UNIVERSITY OF CALIFORNIA,
IRVINE

Identifying selection in differentiated populations through simulation, experimental evolution, and whole genome sequencing

DISSERTATION

submitted in partial satisfaction of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

in Biological Sciences

by

James Baldwin-Brown

Dissertation Committee:
Professor Anthony Long, Chair
Associate Professor Kevin Thornton
Professor Timothy Bradley

2016

DEDICATION

To Laurel, who carried me the whole time.

TABLE OF CONTENTS

LIST OF FIGURES

Page

LIST OF TABLES

ACKNOWLEDGMENTS

I would like to thank the friends and colleagues who have assisted me in this long process. Tony Long, Kevin Thornton, Tim Bradley, Steve Weeks, J.J. Emerson, Mahul Chakraborty, Molly Burke, Stuart Macdonald, and many others contributed in various ways to this work and it would have been impossible without them.

Additionally, I would like to acknowledge the national institute of health, whose grant support funded the laboratory in which this research was conducted, the department of education, whose GAANN grant funded my research for one and a half years, and the HHMI-UCI teaching fellows program, which contributed money for me to travel to conferences to disseminate this work. NIH grants of importance include grant R01-RR024862 and R01-GM085251, awarded to Anthony Long and used to partially fund the results of Chapter 1, NIH grant R01-GM085183, awarded to Kevin Thornton and also used to partially fund Chapter 1. Chapter 2 was made possible, in part, through access to the Genomic High Throughput Facility Shared Resource of the Cancer Center Support Grant (CA-62203) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496-01 and 1S10OD010794-01.

I would also like to thank Nucleic Acids research and Molecular Biology and Evolution, the two journals in which the first two chapters of this research were published. Their open access policies allow the reprinting of articles in the form of a Ph.D. dissertation.

CURRICULUM VITAE

James Baldwin-Brown

EDUCATION

Doctor of Philosophy in Biological Sciences **2016**
University of California, Irvine *Irvine, CA*

Bachelor of Science in Biological Sciences **2011**
University of California, Davis *Davis, CA*

RESEARCH EXPERIENCE

Graduate Research Assistant **2011–2016**
University of California, Irvine *Irvine, California*

TEACHING EXPERIENCE

Teaching Assistant **2011–2015**
University of California, Irvine *Irvine, CA*

REFEREED JOURNAL PUBLICATIONS

Chakraborty M, Baldwin-Brown JG, Long AD, Emerson JJ. 2016 Jul 25. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. Nucl. Acids Res.:gkw654. **2016**
Nucleic Acids Research

Baldwin-Brown JG, Long AD, Thornton KR. 2014. The Power to Detect Quantitative Trait Loci Using Resequenced, Experimentally Evolved Populations of Diploid, Sexual Organisms. Mol Biol Evol 31:10401055. **2014**
Molecular Biology and Evolution

CONFERENCE PARTICIPATION

Poster: Identifying differentiation of populations of the clam shrimp *Eulimnadia texana* through genome assembly and pooled population sequencing. **July 2016**
Allied Genetics Conference

Talk: Identifying differentiation of populations of the clam shrimp *Eulimnadia texana* through genome assembly and pooled population sequencing. **June 2016**
Evolution 2016

Poster: Identifying differentiation of populations of the clam shrimp *Eulimnadia texana* through genome assembly and pooled population sequencing. **July 2015**
SMBE 2015

Poster: Power to Detect QTL using Evolve-and-Resequence: A Simulation of Phenotype-Driven Evolution. **July 2014**
SMBE 2014

Poster: The Ability to Detect Quantitative Trait Loci using Experimental Evolution. **July 2013**
SMBE 2013

Talk: The Ability of Resequenced Experimentally Evolved Populations to Detect Quantitative Trait Loci. **June 2013**
Evolution 2013

SOFTWARE

Quickmerge `https://github.com/mahulchak/quickmerge`
Genome assembly merging program to increase contiguity of assembly.

ABSTRACT OF THE DISSERTATION

Identifying selection in differentiated populations through simulation, experimental evolution, and whole genome sequencing

By

James Baldwin-Brown

Doctor of Philosophy in Biological Sciences

University of California, Irvine, 2016

Professor Anthony Long, Chair

Population differentiation is both one of the central processes underlying the diversity that we observe in the natural world, and a mechanism that can be used to differentiate between evolutionary forces both at the level of the polymorphism, and at the level of the entire genome. Here, I use simulated evolution to analyze the statistical power to detect signals of selection in artificially selected laboratory populations, and use genomic data from wild populations of the clam shrimp *Eulimnadia texana* to identify genomic signals of selection in wild populations. Several loci in the wild populations appear to be under selection, and I analyze the types of genes that appear to contribute to differentiation of these populations. Additionally, I describe an analysis of genome assembly techniques that allowed for the creation of a highly contiguous genome assembly in the clam shrimp. I find that a pipeline that uses custom software to combine the results of several different genome assemblers is capable of producing genomes using long-read genomic sequencing data that are orders of magnitude more contiguous that pre-long-read methods. Simulations of experimental evolution indicated that extremely high levels of replication were necessary in order to achieve high power to detect signals of selection in experimental evolution. To this end, I describe a set of replicate experimentally evolved populations of *E. texana* that can be used to identify regions under selection with much higher power than could be accomplished with earlier experimental evolution schemes.

Chapter 1

1.1 Chapter Description

Introduction

1.2 Understanding population differentiation

Population differentiation is one of the central concepts in the study of evolution. Much of the diversity of life in nature is due to the differentiation of reproductively isolated populations over time, both with regard to within species variation and between species variation. Differentiated populations are especially useful to researchers attempting to identify the genetic mechanisms that underlie complex traits, as populations with phenotypic differences can be genotyped to identify the genotypic changes underlying these differences. In the following document, I will demonstrate the limits of phenotypic differentiation for identifying quantitative trait loci in the context of experimental evolution with resequencing, describe a set of populations that have been experimentally evolved according to these limits, and discuss the result of analyzing a set of naturally differentiated populations. In addition, I will describe the process of adapting a non-model organism (the clam shrimp *Eulimnadia texana*) to use as a model for genomics, and discuss genome assembly techniques generated in service of this goal.

