective association mapping with a large outbred population. The Hybrid Swarm is founded by dozens to hundreds of inbred lines, crossed for a limited number of generations. To reduce genotyping costs of the Hybrid Swarm, we developed and evaluated a pipeline to reconstruct whole genomes using ultra-low coverage sequencing data. We developed and tested our pipeline by reconstructing whole genomes for thousands of simulated Hybrid Swarm individuals. Our simulated genomes draw from natural variation in the *Drosophila melanogaster* Genetic Reference Panel (DGRP), as well as from variation generated from coalescent models representing a broad range of genetic diversity parameters for common model systems. We show that the Hybrid Swarm approach allows for highly accurate genotyping (average 99.9% genotypic accuracy) from ultra-low-coverage (0.005-0.05X) whole-genome individual-based sequencing. We then perform simulated GWAS to describe power and precision of association mapping in the Hybrid Swarm compared to inbred lines, recombinant inbred lines, and a highly outbred (F_{50}) population. Our computational tools are capable of efficiently simulating low-coverage reconstruction and GWAS power analysis of any model system. Together, our results demonstrate the feasibility of cost-effective high-resolution association mapping in a large outbred population.

Methods

Generating and preparing simulated reference panels.

In order to evaluate low-coverage reconstruction for various degrees of genetic diversity, we generated reference panels using haplotypes produced by coalescent models across a range of genetic diversity levels. Haplotypes were generated using the *R* (R Core Team 2016) package *scrm* (Paul R. Staab et al. 2015) and subsequently restructured into VCF file format (Danecek et al. 2011). We generated ten independent panels for each of 18 combinations of population size ($N_e = 10^4, 10^5, 10^6$), mutation rate ($\mu = 10^{-9}, 5 \times 10^{-9}, 10^{-8}$), and number of haplotypes (32, 128). The value for θ for each simulation was defined as $4N_e\mu$. We simulated a chromosome-length locus of 25 Mb with a recombination rate of 1.5 cM/Mb. SNP positions output by *scrm* (a decimal within the range of 0 to 1) were converted to base pair positions by multiplying the decimal by chromosome length (25 $\times 10^6$ base pairs for our simulations) and rounding down to the nearest integer. Any sites with more than two alleles were converted to a biallelic site by discarding tertiary or quaternary alleles. Genotype values were re-coded as polarized signed integers: +1 for reference and -1 for alternate alleles. For every position, reference and alternate alleles were defined by randomly selecting one of the twelve non-repeating pairs of nucleotides. Reference genome FASTA files were created with a custom python script that generated a 25 million length string of nucleotide characters with weighted probability to achieve 45% GC-content, followed by replacing variable positions with their respective reference alleles.

Preparing DGRP haplotype data

As a case study of low-coverage genome reconstruction in a model system, we incorporated wild fruit fly genetic diversity from the *Drosophila melanogaster* Genetic Reference Panel (Mackay *et al.* 2012) DGRP freeze 2 as available from the *Drosophila* Genome Nexus (Lack *et al.* 2015). To minimize missing data, we included the 129 lines (out of 205) which exhibited aligned whole genome FASTA files with less than 50% of

nucleotides indicated by the ambiguity character N. We excluded insertions, deletions, fixed sites, and sites with more than two alleles. Any heterozygous genotype calls were masked as missing data. Diploid genotypes were re-coded as a single signed integer value, with +1 for homozygous reference, -1 for homozygous alternate, and 0 for missing data. This resulted in a polarized VCF file containing only biallelic SNPs and only homozygous (or missing) genotype calls.

Simulating Mapping Populations

To generate simulated populations, we developed a forward-simulator in R that stores ancestral haplotype block maps instead of genotypes. Our analyses necessitated a method of storing genotype information for thousands of individuals across thousands of simulations. To do so, we leveraged information redundancy that exists between related individuals in recombinant populations, generating haplotype block files. We achieved between three and four orders of magnitude of compression relative to a VCF file. For example, for a population containing 5000 diploid genotypes at nearly four million sites, a compressed VCF file is approximately 6.5 GB, compared to approximately 3.5 MB for a haplotype block file. This reduced file size is what allowed us to generate and store 28,000 total independent GWAS simulations (500 each for 56 parameter combinations). When haplotype block ancestry is known and recorded, as is possible with simulations, genotypes must only be recorded once (for the ancestral founders). Recombinant individual genotypes can then be reconstituted by extracting ancestral genotypes from ancestor and base pair position indices.

We simulated Hybrid Swarms through random mating over five non-overlapping generations at a population size of 10,000. Simulations proceeded in the following manner: first, a subset of either 32 or 128 founders was selected. Then, of that founder subset, 10,000 individuals were sampled with replacement. All possible founders were chosen with equal probability, and were assigned male or female sex with a 1:1 ratio, where sex was determined by the presence of a designated sex

chromosome. Sexual reproduction was simulated by random sampling of recombinant gametes from male-female pairs. Once 10,000 recombinant progeny were generated, the parental generation was discarded. Reproduction continued until the F_5 population was achieved. Recombination frequency was modeled as a Poisson process with an expected value $\lambda = \Sigma(Morgans)$ per chromosome. For simulations of *D. melanogaster* populations based on DGRP chromosomes, recombination occurred only in females, with recombination frequency and position based on values from Comeron et al (2012). For populations founded by simulated haplotypes, recombination occurred in both sexes, with recombination occurring uniformly across each chromosome (Supplemental Figure S1).

Simulating and Mapping Sequencing Data

We used *wgsim* (Li 2011) to generate simulated reads. To achieve a desired level of sequencing coverage $C = 0.05$ or 0.005, we generated $N = (C \times S)/(2 \times L)$ reads per chromosome, with read length $L = 100$ bp and chromosome length S bp. We specified a base error rate of 0.001 and an indel fraction of 0. Remaining *wgsim* parameters were left as default.

We assembled paired-end reads using *PEAR* (Zhang *et al.* 2014) and separately aligned the assembled and unassembled groups to a reference genome with *bwa* 0.7.14 using the BWA-MEM algorithm (Li 2013). Reads from DGRP-derived populations were mapped to the *D. melanogaster* reference genome v5.39, and reads from coalescent-derived populations were mapped to their respective simulated reference genomes. After converting mapped reads to compressed BAM format with *samtools* 1.3.1 (Li *et al.* 2009), we removed PCR duplicates with *Picard tools* 2.0.1 (Broad Institute 2015a).

Most Likely Ancestors Selection

To make chromosome reconstructions in the hybrid swarm computationally tractable (Figure 1), we developed a method of accurately selecting a subset of most

likely ancestors for any single chromosome. We then used that ancestor subset to reconstruct haplotype blocks using the *RABBIT* package (Zheng *et al.* 2015) in Mathematica. RABBIT operates as a Hidden Markov Model (HMM) using the Viterbi algorithm to return the most likely series of parental combinations (hidden states) across the genome (SNP positions) given the observations (sequenced alleles). For every position in the genome, the Viterbi algorithm evaluates relative likelihoods of transitioning to any possible hidden state. Because the hidden states in our case are ancestor combinations, there will be $(N^2 + N)/2$ combinations of N haplotypes to evaluate at every site. This number of evaluations is tractable at smaller values of N but grows at a quadratic rate. For example, increasing the number of founding haplotypes from 8 to 128 is a 16-fold increase in haplotypes, but it would incur orders of magnitude increases in computational effort (Figure 1). Thus, in order to make reconstructions in *RABBIT* computationally tractable for hybrid swarm individuals, it is necessary to identify a subset of founders that accurately includes the true ancestors contributing to any given chromosome.

We used the software package *HARP* (Kessner *et al.* 2013) to rank the population founding lines based on likelihood of being a true ancestor of a chromosome to be reconstructed. *HARP* was originally developed to estimate haplotype frequencies from pooled sequence data, and we co-opted it to assess relative likelihood that any founder contributed to a genomic window. We ran *HARP* with non-overlapping 100 kb windows with a minimum frequency cutoff 0.0001, producing output which can be visualized as a heat map of ancestor likelihood across the chromosome. A custom *R* script analyzed this *HARP* output and ranked all possible founders in terms of likelihood of contribution for a given chromosome. Briefly, a chromosome-wide significance threshold was calculated, e.g. the 95% or 99% quantile of all likelihoods across all founders and all chromosome windows. Then, every potential ancestor for each 100 kb window was classified as falling above or below this threshold. Founding lines were then ranked in descending order of the number of

windows passing the threshold. We examined two measures of effectiveness for this method across a range of quantile threshold values (90%, 95%, 99%, and 99.9%) when selecting up to a maximum number of most likely ancestral haplotypes. The first measure is the number of true ancestral founders excluded; the second measure is the fraction of the chromosome derived from ancestors missing from the selected subset.

Chromosome Reconstruction with RABBIT

We used the MAGIC reconstruct method of the Mathematica package RABBIT (Zheng *et al.* 2015) to perform chromosome reconstructions, which has been shown to be accurate for genotype estimation at sequencing coverage at 0.05X for a variety of multiparent populations (Zheng *et al.* 2018). RABBIT requires three inputs: observed genotypes in the individual being reconstructed; map distance (in cM units) of the same loci; and genotypes for the potential ancestors at those same loci. For DGRP-derived simulated populations, we specified map distance based on values reported by Comeron et al. (2012) by performing linear interpolation of cumulative map units (cM) as a function of base pair position. For populations derived from simulated haplotypes, we used a linear function of 37.5 cM over each 25 Mb chromosome. To specify genotype information, we first counted reference and alternate reads using the Genome Analysis Toolkit *ASEReadCounter* (Broad Institute 2015b). Because it is not possible to make confident homozygote genotype calls from low coverage sequencing data where most sites are observed only once or twice, we only included diploid genotype observations for sites where both reference and alternate alleles were observed. As RABBIT allows for an ambiguous allele character, for all sites where only reference or alternate reads were observed (but not both), we included one ambiguous allele.

To minimize memory and runtime requirements, we included at most 5,000 SNPs per chromosome, selected for maximum ancestor-discerning information content. If an observed (sequenced) allele is common, it will only slightly narrow down the possibility of ancestors. If a sequenced allele is rare—at the most extreme, unique

to one individual—it provides greater information from which founder that site is derived. Thus, we designate information-rich sites as those where the frequency of the sequenced allele is the lowest with respect to the pool of most likely ancestors. In order to sample sites with high information content spread throughout the chromosome, we used an iterative approach. First, we included all heterozygous sites (i.e. where reference and alternate alleles are both observed). Then, 10% of all SNPs were randomly sampled, and we retained up to the top 0.2% most informative sites, repeating the random sampling and retention until we designated 5,000 SNPs.

We ran RABBIT independently for each chromosome using the Viterbi decoding function under the joint model, with all other RABBIT parameters left at default. RABBIT output was converted to a phased chromosome haplotype map, which we then used to extract and concatenate genotype information from a VCF file containing founder genotypes. To calculate genotype reconstruction accuracy, we first imported true (simulated) and estimated (reconstructed) genotypes using a custom *R* script. We measured the fraction of all remaining sites where the estimated diploid genotype is identical to the originally simulated diploid genotype, excluding fixed sites with respect to the founding haplotypes, and excluding any sites with missing genotype information. Because male individuals do not possess two copies of the sex chromosome, we only evaluated accuracy for autosomes.

To measure accuracy of estimated frequency of recombination events, true and estimated recombination counts were first summed over both copies of each chromosome in a simulated individual. This removed the possibility of introducing error by comparing the wrong copies of chromosomes. Only detectable recombination events were considered, i.e. those that did not occur between homologous haplotypes. We then used the *epi.ccc* function of the *R* package *epi*R (Stevenson 2018) to calculate Lin's concordance correlation coefficient (ρ) between the true and estimated recombination counts.

Modeling Computational Complexity of Chromosome Reconstruction

To estimate the rate at which computational requirements grows with data input, we performed chromosome reconstructions with varying numbers of potential founders and markers (SNPs). This allows us to extrapolate the runtime and memory for performing the most resource intensive chromosome reconstructions (i.e. those with > 40 founding lines). To generate runtime and memory usage data, we performed 900 reconstructions using varying sizes of RABBIT input for a single example individual 2L chromosome. Reconstructions included the four true ancestors of the simulated individual, plus 0 to 32 additional haplotypes (for a total of between 4 and 36 founders, in steps of 4) and a random selection of marker SNPs (between 500 and 5000 in steps of 500). Ten replicates, each with a unique random set of SNPs, was conducted for each combination of N founding lines and S SNPs using a single core on the University of Virginia computing cluster, with total runtime and peak memory usage as reported from the SLURM workload manager (CPUtime and MaxRSS, respectively). We then modeled the mean runtime and memory usage (averaged across 10 replicates per parameter combination) as a function of number of founding lines and number of SNPs fed into RABBIT. For runtime, simulations involving 8 or fewer founding lines were omitted from the regression model because they ran too quickly to resolve non-zero runtime. Memory was modeled as $Memory(GB) = 7.367 \times 10^{-9} \times SN^4 + 0.0316$, while runtime was modeled as $Runtime(Minutes) = [1.189 \times 10^{-3} \times N^2 + 1.038 \times 10^{-6} \times SN^2 + 2.649 \times 10^{-4} \times S]^2$.

Simulated GWAS

We performed GWAS on mapping populations produced by random sampling and permutation of the previously-described forward-simulated populations. Although the forward simulator we developed is efficient, it would not have been computationally feasible to simulate 500 fully independent mapping populations (per parameter combination) in a reasonable amount of time. Instead, we generated ten independent forward-simulated populations, and for each of those, generated fifty

randomly permuted subsets (Figure 2). For a single simulated mapping population, we began by sampling (with replacement) a random subset of 5,000 individuals, out of 10,000 total individuals generated by forward-simulation. Then, we performed a permutation of haplotype ancestry with a new, randomly-ordered (equally sized) subset of founders. The permutation of ancestry was one-to-one, e.g. all haplotype blocks that were previously derived from founder X would be translated to founder Y, and blocks previously derived from Y would in turn be mapped to founder Z.

In addition to Hybrid Swarm populations, which we ran through the simulated sequencing and mapping pipeline, we generated four additional types of mapping populations for comparing GWAS performance: Highly outbred (F_{50}) populations, similar to sampling wild individuals; Inbred Lines (ILs), similar to mapping with the DGRP); and Recombinant Inbred Lines (RILs), similar to mapping with the DSPR.

The F_{50} populations were generated in same manner as the Hybrid Swarm, except for fifty non-overlapping generations of recombination instead of five generations. The ten resulting forward-simulated populations were resampled and permuted as we did with the Hybrid Swarms.

We simulated ten initial sets of 800 RILs using the same forward-simulator as previously described, each initialized with a random subset of eight DGRP haplotypes. Populations randomly recombined at a population size of 10,000 for fifty non-overlapping generations, after which 800 random male-female pairs of individuals were isogenzied through 25 generations of full-sibling mating. This scenario roughly corresponds to the *Drosophila* Synthetic Population Resource (King *et al.* 2012a). For computational simplicity, after the 25 generations of isogenization we removed any remaining residual heterozygosity by forcing the identity of a second chromosome copy to be identical to the first copy. We then sampled 5,000 draws (with replacement) of the 800 RILs followed by ancestry permutation as described above.

To simulate GWAS on Inbred Lines, no forward-simulation was necessary. For a single simulated population, we first randomly selected 128 DGRP lines with high coverage and low levels of heterozygosity as the set of founders. Then, those 128 lines were randomly sampled with replacement 5,000 times. As with hybrid swarm and RILs, for any parameter combination we generated a total of 500 mapping populations.

Phenotypes were modeled as probabilistic assignment to a case or control group dependent on allele dosage at a purely additive single SNP. We designated a causal locus as a random autosomal biallelic SNP segregating within 0.5% of a desired minor allele frequency (50%, 25%, and 12.5%). We modeled SNPs at 5% and 10% percent variation explained (PVE), where reference allele homozygotes were assigned to the case group with probability $50\% - PVE/2$, and alternate allele homozygotes were assigned to the case group with probability $50\% + PVE/2$. Heterozygotes were equally likely to be assigned case and control.

To perform many replicates of GWAS for many parameter combinations, we performed a simple χ^2 test of independence for reference and alternate allele counts between case and control groups. To do so most efficiently, we developed a method of aggregating allele counts that uses a haplotype map table in conjunction with a single table of founder genotypes (Figure 2). Briefly, haplotype table breakpoints across all individuals were sorted in ascending order. When iterating through ascending unique start and stop positions, between any pair of breakpoints, all SNPs will be comprised of the same number of each founding haplotype. Haplotype IDs could then be counted and sorted in the same column position order as the table containing polarized allele status (-1 for alternate, +1 for reference). Multiplying the genotype table by the haplotype count vector results in final allele counts, polarized negative for alternate alleles and positive for reference alleles. For inbred mapping populations, we corrected for non-independent allele draws by dividing the χ^2 value by two.

To describe the accuracy of simulated GWAS, we measured the likelihood of including a locus that is near the causal site when considering a set of the top N most significant SNPs. Here, 'near' is defined as either exact-SNP resolution, or within 1, 10, or 100 Kb. In the case of 1 kb precision, we first consider the set of SNPs +/- 1 kb from the most significant locus (greatest chi-square statistic). Then, we consider the second set of SNPs as those within +/- 1 kb of the most significant locus outside of the window already accounted for. This selection of significant clusters was repeated iteratively for the top 25 regions, for window sizes of 0 (exact SNP resolution), 1 kb, 10 kb, and 100 kb.

We calculated genomic inflation factor (GIF, λ_{1000}) as the value of $\chi^2_{observed}/\chi^2_{expected}$ with two degrees of freedom. Because GIF increases with sample size, we performed a correction to report the level of GIF expected with a sample size of 1,000 case and 1,000 control individuals (Freedman *et al.* 2004).

Assessing counts of variable sites at appreciable frequency in the DGRP

There is a reduction in power to detect associations with alleles segregating at low minor allele frequencies. When a population is founded by N lines, any SNP will be segregating at a relative frequency of at least 1/N, given that the SNP is not fixed within the population and haplotypes are equally represented. We counted the number of sites on a given chromosome arm segregating above a minor allele frequency threshold of MAF = (0.05, 0.125, and 0.25) for random draws (without replacement) when sampling N = (2, 4, 8, 16, 32, 64, 128) haplotypes of the 129 included DGRP lines We performed this sampling 20 times for each chromosome arm.

Results

Computational Complexity of Chromosome Reconstruction

To determine reasonable limits for numbers of SNPs and haplotypes used for chromosome reconstruction with RABBIT, we modeled peak memory usage and runtime across a range of input sizes. Peak memory grew linearly with number of SNPs used, and at a greater-than-linear rate with haplotypes (Figure 1A, $Memory = 7.367 \times 10^{-9} \times SN^4 + 0.0316)$, $F = 3.534 \times 10^6, df = 1 \& 88,\ R^2 = 1$). The runtime of RABBIT increased at a greater-than-linear rate for both number of SNPs and number of haplotypes, though the *N* parameter dominates (Figure 1B, $Runtime = [1.189 \times 10^{-3} \times N^2 + 1.038 \times 10^{-6} \times SN^2 + 2.649 \times 10^{-4} \times S]$, $F = 4.316 \times 10^4, df = 3 \& 67, R^2 = 0.9995$). These models allowed us to estimate resource requirements at greater numbers of haplotypes (Figure 1, C & D) which would be unfeasible to measure empirically.

Most-Likely-Ancestor Selection

To reduce computational requirements of haplotype reconstructions with RABBIT, we developed and evaluated an algorithm for selecting a minimum representative set of Most-Likely-Ancestors (MLAs) for chromosome reconstruction. We found a HARP threshold of 0.99 (see methods) discerned a minimal subset of founding lines that tended to include a given chromosome's true ancestors (Figure 3). At this threshold, outcomes became asymptotic at the computationally tractable cap of 16 founding lines (Figure 1). Thus, we performed chromosome reconstruction using up to 16 most-likely-ancestors as inferred with a HARP threshold of 0.99.

In all cases, decreasing the HARP threshold from 0.95 to 0.90 further reduced chromosome representation while increasing the number of extraneous founding lines selected for reconstruction. While a higher HARP threshold of 0.999 yielded the smallest and most computationally tractable set sizes of MLAs ($\overline{N}$=2.4-3.5), the strict threshold excluded true ancestors, resulting in a set that is least representative of chromosomes to be reconstructed. For 128-founder populations, a threshold of 0.999 failed to identify founders constituting an average of 3.33% and 13.9% of

chromosomes for DGRP- and Coalescent-founded populations, respectively. In 32-founder populations, the 0.999 threshold missed founders representing an average of 15.5% and 29.7% of chromosomes from DGRP- and Coalescent-founded populations, respectively.

Populations simulated with genetic variation derived from coalescent models described above included the parameters $N_e = 10^6$ and $\mu = 5 \times 10^{-9}$. The effectiveness of most-likely-ancestor selection for populations modeled across extended values of N_e and μ is shown in supplemental Figures S2 and S3, respectively. Similarly, the number of most-likely-ancestors chosen for reconstruction in RABBIT are shown in Figures S5 and S6.

Selected MLA set size is described in Figure S5 for 32-founder populations and Figure S6 for 128-founder populations. Ancestor selection effectiveness for DGRP-derived populations at two levels of sequencing coverage (0.005X and 0.05X) is shown in supplemental Figure S4, and the corresponding number of most-likely-ancestors chosen for reconstruction are shown in supplemental Figure S7.

Reconstruction Accuracy

Chromosome reconstruction of simulated $\mathrm{F_5}$ Hybrid Swarm genomes at 0.05X sequencing coverage yielded highly accurate genotype estimates (Figure 4). The median percent of sites with correctly estimated genotypes was greater than 99.9% whether the population was founded by 32 or 128 founding lines for either DGRP or coalescent ($\mathrm{N_e} = 10^6$ and $\mu = 5 \times 10^{-9}$) haplotypes. We additionally report reconstruction accuracy in coalescent-derived populations across a range of $\mathrm{N_e}$ and μ values in supplemental Figure S8.

For simulations founded by DGRP lines, 80.5% of reconstructed chromosomes from 32-founder populations exhibited > 99.9% accuracy, with the remaining 19.5%

of reconstructions contributing to a long tail with a minimum of 84.37%. Increasing the number of founding lines to 128 resulted in genotype accuracy above 99% for all cases (minimum: 99.4%), with 83% of reconstructed chromosomes achieving greater than 99.9% accuracy.

Although median accuracy for coalescent-derived populations was equivalent to that of DGRP-derived populations (99.9%), coalescent-derived populations with 32 founders exhibited a greater number of low-accuracy reconstructions. While 82.5% of simulations with 32 coalescent haplotypes were at least 99% accurate, the remaining 17.5% of reconstructions contributing to a long tail with a minimum accuracy of 59.7%. Increasing the number of founding lines to 128 resulted in 96.3% of simulations being greater than 99% accurate, with a minimum accuracy of 89.6%.

The number of recombination events estimated from chromosome reconstruction was most accurate for populations founded by 128 lines (Table 1). Reconstructions of DGRP- and Coalescent-derived chromosomes yielded recombination count estimates that were 98.6% and 95.6% concordant with their respective true recombination counts (Lin's concordance correlation coefficient, ρ). When populations were founded by 32 lines, recombination count estimates were more inaccurate, with DGRP- and Coalescent-derived reconstructions achieving 50.2% and 75.9% concordance with their respective true recombination counts. For 32-founder populations, DGRP-derived reconstructions tended to slightly overestimate recombination counts, while the same counts were underestimated for coalescent-derived populations.

Simulations that inferred an unlikely high number of recombination events tended to exhibit reduced accuracy (Figure 4). All DGRP-derived simulated individuals (of 1600 total) exhibited ≤ 8 recombination events, and all but three coalescent-derived simulated individuals (7197 of 7200 total) exhibited ≤ 9 recombination events.

Accordingly, we considered any reconstructions to be 'hyper-recombinant estimates' if the inferred recombination count is greater than 8 for DGRP-derived populations or greater than 9 for coalescent-derived populations.

At 0.05X sequencing coverage, hyper-recombinant estimates did not occur for 128-founder populations, and only rarely resulted from 32-founder populations. Within DGRP-derived 32-founder populations, reconstructions with hyper-recombinant estimates were below the sixth percentile of genotype accuracy (N=6/400 simulations, genotype accuracy range=92.8%-98.8%). For coalescent-derived 32-founder populations, reconstructions estimated as hyper-recombinant fell in the bottom 9% of genotype accuracy (N=3/400 simulations, genotype accuracy range = 92.1%-95.6%). Although hyper-recombinant estimates always fell in the bottom 10% of accuracy, the least accurate reconstructions were not hyper-recombinant. For coalescent-derived 32-founder populations, 4.25% (17/400) of reconstructions without hyper-recombinant estimates exhibited lower genotype accuracy than the least accurate hyper-recombinant simulation (range = 59.7%-92.1%). Similarly, for DGRP-derived 32-founder populations, 2.5% (10/400) of reconstructions without hyper-recombinant estimates exhibited lower genotype accuracy than the least accurate hyper-recombinant simulation (range = 82.3%-92.8%).

Reducing sequencing coverage by an order of magnitude from 0.05X to 0.005X resulted in more frequent hyper-recombinant reconstruction estimates, though overall median genotype accuracy remained above 99% (Figure S9). Hyper-recombinant reconstructed chromosomes exhibited genotype estimates with accuracy below 99%, while the remaining simulations (with lower recombinant counts) achieved above 99% genotype accuracy. For populations founded by 32 DGRP lines and sequenced at 0.005X coverage, 14% of simulations produced hyper-recombinant estimates (N=56/400), of which only 26.8% (N=15/56) surpassed 99% genotype accuracy (median=98.5%). The remaining 86% of simulations (N=344/400) that were not

hyper-recombinant retained greater accuracy, with 89% of simulations resulting in at least 99% genotype accuracy (median=99.5). Increasing the number of founding DGRP lines from 32 to 128 at 0.005X coverage failed to eliminate hyper-recombinant estimates. With 128 founding lines, 14.5% of simulations were hyper-recombinant (N=58/400), of which 24.1% (N=14/58) surpassed 99% genotype accuracy (median=98.6%). The 85.5% of simulations that were not hyper-recombinant (N=342/400), exhibited accurate genotype estimates, with 86.5% (296/342) of simulations achieving over 99% genotype accuracy (median=99.6).

GWAS Simulation Accuracy

To report the power of a GWAS, we must first define a "true positive" result. Consider a putative SNP identified by GWAS that is 50 kb from the causal SNP. Such a result would be considered a false positive if the aim is to identify the exact responsible nucleotide, but may be a true positive with respect to identify an associated gene. To cover both use cases, we describe a true positive in terms of both SNP-level resolution (requiring an exact base-pair match), or region-level resolution (allowing for tolerance up to 100kb between putative hits and the causal SNP). Additionally, it is unrealistic to simply evaluate the single top result of a GWAS. Rather, a set of candidate loci may be chosen for follow-up evaluation in confirmatory studies, and the probability of including the causal SNP will increase as a greater number of putative SNPs are evaluated. Most distinct changes in GWAS power occurred when including between most significant to top 10 most significant candidate loci, after which power increased at a reduced rate, if an asymptote was not already reached.

The estimated power of GWAS using a specific type of mapping population, i.e. the fraction of simulations with a true positive, is shown in Figure 5. For simplicity, we focus on GWAS power when including the top 10 most significant candidate loci—a reasonable number of putative sites that may be investigated in follow-up studies.

F_5 Hybrid Swarms founded by either 32 or 128 founding lines exhibited nearly equivalent power compared to F_{50} outbred populations across all parameter combinations. For common alleles, i.e. those segregating at 50% frequency, all outbred populations achieved approximately 50% power to identify a causal variant with SNP-level precision, and 70% power at the gene-level. Both inbred populations were highly effective at detecting associations at the gene-level (99% and 99.8% for ILs and RILs, respectively). SNP-level power lower than gene-level power for RILs (75.4%), but only marginally reduced for inbred lines (97.4%).

Power to detect associations is reduced when the causal allele is rare (segregating at 12.5% frequency). For such rare alleles, the gene-resolving power of ILs drops by nearly half (to 54.8%), while RILs maintained high power (81.8%). All outbred populations exhibited approximately 20% power to detect rare alleles at the gene-level. Inbred lines were the sole frontrunner for identifying low frequency alleles with SNP-level resolution (37.2%), followed by 128-founder Hybrid Swarm (10.8%), F_{50} outbred (8.2%), 32-founder Hybrid Swarm (6%), and RILs (3.2%).

GWAS Genomic Inflation Factor

If individuals are assigned to case and control groups with equal probability, then the resulting χ^2 statistics should follow the expected distribution. If individuals are not sorted into groups randomly, i.e. allele state at a causal SNP dictates nonrandom group assignment, then χ^2 values for that SNP should be inflated to some extent. Nonrandom associations between a causal SNP and other loci can inflated test statistics across a chromosome, or across a whole genome. The genome-wide inflation factor (λ) can be expressed as the ratio of observed and expected median χ^2 values (Figure 6). Because our simulations model a single causal SNP, λ is a reflection of greater-than-chance associations arising due to linkage with the causal SNP being modeled, which can serve as a proxy for false positive rate.

Because the median expected χ^2 statistic increases with sample size, we report λ_{1000}, a sample-size-corrected value that is comparable across studies (Freedman *et al.* 2004). We calculated λ as aggregated across three groups: linked, including only the autosome arm containing the causal SNP; unlinked, including the unlinked autosome that doesn't contain the causal SNP; and autosomal, for all sites across both autosomes two and three

Inflation factor measured across autosomes two and three was greatest for ILs, followed by 32-founder Hybrid Swarm, RILs, 128-founder Hybrid Swarm, and F_{50} Outbred populations. This order was observed whether the causal allele was common or rare, though with reduced values of λ at the lower allele frequency (Figure 7).

Only inbred populations displayed inflation on unlinked autosomes. When the causal allele is common (50% frequency), inflation on unlinked sites was greater for Inbred lines (median $\lambda = 1.17, Interquartile\ Range\ or\ IQR = 0.11$) than for RILs ($\lambda = 1.02, IQR = 0.07$). There was no inflation for unlinked chromosome in outbred populations, where $\lambda = 1.0$ with varying degrees of dispersion ($IQR = 0.10, 0.06$ and 0.03, respectively, for 32-founder HS, 128-founder HS, and F_{50} outbred populations). Unlinked sites remained inflated for ILs even when the causal allele was rare ($\lambda = 1.07, IQR = 0.09$). Distributions for λ across an extended range of autosome groups, PVE and allele frequencies are shown in Figure S11.

When we dissociated phenotype from genotype with purely random case-control assignment (i.e. PVE was set to 0% in our simulations), λ was centered at 1.0 for all populations. F_{50} outbred populations exhibited the lowest dispersion ($IQR = 0.02$), followed by 128-founder Hybrid Swarms ($IQR = 0.04$), RILs ($IQR = 0.06$), and 32-founder Hybrid Swarms or ILs ($IQR = 0.07$ each).

Frequency of sites segregating at appreciable frequency

The number of SNPs segregating amongst DGRP haplotypes with at least a given MAF strongly depends on the haplotype subset count for a given population (Figure 8). If only considering SNPs segregating at or above a frequency of 12.5% on chromosome arm 2L, a population founded by 8 lines will yield approximately twice as many SNPs compared to a population founded by 128 lines (N=8 lines yields a median of 140K SNPs; N=128 lines yields a median of 71k SNPs). If the minimum MAF threshold is instead set to 5%, then populations with a greater number of lines exhibit a greater number of SNPs—with a maximum number of segregating sites with N=16 lines (median of 231.6k SNPs), nearly as many for 128 lines (median of 194k SNPs), and fewer for N=8 lines (median of 133k SNPs).

Discussion

Herein, we modeled the feasibility and statistical properties of genome-wide association mapping using the Hybrid Swarm, an outbred population derived from limited and random outcrossing of an arbitrary number of founding strains. We show that it is possible to accurately reconstruct whole genomes from Hybrid Swarm populations using ultra-low coverage sequencing data (Figure 5). Genome-wide association mapping using the Hybrid Swarm approach performs as well as mapping in highly outbred F_{50} populations in a case-control GWAS framework (Figures 6, Supplemental Figure S10). While mapping using the Hybrid Swarm approach generally has reduced power compared to mapping using inbred lines (as would any outbred population in general) a limited number of generations of recombination reduces false positives arising from long-distance linkage disequilibrium present in founding strains (Figure 7, Supplemental Figure S11). Together, our results demonstrate the feasibility and potential of using the Hybrid Swarm approach for generating and genotyping outbred mapping populations in a cost-effective and computationally efficient manner (Figure 1).

Benefits of the Hybrid swarm Approach

The Hybrid Swarm approach is applicable to a wide variety of organisms and experimental designs, conferring potential benefits over inbred reference panels. These benefits are realized in three primary ways by: (1) allowing researchers to address questions that require heterozygotes; 2) reducing labor and the influence of cage effects with random mating in a common environment; and 3) breaking down population structure when incorporating individuals from divergent populations. These benefits are possible due to the ability to reconstruct genomes accurately and in a cost-efficient manner for a large number of individuals.

Note that the Hybrid Swarm method is not limited to populations founded by inbred lines, as the technique can be applied to populations where phased genomes are available for all outbred founders. Research systems without inbred reference panels can thus make an up-front investment of fully phasing founder genomes to realize downstream savings of reconstructing progeny from low-coverage sequencing data. Due to the relative ease of generating phased genomes from a variety of long-read sequencing technologies (Pollard *et al.* 2018), the Hybrid Swarm method may enable association mapping in a wide variety of organisms.

One clear difference between inbred and outbred mapping populations is the presence of heterozygotes. On the one hand, the presence of heterozygotes in outbred populations decreases power to detect association relative to inbred lines for an additive allele with a given effect size (Figure 6, Supplemental Figure S10). However, the reduced statistical power of association mapping in outbred populations may be ameliorated by reduced inbreeding depression and by the ability to assess the heterozygous effects of alleles.

The ability to assess heterozygous effects of alleles will provide valuable insights into several interesting aspects of biology, such as the nature of dominance and the

identity of regulatory polymorphisms. An increased understanding of dominance relationships and regulatory polymorphisms is important for advancing our understanding of quantitative trait variation and evolution. For instance, several theoretical models have shown that context dependent dominance of quantitative fitness traits can underlie the stable maintenance of polymorphisms subject to seasonally variable (Wittmann *et al.* 2017) or sexually antagonistic (Connallon and Chenoweth 2019) selection. The ability to efficiently map loci with context dependent dominance relationships will aid in the understanding of the stability and abundance of polymorphisms maintained by these forms of balancing selection. Regulatory polymorphisms are known to underlie genetic variation and play an important role in local adaptation (King and Wilson 1975; Andolfatto 2005; Kopf et al. 2015; Nourmohammad et al. 2017). Because the Hybrid Swarm approach can be used in a variety of organisms, it is a promising design to identify *cis*-eQTL, promoting our generalized understanding of regulatory evolution.

The Hybrid Swarm approach allows a mapping population to be propagated as a single large outbred population via undirected crossing. This design confers benefits over alternatives of either rearing inbred lines separately or performing controlled crosses. First, a single population reduces the influence of random block effects associated with rearing families or closely related individuals in separate enclosures or defined areas. Second, random outbreeding of a single population requires less labor effort compared to performing controlled crosses or separate serial propagation of inbred lines. One drawback of the randomly outbred method is susceptibility to genetic drift, and subsequent loss of a subset of haplotypes. The distribution of haplotypes can also be skewed by line-specific differences in fitness or fecundity, with such differences being observed for DGRP lines (Horváth and Kalinka 2016).

To attenuate haplotype dropout, it may be prudent to seed Hybrid Swarm with a large population of F_1 hybrids produced by round-robin crosses. The F_1 population

would then be followed by four generations of random outbreeding. Initial seeding with round-robin F_1s would also attenuate the impact of drift due to aforementioned line-specific differences in fitness or fecundity.

Recombination between lines in the Hybrid Swarm approach allows for greater dissection of functional polymorphisms segregating between populations with distinct genetic structure—nonrandom patterns of genetic variation within or between populations. If an association study incorporates haplotypes from multiple distinct source populations, causal variants would segregate along with other linked variants. Thus, in order to address questions related to local adaptation, it is necessary to minimize the influence of linked non-causal loci in association studies. While corrections due to relatedness can reduce the type I error rate (Yu *et al.* 2006), genetic structure would be further reduced by increased numbers of founding haplotypes and increased generations of recombination. Reduction in population structure in the Hybrid Swarm is evidenced by genome-wide inflation λ in inbred lines even on chromosomes physically unlinked to a simulated causal SNP, whereas five generations of recombination was sufficient to reduce this inflation (Figure 7).

The Hybrid Swarm method is similar but distinct from advanced intercross population (AIP) design (Mackay and Huang 2018), another option for heterozygous mapping populations. With an AIP, few lines (e.g., 8) are crossed for many generations, as opposed crossing dozens to hundreds of lines for few generations. The choice to use an AIP or hybrid swarm population will influence the number of SNPs segregating at or above a desired minor allele frequency (Figure 8). In order for an association test to detect a causal variant with single-nucleotide precision, that variant must not be fixed within the population, and must be segregating above a minor allele frequency required to detect phenotypic association at a given effect size and sample size. If sample size precludes sites segregating at a minor allele frequency below 1/8, then a population founded by 8 haplotypes would yield the greatest number of variants. If power is

sufficient to detect association with alleles segregating above a frequency of 5%, then populations founded by 16+ lines would yield a greater number of variants. In cases where only few founding haplotypes are available, an AIP may be necessary, as the breakup of linkage disequilibrium can only be accomplished with many generations of crosses instead of leveraging greater haplotype diversity.

Computational Considerations

The simulations conducted for this analysis were made feasible by three primary innovations. First, the haplotype block file format allowed us to leverage information redundancy between related individuals and store highly compressed, lossless genotype information. With nearly 1/2000th the file size of a compressed VCF file, haplotype block files greatly reduced both the disk storage footprint and time required for disk write operations. Second, instead of performing forward-in-time simulations for every single iteration, permuted subsets of simulated populations allowed for more rapid GWAS simulations. The format of haplotype block files facilitated permutations of the ancestry contained within a population's mosaic haplotypes, generating novel population genetic structure while preserving the forward-simulator's influence of drift and meiotic recombination. Third, instead of extracting site-specific genotypes for every individual, we decreased the number of computational operations by performing aggregate counts across all sites between adjacent recombination events in the population.

F_5 hybrid swarm populations performed equivalently to F_{50} outbred population in a case-control GWAS framework. This is likely owed to the large number of unique haplotypes within the Hybrid Swarm population, reducing the influence of long distance LD, and in turn reducing false positive GWAS hits. One interpretation is that only slightly recombinant populations are sufficient representations of highly outbred (or wild) populations in a GWAS framework. Inbred populations did exhibit greater power than outbred populations for identifying a causal locus, although this result is to

be expected. Because we simulated a purely additive trait for which heterozygotes are equally likely to be assigned to either case or control group, heterozygotes contribute no statistical signal of association. Accordingly, for a causal allele segregating at 50% frequency, sample sizes for any outbred populations will be effectively half that of an inbred population.

Applying the Hybrid Swarm approach

At minimum, the Hybrid Swarm approach requires a sequenced set of individuals for founding a recombinant population. Although our simulations presented here were conducted with inbred founding lines, genome reconstructions can similarly be performed with any phased genomes. For example, 16 phased outbred founders could be treated as 32 independent haplotypes. Phased genomes are becoming increasingly accessible with the advent of long-read sequencing platforms and phasing software, allowing this technique to be applied to even more systems. Optionally, a recombination rate map for the population can be provided, otherwise recombination is assumed to occur with equal user-defined probability across any chromosome.

As a first step, power analyses using our rapid association test simulation pipeline will inform choices of sample size and mapping population design (Figure 3). After determining a feasible sample size for a given SNP of minimum percent variation explained, researchers can evaluate the accuracy of low-coverage chromosome reconstructions for a simulated proposed mapping population. Note that while we performed association tests in a case/control framework, the relative power of the Hybrid Swarm is expected to be the same for quantitative traits, which could garner additional power from sampling individuals from phenotypic extremes (D. Li, Lewinger, Gauderman, Murcray, & Conti, 2011).

For our simulations, we parameterized chromosome reconstructions using a maximum of $N = 16$ MLAs and $S = 5000$ SNPs, which required less than 3 GB of memory and completed in under 5 minutes on a single core. However, these values may not be ideal for all mapping population designs or organism systems. It may be necessary to select greater number of MLAs prior to reconstruction if haplotypes are difficult to differentiate due to being less divergent (i.e. exhibiting lower θ_π) than those simulated here. For example, reconstruction accuracy was low for coalescent-derived mapping populations modeled with $\theta = 4 \times 10^{-5}$, which may reflect those of *C. elegans* (Barrière and Félix 2005). Further, 5000 SNPs may be an over- or under-estimate of those required in other systems. Because recombination between haplotypes can only be inferred at sampled variable sites, SNP density directly influences how close inferred breakpoints will be resolved with respect to their actual position. The models described in Figure 1 can be used to estimate the requirements of a proposed input size for chromosome reconstruction. Note that runtime and memory requirement is dependent on the total number of SNPs used, not the length of the chromosome being reconstructed.

To evaluate whether low coverage sequencing data will yield accurate genotype estimates for a given proposed mapping population, researchers can test reconstruction accuracy *in silico*. We provide a convenient forward-simulation R script for this purpose that generates output in the haplotype map format (Figure 2). Simulated individuals can then be ran through the sequencing and mapping pipeline at a desired level of coverage. After generating simulated mapped individuals, researchers can optimize the number of MLAs and HARP threshold that provide most effective MLA selection for their mapping population (Figure 4). This step may reveal haplotypes that are consistently problematic or inaccurately chosen, which can be excluded from further simulations (and when generating the true mapping population). Researchers can then conduct chromosome reconstruction using the optimized MLA selection parameters and evaluate whether accuracy is acceptable (Figure 5).

After performing chromosome reconstructions, a quality controls step may be applied whereby troublesome regions are masked. For example, a reconstructed chromosome with sequence of short recombination blocks could be masked prior to evaluating genotyping accuracy or performing association testing. In our simulations, it was surprisingly difficult to diagnose exact factors contributing to the least accurate reconstructions. However, these highly recombinant reconstructions still achieved 90-99% accuracy, suggesting that accuracy may be achieved even for anomalous hyper-recombinant individuals (Figure 5) After affirming the accuracy of simulated reconstructions, the optimized parameters can be applied to a genuine mapping population akin to the simulated one.

Conclusions

An outbred high-resolution mapping population that can be generated in little time is an attractive option for researchers, but such mapping populations have been prohibited by genotyping costs or computational requirements to impute genotypes from ultra-low sequencing data. Our work includes a computationally efficient framework for GWAS power analysis and demonstrates the feasibility of the Hybrid Swarm as a cost-effective method of fine-scale genetic mapping in an outbred population.

Literature Cited

Barrière A., and M. Félix, 2005 *WormBook: The Online Review of C. elegans Biology [Internet]*. Pasadena (CA).

Bergland A. O., H. Chae, Y.-J. Kim, and M. Tatar, 2012 Fine-Scale Mapping of Natural Variation in Fly Fecundity Identifies Neuronal Domain of Expression and Function of an Aquaporin. PLoS Genet. 8: e1002631. https://doi.org/10.1371/journal.pgen.1002631

Broad Institute, 2015a The Picard toolkit

Broad Institute, 2015b Genome Analysis Toolkit: Variant Discovery in High-Throughput Sequencing Data

Cheng R., J. E. Lim, K. E. Samocha, G. Sokoloff, M. Abney, *et al.*, 2010 Genome-wide association studies and the problem of relatedness among advanced intercross lines and other highly recombinant populations. Genetics 185: 1033–1044. https://doi.org/10.1534/genetics.110.116863

Chesler E. J., D. R. Miller, L. R. Branstetter, L. D. Galloway, B. L. Jackson, *et al.*, 2008 The Collaborative Cross at Oak Ridge National Laboratory: Developing a powerful resource for systems genetics. Mamm. Genome 19: 382–389. https://doi.org/10.1007/s00335-008-9135-8

Chia R., F. Achilli, M. F. W. Festing, and E. M. C. Fisher, 2005 The origins and uses of mouse outbred stocks. Nat. Genet. 37: 1181–1186. https://doi.org/10.1038/ng1665

Comeron J. M., R. Ratnappan, and S. Bailin, 2012 The Many Landscapes of Recombination in Drosophila melanogaster. PLoS Genet. 8: 33–35. https://doi.org/10.1371/journal.pgen.1002905

Connallon T., and S. F. Chenoweth, 2019 Dominance reversals and the maintenance of genetic variation for fitness. PLoS Biol. 17: 1–11. https://doi.org/10.1371/journal.pbio.3000118

Danecek P., A. Auton, G. Abecasis, C. A. Albers, E. Banks, *et al.*, 2011 The variant call format and VCFtools. Bioinformatics 27: 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Darvasi A., and M. Soller, 1995 Advanced intercross lines, an experimental population for fine genetic mapping. Genetics 141: 1199–207.

Freedman M. L., D. Reich, K. L. Penney, G. J. McDonald, A. A. Mignault, *et al.*, 2004 Assessing the impact of population stratification on genetic association studies.

Nat. Genet. 36: 388–393. https://doi.org/10.1038/ng1333

Horváth B., and A. T. Kalinka, 2016 Effects of larval crowding on quantitative variation for development time and viability in Drosophila melanogaster. Ecol. Evol. 6: 8460–8473. https://doi.org/10.1002/ece3.2552

Howie B., J. Marchini, and M. Stephens, 2011 Genotype imputation with thousands of genomes. G3 1: 457–70. https://doi.org/10.1534/g3.111.001198

Huang X., M.-J. Paulo, M. Boer, S. Effgen, P. Keizer, *et al.*, 2011 Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. Proc. Natl. Acad. Sci. U. S. A. 108: 4488–93. https://doi.org/10.1073/pnas.1100465108

Huang B. E., A. W. George, K. L. Forrest, A. Kilian, M. J. Hayden, *et al.*, 2012 A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol. J. 10: 826–839. https://doi.org/10.1111/j.1467-7652.2012.00702.x

Kessner D., T. L. Turner, and J. Novembre, 2013 Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. Mol. Biol. Evol. 30: 1145–58. https://doi.org/10.1093/molbev/mst016

King E. G., S. J. Macdonald, and A. D. Long, 2012a Properties and power of the Drosophila synthetic population resource for the routine dissection of complex traits. Genetics 191: 935–949. https://doi.org/10.1534/genetics.112.138537

King E. G., C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen, *et al.*, 2012b Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. Genome Res. 22: 1558–1566. https://doi.org/10.1101/gr.134031.111

Kover P. X., W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, *et al.*, 2009 A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana, (R. Mauricio, Ed.). PLoS Genet. 5: e1000551. https://doi.org/10.1371/journal.pgen.1000551

Krämer N., N. Ranc, N. Meyer, M. Ouzunova, C. Lehermeier, *et al.*, 2014 Usefulness of Multiparental Populations of Maize (Zea mays L.) for Genome-Based Prediction . Genetics 198: 3–16. https://doi.org/10.1534/genetics.114.161943

Lack J. B., C. M. Cardeno, M. W. Crepeau, W. Taylor, R. B. Corbett-Detig, *et al.*, 2015 The drosophila genome nexus: A population genomic resource of 623 Drosophila melanogaster genomes, including 197 from a single ancestral range population. Genetics 199: 1229–1241. https://doi.org/10.1534/genetics.115.174664

Li R., M. A. Lyons, H. Wittenburg, B. Paigen, and G. A. Churchill, 2005 Combining data from multiple inbred line crosses improves the power and resolution of quantitative trait loci mapping. Genetics 169: 1699–709. https://doi.org/10.1534/genetics.104.033993

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li H., 2011 wgsim (short read simulator)

Li H., 2013 Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM

Mackay T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, *et al.*, 2012 The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178. https://doi.org/10.1038/nature10811

Mackay T. F. C., and W. Huang, 2018 Charting the genotype-phenotype map: lessons from the *Drosophila melanogaster* Genetic Reference Panel. Wiley Interdiscip. Rev. Dev. Biol. 7: e289. https://doi.org/10.1002/wdev.289

MacKay T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, *et al.*, 2012 The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178. https://doi.org/10.1038/nature10811

Monir M. M., and J. Zhu, 2017 Comparing GWAS Results of Complex Traits Using Full Genetic Model and Additive Models for Revealing Genetic Architecture. Sci. Rep. 7: 38600. https://doi.org/10.1038/srep38600

Mott R., C. J. Talbot, M. G. Turri, A. C. Collins, and J. Flint, 2000 A method for fine mapping quantitative trait loci in outbred animal stocks. Proc. Natl. Acad. Sci. 97: 12649–54. https://doi.org/10.1073/pnas.230304397

Nikpay M., A. Goel, H. H. Won, L. M. Hall, C. Willenborg, *et al.*, 2015 A comprehensive 1000 Genomes-based genome-wide association meta-analysis of coronary artery disease. Nat. Genet. 47: 1121–1130. https://doi.org/10.1038/ng.3396

Paul R. Staab, Sha Zhu, Dirk Metzler, and Gerton Lunter, 2015 {scrm}: efficiently simulating long sequences using the approximated coalescent with recombination. Bioinformatics 31: 1680–1682.

Pollard M. O., D. Gurdasani, A. J. Mentzer, T. Porter, and M. S. Sandhu, 2018 Long reads: their purpose and place. Hum. Mol. Genet. 27: R234–R241. https://doi.org/10.1093/hmg/ddy177

R Core Team, 2016 R: A Language and Environment for Statistical Computing

Singh R., I. T. Lobina, M. Thomson, S. McCouch, C. Dilla-Ermita, *et al.*, 2013 Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6: 11. https://doi.org/10.1186/1939-8433-6-11

Spencer C. C. A., Z. Su, P. Donnelly, and J. Marchini, 2009 Designing genome-wide association studies: Sample size, power, imputation, and the choice of genotyping chip. PLoS Genet. 5. https://doi.org/10.1371/journal.pgen.1000477

Srivastava A., A. P. Morgan, M. L. Najarian, V. K. Sarsani, J. S. Sigmon, *et al.*, 2017 Genomes of the Mouse Collaborative Cross. Genetics 206: 537–556. https://doi.org/10.1534/genetics.116.198838

Stevenson M., 2018 epiR: Tools for the Analysis of Epidemiological Data

Wittmann M. J., A. O. Bergland, M. W. Feldman, P. S. Schmidt, and D. A. Petrov, 2017 Seasonally fluctuating selection can maintain polymorphism at many loci via segregation lift. Proc. Natl. Acad. Sci. 114: E9932–E9941. https://doi.org/10.1073/pnas.1702994114

Wu Y., Z. Zheng, P. M. Visscher, and J. Yang, 2017 Quantifying the mapping precision of genome-wide association studies using whole-genome sequencing data. Genome Biol. 18: 1–10. https://doi.org/10.1186/s13059-017-1216-0

Yu J., G. Pressoir, W. H. Briggs, I. V. Bi, M. Yamasaki, *et al.*, 2006 A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat. Genet. 38: 203–208. https://doi.org/10.1038/ng1702

Zhang J., K. Kobert, T. Flouri, and A. Stamatakis, 2014 PEAR: A fast and accurate Illumina Paired-End reAd mergeR. Bioinformatics 30: 614–620. https://doi.org/10.1093/bioinformatics/btt593

Zheng C., M. P. Boer, and F. A. van Eeuwijk, 2015 Reconstruction of genome ancestry blocks in multiparental populations. Genetics 200: 1073–1087. https://doi.org/10.1534/genetics.115.177873

Zheng C., M. P. Boer, and F. A. van Eeuwijk, 2018 Accurate genotype imputation in multiparental populations from low-coverage sequence. Genetics 210: 71–82. https://doi.org/10.1534/genetics.118.300885

Tables & Figures

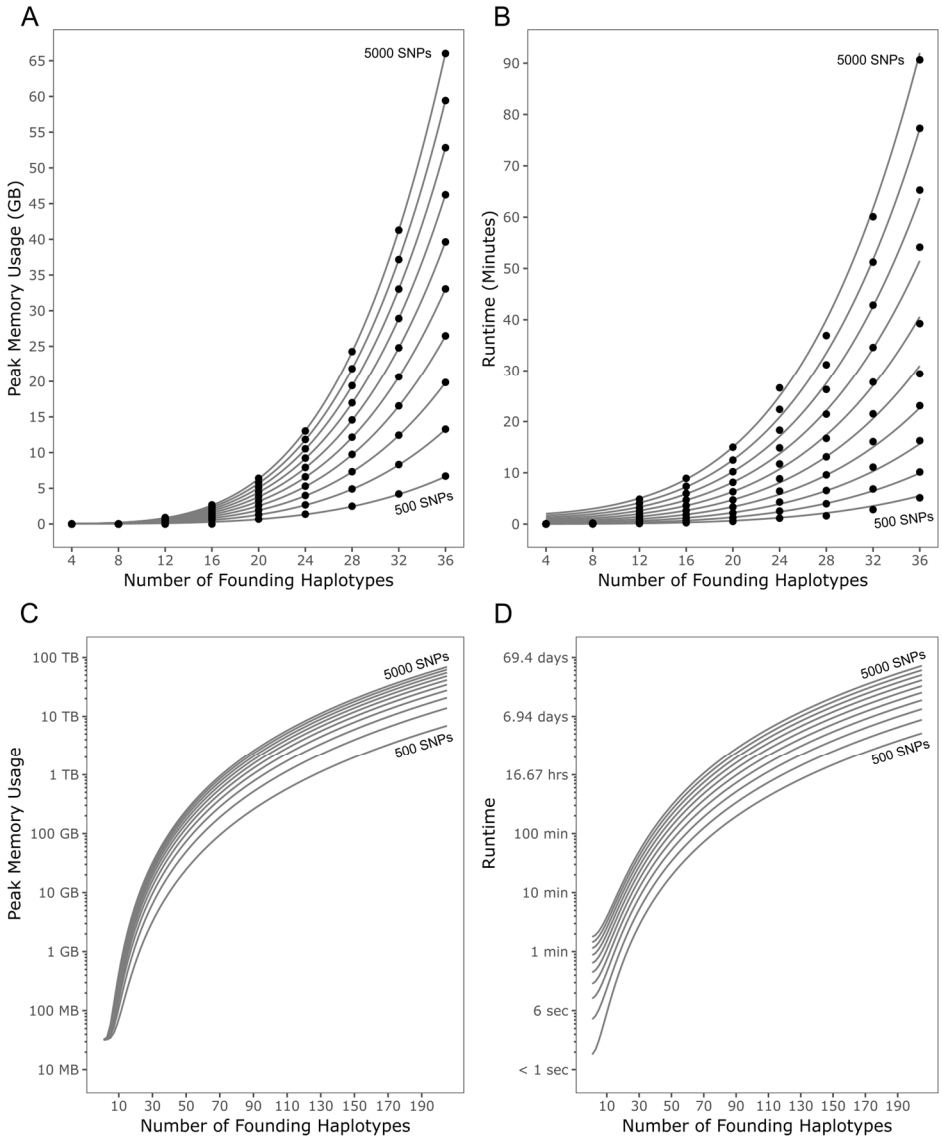


Figure 1. Resource usage of RABBIT during haplotype reconstruction. All reconstructions involve the same simulated 2L chromosome arm comprised of four haplotypes. Simulations included varied numbers of founding haplotypes (N) and a randomly selected set of markers (number of SNPs, S, incremented in steps of 500). All simulations included, at minimum, the four true haplotypes for the simulated individual. In A and B, points depict the mean of empirical values (over 10 replicates) and gray lines depict the defined regression models. Predicted peak memory usage and runtime are displayed on a log scale over a greater range for number of founding haplotypes in C and D, respectively.