1.3 Evolve-and-resequence

Experimental evolution seems naturally well suited to the discovery of quantitative trait loci. If a set of populations descended from a common ancestor are split into two selection regimes, with one regime selecting upon a trait of interest, it follows that allele frequencies at locations in the genome that influence the trait of interest will change in a way that is consistently different between populations. These allele frequency differences can be detected and the loci of interest can be identified. This type of study has been performed by several groups [153, 27, 73, 157, 130, 156], but relatively little progress has been made in understanding whether the methodology should be effective. In practice, genetic drift due to small population sizes may lead to false positives or low power, as may a lack of statistical power due to limited replication. I used a set of simulation data generated by Kevin Thornton to identify the potential problems with this type of study. Use of simulation data allowed me to compare the true causative site loci to those that could be detected via various statistical tests under a variety of experimental conditions. After carefully controlling for false positives, I identified a range of experimental conditions that would produce relatively high power ($> 50\%$, $> 80\%$) to localize QTL effectively. The level of replication necessary for effective power was well above that used in most studies that occurred on macroscopic animals. This study left open the possibility that experimental evolution and resequencing could be performed successfully if replication could be increased.

1.4 Clam shrimp as a model organism

I chose to attempt to experimentally evolve populations of *E. texana* for salinity tolerance using the experimental conditions identified by my simulation work. We chose the clam shrimp as a model organism because it is physically small, can be raised in the laboratory easily, has short generation times, and has a small genome. We chose salinity tolerance both because it is likely that salinity varies from population to population in nature because of drying of ponds over time, and because it is easy to modify in the lab.

E. texana, like all clam shrimp, is a vernal pool shrimp. Vernal pool shrimp live in small, temporary pools in a variety of climates. All vernal pool shrimp require fresh water to live, and so must cope with the drying out of vernal pools in the summer. They persist by laying eggs that enter a state of diapause when dried. The eggs hatch upon being rehydrated in the spring. Vernal pool organisms have been noted for having a variety of modes of reproduction, including hermaphroditism, androdioecy, and standard obligate dioecy [164]. Some have speculated [164] that these unique reproductive modes are due to the isolated nature of the shrimp, being bound in land-locked pools where mates are not always available. Clam shrimp have not historically been used for studying genomics or evolution. The primary research on *E. texana* has focused on its reproduction and sex determination. Clam shrimp are known to be androdioecious —that is, they have males and hermaphrodites, but no true females [140]. Hermaphrodites can self-fertilize or be fertilized by males, but cannot reproduce with other hermaphrodites because they lack specialized reproductive claspers only observed in males [140]. The sex determining system is known to be genetic [140], but the sex determining locus has not been localized. Clam shrimp go through one generation every time their pools are hydrated and are obligately sexual [165, 140], in contrast to other vernal pool systems such as *Daphnia pulex*, which reproduce continually and via both parthenogenesis and sexual reproduction. This allows the number of generations to be precisely controlled. Additionally, the fact that the shrimp are trapped within small pools of water allows individuals that are not highly geographically distant to become genetically isolated, allowing for observation of the genetic differentiation of a large number of populations. We show here that *E. texana* has a relatively small genome of about 150 million basepairs, which makes whole genome sequencing highly affordable. All of these traits lead to a system that is well suited to use as a genetic model system, but which has no resources in place for performing genetic analysis. We set up a set of experimentally evolved populations of *E. texana* and simultaneously developed a set of genomic tools for analyzing the clam shrimp, which we then used to analyze a set of natural clam shrimp populations, as documented below.

1.5 Experimental evolution of clam shrimp

Based on the results of our simulations, we needed to grow large number of shrimp quickly in a large number of replicated populations. We set up a laboratory environment that would allow us to grow up to 36 populations at a time with population sizes of up to 1,000 shrimp each, then divided those 36 populations into 18 “experimental” populations (with salt added to the water) and 18 “control” populations with very low salt levels. We reared these populations in a consistent set of environments on a 3-week cycle. Tests of fitness under different salt conditions indicate that the “experimental” shrimp were adapted to high salinity within 7 generations. Further experiments on these shrimp lines should involve sequencing pools of individuals from each of the populations and statitical comparison of allele frequencies to identify regions under selection in a replicated way in the "experimental" population.

1.6 The necessity of genome assembly

In order to use clam shrimp as a model organism, I needed a set of genomic tools and data that I could analyze. Chief among these is a high quality genome assembly. A genome assembly is used by a large portion of modern genomics tools, but genome assembly is a difficult problem. Most genome assembly techniques receive only cursory validation before use, so I endeavored to empirically test genome assembly techniques using the most up to date whole genome sequencing methods [33]. I ultimately produced a "roadmap" to genome assembly that could be used to identify the combinations of data that could produce high quality assemblies. Traditional genome assemblies have either relied on scaffolding of relatively non-contiguous assemblies derived from Illumina short read data [146] using mate pair libraries (that is, libraries in which the ends of long DNA fragments are sequenced), or on large numbers of very long reads. I used two pipelines to generate assemblies: first, a collaborator generated a set of assemblies using very long read data and the program *PBcR*, part of the *Celera* assembler package [126]. At the same time, I generated a set of "hybrid" assemblies using the program *DBG2OLC* [171]. These hybrid assemblies were generated using both short and long read data. We found that a tradeoff exists between hybrid assemblies and long read only assemblies. Specifically, hybrid assemblies work well at low levels of long read coverage, but are quickly surpassed by long read only assemblies when the long read coverage is high.

Knowing this, we developed a pipeline where we would generate a hybrid assembly and a long read only assembly, then merge the assemblies together. My collaborator wrote the merging program, and I tested it and wrote the wrapper program associated with it. We found that the merged assemblies were as good as or better than the constituent assemblies in all cases. Finally, we laid down a set of recommendations for genome assemblies using our pipeline.

1.7 Differentiation of natural populations

Much effort has gone into identifying the differentiation of both natural populations and laboratory populations, going back at least to the use of pairwise F_{ST} to identify populations with more among-population differentiation than expected [168]. Populations can differ in allele frequencies for the standard evolutionary reasons: natural or artificial selection, genetic drift, mutations, or migration. In the presence of drift alone, allele frequencies are expected to remain constant on average, with any differentiation being unique to individual populations. On the other hand, if natural selection is acting differently on two populations due to differences in the populations' environments, allele frequencies at the loci under selection should differ more among these populations than we would expect in the drift-only case. Assuming the correctness of these statements, and our ability to correct for relationships that already exist amongst a set of natural populations, we should be able to use differences in allele frequency between populations as a way to test for natural selection at each polymorphic location in the genome. Numerous statistical methods [168, 68, 50] exist to attempt to detect natural selection in this way. I gathered a set of sequencing data from natural populations of *E.*

texana in order to identify regions under selection using these statistics. Before that, however, I generated a highly contiguous genome assembly using a deeply sequenced inbred strain of shrimp, then annotated it using RNAseq [160] data.