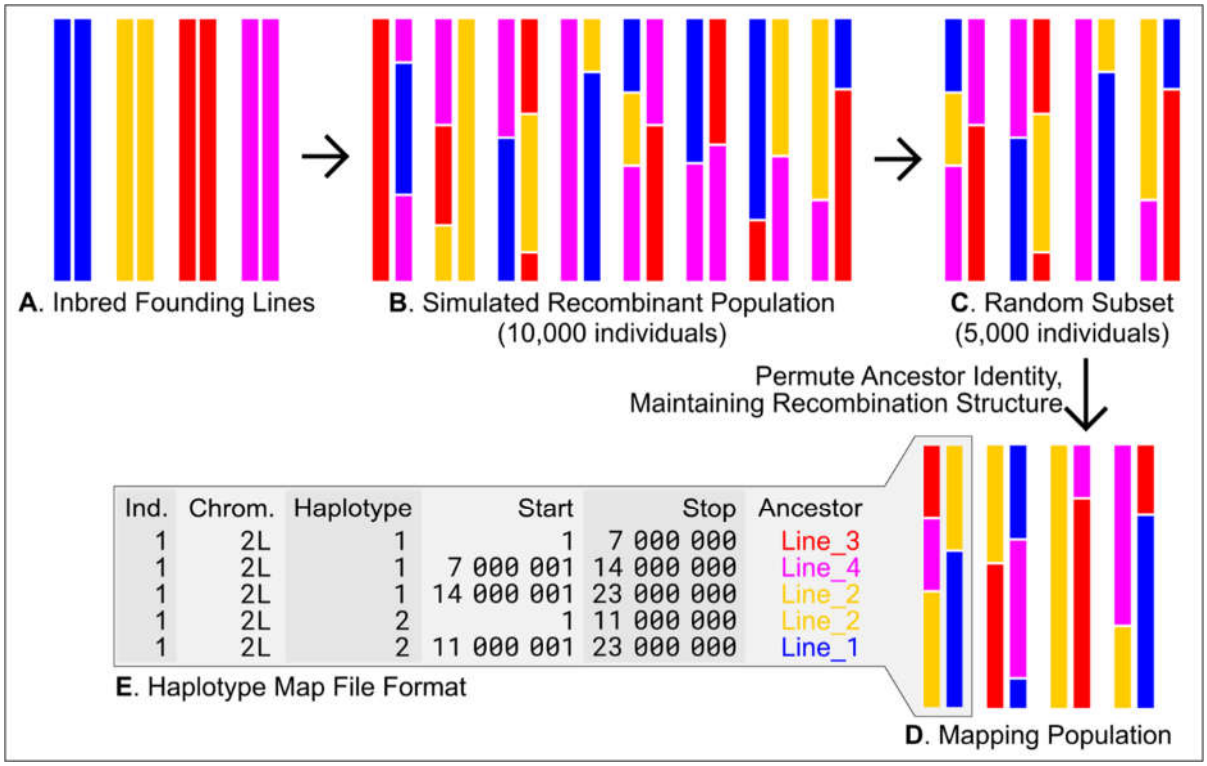


Figure 2. Basic structure of the forward simulator pipeline. Inbred founding lines (A) are randomly intercrossed to produce a recombinant population (B). Rapid generation of independent mapping populations is achieved by random down-sampling (C) and permutation of ancestry (D). Population genetic data is encoded in a highly compressed format (E) that references the positions of haplotype blocks instead of genotypes at every site, enabling us to generate 500 mapping populations for a given parameter combination. Individuals are probabilistically assigned to case or control groups based on genotype at a randomly chosen causal SNP segregating at a specified frequency.

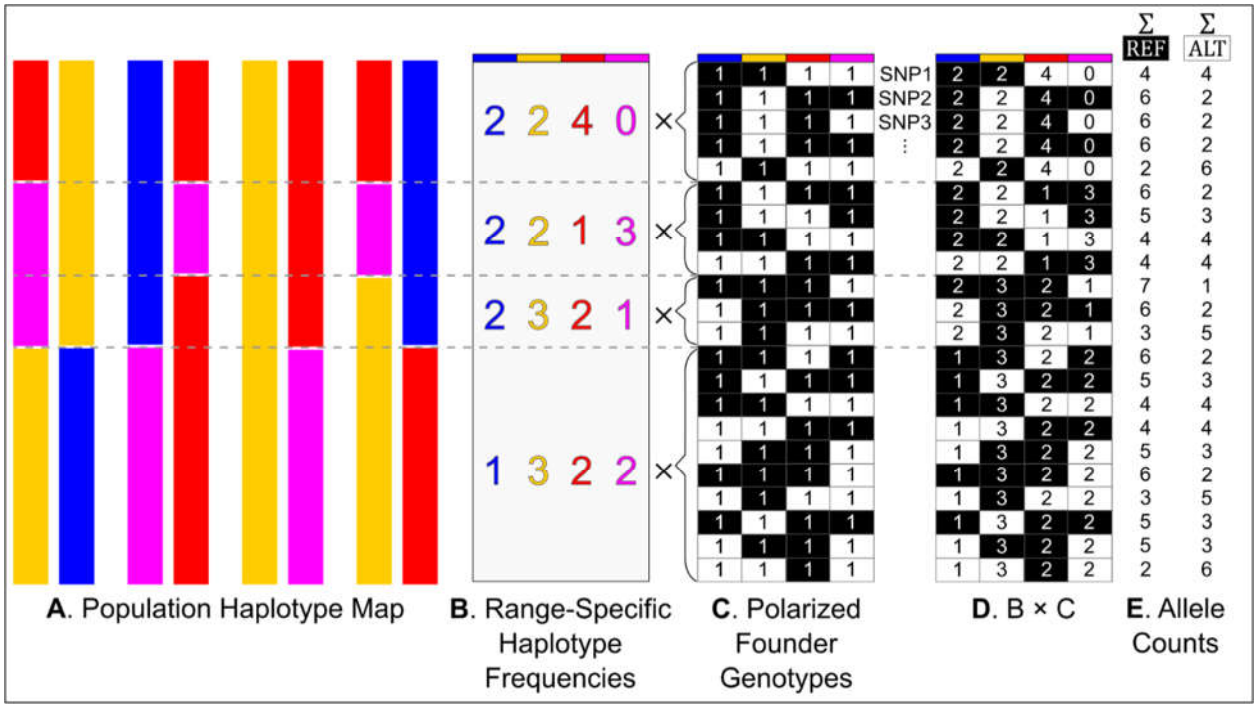


Figure 3. Schematic for rapid association testing with the haplotype block files. For a given population represented by a haplotype map file (A), all SNPs between sorted breakpoints (indicated by dashed lines) will share identical aggregated haplotype frequencies (B). Haplotype frequencies are multiplied by a founder genotype matrix (C) where alleles are coded reference (black cells) and alternate (white cells). Conditional row sums of the resulting matrix (D) yields reference and alternate frequencies at each locus (E), to be used for χ^2 tests of independence.

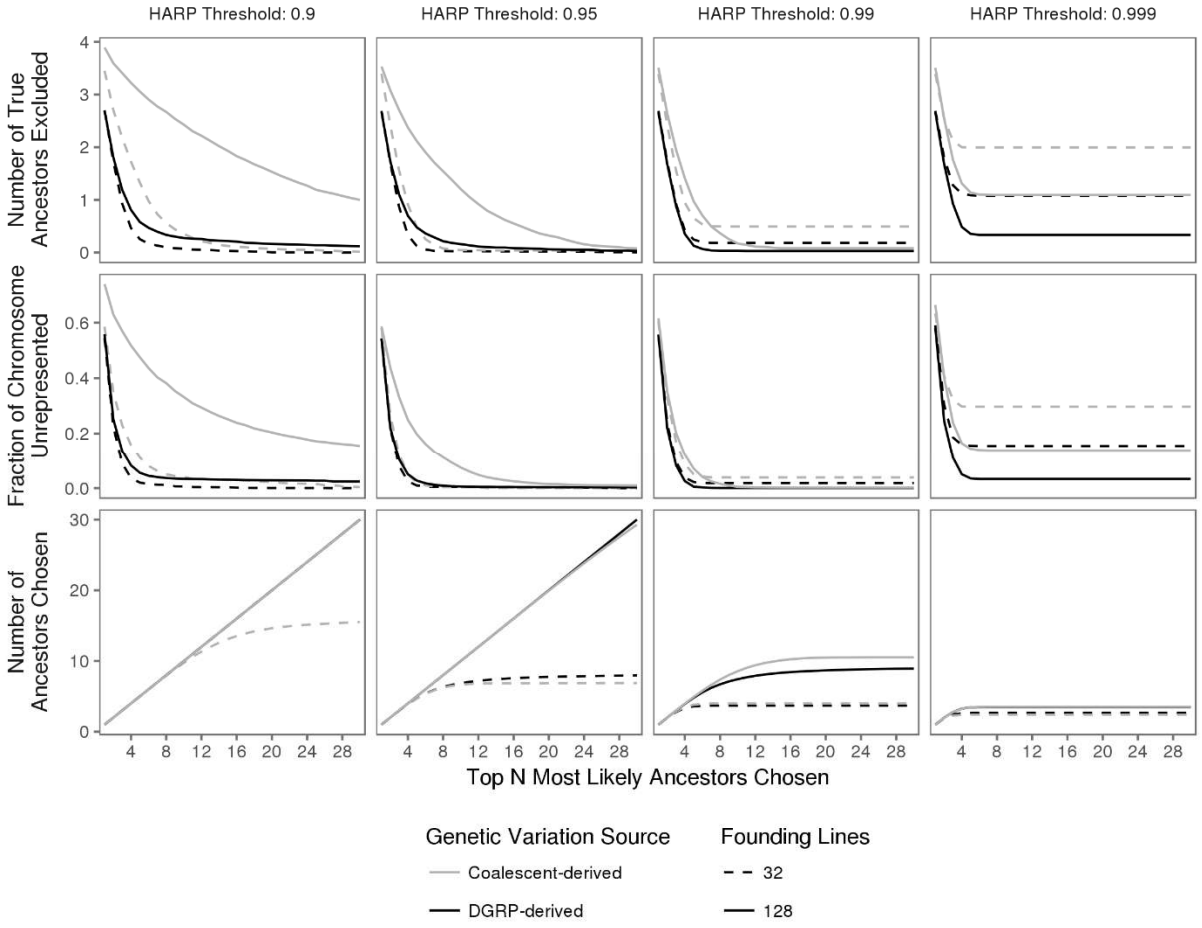


Figure 4. Optimization curves for Most-Likely-Ancestor (MLA) selection. Increasing the upper limit for the number of MLAs chosen reduces the number of true ancestors missed, similarly reducing the fraction of a given chromosome that is not represented within the selected set of MLAs. Ancestors that fail to pass the HARP threshold across all genomic windows are not selected, resulting in realized sets of MLAs (Number of Ancestors Chosen) below the upper-limit allowed (x-axis). Data shown reports means across 400 replicates made up of 100 simulated individuals (4 autosomes each for coalescent simulations, 4 autosome arms each for DGRP simulations) per parameter combination. Coalescent-derived populations described here were simulated with $\boldsymbol{N_e = 10^6}$ and $\boldsymbol{\mu = 5 \times 10^{-9}}$.

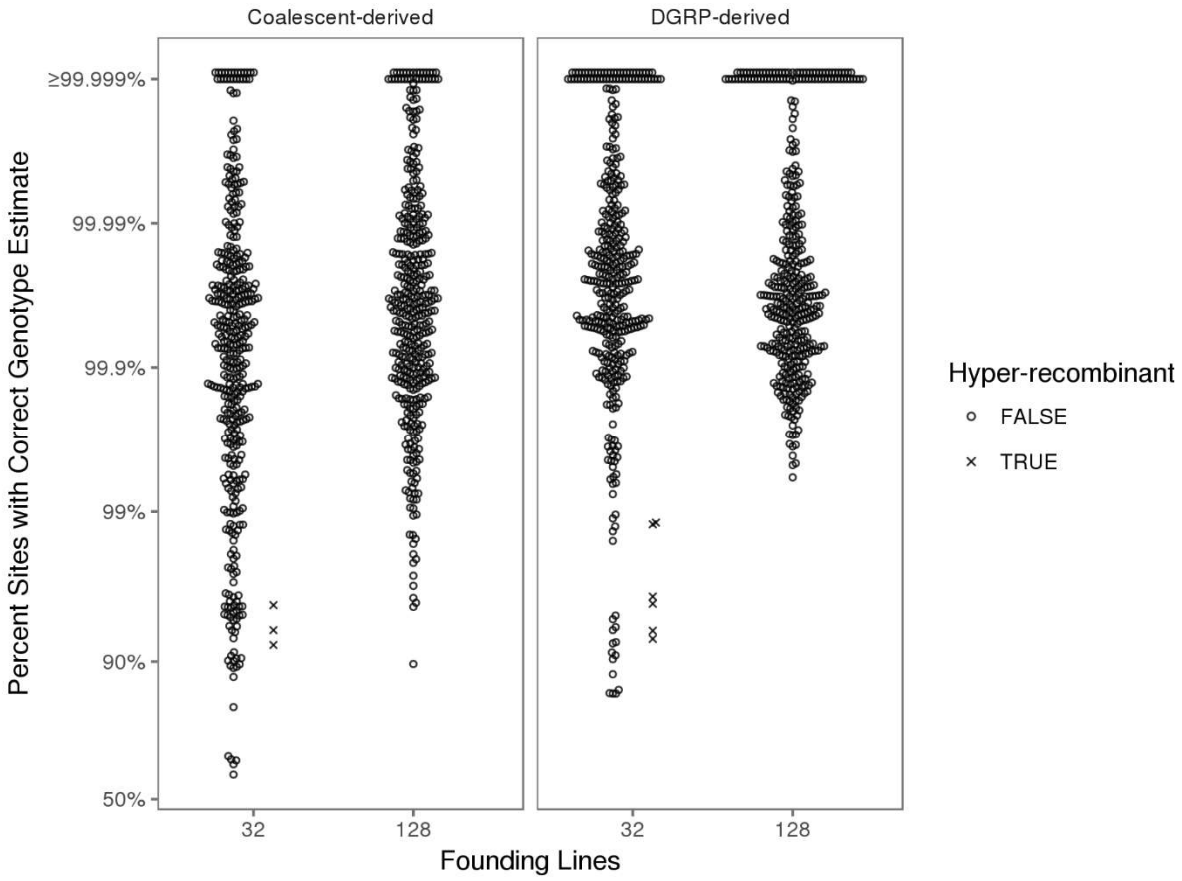


Figure 5. Accuracy of genome reconstruction pipeline for simulated F_5 Hybrid Swarm individuals. Reconstructions were performed for populations simulated as being founded by either 32 or 128 inbred lines at 0.05X sequencing coverage with up to 16 MLAs as determined with a HARP threshold of 0.99. Accuracy, calculated as the per-chromosome fraction of variable sites with a correct diploid genotype estimate, is shown on logit-transformed scale. Values are coded depending on the number of estimated recombination events, with highly recombinant estimates ($\geq$10 recombination events) displayed as an $\times$. Each parameter combination includes 400 reconstructed autosomes (individual circles) for 100 simulated individuals. The coalescent-derived individuals displayed here were simulated with an effective population size of $\boldsymbol{N_e = 1 \times 10^6}$ and mutation rate $\boldsymbol{\mu = 5 \times 10^{-9}}$.

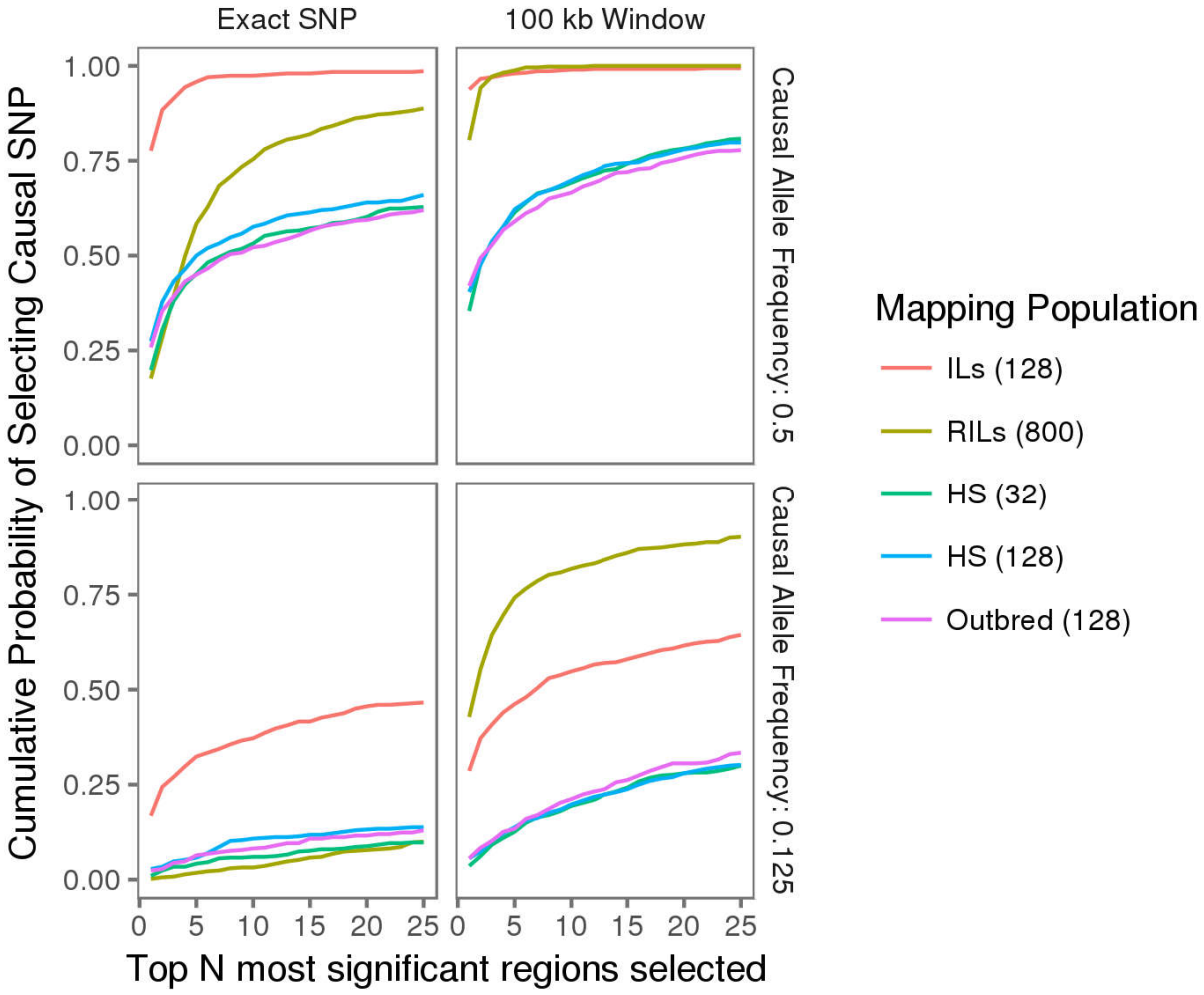


Figure 6. Accuracy of simulated GWAS for various mapping populations. Plots display the cumulative probability of including a causal SNP when selecting the top N most significant SNPs, or 100kb windows around those SNPs, out of 500 simulated GWAS (each comprised of 5000 individuals phenotypically assigned in a case-control framework). Homozygotes for the reference allele were assigned to the case group with 45% probability, while homozygotes for the alternate allele were assigned to the case group with 55% probability (a difference of 10%), and heterozygotes are assigned to case and control groups with equal probability. ILs: inbred lines. RILs: recombinant inbred lines. HS: Hybrid Swarm populations founded by 32 or 128 lines. Outbred: An F_{50} population founded by 128 lines.

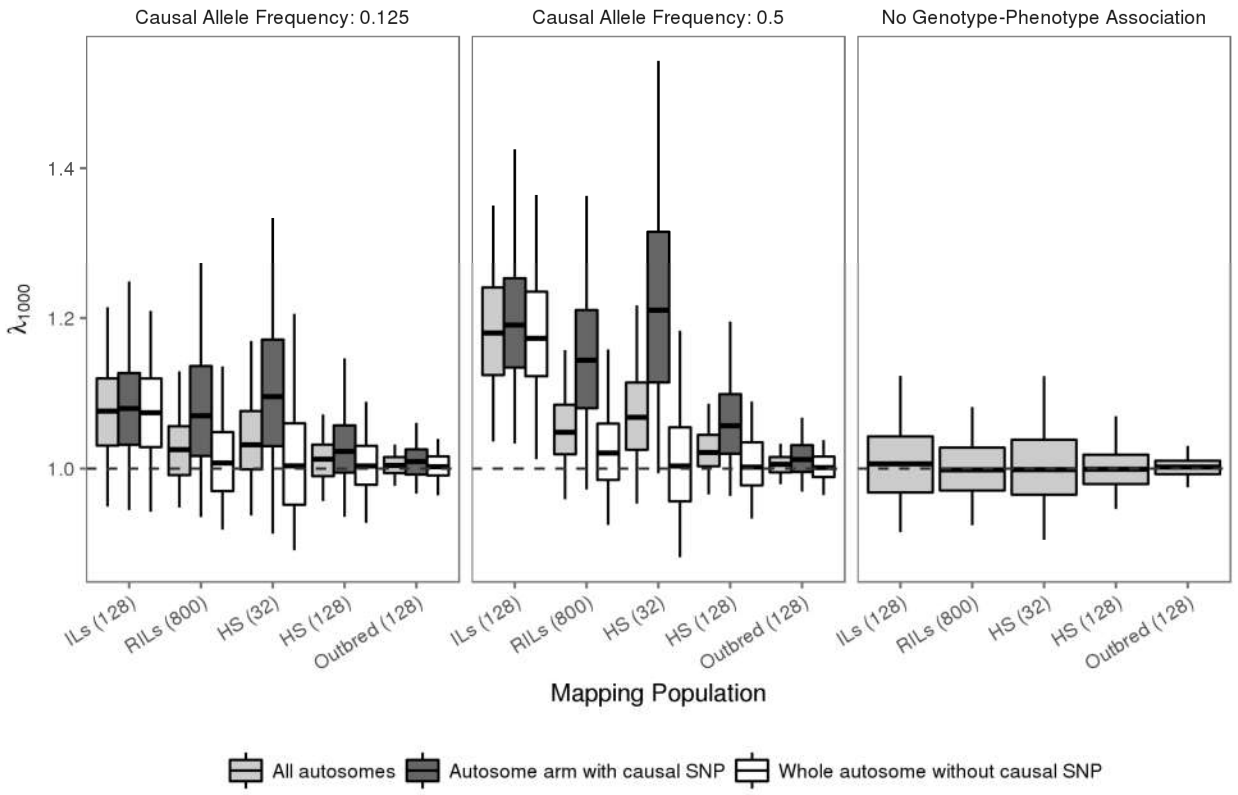


Figure 7. Genomic Inflation Factor (GIF, $\boldsymbol{\lambda_{1000}}$) for simulated GWAS with a causal allele segregating at a specified frequency. $\boldsymbol{\lambda}$ is calculated as the ratio of observed to expected $\boldsymbol{\chi^2}$ values, and a correction is performed to produce the null expectation given the sample size had actually been 1000 individuals (see Materials and Methods for details). Data are averaged over 500 simulated GWAS (each comprised of 5000 individuals phenotypically assigned in a case-control framework). Homozygotes for the reference allele were assigned to the case group with 45% probability, while homozygotes for the alternate allele were assigned to the case group with 55% probability (a difference of 10%), and heterozygotes are assigned to case and control groups with equal probability. Boxes represent the median and interquartile range; whiskers extending to the lower and upper bounds of the 95% quantiles.

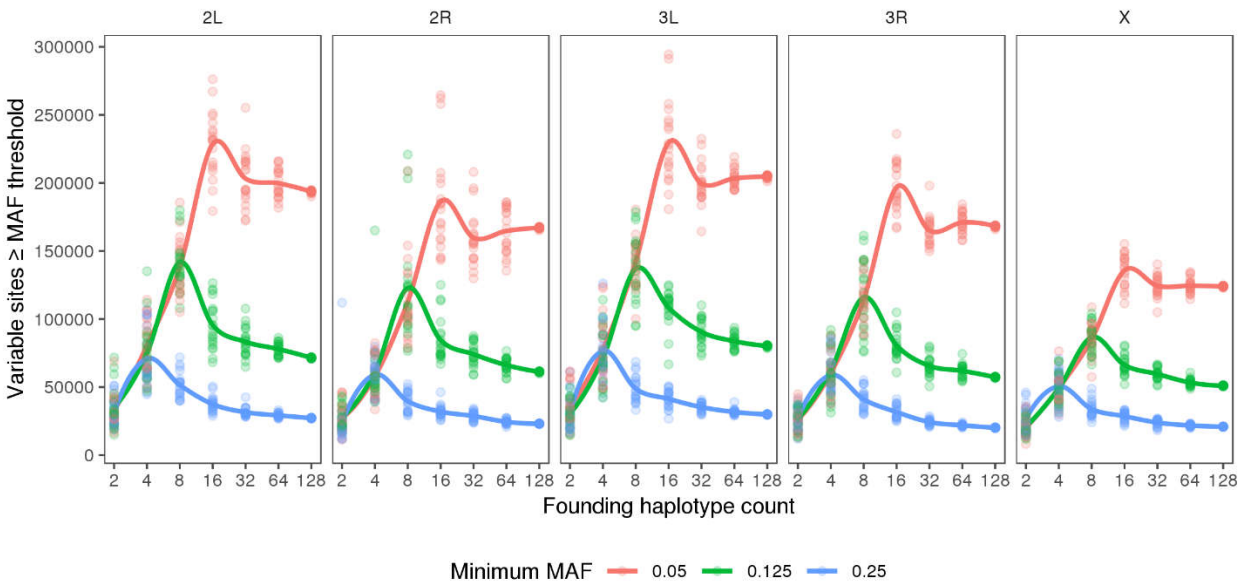


Figure 8. Counts of variable sites depending on number of founding DGRP haplotypes. Each point represents the number of sites segregating at or above a given minor allele frequency threshold when drawing *N* haplotypes, with 20 replicates per parameter combination. With a minimum minor allele frequency (MAF) of 12.5%, a population founded by eight haplotypes exhibits approximately double the number of variable sites compared to a population founded by 128 haplotypes. With a minimum MAF of 5%, populations with eight founding haplotypes present with fewer SNPs compared to populations founded by 16 or more haplotypes.

Population	N Founders	$Lin's\ \rho$	$\underline{\Delta}$	σ_Δ
DGRP	128	0.986	-0.015	0.25
DGRP	32	0.502	0.17	2.15
Coalescent	128	0.956	-0.17	0.44
Coalescent	32	0.759	-0.31	1.26

Table 1. Accuracy of estimated number of recombination events following chromosome reconstruction. A high concordance correlation coefficient (Lin's $\boldsymbol{\rho}$) indicates agreement between estimated and true recombination counts for 400 reconstructed chromosomes (coalescent-derived populations) or chromosome arms (DGRP-derived populations). Coalescent-derived populations are described across a range of values for effective population size $\boldsymbol{N_e}$ and mutation rate $\boldsymbol{\mu}$. $\boldsymbol{\underline{\Delta}}$ and $\boldsymbol{\sigma_\Delta}$ denote mean and standard deviation, respectively, of difference between estimated and true recombination counts. Reconstructions were performed with a maximum of 16 most-likely-ancestors with a HARP threshold of 0.99 (see methods for more details).

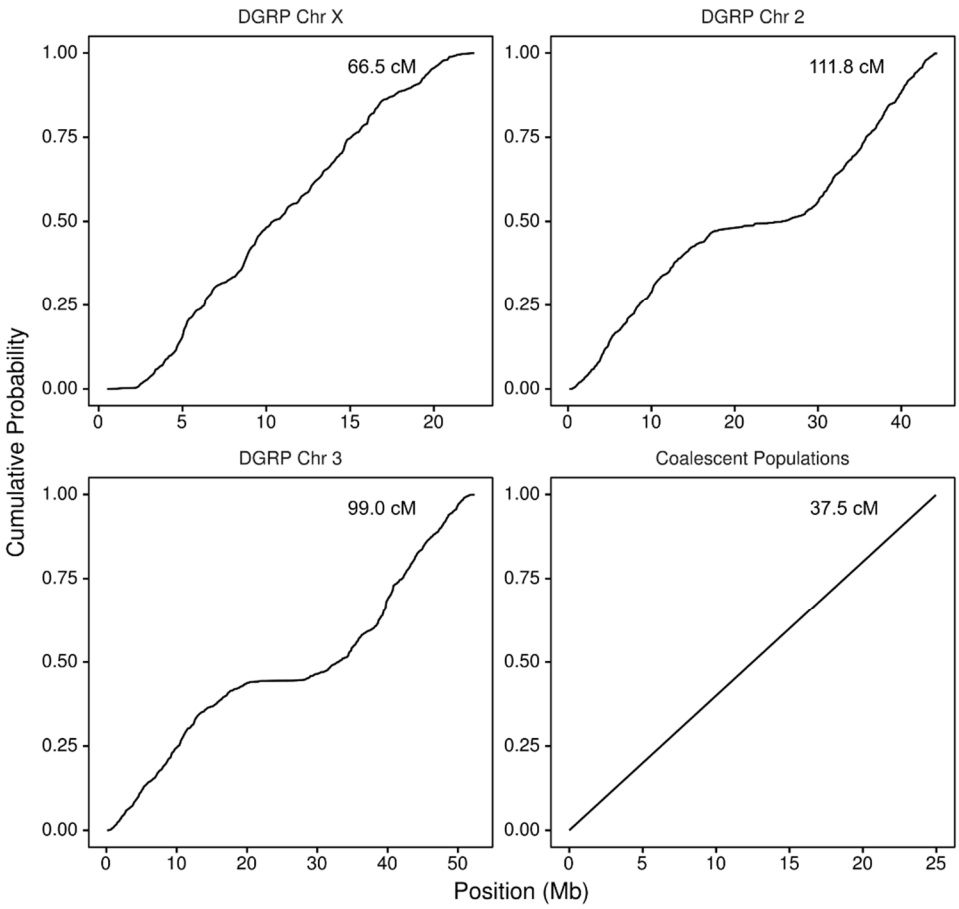


Figure S1. Recombination probability functions used for simulated individuals. Recombination is modeled as a Poisson process, with position sampled from linear interpolation of recombination rates measured in *D. melanogaster* by Comeron et al. (2012). The frequency of recombination samples cumulative map distance (inset, e.g. a 99 cM chromosome is modeled as a Poisson variable with an expected value of $\boldsymbol{\lambda = 0.99}$). For DGRP-derived individuals, recombination was simulated for full chromosomes two and three, and reconstructions were then conducted independently for arms 2L, 2R, 3L, and 3R.

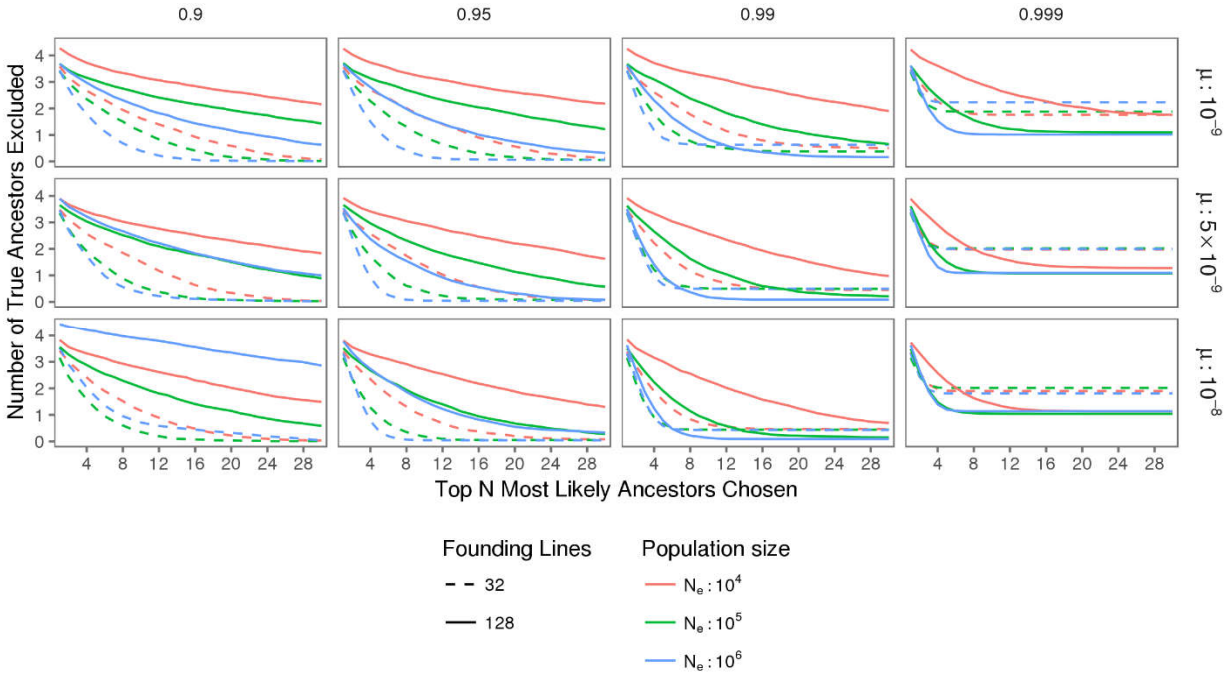


Figure S2. Optimization curves for Most-Likely-Ancestor inclusion, by count, in SCRM-derived F_5 hybrid swarm individuals. The number of missed true ancestors is shown as a function of the number of ancestors chosen across a range of HARP threshold values (0.9 to 0.999), effective population sizes ($\boldsymbol{N_e}$) and mutation rates ($\boldsymbol{\mu}$). Values are averaged across 400 reconstructed autosomes (from 100 individuals) per parameter combination.

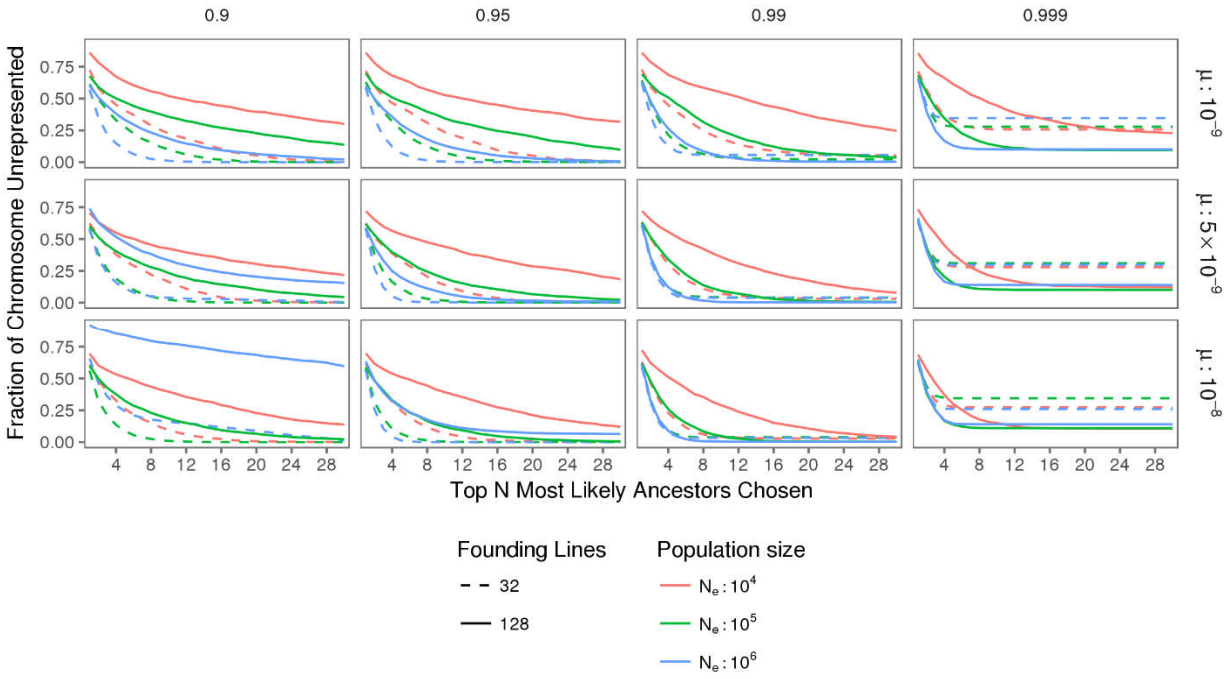


Figure S3. Optimization curves for Most-Likely-Ancestor inclusion, by chromosome representation, in simulated SCRM-derived F_5 hybrid swarm individuals. The proportion of the chromosome not covered by the chosen ancestors is shown as a function of the number of ancestors chosen for populations founded by either 32 or 128 inbred founding lines across a range of HARP threshold values (0.9 to 0.999), effective population sizes ($\boldsymbol{N_e}$) and mutation rates ($\boldsymbol{\mu}$). Each line summarizes the arithmetic mean fraction of sites where the true ancestor is not included within the inferred set of Most-Likely-Ancestors. Values are averaged across 400 reconstructed autosomes (from 100 individuals) per parameter combination.

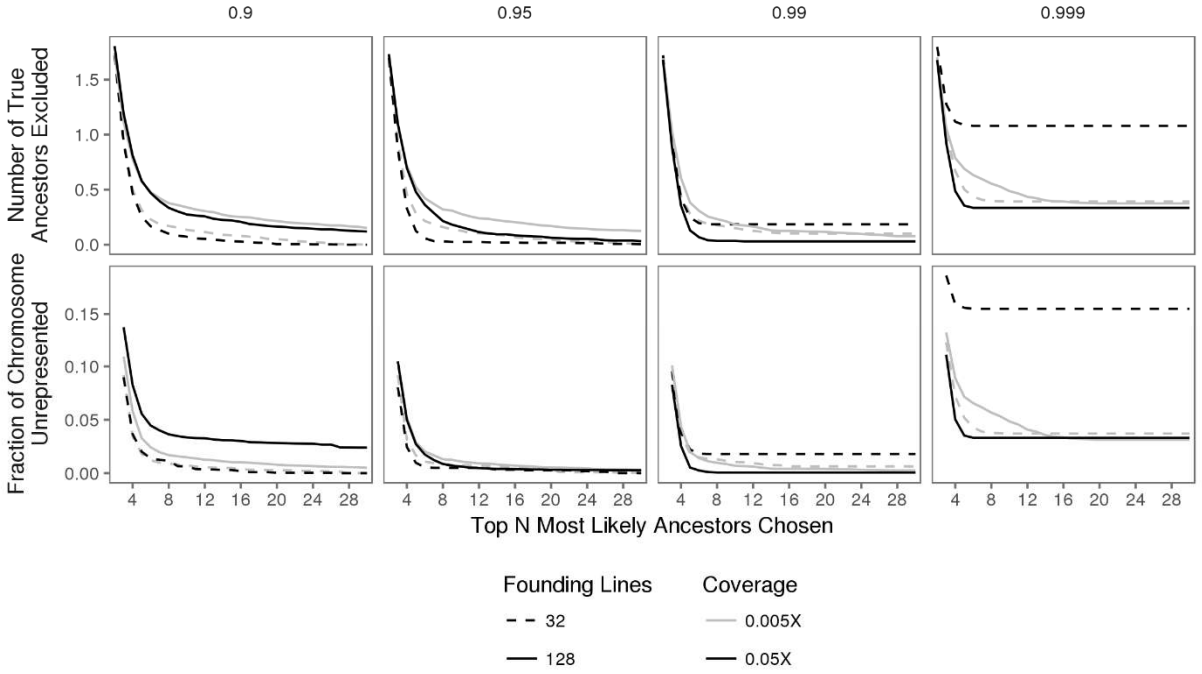


Figure S4. Optimization curves for Most-Likely-Ancestor (MLA) selection for DGRP-derived F_5 hybrid swarm individuals. Effectiveness is shown for populations founded by either 32 or 128 inbred founding lines across a range of HARP threshold values (0.9 to 0.999), for two levels of sequencing coverage. Values are averaged across 400 reconstructed autosomes (from 100 individuals) per parameter combination.

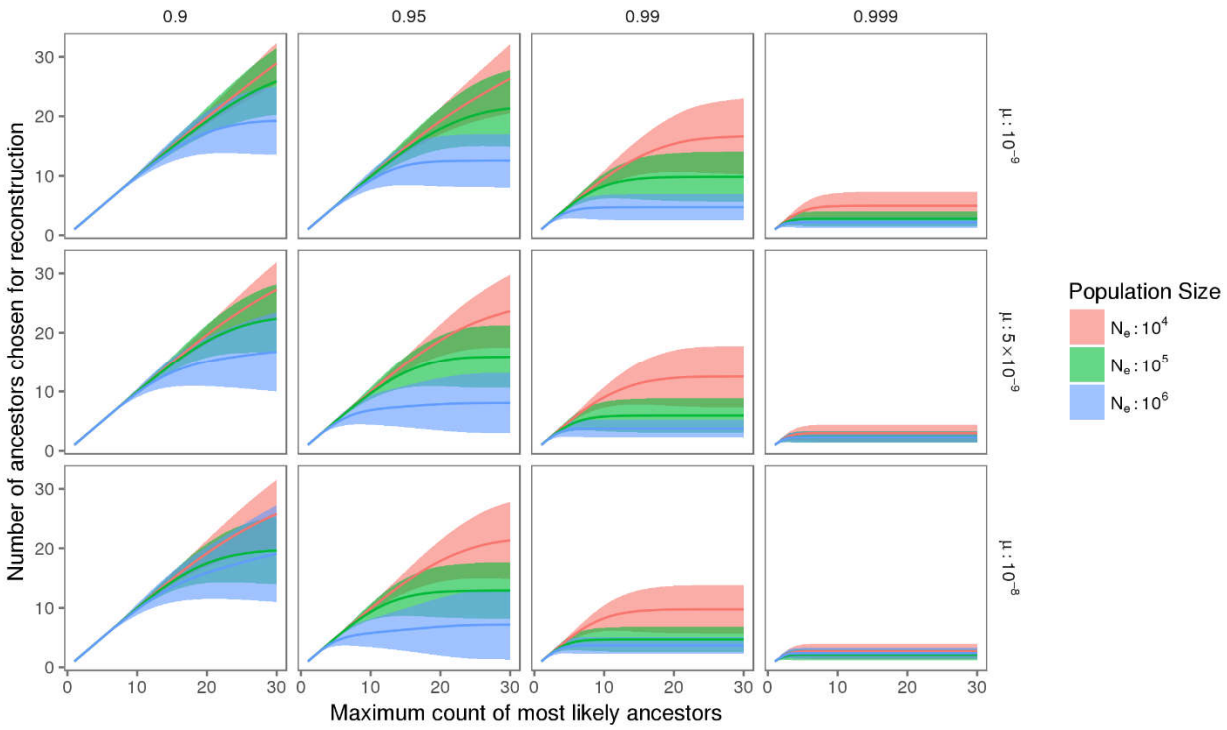


Figure S5. Distribution of Most-Likely-Ancestor counts for simulated, Coalescent-derived, 32-founder F_5 hybrid swarm individuals. The mean value ± 1 standard deviation is shown by the solid line and ribbon, respectively, across a range of HARP threshold values (0.9 to 0.999), effective population sizes ($\boldsymbol{N_e}$) and mutation rates ($\boldsymbol{\mu}$). The number of Most-Likely-Ancestors dictates the computational complexity (runtime and memory requirements) of chromosome reconstruction. Each parameter combination includes 400 chromosomes (from 100 simulated individuals) simulated with 0.05X sequencing coverage.

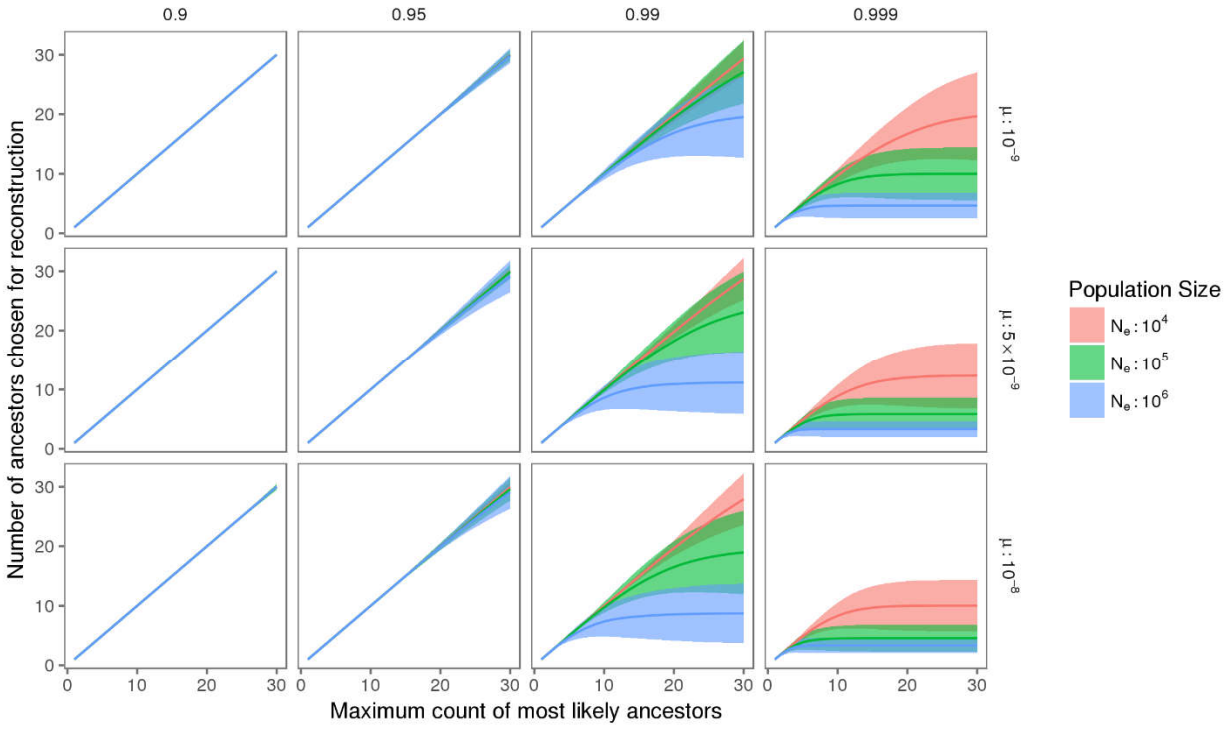


Figure S6. Distribution of Most-Likely-Ancestor counts for simulated, Coalescent-derived, 128-founder F_5 hybrid swarm individuals. The mean value ± 1 standard deviation is shown by the solid line and ribbon, respectively, across a range of HARP threshold values (0.9 to 0.999), effective population sizes ($\boldsymbol{N_e}$) and mutation rates ($\boldsymbol{\mu}$). The number of Most-Likely-Ancestors dictates the computational complexity (runtime and memory requirements) of chromosome reconstruction. Each parameter combination includes 400 chromosomes (from 100 simulated individuals) simulated with 0.05X sequencing coverage.

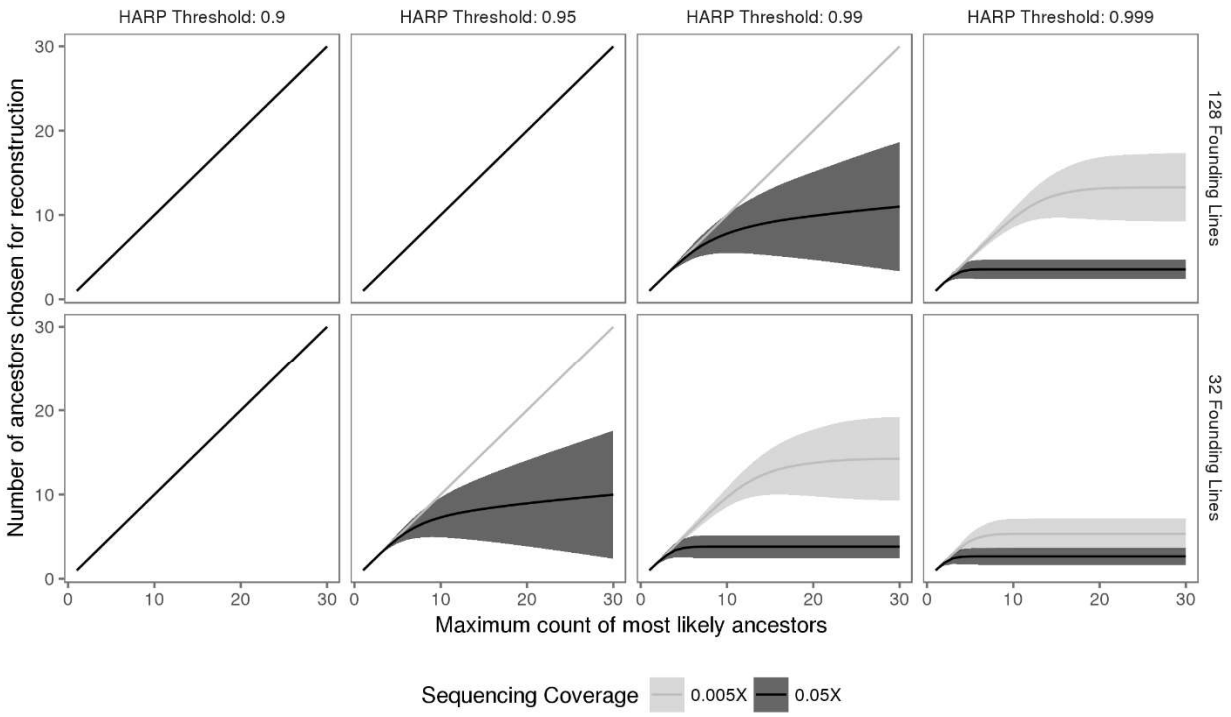


Figure S7 Distribution of Most-Likely-Ancestor counts for simulated, DGRP-derived F_5 hybrid swarm individuals. The mean value ± 1 standard deviation is shown by the solid line and ribbon, respectively, for populations founded by either 32 or 128 inbred lines, across a range of HARP threshold values (0.9 to 0.999) and two levels of sequencing coverage. Each parameter combination includes 400 chromosomes (from 100 simulated individuals) simulated with 0.05X sequencing coverage.