I analyzed the allele frequencies of the natural populations using several statistics: F_{ST}, *Bayenv*'s $X^T X$ and Bayes Factor statistics, *LFMM*'s z-values, and *SweeD*'s composite likelihood scores. I found putative evidence of selection at several loci across the genome. Gene ontology terms associated with these regions were mostly related to vision, leading me to hypothesize that the clarity of the water in different ponds may be a driver of selection. Additionally, manual examination of the most significant loci lead to the identification of a number of genes that appear to be under selection and correlated with specific environmental variables. One such example is the clam shrimp ortholog of *Drosophila CG10413*, which appears to be associated with latitude and is predicted to have sodium/potassium/chloride symporter activity.

1.8 The following documents

What follows here are a set of documents, two of which are currently published (chapters 1 and 2), one of which is being prepared for publication (chapter 3), and one of which is unpublished (chapter 4). These documents discuss the above topics in detail and cover the research that I have performed in my time as a graduate student at the University of California, Irvine.

Chapter 2

2.1 Article

The Power to Detect Quantitative Trait Loci Using Resequenced, Experimentally Evolved Populations of Diploid, Sexual Organisms

James G. Baldwin-Brown, Anthony D. Long, Kevin R. Thornton

Department of Ecology and Evolutionary Biology, University of California, Irvine, CA, 92697

Corresponding Author: James G. Baldwin-Brown

Email: jbaldwi1@uci.edu

2.2 Preface

This chapter was originally published in *Molecular Biology and Evolution* under the title "The Power to Detect Quantitative Trait Loci using Resequenced, Experimentally

Evolved Populations of Diploid, Sexual Organisms" [14]. It is reprinted here in its original form. The simulation machinery used to generate the dataset was written and run by Kevin Thornton with input from Anthony Long. I performed the statistical analysis of the dataset, drew the primary conclusions, and wrote the text of the paper.

2.3 Abstract

A novel approach for dissecting complex traits is to experimentally evolve laboratory populations under a controlled environment shift, resequence the resulting populations, and identify SNPs and/or genomic regions highly diverged in allele frequency. To better understand the power and localization ability of such an evolve and resequence approach, we carried out forward-in-time population genetics simulations of 1Mb genomic regions under a large combination of experimental conditions, then attempted to detect significantly diverged SNPs. Our analysis indicates that the ability to detect differentiation between populations is primarily affected by selection coefficient, population size, number of replicate populations, and number of founding haplotypes. We estimate that evolve and resequence studies can detect and localize causative sites with 80% success or greater when the number of founder haplotypes is over 500, experimental populations are replicated at least 25-fold, population size is at least 1000 diploid individuals, and the selection coefficient on the locus of interest is at least 0.1. More achievable experimental designs (less replicated, fewer founder haplotypes, smaller effective population size, and smaller selection coefficients) can have power of greater than 50% to identify a handful of SNPs of which one is likely causative. Likewise, in cases where s ≥ 0.2, less demanding experimental designs can yield high power.

2.4 Introduction

Quantitative traits are of special interest to biologists. The variation in many traits of medical, agricultural and evolutionary relevance is due to the concerted action of several genes and the environment. QTL mapping has been effective at explaining the majority of the heritability of a trait, but is poorly suited to resolving the location of QTL beyond several cM [115]. More recently, several groups have attempted to increase the resolution of QTL mapping using advanced generation recombinant inbred lines (c.f., [95, 11, 84]), but resolution is still limited to cM scales. Recently, genome wide association studies (GWAS) have become a major method for investigating the genetic basis for quantitative traits ([29, 28, 38]). Although GWAS studies have identified replicable associations between SNPs and complex traits, associated SNPs tend to explain only a small fraction of the heritable variation in the study trait [116], a problem that cannot be solved by increasing sample sizes to tens of thousands of individuals [145] or replacing SNPchips with complete resequenced genomes [150]. Clearly, it is of value to explore novel methods for dissecting complex traits.

In systems that have short generation times and that can easily be reared in the lab in large numbers, an alternative experimental approach to dissecting complex traits has been to "evolve and resequence" (E&R) populations of organisms. E&R studies have been performed with both asexual [136, 15, 86, 152, 131] and sexual [153, 27, 73, 157, 130, 156] populations. Because asexual experimental evolution lacks recombination and standing variation in the base population, the footprints of selection in the genome and the means by which an investigator may hope to identify causal variants are different in sexual and asexual systems. Thus, we limit our focus to E&R studies in sexual systems. Under the E&R paradigm, a base population is divided into several replicate populations, half of which are subjected to a well-defined selection pressure, and the other half of which are maintained without selection. Next, the DNA pools from each population are re-sequenced using NextGen technology and allele frequencies in each pool are estimated. SNPs and/or genomic regions showing consistent differentiation between selected and control population are candidates for harboring causative variants. Studies using this design have claimed to detect numbers of candidate causative sites from 662 [27] to almost 5000 [130] for various quantitative traits. Currently, the causative sites detected by E&R methods have not been validated.

To date, the field of E&R has been almost entirely empirically motivated. Study designs have varied greatly in terms of the number of replicate populations, the population sizes maintained, the number of generations over which the experiment was carried out, and the number of haplotypes in the base population from which selection was initiated. For example, Burke et al. (2010), Teotónio et al. (2009), Turner and Miller (2012), and Orozco-Terwengel et al. (2012) maintained population sizes in excess of 1000 individuals, while Turner et al. (2011) used population sizes of around 225, and Johansson et al. (2010) used effective population sizes of 27 to 44 individuals. The number of founder haplotypes is often not precisely known, but can vary from a few dozen individuals [73] to 113 isofemales [130] up to 173 inbred lines [156]. The number of generations of evolution also varies widely between experiments: Turner and Miller (2012) used 14 generations of selection, Orozco-Terwengel et al. (2012) used 37, Teotónio et al. (2009) and Johansson et al. (2010) used 50, Turner et al. (2011) used 100, and Burke et al. (2010) used 600. Replication varies as well: Turner and Miller (2012) sequenced two replicate populations each for two experimental treatments, Turner et al. (2011) sequenced two replicate populations each for two experimental treatments and one control, Orozco-Terwengel et al. (2012) sequenced three replicate populations undergoing domestication, Johansson et al. (2010) sequenced two populations selected for divergence, and Teotónio et al. (2009) sequenced 29 total populations – five control populations, four replicate populations for each of three treatments, and four reverse-evolved populations for all three treatments. Burke et al. (2010) sequenced five experimental and five control populations, but each treatment was sequenced as a single pool because of technological constraints. It is of value to quantify the extent to which these experimental design decisions impact the power to detect causative sites and contribute to false positives.