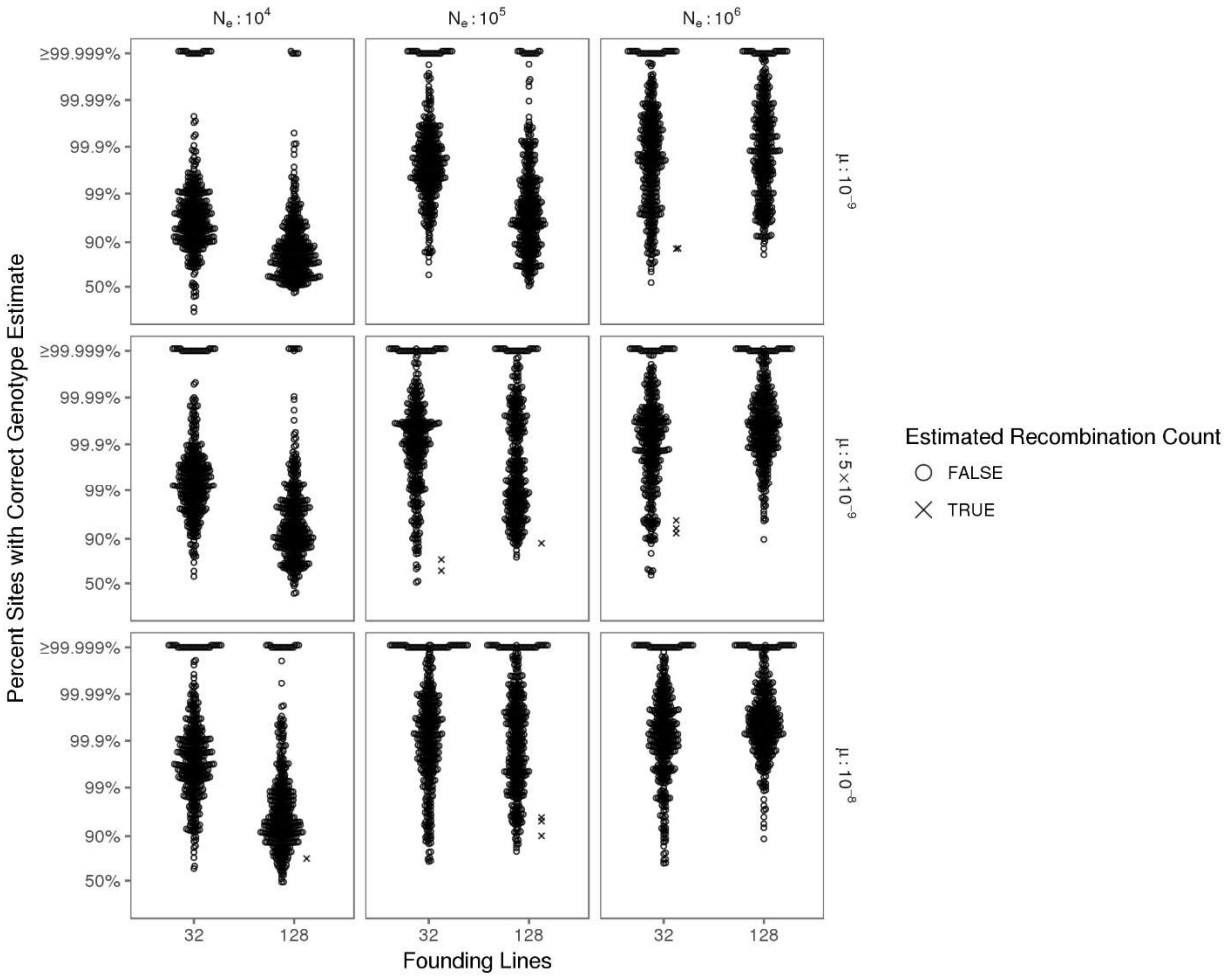


Figure S8. Accuracy of genome reconstruction for simulated, coalescent-derived F_5 hybrid swarm individuals. Reconstructions were performed for populations simulated as being founded by either 32 or 128 inbred lines for various effective population sizes ($\boldsymbol{N_e}$) and mutation rates ($\boldsymbol{\mu}$). Accuracy is represented on a logit scale, as most points occur above 90%. Reconstructed chromosomes that are predicted to exhibit $\geq$10 recombination events are denoted by an X. Each parameter combination includes 400 reconstructed chromosomes (from 100 simulated individuals).

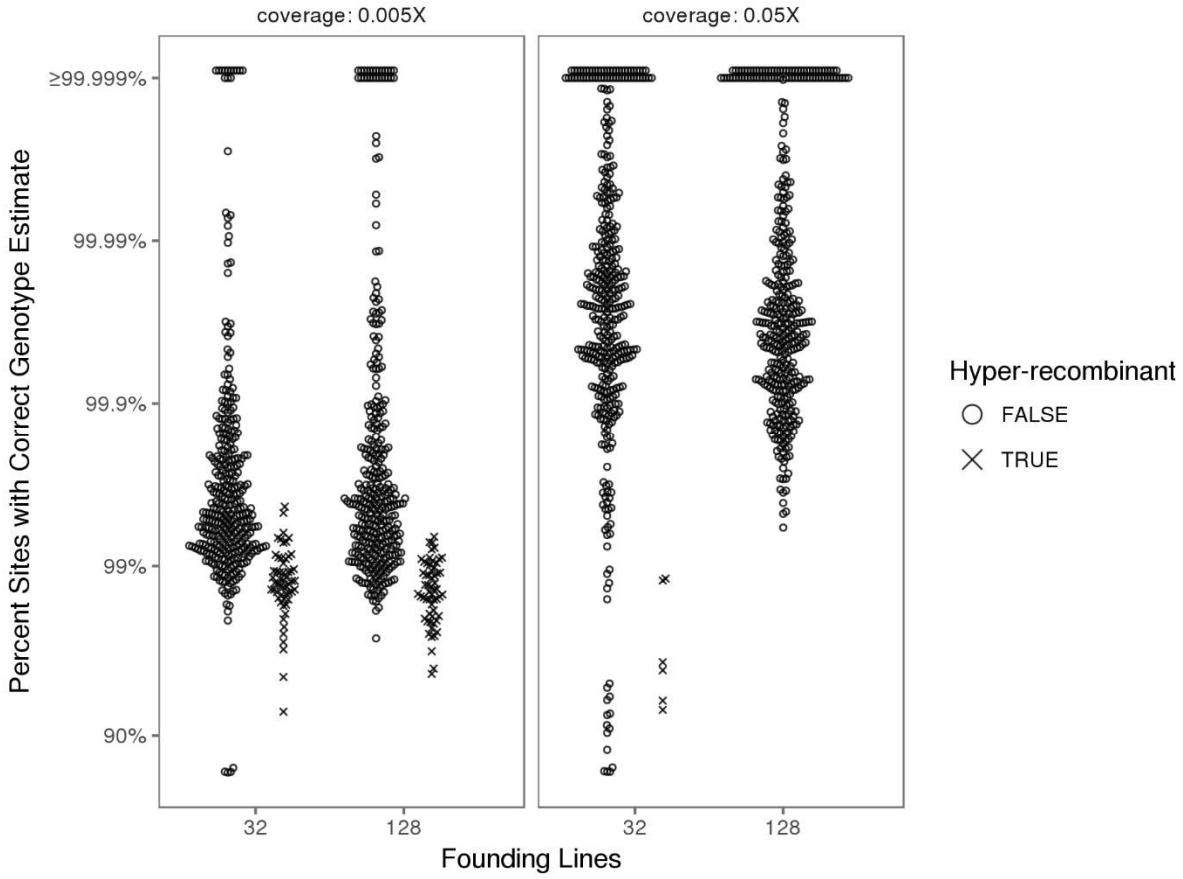


Figure S9 Accuracy of genome reconstruction for simulated, DGRP-derived F_5 hybrid swarm individuals. Reconstructions were performed for populations simulated as being founded by either 32 or 128 inbred lines for two levels of ultra-low sequencing coverage. Accuracy is represented on a logit scale, as most points occur above 90%. Accuracy values are marked depending on the number of estimated recombination events, with highly recombinant estimates ($\geq$10 recombination events) displayed as an ×. Each parameter combination includes 400 reconstructed chromosomes (from 100 simulated individuals).

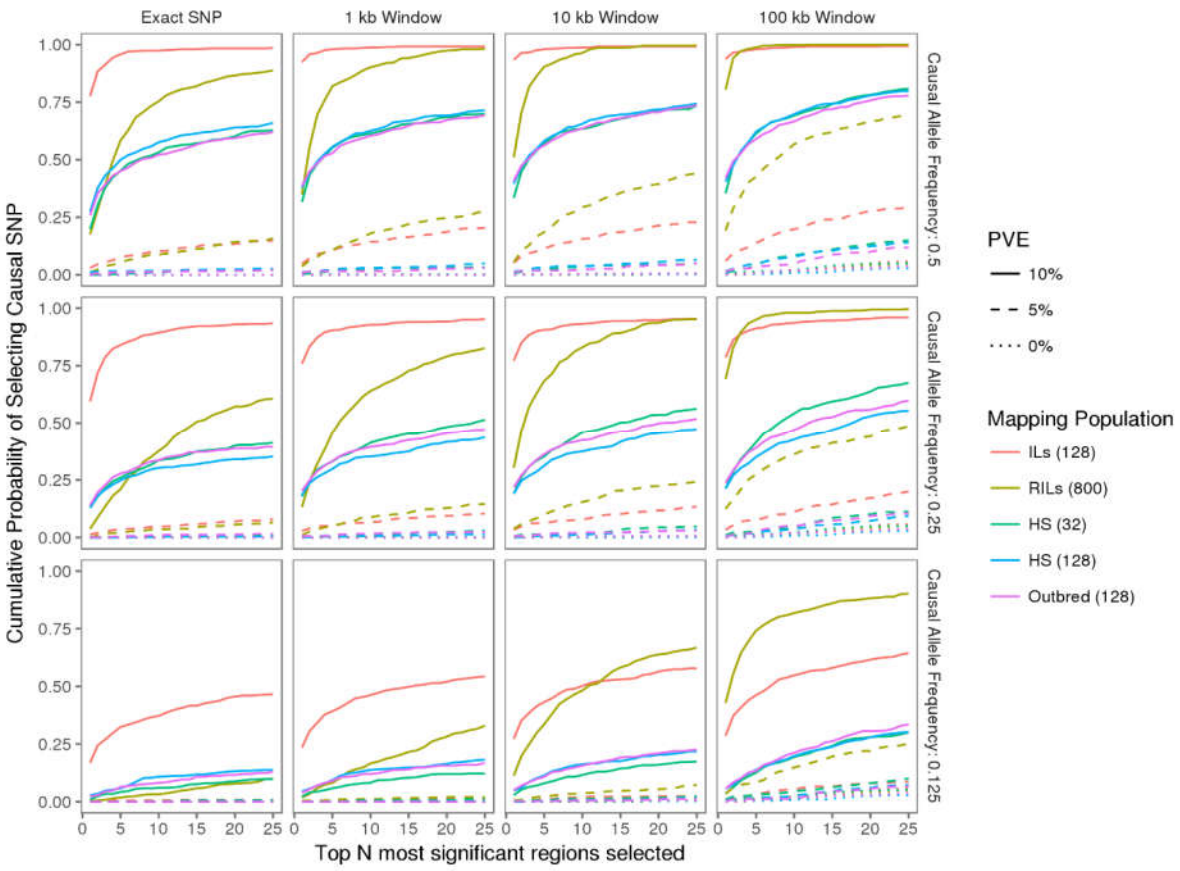


Figure S10. Probability of selecting a causal SNP (or a nearby neighbor) in simulated GWAS. Each line represents the fraction of GWAS simulations (out of 500 total GWAS, each comprised of 5000 individuals in a case-control framework) where the causal SNP is selected within N most significant regions. Case-control status was assigned based on reference allele dosage at a randomly selected causal SNP segregating with frequency of 50%, 25%, or 12.5%. For 10% PVE, homozygotes for the reference allele were assigned to the case group with 45% probability, while homozygotes for the alternate allele were assigned to the case group with 55% probability (a difference of 10%). For 5% PVE, homozygotes for the reference allele were assigned to the case group with 47.5% probability, while homozygotes for the alternate allele were assigned to the case group with 52.5% probability (a difference of 5%). All heterozygotes (and any individuals modeled with 0% PVE, irrespective of genotype) were equally likely to be assigned to case or control group (a difference of 0%).

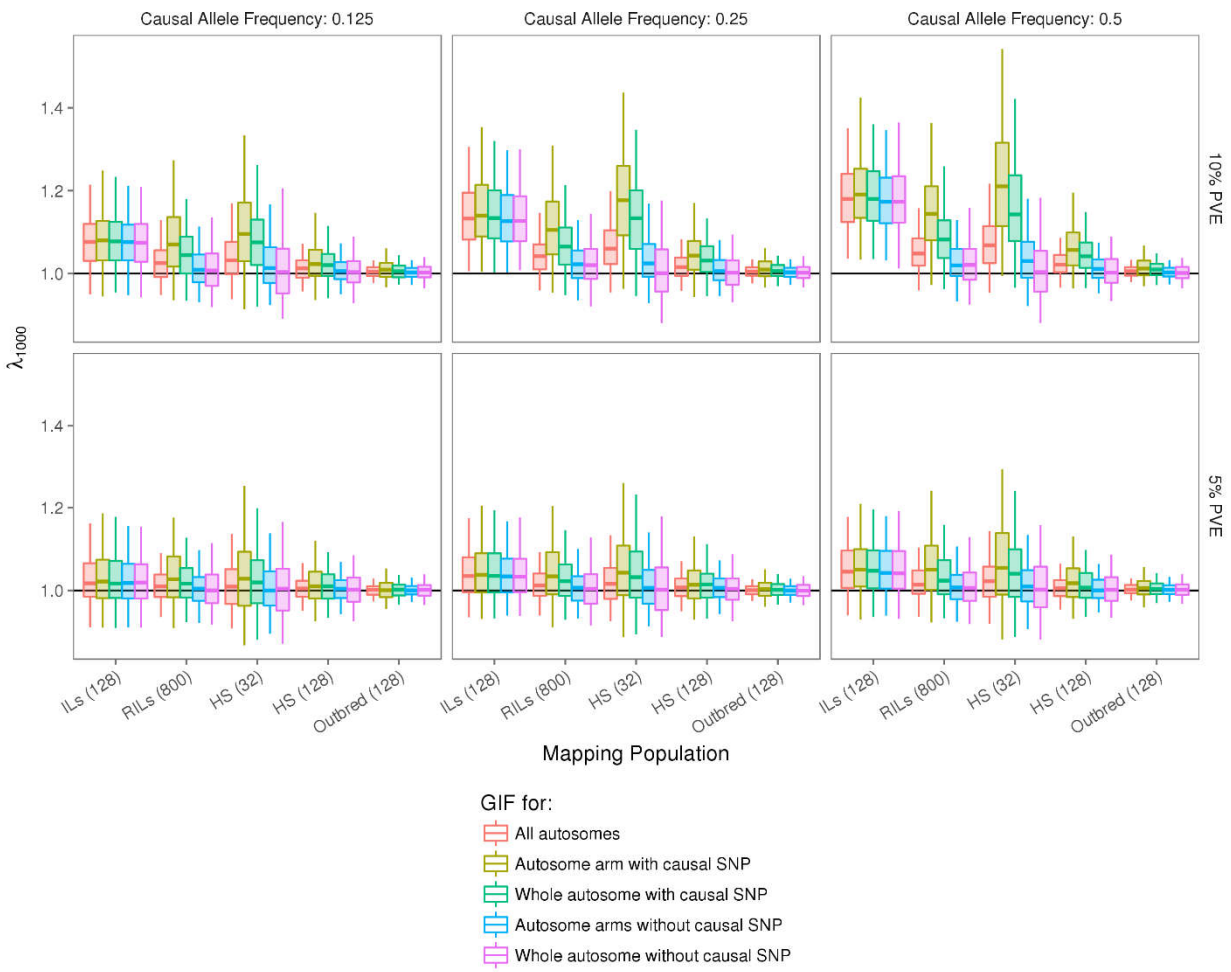


Figure S11. Genomic Inflation Factor (GIF, λ_{1000}) in simulated DGRP GWAS as a function of minor allele frequency and percent variation explained. GIF is greatest on the arm which contains (and is directly and most strongly linked to) the SNP associated with case and control status. When PVE is 0 (and thus, case and control status is randomly assigned), GIF centers around 1.0 as expected.

Population	Coverage	Founders	N_e	μ	$Lin's\ \rho$	$\underline{\Delta}$	σ_Δ
DGRP	0.005X	128	-	-	0.201	3.06	3.11
DGRP	0.005X	32	-	-	0.201	3.15	2.81
DGRP	0.05X	128	-	-	0.986	-0.015	0.25
DGRP	0.05X	32	-	-	0.502	0.17	2.15
Coal.	0.05X	128	10^4	1×10^{-9}	0.029	-2.72	1.60
Coal.	0.05X	32	10^4	1×10^{-9}	0.219	-1.84	1.50
Coal.	0.05X	128	10^5	1×10^{-9}	0.288	-1.57	1.64
Coal.	0.05X	32	10^5	1×10^{-9}	0.761	-0.53	0.99
Coal.	0.05X	128	10^6	1×10^{-9}	0.742	-0.46	1.09
Coal.	0.05X	32	10^6	1×10^{-9}	0.754	-0.42	1.22
Coal.	0.05X	128	10^4	5×10^{-9}	0.203	-2.03	1.73
Coal.	0.05X	32	10^4	5×10^{-9}	0.622	-0.89	1.17
Coal.	0.05X	128	10^5	5×10^{-9}	0.652	-0.64	1.41
Coal.	0.05X	32	10^5	5×10^{-9}	0.782	-0.26	1.16
Coal.	0.05X	128	10^6	5×10^{-9}	0.956	-0.17	0.44
Coal.	0.05X	32	10^6	5×10^{-9}	0.759	-0.31	1.26
Coal.	0.05X	128	10^4	1×10^{-8}	0.238	-1.65	1.86
Coal.	0.05X	32	10^4	1×10^{-8}	0.745	-0.66	1.05
Coal.	0.05X	128	10^5	1×10^{-8}	0.776	-0.34	1.12
Coal.	0.05X	32	10^5	1×10^{-8}	0.846	-0.25	0.93
Coal.	0.05X	128	10^6	1×10^{-8}	0.937	-0.22	0.56
Coal.	0.05X	32	10^6	1×10^{-8}	0.833	-0.32	0.95

Table S1. Accuracy of estimated number of recombination events following chromosome reconstruction. A high concordance correlation coefficient (Lin's $\boldsymbol{\rho}$) indicates agreement between estimated and true recombination counts for 400 reconstructed chromosomes (coalescent-derived populations) or chromosome arms (DGRP-derived populations). Coalescent-derived (Coal.) populations are described across a range of values for effective population size $\boldsymbol{N_e}$ and mutation rate $\boldsymbol{\mu}$. $\boldsymbol{\underline{\Delta}}$ and $\boldsymbol{\sigma_\Delta}$ denote mean and standard deviation, respectively, for difference between estimated and true recombination count. Reconstructions were performed with a maximum of 16 most-likely-ancestors with a HARP threshold of 0.99 (see methods for more details).

CHAPTER THREE

Evaluating the genetic basis of *Drosophila melanogaster* starvation and desiccation tolerance in a Hybrid Swarm population

Abstract

By identifying genetic variants associated with a trait of interest, as in genome-wide association studies (GWAS), researchers can bridge the gap from genotype to phenotype. Because a phenotype can be modulated through nuanced, heritable variation in expression, our understanding of adaptive mechanisms can be strengthened by combining DNA GWAS and RNA expression data. Here, we conduct a holistic investigation into the genetic bases of a complex life history trait—tolerance to starvation and desiccation conditions—by combining GWAS with differential expression and allele-specific expression data. Using the Hybrid Swarm method, we generate and inexpensively genotype a *D. melanogaster* mapping population. Through GWAS, we identify variants associated with survival, including mitochondrial proteins, which have been implicated in stress tolerance and lifespan extension. Differential expression analysis confirms these mitochondrial proteins are downregulated in starvation conditions, along with oogenesis and protease genes. For accurate detection of Allele-Specific Expression (ASE), we evaluate a bias-free method of expression data mapping. We did not observe increased representation of ASE for SNPs associated with seasonality. Our results suggest there is variable propensity of ASE across tissue classes, with a reduction of ASE for genes primarily expressed in the ovary and increased ASE for genes expressed in the carcass, potentially indicating tissue-specific targets of selection. Together, our results indicate the potential of bias-free expression quantification and utility of Hybrid Swarm populations for the study of complex traits.

Introduction

Natural populations experience selective pressures that vary across space and time, and these pressures govern the distribution, phenology, and physiology of organisms (Bellard *et al.* 2012). Climate is a major determinant of species distributions, and climate instability may produce increasingly variable environments (MacMillan and Sinclair 2011) with "worst-case scenario" predictions of rapid, widespread extinction of animal species (Bellard *et al.* 2012). While species distribution models predict how a

population is currently suited for a given environment, habitat preference is not necessarily static over time. In order to understand how a population may evolve in response to a changing climate, it is useful to incorporate genetic mechanisms of local adaptation (reviewed in Savolainen *et al.* 2013).

An organism's phenotype is influenced by protein coding and non-coding regulatory genetic variation. King and Wilson (1975) argued that the phenotypic differences between humans and chimpanzees were too great to be solely due to protein differences, but could be explained by large effects from non-coding regulatory mutations. It is now known that regulatory genetic variation plays a critical role in adaptation (Andolfatto 2005; Abzhanov *et al.* 2006; Kopf *et al.* 2015; Nourmohammad *et al.* 2017; Mack *et al.* 2018), with one of the first examples of adaptive expression variation in the teleost fish *Fundulus heteroclitus* conferring local adaptation to temperature (Crawford and Powers 1992). Despite our knowledge that phenotypes vary across space and time, and that both coding and regulatory variation are drivers of phenotypic variation, the forces that maintain genetic variation and precise genetic mechanisms that underlying adaptation in natural populations remain uncertain.

To elucidate the relationships between protein coding and expression variation in natural populations, we studied the genetic bases of starvation and desiccation tolerance in a Hybrid Swarm mapping population of *Drosophila melanogaster*. By sampling individuals both during population expansion and after an extended period without food, studied periods of bountiful resources and reproduction, as well as end-of-season resource scarcity. Starvation and desiccation tolerance in *D. melanogaster* is an ideal system to study adaptation for many reasons: *D. melanogaster* exhibit large population sizes, reproduce with rapid generation time, and display phenotypically and genetically variable distributions across space and time. The environmental stresses of starvation and desiccation vary both geographically and temporally, and are expected to change along with a changing climate.

Using our Hybrid Swarm population, we first identify variants contributing to desiccation and starvation tolerance during 23 hours without food in a GWAS framework. Second, we evaluate the extent to which expression modulation contributes to response to starvation and desiccation, and describe the physiological context of these expression changes. Third, we provide evidence for genes that undergo changes in allele-specific expression, a potential mechanism for maintenance of genetic variation via reversals of dominance. Fourth, we intersect our data with previously characterized seasonal and clinal genetic variants to determine whether allele-specific expression is enriched for seasonal and clinal loci. Fifth, we determine whether genes primarily expressed in specific tissues exhibit different propensity for allele-specific expression. Together, our results provide a holistic view of the genetic mechanisms—coding and noncoding—contributing to temporal and spatial life-history variation in natural populations.

Methods

Generating the Hybrid Swarm Population

We generated a hybrid swarm mapping population through undirected mating of *Drosophila melanogaster* Genetic Reference Panel (DGRP) inbred lines (Mackay *et al.* 2012) within four replicate cages (6' x 6' x 6'). Populations expanded for four generations on cornmeal-molasses media served in 9"x13" aluminum baking trays. Generations were discrete: we removed egg-laden media from the cages, and replaced egg-laden media in emptied cages after removal of the previous generation. We collected pre-starvation samples of the population during peak reproduction at the end of the 4th generation. All food was removed from the cages to initiate the starvation/desiccation treatment, and we sampled post-starvation samples after 23 hours, when over 90% of the population had died.

Genomic DNA/mRNA Isolation

We homogenized individual flies in 350 μL Lysis Buffer RLT Plus using 2-3 1 mm beads in a bead shaker for 2 minutes on 96 well plates. DNA and RNA were extracted using the AllPrep DNA/RNA Micro Kit (Qiagen product number 80284). Following DNA extraction, we purified samples with Ampure XP Beads (Beckman Coulter product number A63880) in a 1.8x ratio of beads to DNA and eluted to a final volume of 5uL. Following RNA extraction, we purified samples and eluted to 50 μL with Ampure XP Beads in a 1.8x ratio of beads to DNA. Following purification, we isolated mRNA with NEBNext® Poly(A) mRNA Magnetic Isolation Module (product number E7490L) and eluted to a final concentration of 11 μL.

Library Preparation

We prepared 1 μL of DNA/cDNA at ~2.5 ng/μL sequencing using a modified Nextera protocol developed by Baym *et al.* (2015), indexing samples with custom primers. Following library preparation, we quantified DNA/RNA concentrations using a Life Technologies Qubit spectrophotometer, and then pooled each plate of ~96 samples equimolarly. To purify each set of pooled libraries, we purified the pool using Ampure XP Beads and eluted to 160 μL. Following purification, all pooled samples were loaded into a 2%, pre-cast SizeSelect E-Gel (Life Technologies product number G661002). We removed pooled samples at 450-500 bp length into a volume of 15 μL nuclease free water. As a final step to amplify the prepared DNA sequencing libraries, we ran all size selected samples through additional 5 rounds of PCR. Each PCR reaction used 5 μL template DNA, 0.6 μL of 100 mM forward and reverse primers (custom synthesized by IDT), 10μL of KAPA HiFi Ready Mix (KAPA Biosystem [KAPA] product number KK2611/2612), and 3.8 μL nuclease free water. Our PCR protocol included 5 minutes of initial denaturation at 95°C followed by 4 rounds of 20 seconds denaturation (98°C), 20 seconds annealing (62°C), 30 seconds elongation (72°C), followed by a final elongation at 72°C for 2 minutes. Following PCR amplification, we purified DNA libraries using Ampure XP beads and quantified concentrations on a Life Technologies Qubit spectrophotometer and Agilent

Bioanalyzer. We diluted each of the pooled libraries to the appropriate concentration for sequencing on a HiSeq 3000 unit.

cDNA Synthesis

We used a modified version of First Strand cDNA Synthesis (Standard Protocol) (NEB product number M0368). Following mRNA isolation, we denatured 11 μL of RNA at 65°C for 5 minutes and then placed RNA on ice. After icing, we added 4μL of 5X First strand buffer (Invitrogen catalog number 18080093), 2 μL of 0.1MM DTT (Invitrogen catalog number 18080093), and 1 μL of RNAseOut (Invitrogen catalog number 10777019), then incubated for 2 mins at 42°C. To generate cDNA, we added 1 μL of Superscript III and incubated for 50 minutes at 42°C, then an additional 15 minutes at 72°C. To perform second strand cDNA synthesis, we added 6 μL of water, 10 μL of 10X FS Buffer, and 3 μL of dNTP mix were added and incubated on ice for 5 minutes. After icing, we added 1 μL of RNAse H (NEB product M0297L) and 5 μL of DNA Pol I (Invitrogen product number 18010017), then samples were incubated for 2.5 hours at 16°C.

Genome Reconstructions

We performed reconstructions of whole genomes from low coverage DNA sequencing data as described in chapter two. Briefly, we mapped low-coverage DNA reads using *bwa* version 0.7.17-r1188 (Li and Durbin 2009). We then used HARP (Kessner *et al.* 2013) to estimate the most-likely-ancestors for any individual chromosome, and performed haplotype reconstruction in RABBIT with this reduced set of ancestor haplotypes. We then performed additional quality control on the estimated genome reconstructions. Because our hybrid swarm populations had undergone 4 rounds of meiosis, short haplotype blocks (one Mb or shorter) should be rare and are most likely errors in haplotype estimation. To reduce the effect of these reconstruction errors, we bridged across short haplotype segments flanked on both ends with the same haplotype identity. The result of this bridging process produced a large continuous haplotype block of shared identity. We then collapsed consecutive

haplotype blocks with shared identity, to reflect contiguous haplotype blocks where no recombination was estimated. Lastly, we masked any remaining short haplotype blocks as missing data, and ignored in downstream accuracy or GWAS analyses. Using the quality-controlled reconstructions, we extracted corresponding genotype information from a file using TABIX (Li *et al.* 2009), combining each reconstructed genome into a single phased VCF file. Our VCF file contained haplotypes from the DGRP Freeze 2 (http://dgrp2.gnets.ncsu.edu/data/website/dgrp2.vcf), with additional masking of residual heterozygosity and repetitive regions as missing genotypes. See supporting information for additional information on generating or accessing our VCF input.

For any individual reconstructed chromosome, we report the uncorrected diploid recombination count (calculated as the number of haplotype blocks minus two. The expected distributions for diploid recombination count and recombination block length are drawn from forward-simulations conducted by Weller & Bergland (2019, in prep).

To assess the extent of missing data within individual reconstructed genomes that arose due to masking short haplotype blocks, we calculated the number fraction of sites excluded by masking per-chromosome. Similarly, to assess whether certain genomic regions presented systematic difficulties with reconstruction, we calculated missing data rates in sliding windows across each chromosome (10 kb non-overlapping windows), and include the regions occupied by large cosmopolitan inversions as described by Corbett-Detig and Hartl (2012).

Using reconstructed and filtered genomes, we created genetic relatedness matrices (GRM) with the packages *SNPRelate* and *gdsfmt* (Zheng *et al.* 2012) in *R* version 3.5.1 (R Core Team 2016). We included biallelic variants pruned with and LD threshold of 0.2, maximum slide length of 5000 base pairs, while filtering sites with a missing rate below 0.15 or a minor allele frequency below 0.05. We generated a GRM for the whole-

genome; a GRM for each chromosome by itself; and a GRM excluding each chromosome for leave-one-chromosome-out (LOCO) analysis (Yang *et al.* 2014). For example, GWAS of chromosome included a GRM generated from variation on chromosomes 3L, 3R, and X. We generated scaled principal components of the whole-genome GRM using the *prcomp* function in *R*.

Genome-Wide Association Study (GWAS)

To identify loci associated with starvation & desiccation tolerance, we conducted a genome-wide association study. We used the package *GENESIS* (Conomos *et al.* 2019) using *R* version 3.5.1. We performed the GWAS in a case-control framework, modeling phenotype (pre- or post-starvation) as a binomial outcome. We evaluated each chromosome (2, 3, and X) independently, using cage and the corresponding LOCO GRM as covariates. We calculated our significance threshold from permutations, randomly shuffling pre- or post-starvation assignment within cage. We performed 1000 GWAS permutations, taking the 5% quantile of the 1000 genome-wide minimum p-values as our significance threshold.

Mapping RNA reads

We evaluated reference allele mapping bias and the number of reads mapped for *RSubread* (Liao *et al.* 2019) and *iMapSplice* (Liu *et al.* 2018) read aligners. These aligners require different formats of gene annotation files—GTF2.2 format for *RSubread* and UCSC Gene Prediction format for *iMapSplice*. To ensure differences in RNA mapping were not due to differently-sourced gene annotation files, we used a Gene Prediction table from UCSC Table Browser and generated its corresponding GTF file using the UCSC *genePredToGtf* tool.

(http://hgdownload.soe.ucsc.edu/admin/exe/linux.x86_64/)

Because this tool did not collapse gene IDs across various isoforms of single genes, we collapsed gene IDs using a custom script.

We mapped RNA reads to the *Drosophila melanogaster* Release 5.57 reference genome (ftp://ftp.flybase.net/releases/FB2014_03/dmel_r5.57). For *iMapSplice*, we also provided a file containing biallelic SNPs present across the inbred lines used to found the Hybrid Swarm, and sampled a maximum of 5 SNP-mers, with a SNP-mer length of 201. To tabulate reference and alternate read counts, we used the *ASEReadCounter* tool from the *GenomeAnalysisToolKit* version 4.0.0 (Broad Institute 2015). To count total mapped reads per chromosome, we used the *idxstats* tool from *samtools* version 1.9 with *htslib* version 1.9 (Li *et al.* 2009).

Quantifying Reference Allele Mapping Bias

After mapping RNA reads with both *RSubRead* and *iMapSplice*, we calculated the number of mapped and unmapped reads with the *idxstats* tool of *samtools*. To tabulate reference and alternate read counts, we used *ASEReadCounter* of the *GenomeAnalysisToolKit*. Using reconstructed genome estimates (see methods: Genome Reconstructions) and a custom *R* script, we imported read counts and excluded any sites that were not predicted to be heterozygous. Of the heterozygous sites, we excluded any loci that did not overlap within at least one exonic coding sequence range of the same GTF file used to map RNA reads, leaving us with a table of reference and alternate read counts at exonic heterozygous sites, which was also used for quantifying allele-specific expression. After excluding low-count observations (defined as fewer than 15 total mapped reads, or fewer than 3 reference and 3 alternate reads per individual), we calculated reference allele mapping bias as the proportion of reads identified as the reference allele.

Quantifying Differential Expression

We used *DESeq2* to generate feature counts for *iMapSplice* mapped reads. We assigned reads to a corresponding gene specifying the 'LargestOverlap' option. Using the FlyBase web browser tool (http://flybase.org/convert/id) we converted Annotation IDs to FlyBase IDs. In the case where IDs did not convert one-to-one (e.g.

one Annotation ID to multiple FlyBase IDs, or the converse), we selected the ID with the greatest number of feature counts across all samples. We modeled feature counts as a function of cage and treatment (pre- or post-starvation) using the default options of the *DESeq* function, excluding genes with fewer than 50 counts across all samples. *DESeq* results were shrunk with the *lfcShrink* function, specifying the treatment as the coefficient. Prior to principle component analysis, *DESeq* results were variance-stabilizing transformed with the *vst* function. Using these variance-stabilized transformation of expression data, we conducted PCA for expression of all genes, and for specific subsets of gene groups. For PCA of egg coat production gene expression, we included the twelve *D. melanogaster* genes within the gene ontology set GO:0035803. For PCA of Vitelline-membrane genes, we included the six genes resulting from searching "Vitelline" on FlyBase: Vitelline membrane 26Ab (Vm26Ab, FBgn0003980); Vitelline membrane 34Ca (Vm34Ca, FBgn0003983); Vitelline membrane 32E (Vm32E, FBgn0014076); Vitelline membrane 26Aac (Vm 26Ac, FBgn0086266); Vitelline membrane 26Aa (Vm 26Aa, FBgn0003979); Vitelline membrane-like (Vml, FBgn0085362).

Overrepresentation Enrichment Analysis (ORA)

To test for functional enrichment of up- or down-regulated genes, we first selected a base set of genes with a base mean of less than 1.0. We then selected the 5% of genes within the tailed extremes of log2 fold change, and separately compared these sets to our base set. We conducted the Overrepresentation Enrichment Analysis (ORA) via *WebGestalt 2019* (http://www.webgestalt.org/) using the *geneontology* database, reporting the top 50 most significant categories. All other parameters were left at default.

Quantifying Allele-Specific Expression

To assess allele-specific expression, we used the read counts at exonic heterozygous sites as described earlier (see methods: Quantifying Reference Allele Mapping Bias). Because any gene can include read counts at multiple loci, reads at neighboring loci may not be independent. To ensure independence of reads, we selected a single SNP locus per gene, choosing the position with the maximum total mapped reads across all samples. For the remaining genes, we tested for allele-specific expression in a generalized linear model framework. We modeled reference and alternate read counts as a quasibinomial outcome of environment (pre- or post-starvation) to acquire intercept estimates and associated p-values. We converted the intercept p-values from the GLM to false discovery rate (FDR) using the *p.adjust* command in *R*, and designate SNPs with a FDR of less than 0.05 in either treatment as true positives for ASE. The resulting ASE table was then used for both clinal/seasonal intersections, and for tissue-specific expression analysis.

Clinal and Seasonal Intersections with ASE

To determine whether ASE correlates with patterns of seasonality or clinality, we merged our ASE table with data by Machado *et al.* (2018). For this merged data set, we designate SNPs as seasonal and clinal using a p-value threshold of 0.05. We generated an expected distribution of clinal and seasonal intersection by performing 100,000 random draws of size N, where N is the observed number of SNPs exhibiting ASE. We then evaluated enrichment of clinal and seasonal SNPs with ASE.

Tissue-Specific Expression

To evaluate whether tissues exhibit different propensity for ASE, we used expression data from FlyBase

(ftp://flybase.org/flybase/associated_files/Gelbart.2010.10.13.tar.gz).

We assigned each gene to the adult tissue group with greatest mean expression for that gene. After merging tissue assignments with ASE estimates, we calculated the odds ratio for observing ASE in a specific tissue with Fisher's Exact Test.

Results

Whole Genome Reconstructions & GWAS

Diploid recombination count tended to be overestimated in raw reconstruction output, but quality control measures resulted in observations tightly matching expectations from forward-simulations (Figure 1A). Uncorrected haplotype blocks tended to be shorter than expected, with marginal improvement by quality control measures. Masking of short haplotype blocks resulted in minimal data loss, removing less than 5% of informative sites for over 50% of individuals, distributed equally across all chromosomes (Figure 2). There was no increased prevalence of short haplotype blocks within the interiors of or at the boundaries of large cosmopolitan inversions *In(2L)t*; *In(2R)NS*; *In(3L)P*; *In(3R)P*; and *In(1)A* (Figure 3), though there was increased masking frequency at the ends and centromeres of chromosomes.

GWAS of starvation and desiccation tolerance yielded 26 SNPs in 14 clusters that passed our permutation significance threshold (Figure 6, Table 1). Associated variants were spread across chromosomes 2 and 3, with no significant associations on the X chromosome. The most significant GWAS result included a cluster of four SNPs on chromosome 3R. These were 3' of ATP Synthase Subunit D (ATPsynD, FBgn0016120) and 5' of the mitochondrial ribosomal protein L55 (mRpL55, FBgn0038678). A second cluster of SNPs on 3R were approximately 13 kb 5' of ribosomal protein L10Aa. Most SNPs were gene-adjacent, not within coding regions. Variants within protein-coding regions included Dystrophin (FBgn0260003), a Fatty acyl-CoA reductase (FBgn0039131), a Cyclin-dependent kinase (FBgn0031730), and WASp (FBgn0024273).

iMapSplice Reduces Reference Allele Mapping Bias

Diploid organisms can be heterozygous for elements that modulate expression of DNA, such as trans-acting transcription factors or cis-acting enhancers or repressors. While trans-acting elements are expected to bind indiscriminately to either allele and produce equal quantities of transcript, cis-acting elements may result in biased or allele-specific expression, ASE. (Ge *et al.* 2009). In order to attribute RNA read counts to differences in allele expression, it is imperative that the RNA is mapped without bias (Castel *et al.* 2015). One source of mapping bias is introduced when read aligners reject sequences containing variants that differ from the reference genome. Mapping score penalties thus result in reads carrying true polymorphisms being discarded more than reads matching the reference allele. As a result, heterozygous sites will tend to map more reference than alternate reads (Panousis *et al.* 2014; Brandt *et al.* 2015). While mismatches at known variable loci can be ignored, such masking leads to a reduction in mapping quality by ignoring true mismatches, i.e. alleles that are not segregating in the population.

We evaluated the reference-allele mapping bias of two modern RNA sequence read aligners: the *SubRead* aligner as implemented in the *R* package *RSubRead* (Liao *et al.* 2019) and the command-line tool *iMapSplice* (Liu *et al.* 2018), which is an updated version of *MapSplice* aligner with variant-aware mapping (Wang *et al.* 2010). Both *RSubRead* and *iMapSplice* use similar seed-and-vote alignment strategies for splice junction discovery, which is important because a significant fraction of reads—up to 20-30%—span exon junctions (Liu *et al.* 2018; Liao *et al.* 2019). Due to their similar seed-and-vote mapping strategies, we expect differences in reads mapped to be driven by reference allele mapping bias. We observed significant reference allele bias with the *SubRead* aligner that was nearly eliminated by using *iMapsplice*'s variant-aware alignment (Figure 7A). We attribute the asymmetry of reference allele bias to misidentification of heterozygous sites during genome reconstruction. Because the reference allele is designated as the more frequent allele, sites mis-called as heterozygous will tend to be true reference allele homozygotes. The distribution at sites more likely to be truly

heterozygous can be visualized with greater minimum read count thresholds, at the cost of reduced representation.

Differential Expression in Desiccation & Starvation Conditions

Flies in fed conditions exhibited abundant and significant differences in expression patterns when compared to those in starvation & desiccation conditions (Figure 8). Over 65% of the genes passing quality filters were differentially expressed between treatments (Bonferroni-corrected p-value < 0.05). Of genes exhibiting significant differential expression, 97% show decreased expression in the starvation treatment. The genes with the greatest extent of differential expression were vitelline membrane proteins, phospholipase, lysozyomes and proteases. Of the genes upregulated during the starvation treatment, most are associated with nucleic acid biosynthesis or metabolism. See supporting information for a full list of differential gene expression.

Over-representation Enrichment Analysis (ORA) of genes up- and down-regulated during starvation A suite of biological processes differ between treatments (Figure 9). In fed conditions, there is an overrepresentation of genes involved in three main categories: cell cycle regulation (establishment of chromosome location, DNA packaging, chromosome organization and segregation, nuclear division/ mitotic cell cycle); modulation of gene expression (chromatin assembly and disassembly, histone methylation); and egg production (single fertilization, oogenesis, egg coat formation, vitelline membrane formation, chorion-containing eggshell formation). Conversely, in starvation & desiccation conditions, upregulated genes show enrichment for various metabolic processes (disaccharides, amino acids, nucleobases, neurotransmitters, carboxylic acid, ribose phosphate, antibiotics). All enriched gene ontology groups exhibited a false discovery rate of less than 1%.

Genome-wide expression values were analyzed using principal components. Pre- and post-starvation treatments primarily separated along principal component 1 (PC1), which explains 53% of expression variation between individuals. PC2 (18% of variation) primarily separated replicate cages. While most individuals tend to cluster within cage pre-starvation, or as a single group post-starvation, there remains a handful of individuals exhibiting expression patterns from the paired treatment. Separation of treatments by principal components is strengthened to 74% of expression variance along PC1 by subsetting the expression of 12 egg coat production genes (Figure 11). Variance explained for expression is further strengthened to 91% along PC1 for the six vitelline membrane proteins (Figure 12).

Allele-Specific Expression

Our generalized linear model of reference and alternate read counts detected significant, allele biased expression (ASE, i.e. deviation from the expectation of equal representation) for 41.1% (of 4218) genes. Power to detect ASE was greater in the pre-starvation treatment (40.6% of genes with significant ASE) than the post-starvation treatment (1.7% of genes with significant ASE). Because no genes changed direction of expression bias between treatments, we considered ASE genes as any gene with significant ASE in either treatment. In testing whether ASE SNPs are associated with seasonality or clinality, we observed an intersection of 9.48% for ASE in seasonal SNPs and 27.95% for ASE in clinal SNPs (Figure 13), and these intersections did not differ significantly from the expected distributions from permutations ($p = 0.63$ and $p = 0.46$ for seasonal and clinal intersections, respectively). Occurrence of ASE differed depending on a gene's tissue of primary expression (Figure 14). Increased propensity of ASE was detected with high significance for genes primarily expressed in the carcass (63.6% or 75/118 genes, $OR = 2.58,\ p = 8.07 \times 10^{-7}$) and salivary glands (52% or 170/326, $OR = 1.64,\ p = 3.2 \times 10^{-5}$). Similarly, ASE was more highly represented with marginal significance for genes primarily expressed in the fat body (51% or 71/137 genes, $OR = 1.57, p = 0.01$) and the heart (52% or 61/117 genes, $OR = 1.58, p =$

0.017). The ovary was the only tissue with reduced ASE representation (30.9% or 243/786 genes, $OR = 0.56, p = 1.85 \times 10^{-11}$). Propensity for ASE did not significantly differ from the overall rate of 41% for the remaining tissue categories.

Discussion

We applied the Hybrid Swarm genome reconstruction method to inexpensively genotype a *Drosophila melanogaster* mapping population subjected to starvation and desiccation conditions. By coupling reconstructed genomes with RNA expression data for nearly 700 individuals, we were able to evaluate the genetic basis of a complex life history trait, differential expression, and allele-specific expression. Our genome-wide association study (GWAS) detected a number of variants passing a conservative permutation-based significance threshold, including a cluster of SNPs adjacent to two mitochondrial proteins which have been implicated in lifespan extension via interactions with the Target of Rapamycin (TOR) pathway and proteostasis. Differential expression analysis for starvation and desiccation conditions indicates a downregulation of oogenesis and protease genes, with increased expression of various metabolic genes, chiefly those involved in purine biosynthesis. We observed variable propensity of allele-specific expression (ASE) across tissue classes, with increased representation of ASE in genes expressed in the carcass, salivary glands, heart, and fat body—and a reduction of ASE for genes primarily expressed in the ovary. Our results suggest that survival in desiccation and starvation conditions, which is negatively correlated with reproductive effort, may not be due to genetic variation within reproductive genes themselves.

GWAS of Starvation and Desiccation Tolerance implicates Mitochondrial Proteins

The most significant GWAS hit on chromosome 3R is 3' of mitochondrial ATP synthase subunit D (ATPsynD) and 5' of mitochondrial ribosomal protein L55 (mRpL55), and this cluster the highest of all 1,000 of our permutations (p=0.001). A

second cluster of SNPs on 3R are adjacent to Ribosomal protein L10Aa. Mitochondrial genes have commonly been implicated in extension of lifespan and stress tolerance. Lifespan extension in yeast, flies, worms, and mice is thought to occur when fewer reactive oxygen species (ROS) are produced due to dietary restriction or mutations that downregulate the TOR pathway (reviewed by Fontana *et al.* 2010), which was first implicated in *C. elegans* by Vellai *et al.* (2003). Inhibition of mitochondrial respiration improves lifespan in *C. elegans* (Kayser *et al.* 2004; Cristina *et al.* 2009) and *D. melanogaster* (Copeland *et al.* 2009; Linford and Pletcher 2009), though *D. melanogaster* life extension may require inhibition of mitochondrial respiration during both development and adulthood (Rera *et al.* 2010). RNAi knockdown of ATPsynD, 5' of our cluster of most significant GWAS hits, has been shown to reduce oxidative damage and extend lifespan in *D. melanogaster* via interactions with the TOR signaling pathway (Sun *et al.* 2014). Mitochondrial ribosome mutants have been implicated in extension of lifespan (Heeren *et al.* 2009), potentially due to decreased expression maintaining proteostasis (Steffen and Dillin 2016). Both ATPsynD and mRpL55 show significantly lower expression in the post-starvation treatment, with respective reductions of 16% and 66%. The discovery of a genetic association between desiccation and starvation survival with mitochondrial proteins, which are commonly known to influence lifespan, is a promising result for the Hybrid Swarm method of association mapping. Follow-up eQTL analysis can shed light on whether the observed variants are causally related to differences in expression.

Starvation and Desiccation Regulate Reproduction & Metabolism

Our starvation and desiccation treatment profoundly influenced gene expression patterns, resulting in significantly reduced expression for roughly two-thirds of genes in the genome. Reproductive genes involved in egg production, e.g. chorion and vitelline membrane proteins, have been previously implicated for differential expression during starvation (Matzkin and Markow 2009). These vitelline and chorion protein genes display the strongest signature of differential expression in our analysis, with expression reduced by 99% in the starvation treatment. The reduction in vitelline

membrane expression is so drastic that pre- and post-starvation conditions can be delineated by PC1 of the vitelline gene expression matrix, which accounts for 91% of expression variance (Figure 12). Interestingly, the pre-starvation individuals are relatively spread along PC1, while post-starvation individuals are tightly clustered. This may suggest that PC1 of vitelline gene expression reflects the extent of reproductive effort that various individuals are undergoing, which is expected to have some degree of variance when fed, but little variance in the starvation treatment.

ASE Interacts with Starvation Treatment

Environment-specific ASE could act as a mechanism of reversing dominance and subsequently maintaining genetic variation. For example, consider a locus which codes for a summer-favored allele and a spring-favored allele. If this locus exhibited environment-specific ASE, a heterozygote could produce the beneficial allele regardless of environment. Although our model did not identify any genes with significant signals for reversal of dominance, we suspect this is due to limited power to detect ASE in the post-starvation treatment. We detected an abundance of ASE in the pre-starvation treatment (approximately 41% or 1384/3385 genes), yet only 1.2% of genes (42/3385) in the post-starvation treatment deviated from equal expression at a FDR of 5%. The post-starvation treatment suffered from reduced sample sizes of genes with adequate read depth, likely resulting from genome-wide downregulation of most genes. While there is no *a priori* reason to expect ASE to be more or less prevalent in the pre-starvation treatment, we cannot rule out this possibility—e.g., allelic gene expression may be less sensitive in well-fed conditions or more sensitive to starvation and desiccation conditions.

ASE Varies by Tissue, but not Seasonally or Clinally

While we did not detect any enrichment of ASE for SNPs previously characterized as seasonal or clinal, we did detect differences in proportions of genes exhibiting ASE for different tissues, where genes primarily expressed in the ovaries

were depleted for ASE, and genes primarily expressed in the carcass, salivary glands, heart, and fat body were enriched for ASE.

For ASE to exist, there must be genetic variation in a cis-acting element to modulate expression. For us to detect ASE, and for that ASE to have phenotypic consequences, there must also be allelic differences in the resulting transcript. Accordingly, failure to detect ASE may a signal of reduced genetic variation at one or both of these levels. A biological interpretation of this observation could be strong purifying selection for genes expressed in the ovaries (Story and Steitz 1992), which is consistent with prior work. Langley *et al.* (2012) showed that male and female reproduction genes are present within "diversity valleys" associated with adaptive protein evolution, and Panhuis and Swanson (2006) measured *dN/dS* ratios less than 1 for candidate female *D. melanogaster* reproductive genes. Conversely, genes primarily expressed in the carapace and salivary glands (and to a lesser extent, the heart and fat body) exhibited an overabundance of ASE. PK1-R, a receptor for *capa* neuropeptides that is highly expressed in the carapace and salivary glands, has been implicated along with anti-diuretic peptides in tolerance to desiccation via water balance homeostasis (Davies *et al.* 2012; Terhzaz *et al.* 2015). Langley *et al.* (2012) suggest that fungal host-pathogen interactions could explain a common nonsynonymous mutation shared by *D. melanogaster* and *D. simulans* within the chitin metabolism gene *Muc11A*. Other host-pathogen interactions could be involved in the various genes expressed in the carapace, a primary point of entry for pathogens.

The Hybrid Swarm Method to Genetic Association Mapping

This work is the first assessment of bias-free RNA sequence mapping using *iMapSplice* for analysis of allele-specific expression. This work also constitutes one of the first GWAS using a Hybrid Swarm population. The discovery of associations with mitochondrial proteins and survival, which have been previously implicated in stress tolerance and lifespan extension, lends credence to the power of the Hybrid Swarm

method of association studies in a randomly outbred population. By using the Hybrid Swarm method, we were able to generate and genotype a mapping population of nearly 700 individuals. Because larger sample sizes are required to detect variants that contribute small effects or segregate at low frequencies, such an experiment may not be feasible with whole genome sequencing at typical levels of coverage (~10-15X).

Without high coverage sequencing, we acknowledge that the underlying genotypes of our mapping population can only be inferred to some degree of accuracy. To ensure confidence in genotype estimates, we performed quality control of reconstructed genomes by bridged or masked short haplotype blocks (defined as less than one megabase in length, see methods). These short haplotype blocks are likely to arise when RABBIT is unable to impute the correct diplotype (Weller & Bergland 2019, in prep), thus the distribution of post-masking data missingness can serve as a proxy for reconstruction accuracy. Because masking resulted in less than 5% of genotypes across all samples, with a most frequent bin of 0% masking, we are confident that reconstructed genomes accurately reflect the underlying genotypes. Resultingly, we expect most genotyping errors to arise from imperfect detection of recombination breakpoints. For researchers who require even greater confidence in genotype estimates, sites adjacent to imputed recombination breakpoints (e.g. $\pm$ 100 kb) could also be masked.

Contributions

The project was led by PS and AOB. Paul Schmidt, Dmitri Petrov, and Alan O. Bergland are credited for project concept and design. Subhash Rajpurohit and Susanne Tilk generated the data. Data analysis and manuscript writing was done by myself and AOB.

Literature Cited

Abzhanov A., W. P. Kuo, C. Hartmann, B. R. Grant, P. R. Grant, *et al.*, 2006 The calmodulin pathway and evolution of elongated beak morphology in Darwin's finches. Nature 442: 563–567. https://doi.org/10.1038/nature04843

Andolfatto P., 2005 Adaptive evolution of non-coding DNA in Drosophila. Nature 437: 1149–52. https://doi.org/10.1038/nature04107

Baym M., S. Kryazhimskiy, T. D. Lieberman, H. Chung, M. M. Desai, *et al.*, 2015 Inexpensive multiplexed library preparation for megabase-sized genomes. PLoS One 10: 1–15. https://doi.org/10.1371/journal.pone.0128036

Bellard C., C. Bertelsmeier, P. Leadley, W. Thuiller, and F. Courchamp, 2012 Impacts of climate change on the future of biodiversity. Ecol. Lett. 15: 365–377. https://doi.org/10.1111/j.1461-0248.2011.01736.x

Brandt D. Y. C., V. R. C. Aguiar, B. D. Bitarello, K. Nunes, J. Goudet, *et al.*, 2015 Mapping Bias Overestimates Reference Allele Frequencies at the *HLA* Genes in the 1000 Genomes Project Phase I Data. G3: Genes|Genomes|Genetics 5: 931–941. https://doi.org/10.1534/g3.114.015784

Broad Institute, 2015 Genome Analysis Toolkit: Variant Discovery in High-Throughput Sequencing Data

Castel S. E., A. Levy-Moonshine, P. Mohammadi, E. Banks, and T. Lappalainen, 2015 Tools and best practices for data processing in allelic expression analysis. Genome Biol. 16: 1–12. https://doi.org/10.1186/s13059-015-0762-6

Conomos M., S. Gogarten, L. Brown, H. Chen, K. Rice, *et al.*, 2019 GENetic EStimation and Inference in Structured samples (GENESIS): Statistical methods for analyzing genetic data from samples with population structure and/or relatedness.