Furthermore, there are no agreed upon statistical approaches for analyzing the sets of pooled allele frequency estimates obtained from E&R studies. For example, Burke et al. (2010), Johansson et al. (2010), and Teotónio et al. (2009), respectively, used Fisher's exact test, a Chi-squared test, and an a posteriori Dunnett test to detect significant allele frequency differences between treatments, while Orozco-Terwengel et al. (2012) and Turner et al. (2011) used, respectively, the Cochran-Mantel-Haenszel test and a statistic referred to as "DiffStat" to determine if allele frequencies differed significantly from simulated allele frequencies subject only to drift. Burke et al. (2010) favored sliding windows of allele frequency change. Turner and Miller (2012) used a graphical approach in which the divergence within treatments was used to establish a null expectation, and divergence between treatments was considered significant if it fell outside this null range. The lack of a consistent standard for statistics and experimental conditions prevents us from confirming the numerous candidate causative sites that these studies claim to have detected.

The exact prediction of allele frequency change at even a single locus is challenging when both selection and genetic drift affect allele frequency. The approximation of allele frequency probability distribution over time that is best suited to this problem, the Kolmogorov forward diffusion equation [49, 83], is a second-order partial differential equation that can only be solved by numerical integration in many cases [67]. This equation is advantageous in that, unlike the binomial sampling method [47] it does not make the assumption that Hardy-Weinberg equilibrium is maintained, which is crucial when modeling experimental evolution because of the small population sizes and large selection coefficients involved. The fact that time-dependent diffusion equations often have no closed-form solutions and make the strong assumption of very large population size and weak evolutionary forces (e.g., small s in the case of selection) motivates the use of simulation in this work. In order to accurately predict the results of E&R experiments without an exact theoretical solution, we chose to quantify the power and false positive rate of E&R studies via forward-in-time population genetic simulations of evolving one Megabase (Mb) regions. We generated base populations with defined numbers of preexisting haplotypes via coalescent simulation (analogous to establishing a laboratory population from a wild caught sample), expanded the base population, and chose diploid individuals to initiate an experimental evolution experiment. Our simulations focused on a single causative SNP, embedded in a 1Mb region filled with neutral SNPs, under constant selection during the course of the experiment. We designated a single SNP to have a positive selection coefficient in the selected population and a selection coefficient of zero in the control populations, and then allowed each population to evolve with selection, recombination, and drift. By simulating replicate E&R studies and then carrying out appropriate statistical tests on the replicated data sets, we obtained an estimate of the proportion of times that a similar experiment would detect a region (CR for causative region) harboring at least one causative SNP and, potentially, identify a causative SNP (CS) embedded in such a region. Because of the existence of linkage disequilibrium and strong selection during experimental evolution, it may be easier to detect CRs than CSs. We carried out these simulations under a variety of conditions: we varied population size (n), number of founder haplotypes (h), selection coefficient on the CS of interest (s), number of replicated populations (r), and number of generations of evolution (g) (Table 1). We termed these parameter combinations "Θ". We included control simulations in which the selection coefficient at the causative SNP in the selected population was zero, which allowed us to determine a Type I error rate for CR and CS detection.

We observed that the false positive rate for CR detection (even when using a very stringent criterion of significance) was extremely high using standard single-marker tests under a minority of conditions when ten replicate populations were used. The power to detect CRs was determined primarily by population size, replication, selection coefficient, and number of generations, with an intermediate number of generations being ideal. The power to localize CSs was similar to the power to detect CRs but was strongly affected by the number of founder haplotypes. Achieving a total power to detect CRs and localize CSs of 80% required almost all parameters to be at their ideal values (1000 individuals per population, 500 founder haplotypes, 25 replicate populations) for the case of a selection coefficient of 0.1, but reasonable levels of power can also be achieved with less costly experimental designs or higher selection coefficients. Our simulations suggest that the experimental designs that could be most effectively utilized for detecting CRs and localizing CSs under the E&R paradigm are not currently widely employed, and likely require considerable experimental effort. Still, the parameter space that provides reasonable power levels is not outside the realm of possibility for E&R studies using macroscopic organisms.

2.5 Results

2.5.1 The False Positive Rate

From the perspective of a naïve observer, any given simulated 1Mb genomic region might or might not contain a CS. In order to determine the fraction of times that we falsely identified a CR, we calculated for every parameter combination (Θ) where s=0 the fraction of cases in which at least one SNP was found to have a p-value of less than 10^{-1}, 10^{-2}, 10^{-3}, etc. through 10^{-14}. We referred to this as the false positive CR detection rate (Fig. 1); that is, the fraction of neutrally evolving regions that are nonetheless flagged as "significantly diverged". It is apparent from the figure that the false positive rate is quite high for certain parameter combinations regardless of the statistical threshold employed. False positives are especially frequent in the specific case in which all of the following are true: there are ten experimental replicates, the population size is only 100 individuals, and there are between 32 and 100 founder individuals. This elevated false positive rate is likely due to the t-statistic used to assess significance not being distributed as a t-distribution, especially in the tails, when the number of replicates is small (Sup. Fig. 1). It is important that our 1Mb false positive rate is essentially zero; otherwise, there is a high likelihood of identifying a false positive CR somewhere in a genome that is several hundred Mb in size. In a genome the size of Drosophila melanogaster (122Mb), the false positive CR detection rate necessary to achieve a genome-wide false positive rate of 0.05 is 0.05/122 = 0.00041. This corresponds to approximately 1 false positive in every 2439 regions tested. In order to accurately measure the false positive rate at low values, we generated 10,000 replicate simulations at each Θ where $s = 0$. With this number of replicate simulations, any Θ with 4 or fewer false positives has an acceptable error rate. At each Θ, we found the most lenient of our chosen significance thresholds that produced 4 or fewer false positives and used it in power calculations for the remainder of the experiment (Sup. Fig. 2, Sup. Table 1). This is a more fair comparison than choosing a single significance threshold that is applied to all Θ because the false positive rate varies widely between Θ such that a significance threshold that is reasonable for some Θ is unnecessarily strict for other Θ, and would not provide a reasonable estimate of the maximum power achievable in those Θ. Θ in which an acceptable false positive rate was not achieved by our most strict significance threshold, 10^{-14}, were discarded. This included all experimental designs where $r = 10$, $n = 100$, $h = 100$, and $g = 500$ or 1000 were simultaneously true. Of the 208 chosen thresholds (one for each combination of n, h, r, and g), the distribution was as follows, with the first item in the list corresponding to 10^{-1}, the second corresponding to 10^{-2}, and so on: 0, 0, 42, 0, 7, 25, 42, 62, 20, 5, 1, 1, 1, and 2. The large number of significance thresholds set to 10^{-3} corresponds to the Θ in which $r = 2$; in these Θ, power and false positive rates are both extremely low, so the selecting of a lenient significance threshold is unsurprising. The mean of the -log10 of the significance thresholds is 6.71.