Copeland J. M., J. Cho, T. Lo, J. H. Hur, S. Bahadorani, *et al.*, 2009 Extension of Drosophila Life Span by RNAi of the Mitochondrial Respiratory Chain. Curr. Biol. 19: 1591–1598. https://doi.org/10.1016/j.cub.2009.08.016

Corbett-Detig R. B., and D. L. Hartl, 2012 Population Genomics of Inversion Polymorphisms in Drosophila melanogaster. PLoS Genet. 8. https://doi.org/10.1371/journal.pgen.1003056

Crawford D. L., and D. A. Powers, 1992 Evolutionary adaptation to different thermal environments via transcriptional regulation. Mol. Biol. Evol. 9: 806–13. https://doi.org/10.1093/oxfordjournals.molbev.a040762

Cristina D., M. Cary, A. Lunceford, C. Clarke, and C. Kenyon, 2009 A regulated response to impaired respiration slows behavioral rates and increases lifespan in Caenorhabditis elegans. PLoS Genet. 5. https://doi.org/10.1371/journal.pgen.1000450

Davies S. A., P. Cabrero, M. Povsic, N. R. Johnston, S. Terhzaz, *et al.*, 2012 Signaling by Drosophila capa neuropeptides. Gen. Comp. Endocrinol. 188: 60–66. https://doi.org/10.1016/j.ygcen.2013.03.012

Fontana L., L. Partridge, and V. D. Longo, 2010 Extending healthy life span-from yeast to humans. Science (80-.). 328: 321–326. https://doi.org/10.1126/science.1172539

Ge B., D. K. Pokholok, T. Kwan, E. Grundberg, L. Morcos, *et al.*, 2009 Global patterns of cis variation in human cells revealed by high-density allelic expression analysis. Nat. Genet. 41: 1216–1222. https://doi.org/10.1038/ng.473

Heeren G., M. Rinnerthaler, P. Laun, P. von Seyerl, S. Kössler, *et al.*, 2009 The mitochondrial ribosomal protein of the large subunit, Afo1p, determines cellular longevity through mitochondrial back-signaling via TOR1. Aging (Albany. NY). 1: 622–636.

Kayser E. B., M. M. Sedensky, P. G. Morgan, and C. L. Hoppel, 2004 Mitochondrial oxidative phosphorylation is defective in the long-lived mutant clk-1. J. Biol. Chem. 279: 54479–54486. https://doi.org/10.1074/jbc.M403066200

Kessner D., T. L. Turner, and J. Novembre, 2013 Maximum likelihood estimation of frequencies of known haplotypes from pooled sequence data. Mol. Biol. Evol. 30: 1145–58. https://doi.org/10.1093/molbev/mst016

King M.-C., and A. C. Wilson, 1975 Evolution at Two Levels in Humans and Chimpanzees. Science (80-.). 188: 107–116. https://doi.org/10.1038/098448b0

Kopf M., S. Klähn, I. Scholz, W. R. Hess, and B. Voß, 2015 Variations in the non-coding transcriptome as a driver of inter-strain divergence and physiological adaptation in bacteria. Sci. Rep. 5. https://doi.org/10.1038/srep09560

Langley C. H., K. Stevens, C. Cardeno, Y. C. G. Lee, D. R. Schrider, *et al.*, 2012 Genomic variation in natural populations of Drosophila melanogaster. Genetics 192: 533–598. https://doi.org/10.1534/genetics.112.142018

Li H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760. https://doi.org/10.1093/bioinformatics/btp324

Li H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The Sequence

Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liao Y., G. K. Smyth, and W. Shi, 2019 The R package *Rsubread* is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res. 47. https://doi.org/10.1093/nar/gkz114

Linford N. J., and S. D. Pletcher, 2009 Aging: Fruit Flies Break the Chain to a Longer Life. Curr. Biol. 19: R895–R898. https://doi.org/10.1016/j.cub.2009.08.050

Liu X., J. N. MacLeod, and J. Liu, 2018 IMapSplice: Alleviating reference bias through personalized RNA-seq alignment. PLoS One 13: 1–14. https://doi.org/10.1371/journal.pone.0201554

Machado H., A. O. Bergland, R. Taylor, S. Tilk, E. Behrman, *et al.*, 2018 Broad geographic sampling reveals predictable and pervasive seasonal adaptation in *Drosophila*. bioRxiv 337543. https://doi.org/10.1101/337543

Mack K. L., M. A. Ballinger, M. Phifer-Rixey, and M. W. Nachman, 2018 Gene regulation underlies environmental adaptation in house mice. Genome Res. 28: 1636–1645. https://doi.org/10.1101/gr.238998.118

Mackay T. F. C., S. Richards, E. A. Stone, A. Barbadilla, J. F. Ayroles, *et al.*, 2012 The Drosophila melanogaster Genetic Reference Panel. Nature 482: 173–178. https://doi.org/10.1038/nature10811

MacMillan H. A., and B. J. Sinclair, 2011 Mechanisms underlying insect chill-coma. J. Insect Physiol. 57: 12–20. https://doi.org/10.1016/j.jinsphys.2010.10.004

Matzkin L. M., and T. A. Markow, 2009 Transcriptional regulation of metabolism associated with the increased desiccation resistance of the cactophilic Drosophila mojavensis. Genetics 182: 1279–1288. https://doi.org/10.1534/genetics.109.104927

Nourmohammad A., J. Rambeau, T. Held, V. Kovacova, J. Berg, *et al.*, 2017 Adaptive Evolution of Gene Expression in Drosophila. Cell Rep. 20: 1385–1395. https://doi.org/10.1016/j.celrep.2017.07.033

Panhuis T. M., and W. J. Swanson, 2006 Molecular evolution and population genetic analysis of candidate female reproductive genes in Drosophila. Genetics 173: 2039–2047. https://doi.org/10.1534/genetics.105.053611

Panousis N. I., M. Gutierrez-Arcelus, E. T. Dermitzakis, and T. Lappalainen, 2014 Allelic mapping bias in RNA-sequencing is not a major confounder in eQTL studies. Genome Biol. 15: 467. https://doi.org/10.1186/s13059-014-0467-2

R Core Team, 2016 R: A Language and Environment for Statistical Computing

Rera M., V. Monnier, and H. Tricoire, 2010 Mitochondrial electron transport chain dysfunction during development does not extend lifespan in Drosophila melanogaster. Mech. Ageing Dev. 131: 156–164. https://doi.org/10.1016/j.mad.2010.01.004

Savolainen O., M. Lascoux, and J. Merilä, 2013 Ecological genomics of local adaptation. Nat. Rev. Genet. 14: 807–820. https://doi.org/10.1038/nrg3522

Steffen K. K., and A. Dillin, 2016 A Ribosomal Perspective on Proteostasis and Aging. Cell Metab. 23: 1004–1012. https://doi.org/10.1016/j.cmet.2016.05.013

Story R. M., and T. A. Steitz, 1992 Levels of natrually occurring DNA polymorphism correlate with recombination rates in D. melanogaster. Nature 355: 242–244. https://doi.org/10.1038/355242a0

Sun X., C. T. Wheeler, J. Yolitz, M. Laslo, T. Alberico, *et al.*, 2014 A mitochondrial ATP synthase subunit interacts with TOR signaling to modulate protein homeostasis and lifespan in Drosophila. Cell Rep. 8: 1781–1792. https://doi.org/10.1016/j.celrep.2014.08.022

Terhzaz S., N. M. Teets, P. Cabrero, L. Henderson, M. G. Ritchie, *et al.*, 2015 Insect capa neuropeptides impact desiccation and cold tolerance. Proc. Natl. Acad. Sci. 112: 2882–2887. https://doi.org/10.1073/pnas.1501518112

Vellai T., K. Takacs-Vellai, Y. Zhang, A. L. Kovacs, L. Orosz, *et al.*, 2003 Influence of TOR kinase on lifespan in C. elegans. Nature 426: 620–620. https://doi.org/10.1038/426620a

Wang K., D. Singh, Z. Zeng, S. J. Coleman, Y. Huang, *et al.*, 2010 MapSplice: accurate mapping of RNA-seq reads for splice junction discovery. Nucleic Acids Res. 38: e178. https://doi.org/10.1093/nar/gkq622

Yang J., N. A. Zaitlen, M. E. Goddard, P. M. Visscher, and A. L. Price, 2014 Advantages and pitfalls in the application of mixed-model association methods. Nat. Genet. 46: 100–106. https://doi.org/10.1038/ng.2876

Zheng X., D. Levine, J. Shen, S. M. Gogarten, C. Laurie, *et al.*, 2012 A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28: 3326–3328. https://doi.org/10.1093/bioinformatics/bts606

Tables & Figures

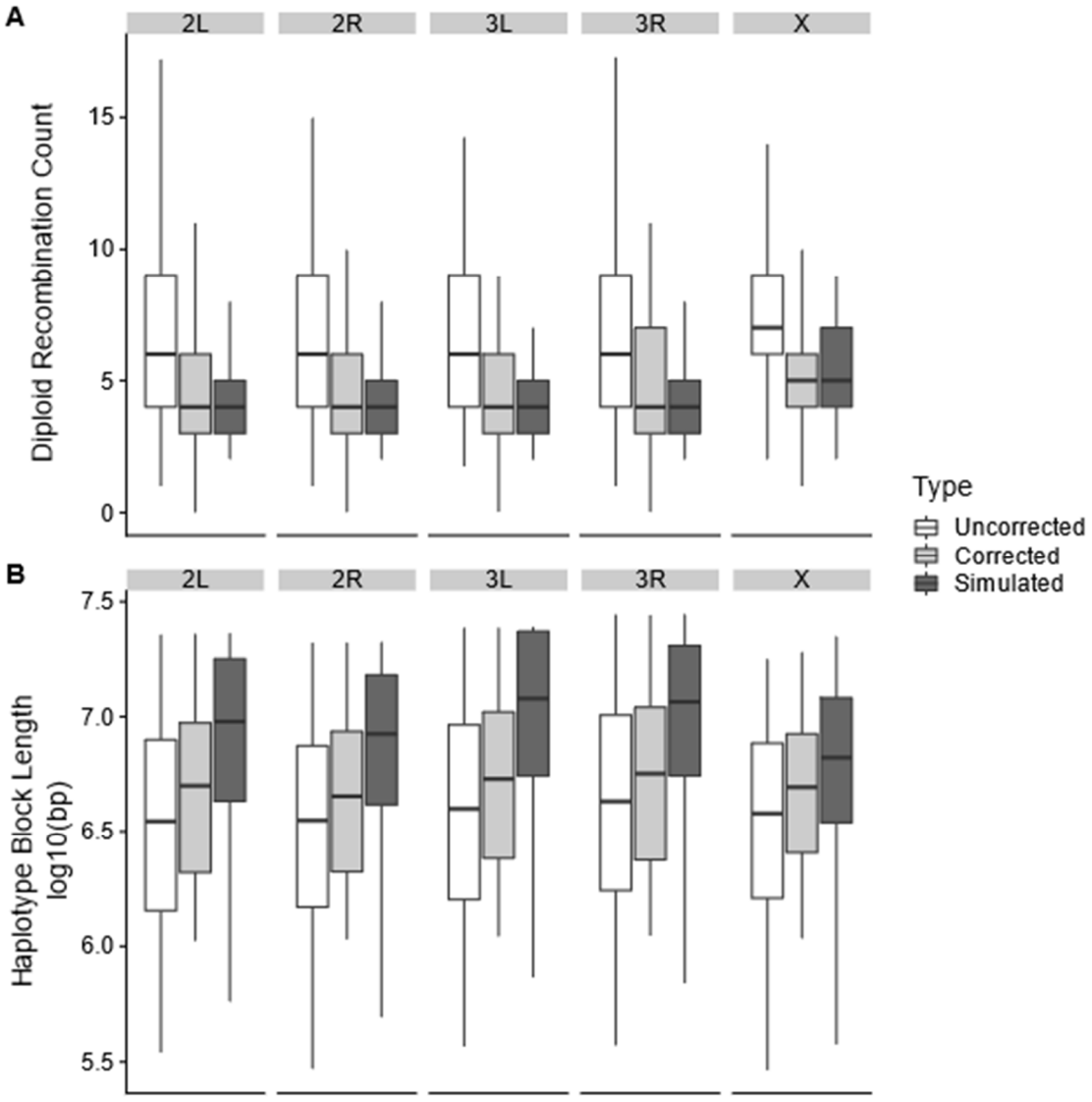


Figure 1. Metrics of reconstructed genomes before and after quality control corrections, as compared to expectations from simulated data. The observed number of recombination events in corrected genome reconstructions closely reflect that from simulated expectations (A). Recombination block length is partially improved by quality control measures (B). Boxes represent interquartile range, with whiskers extending to 95% of the range.

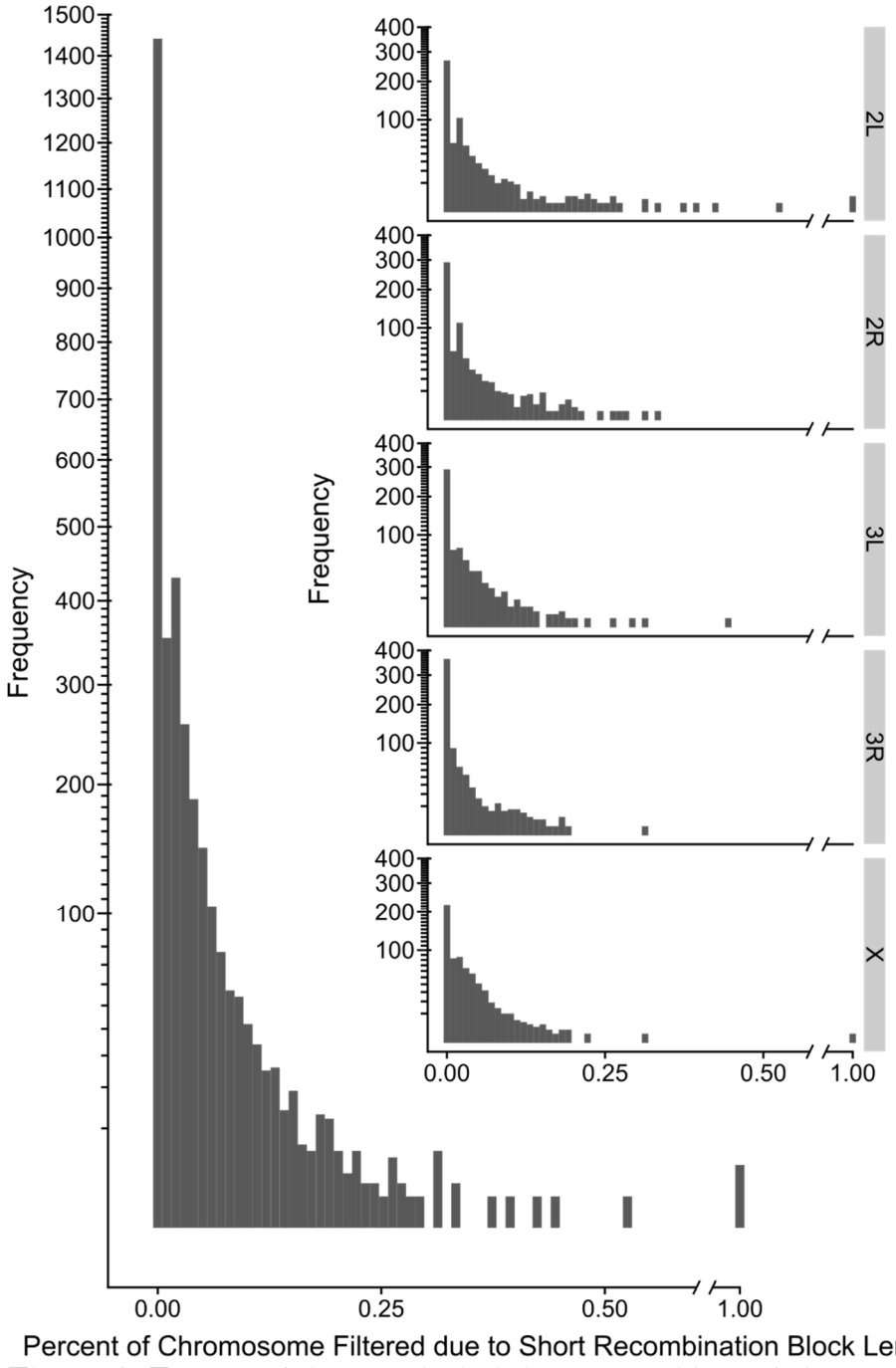


Figure 2. Extent of data excluded due to masking of short (< 1 Mb) estimated recombination blocks, as a percent of nucleotides. Frequencies are shown on a square-root transformed axis.

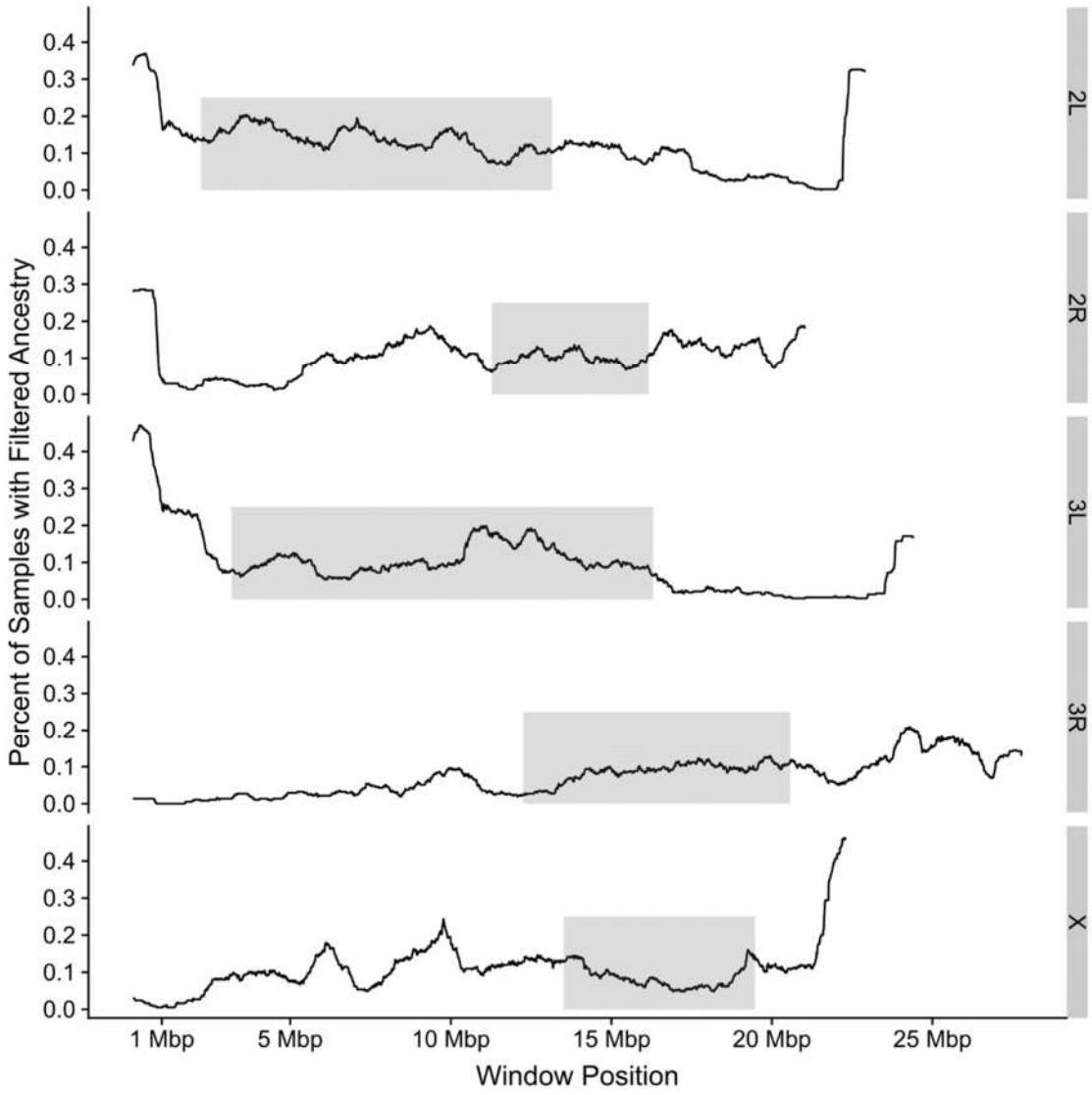


Figure 3. Position-specific extent of data exclusion due to masking of short (< 1 Mb) recombination block. Values represent the percent of samples with at least one filtered haplotype in 10 kb non-overlapping windows for each chromosome. Gray boxes represent locations of large cosmopolitan inversions *In(2L)t*; *In(2R)NS*; *In(3L)P*; *In(3R)P*; and *In(1)A* (on chromosome X).

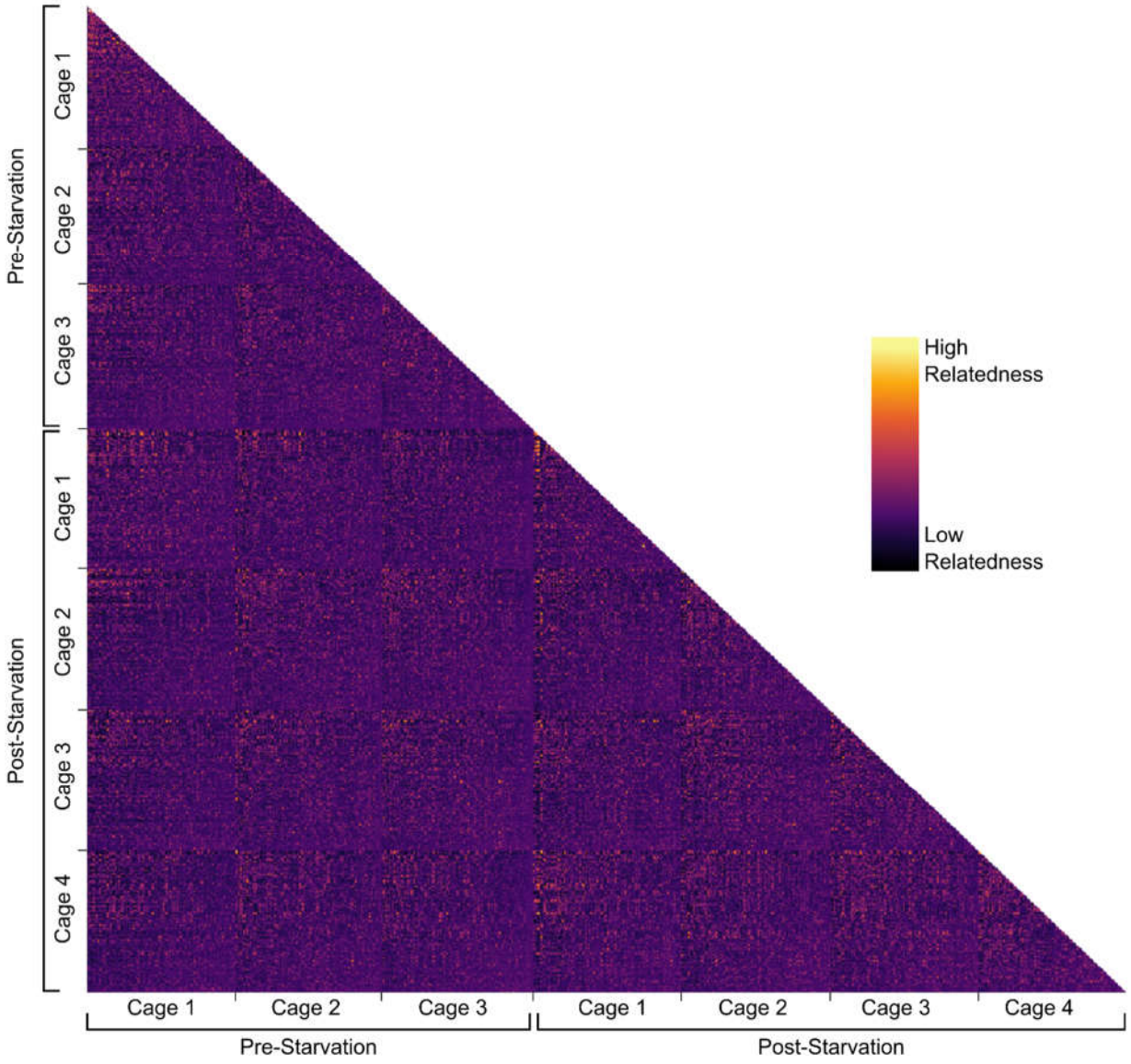


Figure 4. Genetic relatedness matrix for 672 reconstructed genomes. Comparisons to one's self, for which relatedness should be 100%, are excluded.

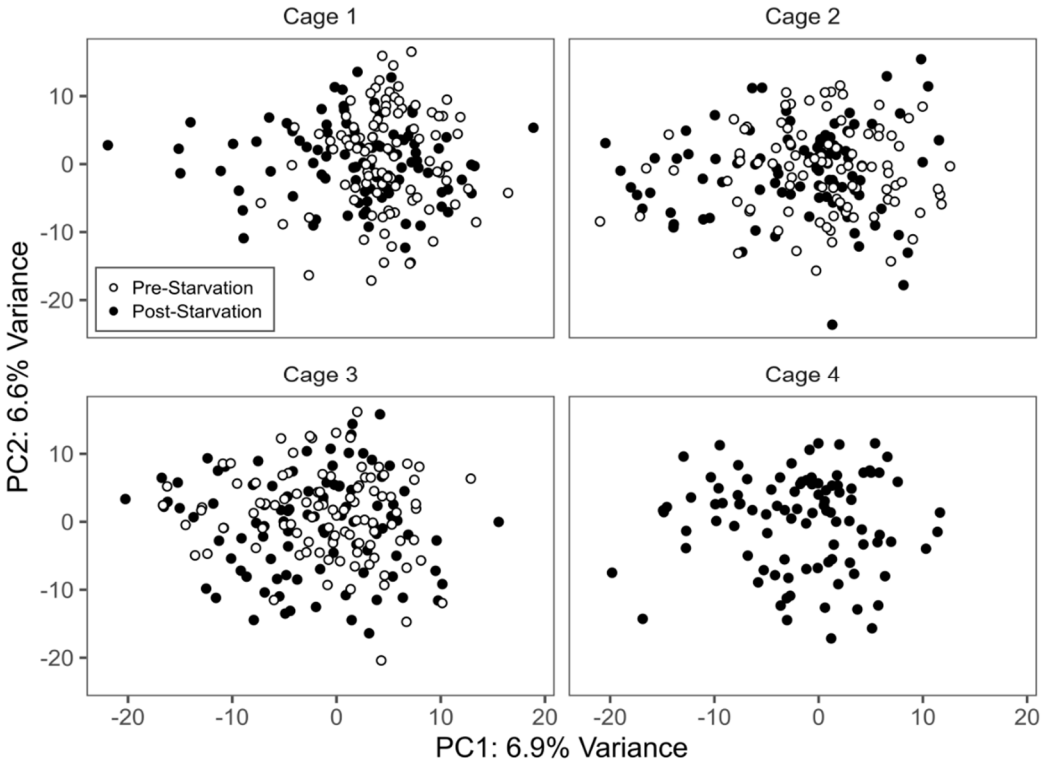


Figure 5. Transformation of the genetic relatedness matrix to its first two principal components. Note that there is no pre-starvation data for cage 4 due to sequencing failure.