2.5.2 Power to Detect a Causative Region (CR)

Having controlled the false positive rate via an individualized statistical threshold, we examined the ability to detect CRs. As in traditional QTL literature, there are two issues at hand. First, is it possible to find an association between genetic features and experimental treatment? This is analogous to CR detection as discussed in this section. Second, if an association is found, to what level of precision can the polymorphism underlying the trait be localized? This is analogous to CS localization in the following sections. As above, we considered a CR detected if it contained at least one significantly diverged SNP (p $\leq$ significance threshold). Because we only used 500 simulations per parameter combination where s $>$ 0, we estimated the amount of error in estimates of power due to limited sampling by finding the 95% confidence interval around each power estimate using binomial sampling. We found that the mean width of the 95% confidence interval for all nonzero power estimates was 4.64%, the standard deviation of these widths was 3.31%, and the range of widths was 0.052% to 8.94%. The power to detect a CR increased with increasing r, n, and s, slowly decreased with increasing h, and was maximized at $g = 500$ when $s = 0.05$ and at $g = 100$ when $s \geq 0.1$ (Fig. 2). When $r = 2$, we observed a power of near zero in all cases. For several simulated parameter combinations, power was quite high, especially when n was large and the s associated with the CS was equal to or greater than 0.05. As expected from standard population genetic theory, as decreasing s approached the reciprocal population size, power to detect a CR decreased substantially. Interestingly, although smaller numbers of starting haplotypes are associated with the greatest power to detect a CR, this effect was weak (a feature of E&R experiments that will be important in identifying CSs). Below, we disregarded parameter values where CR power with that parameter value was always below 35%; specifically, we disregarded all Θ where $s \leq 0.005$, $r = 2$, $n \leq 100$, or the specific case where $n \leq 250$ and $r \leq 5$.

2.5.3 Power to Identify a Causative Site (CS)

The goal of an E&R study is CR detection followed by the identification of a CS within the detected CR. In order to determine most effective method of CS localization, we examined the distance from the most significant marker to the causative SNP (MSM-CS distance) in each simulated region in which at least one SNP was significant (Fig. 3). A large fraction of MSM-CS distances were equal to zero for cases of Θ where

CR detection power was high, indicating that precise localization is possible under some circumstances. The nonzero MSM-CS distances appeared to be skewed such that a large fraction of MSMs were within 100kb of the CS, indicating that these MSMs are likely driven to high levels of divergence by linkage to the CS, rather than drift. Indeed, if we take, for example, the (relatively moderately powered, drift-heavy) case in which $s = 0.05$, $n = 500$, $h = 32$, $r = 10$, and $g = 500$, 95% of all non-zero MSM-CS distances were less than 59kb when only significant regions were considered. For a large portion of the Θ cases with high CR detection power (i.e., $n = 500$, $s \geq 0.05$, $r \geq 10$, $h \geq 100$, except where n = 500, h = 32, r = 10, and g = 1000), the median MSM-CS distance is zero, while the mean is a non-zero value. We observed a similar pattern in the CS rank (Sup. Fig. 3). Although selective sweeps are clearly visible in the raw significance scores (Sup. Fig. 4), the fact that a large majority of the MSMs in most regions with high power have an MSM-CS distance of 0 seems to indicate that a sliding window analysis would be no better than a single-SNP analysis at localizing CSs. Indeed, our attempts to use a sliding window for CS localization by identifying the sliding window with the largest summed -log(p) values in each region produced lower power than single-SNP analyses (Sup. Fig. 5). Thus, we chose to localize CSs through single-SNP analyses.

We examined several methods of precisely localizing a CS, conditional upon identifying its CR as significant. Most strictly, we may identify a CS as being correctly localized only if the most significant marker (MSM) in a CR is the CS. Alternatively, many would consider any analysis that restricts the likely location of a CS to a small region (i.e., 10kb), a small number of SNPs (i.e., 25), or within a small LOD drop of the MSM (i.e., within 2 LOD) to have utility, as additional experiments may be capable of identifying the CS. Figure 4 summarizes the power to localize the CS to: an exact location, within 10kb of the MSM, within the top 25 most significant SNPs in a region, or within a 2 LOD drop of the MSM, all conditional on CR detection and s = 0.1. g is set to 500 in all plots below except where specified for ease of viewing, and because the effect of g on power was relatively small in the parameter space where power is high. The primary factor that affected CS localization was h (the number of founding haplotypes). When h is small, it appears that high linkage disequilibrium results in significant allele frequency divergence at SNPs near the CS, making it difficult for the CS to be differentiated from neighboring SNPs. From figure 4 it is apparent that the localization power was quite high provided that h was high. h negatively affected CR detection power, yet positively affected CS localization power. As discussed below, the overall effect of h on power was positive due to the extreme effect of h on CS localization power. In a best case scenario where $n = 1000$, $s = 0.05$, $g = 500$, and $r = 25$, an h of 4 produces an exact location power of only 4.0%, while an h of 100 produces an exact location power of 76.4%.