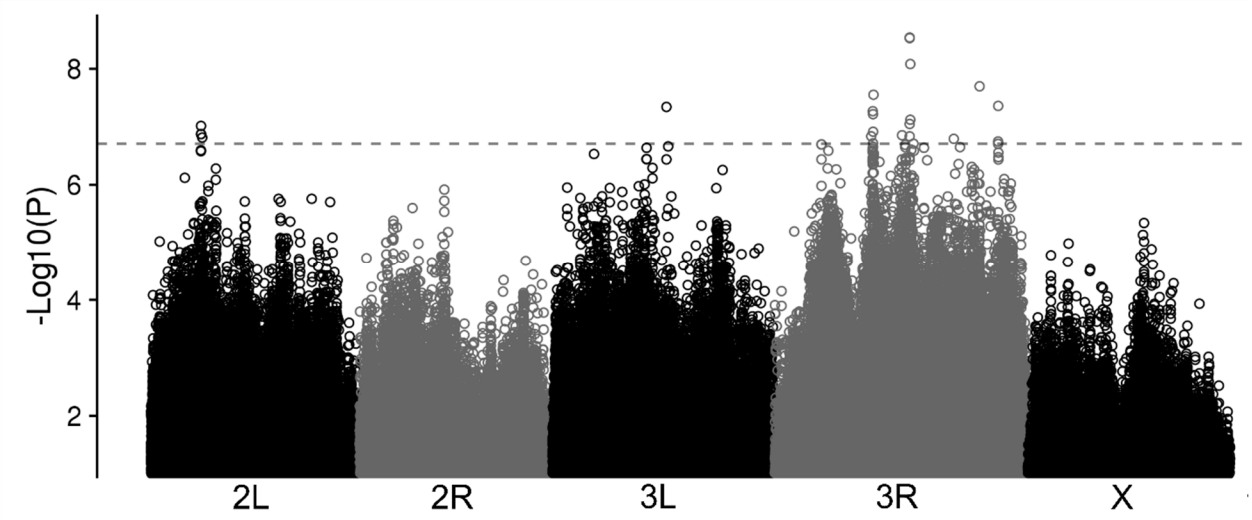


Figure 6. Manhattan plot of Genome Wide Association Study (GWAS) for survival in starvation & desiccation conditions. Dashed line represents our significance threshold, calculated as the 5% quantile of most-significant P-values from 1,000 permutations where treatment was randomly assigned within each cage. The loci with the most significant p-values include a cluster of 4 SNPs on chromosome 3R, 3' adjacent to ATP Synthase Subunit D.

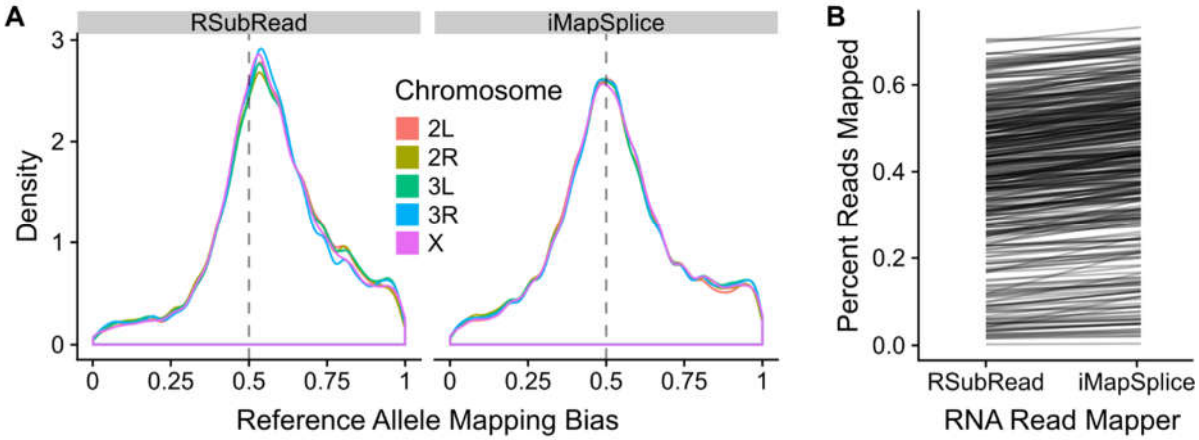


Figure 7. Mapping of RNA data is improved by using iMapSplice. The *RSubRead* mapper resulted in reference allele bias, with density centered above 50% (A). Reference allele bias was not present when mapping RNA reads with *iMapSplice*. The *iMapSplice* mapper also resulted in fewer unmapped RNA reads (B). Both mappers used a gene annotation file modified from the same source file.

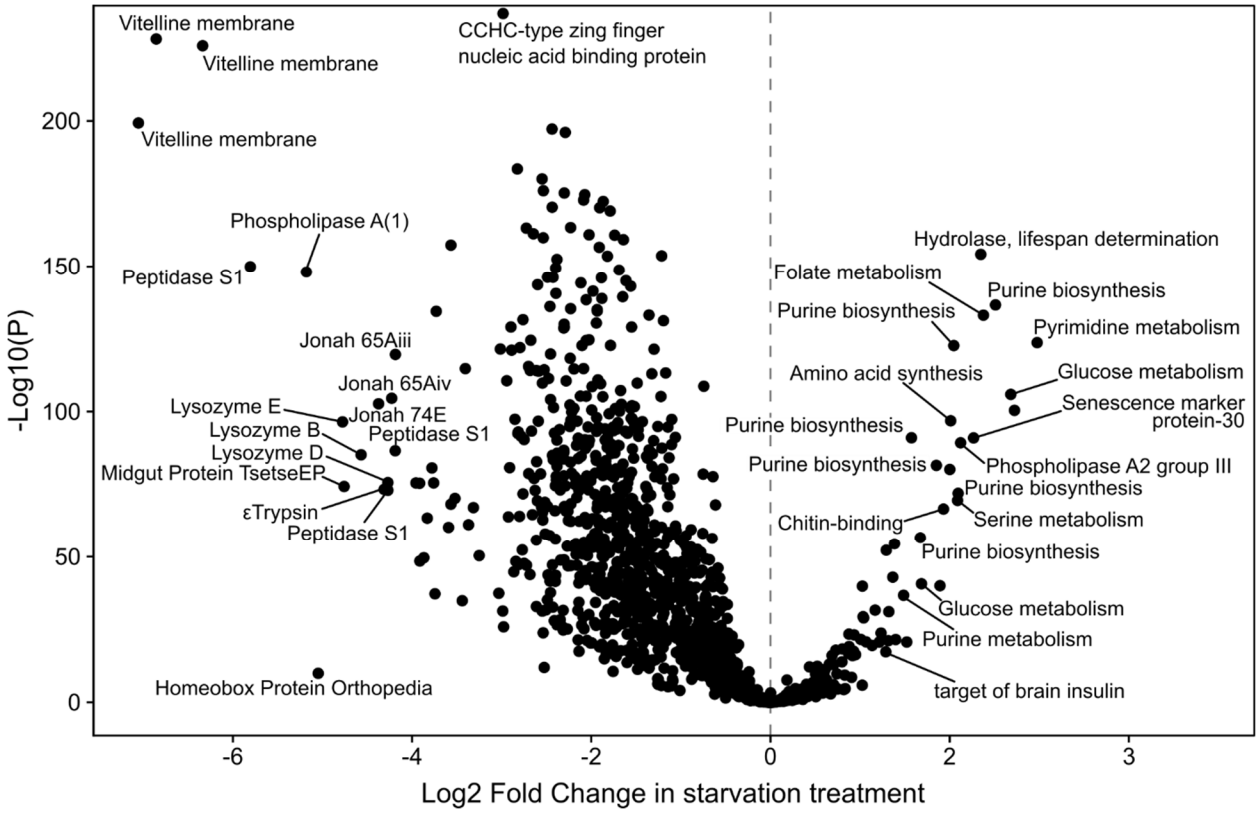


Figure 8. Differential expression between flies sampled before and after starvation & desiccation conditions. Over 65% of genes pass the Bonferroni-corrected significance threshold for differential expression between groups. We excluded more lowly-expressed genes, i.e. those determined by DESeq2 to have base mean of less than 50.

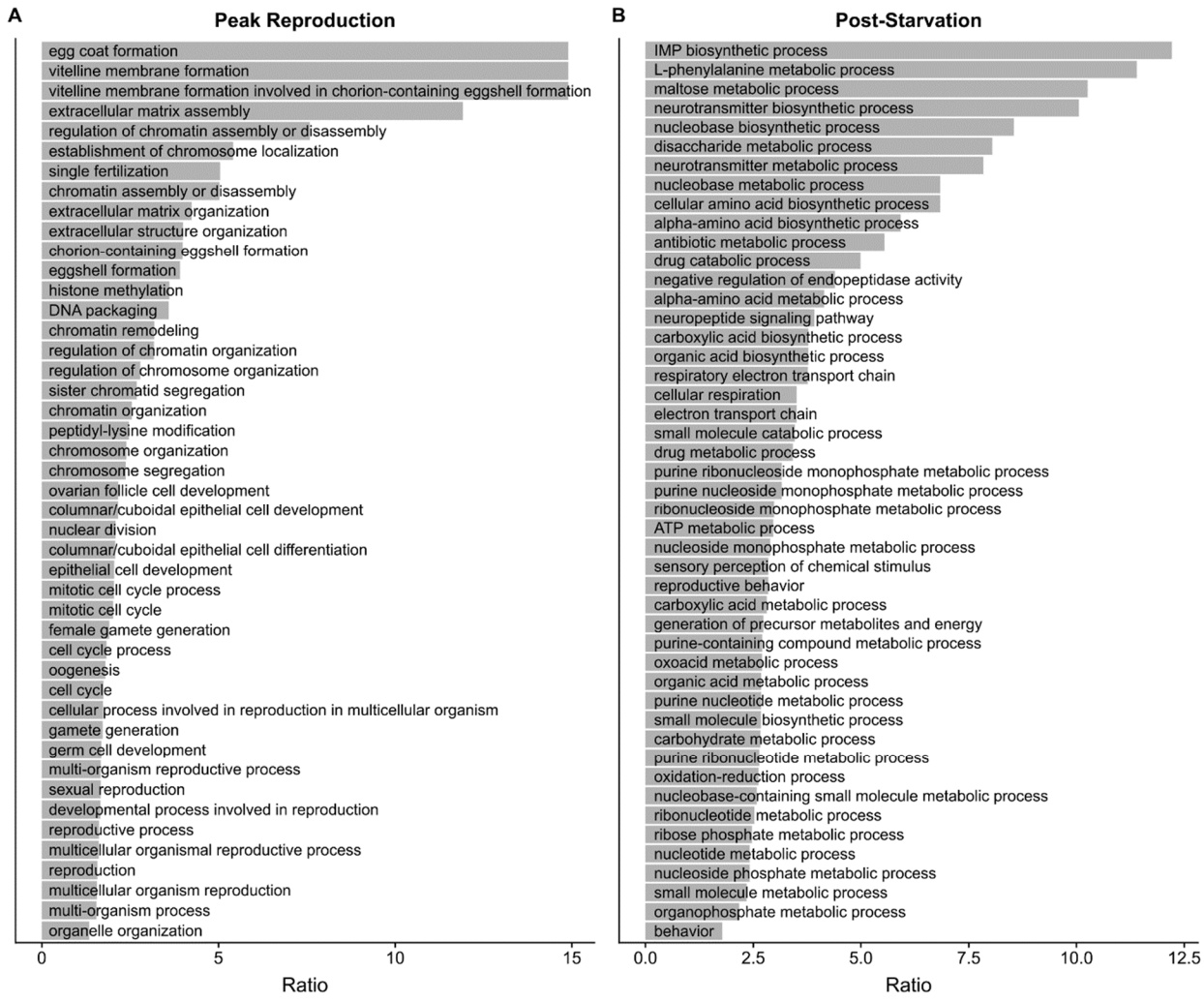


Figure 9. Over-representation Enrichment Analysis (ORA) for differentially-expressed genes. Our base set includes protein-coding genes with a base mean of expression > 1. Our genes of interest included either the 5% of our base genes most downregulated (left) or upregulated (right) in starvation & desiccation conditions. All gene ontology groups exhibit a FDR of less than 1%.

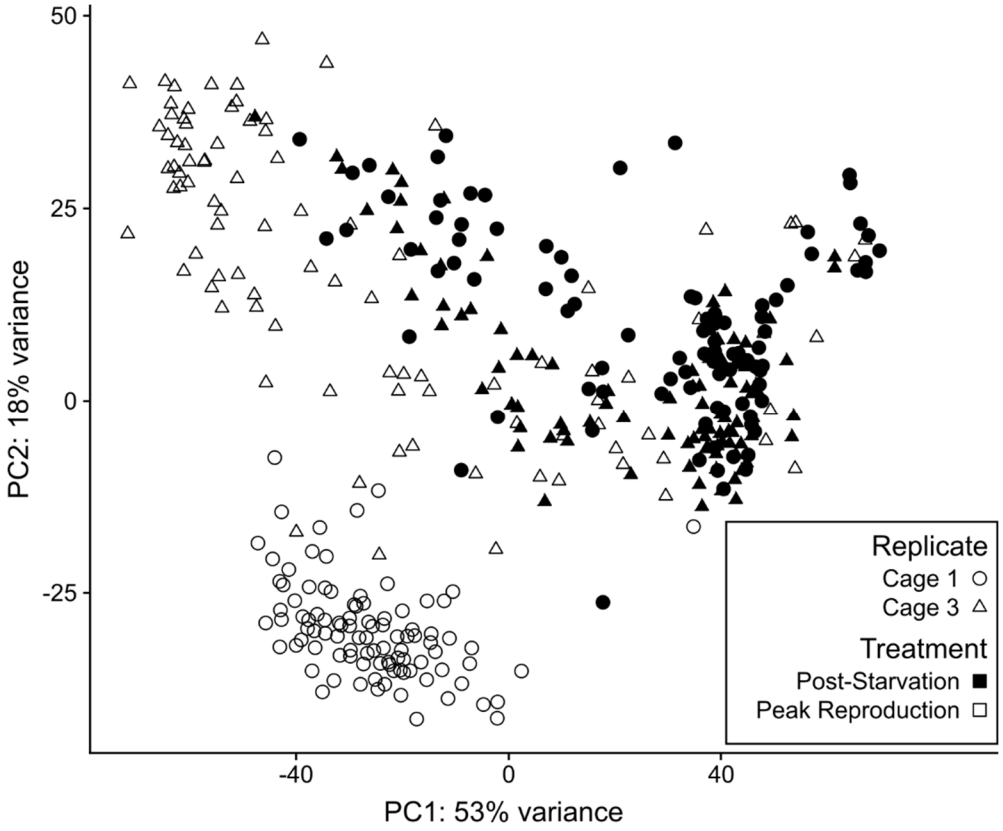


Figure 10. Transformation of the gene expression matrix into its first two principal components. PC1 primarily separates pre- and post-starvation treatments, explaining 53% of variation in expression. PC2 primarily separates the two replicates cages, explaining 18% of variation in expression.

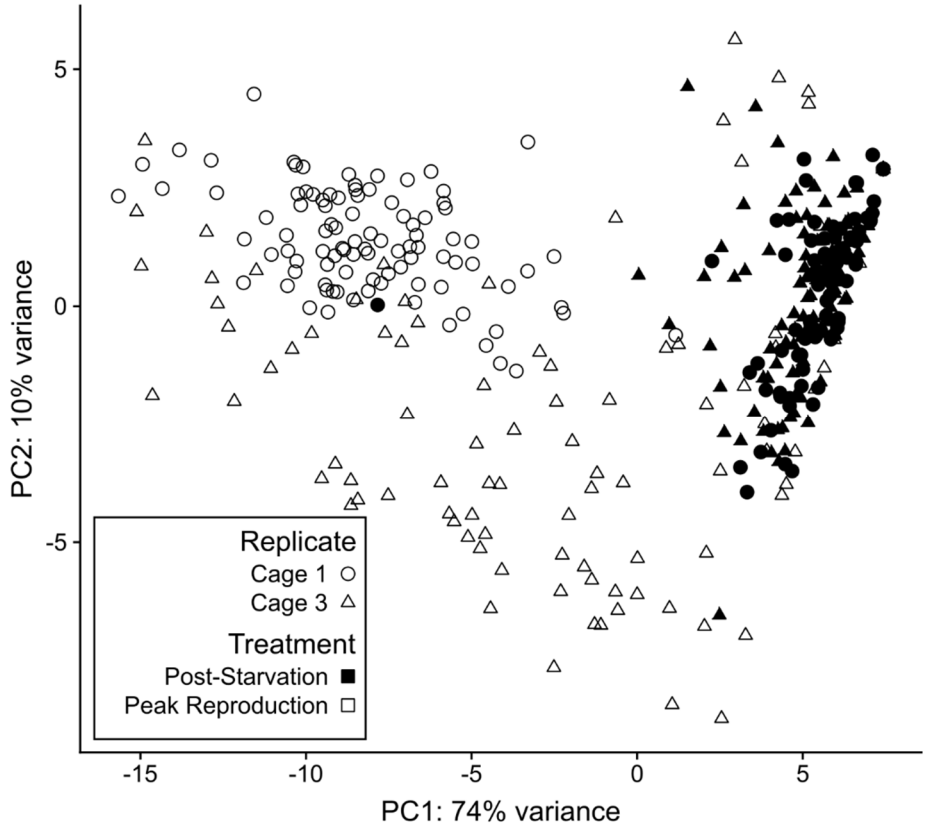


Figure 11. Principal components plot for expression of 12 *Drosophila melanogaster* Egg Coat Production genes. This includes genes from the gene ontology set GO:0035803

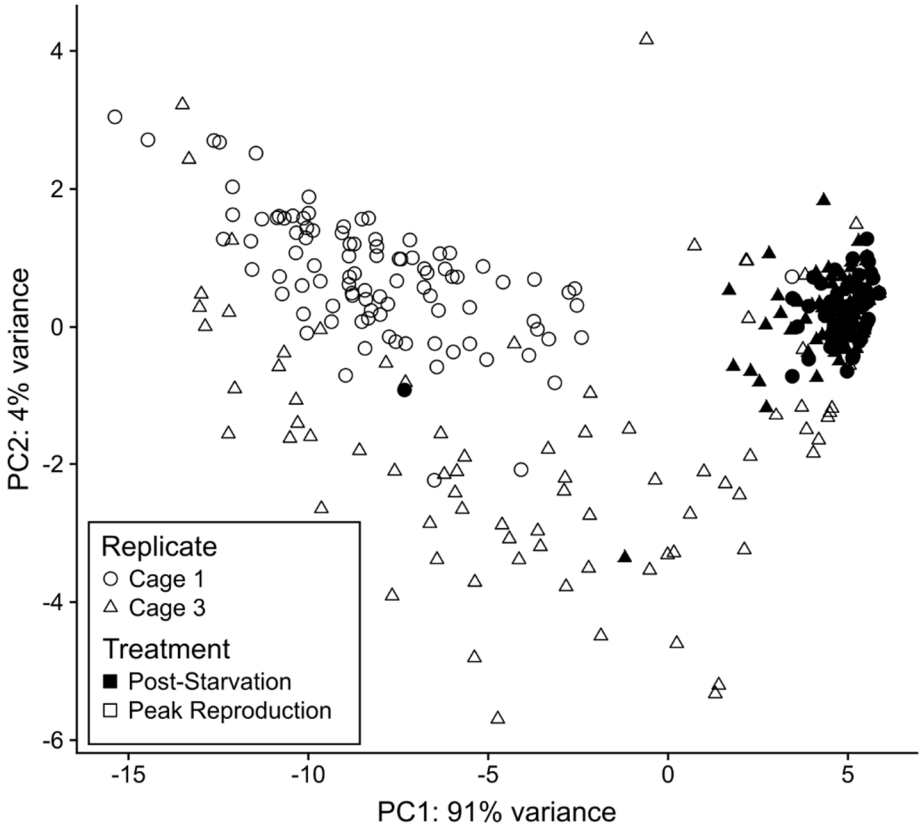


Figure 12. Principal components plot for expression of six Vitelline Membrane proteins. This includes the genes Vm26Ab, Vm34Ca, Vm32E, Vm26Ac, Vm26Aa, and Vml.

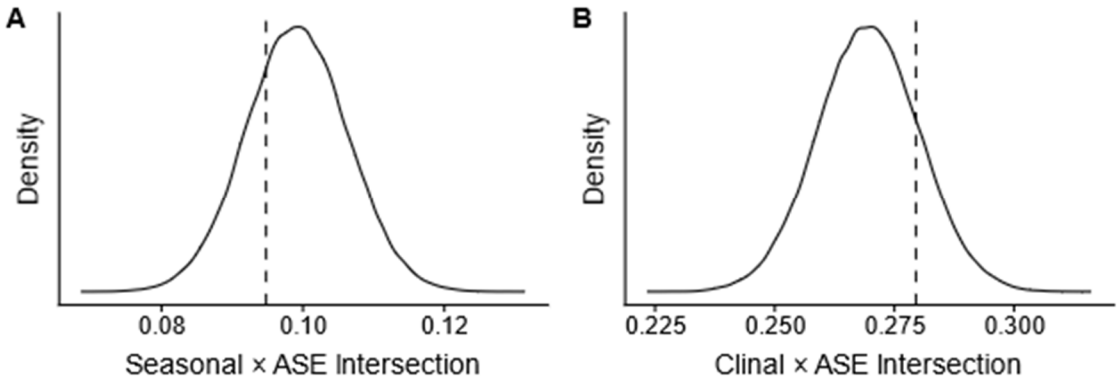


Figure 13. Intersections allele-specific expression (ASE) and seasonality or clinality. Solid and dashed lines represent the expected and observed fraction of ASE SNPs that are also seasonal or clinal. There is an insignificant depletion of Seasonal SNPs with ASE, and an insignificant enrichment of Clinal SNPs with ASE. Expected distributions come from 10,000 random samples of the same size as observed ASE SNPs.

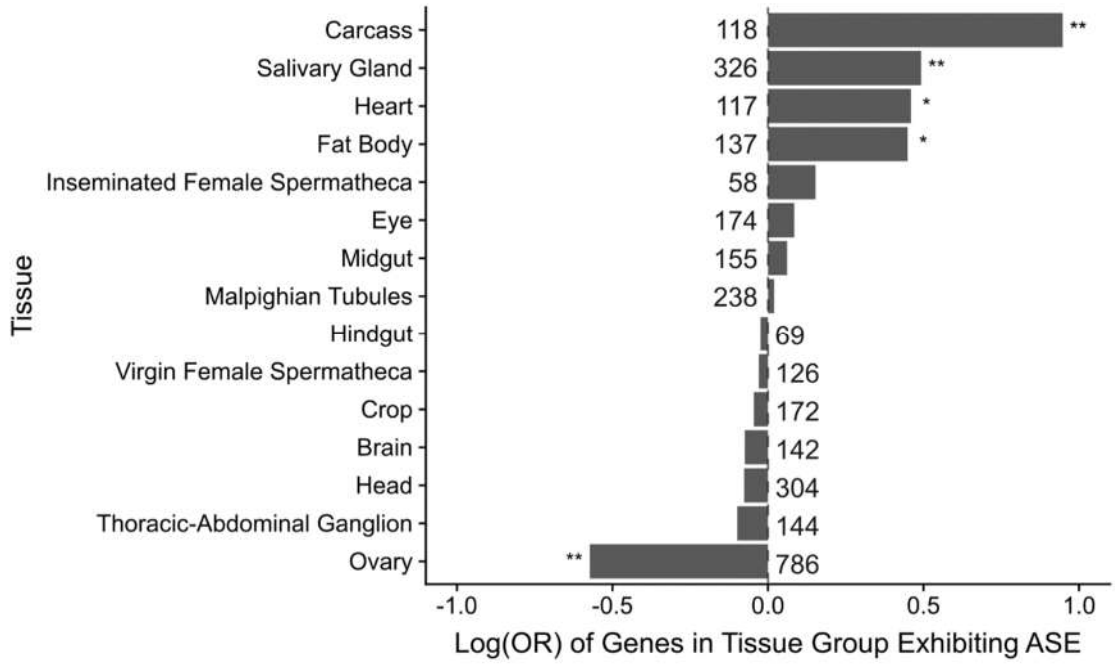


Figure 14. Log odds ratios for ASE in specific tissues. Genes primarily expressed in the carcass have the greatest ASE, whereas genes primarily expressed in the ovaries is the only group with reduced representation of ASE. Counts of genes within each tissue class are shown adjacent to the center line. Genes are classified by the tissue where it is most highly expressed.

Cluster	Chr	Position	Z	-log10(P)	Neighboring Gene(s)	Gene Name (or description)
1	2L	5581150	5.33	7.02		
	2L	5581171	5.27	6.87	FBgn0031722	Coiled-coil domain-containing protein 34
	2L	5581176	5.27	6.87		
2	2L	5713484	-5.25	6.81	FBgn0031730	Cyclin-dependent kinase
3	3L	12676247	5.47	7.34	FBgn0014343	mirror
4	3R	10689620	-5.26	6.84	FBgn0266756	bitesize
5	3R	10858358	-5.44	7.27	FBgn0038281	Ribosomal protein L10Aa
	3R	10883642	-5.42	7.22		
6	3R	10890771	-5.21	6.72	FBgn0038282	defective proboscis extension response 9
	3R	10895072	5.29	6.91		
7	3R	10936003	-5.55	7.55	FBgn0261859	FERM/acyl-CoA-binding protein superfamily
8	3R	14081436	5.27	6.85	FBgn0261019	modigliani
					FBgn0266195	Trimethylguanosine synthase 1
9	3R	14921898	-5.35	7.05		
	3R	14921970	-5.26	6.84	FBgn0016120	ATP synthase, subunit D
	3R	14922137	-5.93	8.53	FBgn0038678	mitochondrial ribosomal protein L55
	3R	14922180	-5.94	8.54		
10	3R	14988398	-5.76	8.08	FBgn0063127	long non-coding RNA:CR33938
	3R	14988418	-5.38	7.12		
11	3R	15336940	5.20	6.71	FBgn0260003	Dystrophin
12	3R	19774105	5.24	6.79	FBgn0039131	Fatty acyl-CoA reductase
13	3R	22618606	5.61	7.70	NA	NA
14	3R	24652834	5.22	6.75	FBgn0024273	WASp
	3R	24665275	5.48	7.36		

Table 1. Loci from GWAS of survival in starvation and desiccation conditions passing permutation threshold. Neighboring genes adjacent to significant loci are named by FlyBase gene ID.

Supporting Information

GitHub Repositories for genome reconstructions:
https://github.com/cory-weller/low-coverage-genome-reconstruction

GitHub Repository for desiccation tolerance analysis:
https://github.com/cory-weller/desiccation_tolerance_cages