The false positive localization rate, equal to $1-$(localization power), can be considered the fraction of CRs in which the CS is not correctly localized. At least one of the localization false positive rates calculated is below 5% in 123 of our simulated Θ, including but not limited to the entire simulated parameter space where $h \geq 500$, $n \geq 500$, $r \geq 10$, and $s \geq 0.05$. It is not possible to calculate a genome-wide false positive localization rate because the number of expected CRs in a genome is unknown. Note that this false positive rate is distinct from the false positive CR detection rate. The false positive CR detection rate indicates specifically the frequency with which CRs are detected where they do not exist, while the false positive localization rate indicates the fraction of the time that a true CR has its CS incorrectly localized. This value may be of special interest to researchers attempting to assess the chance that a significant SNP in a study is likely to be a CS, or merely a neighbor of a CS.

2.5.4 Total Power

Total power is the product of CR detection power and CS localization power given CR detection. Total power is then the fraction of all CSs that, starting from no prior knowledge about the data, can be detected and localized successfully. Figure 5 gives the total exact power, total top 25 power, and CR detection power as functions of Θ when $g = 100$. The range where $h = 4$ is excluded because the CS localization power conditional upon CR detection in these Θ is less than 80% for all statistics except the within 2 LOD power, and few E&R experiments use only 4 founder haplotypes. Total power is highest when s, n, h, and r are maximized and g is at a value of 100. The parameters necessary to achieve at least 80% exact location power for the $s = 0.1$ case are $n \geq 1000$, $r \geq 25$, and $h \geq 500$ (Fig. 5, Sup. Fig. 6). This is a sobering result because it is experimentally difficult (in a system like Drosophila) to achieve values of Θ that reach a total exact location power above 80%. On the other hand, in the cases where $s \geq 0.1$, the same goal of 80% exact location power is much more achievable: 21 of our simulated Θ, including but not limited to all cases in which $s \geq 0.1$, $h \geq 500$, $r \geq 15$, $n \geq 1000$, and $100 \leq g \leq 500$ produce a total exact location power greater than 80%. Thus, exact localization requires relatively strict experimental conditions, but strongly selected SNPs are more easily localized. Unsurprisingly, within 10kb power, top 25 power, and within 2LOD power were consistently higher than exact location power, and were higher than 80% when $s \geq .05$, $n \geq 1000$, $r \geq 25$, $g = 500$, and $h \geq 32$, except in the case where $s = 0.05$ and $h = 100$, suggesting that ambitious, yet achievable, experimental designs are capable of localizing causative sites to a few dozen or even fewer candidate SNPs.

In many experimental systems, there is a direct tradeoff between n and r when setting up an E&R study because both of these parameters are space and resource limited. Both affect the total power differently: increasing replication (r) improves the number of degrees of freedom during statistical analysis, while increasing the population size maintained during the experiment (n) decreases the effect of genetic drift on allele frequencies. Both parameters are subject to diminishing returns as their values are increased. For example, in the case where $s = 0.05$, $g = 500$, $r = 10$, and $h = 100$, a doubling of n from 250 to 500 increases exact location power from 5% to 27%, while a doubling of n from 500 to 1000 increases exact location power from 27% to 46%. With the same parameters and an n of 500, an increase of r from 5 to 10 increases exact location power from 8% to 27%, but a similar increase in r from 10 to 15 only increases exact location power from 27% to 43%. Because diminishing returns occur, the ideal r and n values for a given laboratory size should be balanced, with the specific values depending on the specific conditions of the experiment (Sup. Fig. 7). Unfortunately, it is difficult to determine an ideal r:n ratio because multiple costs are involved: the cost of more replicates vs. more generations, the cost of sequencing vs. rearing, and so on.

As noted above, the effect of g on the total power to detect and localize CSs was small in the parameter space where power was high, so g was omitted from several plots for simplicity. It was apparent that there was a strong interaction between g and s with regard to power. At $s = 0.05$, an intermediate g (500) appeared to be superior to either high (1000) or low (100) g values in terms of the power to detect CS-containing regions and the total power to localize CSs (Fig. 6); at $s = 0.1$ and $s = 0.2$, the relationship between g and power was generally negative. One possible explanation for this result is that, when $s = 0.05$, selection had largely fixed any CS's by generation 500, but drift continued to influence allele frequencies at linked markers past generation 500 resulting in increased noise after 500 generations, whereas CSs with higher selection coefficients, i.e. $s = 0.1$ or 0.2, were mostly fixed by generation 100, causing power to decrease when s>100 due to genetic drift. We found that the number of fixed or lost CS alleles in populations where $s = 0.05$ increased from approximately 25% fixed or lost when $g = 100$ up to approximately 100% fixed or lost when $g = 500$ (Fig. 7; see Sup. Fig. 8 for allele frequencies), but that the total number of fixed alleles continued to increase even when $g = 1000$, implying that functional standing genetic variation in fitness was largely exhausted by generation 500, but that drift at linked neutral markers continued to occur. This result seems to confirm that rapid selection and slow drift cause intermediate numbers of generations to be ideal for CS detection and localization.

We used multiple linear regression to attempt to create a model that predicts exact location power as a function of the s, r, g, h, and n (Sup. Table 2, Sup. Fig. 9). We generated a table of total exact location power and the five experimental design variables of interest, then censored it in R to only contain the Θ where $10 \leq r \leq 25$, $250 \leq n \leq 1000$, $32 \leq h \leq 500$, $0.05 \leq s \leq 0.2$, and $100 \leq g \leq 1000$ in order to focus on modeling the power curve in the area where power is highest. We then used the `lm` function in *R* to fit the linear model below:

$$\begin{aligned} 0.245 \times log_{10}(s) + 0.668 \times log_{10}(r) + 0.437 \times log_{10}(n) \\ +0.179 \times log_{10}(h) - 0.0001559 \times g - 1.594 = \text{total exact location power} \end{aligned}$$

Before calculating the slopes, we modified s, r, h, and n by applying the `log10()`

function to them as this improved the fit of the model. In the limited parameter space examined, the linear model explains 86.9% (adjusted R^2) of the variation in total exact location power and has a standard error of 0.07782. Values produced by this formula that are above 1 or below 0 should be assumed to be, respectively, 1 or 0. Although this equation does not take interactions between experimental conditions into account, it produces a relatively accurate power estimate in the aforementioned parameter range.

2.5.5 Multiple Causative SNPs

We simulated the possible case of a 20Mb chromosome containing 6, 26, or 51 CSs in order to test the effect of multiple CSs on CR detection power and total power. Specifically, we simulated $s = 0.05$, 0.1, and 0.2 (the selection coefficients that produced reasonable power levels in the previous simulation), $g = 100$, 500, and 1000, and $r = 2$, 5, 10, 15, and 25. We simulated two different combinations of h and n: 1) the highest power level that we simulated ($n = 1000$, $h = 500$), and 2) a moderate power level ($n = 500$, $h = 100$). We generated 250 replicate experiments under each of these parameter combinations. In each replicate, one of the CSs was placed at the center of the chromosome and the others were randomly distributed throughout the chromosome but not within the 1Mb region surrounding the central CS. The external CSs always had the same s as the central CS. Over the course of the forward simulation, the allele frequencies of all SNPs in the 1Mb region surrounding the central CS were recorded and used to calculate p values. We analyzed the resulting p values according to the same framework used in the previous simulation. When compared with the single CS simulation, the multiple CS simulations almost universally produced higher CR detection power and lower CS localization power (Sup. Fig. 10). This result is expected. A larger number of neighboring CSs should increase the average significance of SNPs in the region of interest by increasing the probability that an external CS is adjacent to the focal region. This should increase CR detection power by increasing the probability that at least one SNP will be significant, but decrease total power by decreasing the probability that the CS will be the most significant SNP in the region. For the three cases where the number of external CSs was equal to 5, 25, and 50, he average shortest distances from the focal CS to the closest external CS were, respectively, 2.05Mb, 0.87Mb, and 0.68Mb.

One consequence of an increased number of CSs coexisting in a population is an increase in the variance of that population's fitness. We calculated the fitness of each possible haplotype in each population and found its corresponding frequency in the population in order to find a distribution (Sup. Fig. 11) of fitnesses at each simulated Θ in which more than one CS is present. As expected, the distribution of fitness becomes broader as s and the number of CSs increase. The variances that we observed (Sup. Fig. 12) when the number of external CSs was 25 or larger seem much higher than those observed in natural populations (Endler 1986, p. 207) and likely much higher than those observed in laboratory experimental evolution (cf. ovary weight in Roff and Fairbairn 2007), indicating that the presence of a large number of CSs with large s values is not realistic under these conditions.

2.6 Discussion

This study provides insight into the experimental designs and genome-wide significance thresholds necessary to detect CRs and localize CSs in E&R studies. Importantly, the experimental parameters necessary for CS detection and localization are more difficult to achieve than most experimentalists likely imagine. Researchers wishing to detect more than 80% of the CRs in which $s = 0.05$ are advised to have $r{\geq}25$, $n{\geq}1000$, and $g{\geq}500$. Researchers wishing to successfully both detect and localize more than 80% of CSs should have $s{\geq}0.01$, $n{\geq}1000$, $r{\geq}25$, and $h{\geq}500$. No value of n simulated in this study was large enough to allow for detection of a useful number of CRs where $s{\leq}0.005$. The low power to detect CSs with small fitness effects is important if we consider than many traits of interest in E&R experiments are quantitative and may have many loci of small effect contributing to standing variation. We have shown that CR detection and CS localization are both improved by high values of n, r, and s. A low h value improved CR detection but negatively affected localization, likely because high linkage disequilibrium limited our ability to distinguish between neighboring SNPs. We found that intermediate g values provided the highest CR detection and CS localization because most detectable CSs have reached fixation by generation 500, or generation 100 in cases where $s{\geq}0.1$. The large effects that n, h, and r have on power seem to indicate that the most efficient method to increase power is to increase these parameters, especially r, which seems to increase total power at a nearly linear rate, at the expense of g, which appears to have a small effect on power when other conditions, such as replication, are kept high.

Although the Θ required for very high power is difficult to achieve in practice, we find evidence that reasonable power levels can be achieved fairly easily. For instance, the parameter space in which the total top 25 power was over 50% was quite large (176 Θ). Indeed, all Θ in which $s{\geq}0.05$, $h{\geq}32$, $r{\geq}15$, and $n{\geq}1000$ produced at least this power level, as did numerous other Θ, such as the Θ where $s{\geq}0.2$, $h{\geq}32$, $r{\geq}5$, $n{\geq}500$, and $g = 100$, except when $h = n = 500$, $s = 0.05$, and $r = 5$. These Θ are perhaps more realistically approached than the Θ necessary to achieve 80% total exact location power. This suggests that ambitious but realistic evolve and resequence experiments can narrow down CSs to a handful of SNPs in a small genetic region. Such SNPs could be validated via additional experiments such as targeted gene knockout/knockin [64]. Given that the primary interest in evolve and resequence experiments is to identify the relationship between genotype and phenotype, meaning that validation experiments will be necessary follow-ups to such experiments, we would argue that it is critical to design experiments with high power to localize CSs.

Our results allow us to reflect on the validity of some of the conclusions drawn in published E&R studies by examining the CR detection power and the false positive rate that we calculated at the Θ that most closely match published experiments. Supplementary Figure 13 shows the CR detection power and false positive rate when s = 0.05 at the simulated Θ that most closely match the Θ used by existing studies. The power levels at our chosen significance thresholds and when s = 0.05, conditional on using our modified t-statistic, were 19%, 1.4%, 0,0,0, and 0 for Burke et al. (2010), Teotónio et al. (2009), Johansson et al. (2010), Turner et al. (2011), Turner and Miller (2012), and Orozco-Terwengel et al. (2012), respectively. Respective powers for s = 0.1, generated using our chosen significance thresholds, were 27.8%, 44.6%, 0, 0, 0.2%, and 0.2%, while respective powers for s = 0.2 were 34%, 87%, 0, 0, 20.4%, and 20.4%. Respective significance thresholds chosen by our system (see "The False Positive Rate" in Results) were 10^{-7}, 10^{-6}, 10^{-3}, 10^{-3}, 10^{-3}, and 10^{-3}. For the same studies, we estimate that CR detection false positive rates were all equal to 0, again conditional upon our statistical test. Notably, although we estimate a high CR detection rate for Teotonio et al. (2009), that study genotyped only 55 loci, so the odds that any of those loci happened to be close enough to causative SNPs to generate a detectable signal of selection are likely low. Therefore, it may not be reasonable to conclude that Teotonio et al. (2009) is more likely than other papers to have produced a true positive result. It should be noted that several of these studies [27, 130] claim that all of the SNPs that have been detected as significant should be treated as candidate CSs, to the extent that Burke et al. (2010) claim that they have detected, on average, a candidate CS every 175 bp. Given that our simulation shows that it is often more difficult to precisely localize a CS than to detect a CS-containing region, and that SNPs up to 100 kb away from a CS can be brought to significant levels of divergence by said CS, it may be more realistic to say that these studies have detected numerous CRs but have limited ability to precisely localize CSs or to determine the number of CSs present in the genome. Admittedly, all of these studies used different test statistics and different significance thresholds than our study, so it is not entirely fair to directly compare the power levels that we estimated from our simulation to the studies in question. That being said, the above studies tended to use a much more aggressive marginal threshold for significance than the ones that we find properly control the false positive rate. A more fair comparison between this simulation and former studies would require the re-analysis of our simulated allele frequencies and the allele frequency data from each experiment using the statistical methods used by the original investigators. Although this is possible using our simulated data set, it is outside the scope of this investigation.

Despite our simulations suggesting low power and high false positive rate, several factors prevent the outright dismissal of published studies. First, gene ontology analysis of genes in regions enriched for change in published studies are consistent with the characters being selected upon. For example, the top 5 enriched gene ontology terms from Burke et al. (2010) were imaginal disc development, smoothened signaling pathway, larval development, wing disc development, and larval development - all have clear causal connections to the “accelerated development” character that was selected. Second, our simulation does not take into account selection coefficients larger than 0.2.

Cases of very strong selection on individual CSs could therefore still allow for high power. Even if the majority of the candidate CSs in a given study are false positives, it is still possible that some of them are true CSs. For example, Johansson et al. (2010) examined a small population of artificially selected chickens. Although their n of 27 to 44 should make even the detection of CSs with a selection coefficient of 0.2 difficult in this case, artificial selection usually involves very high selection coefficients that may be high enough to override the force of genetic drift. Johansson et al. (2010) note that, in the candidate QTLs detected in previous studies, estimated selection coefficients (selection against the unfit allele at the candidate QTL of interest) lie in the range of 0.19 to 0.93, well above the 0.2 simulated here. QTL mapping experiments routinely detect a small number of CSs of relatively large effect (e.g. [84]), so it follows that under strong selection of the type used in experimental evolution, some CSs should have selection coefficients above 0.05, which could account for Johansson et al.'s (2010) ability (and the ability of other E&R experiments) to detect apparently true CSs. A caveat of this line of reasoning is that routine detection implies selection response is due to a handful of genes of large effect as opposed to dozens to hundreds of genes of much more subtle effect as claimed in the recent evolve and resequence literature [153, 27, 73, 157, 130, 156]. Thus, the claims of the literature of localization of causative sites and dozen to hundreds of sites responding to selection seem mutually exclusive given the experimental designs employed.

This study makes a number of simplifying assumptions, all of which we believe to be realistic when describing the case of experimental evolution of small populations. Our simulation machinery operates based on the Wright-Fisher model of population genetics, in which the gametes of each generation are aggregated into a gene pool to generate the next generation. The assumptions of this model, that generations are discrete and mating is random, are realistic for experimental evolution. A further assumption is that all heritable variation is additive within and between loci. Although it is certainly true that non-additive variation exists, the majority of heritable variation is likely additive in nature [69]; therefore, the omitting of non-additive variation in our power analysis should not dramatically affect our power estimates. Our simulations were limited to 1Mb gene regions instead of complete genomes, and all simulated regions have one or no selected loci. We detected CRs by determining if a region contained a significant SNP, then localized by identifying the MSM in the region as the CS. Importantly, we observed that the MSM could be quite distant from the CS: even for parameter combinations with high power to detect a region as significant, a small portion of MSMs were up to 100kb away from the CS, though few were more distant than that (Fig. 3). While our simulations assume that the density of CSs in the genome is relatively low (at most 1 per Mb), our observation that peaks of significant allele frequency change may be quite distant from CSs suggests that the number of significant markers may not be a reliable proxy for the true number of CSs in the genome and call into question whether it is reasonable to deem any significant SNP a candidate CS, especially when h is low. In our simulations, when h was less than or equal to 32 in a population and power was greater than 0, the average exact localization power across our simulations was only 0.244. Although the selection of a 1Mb region for our simulations was somewhat arbitrary, we believe it is an appropriately sized region to consider. The selective sweeps that occurred in our simulated populations appeared to extend less than 500kb from the CS under most circumstances (Sup. Fig. 4), and few replicate simulations generated an MSM-CS distance greater than 100kb. Indeed, in the parameter space where $s \geq 0.05$, 95% of all simulated regions that contained at least one SNP significant at a 10^{-8} threshold had an MSM-CS distance less than or equal to 228kb. Thus, any SNPs simulated further from the CS would resemble SNPs in neutral regions. Similarly, our decision to use the entire simulated region as a candidate CR instead of using only a limited area (say, 100kb) is justified in that the various powers simulated here are virtually unchanged when one compares the power using a 1Mb region and a 100kb region. The mean difference between the 1Mb CR detection power and the 100kb CR detection power is 0.3%, indicating that nearly all CSs that can be detected can be localized to an area the size of a selective sweep around the MSM; in other words, it is reasonable to conclude that CR detection power = total within 100kb power.

Notably, our goal in simulating 1Mb regions was not to test the efficacy of the particular CS-detection technique used here; researchers attempting to adapt this technique to empirical use would need to first divide their genome of interest into arbitrary 1Mb blocks in order to perform our CR detection step, which would be an unnecessarily arbitrary method of subdividing a genome. Rather, our reasoning for choosing to simulate 1Mb blocks was to be certain to capture all of the genetic change due to linkage to the CS in each simulated region. Our CR detection power is thus an upper bound on the ability of a study with a certain set of experimental parameters to detect the presence of any particular CS. It gives no indication as to the ability of that study to localize that CS, except perhaps to say that if a significant SNP is located, the CS that drove its divergence must be close enough to it to have affected its allele frequency via a selective sweep. Some problematic effects that could occur if our CR detection method were applied as-is to real life data, such as the possibility that a CS could be immediately adjacent to a 1Mb focal region and could thus drive a SNP to significance in a non-CS-containing region, are not considered further here.

It is possible to imagine much more complicated models and significance tests than the ones we used. For example, we did not attempt to use the combined p-values of multiple insignificant SNPs to determine the significance of a region because there is no simple way to determine the probability of observing any particular set of multiple p-values if, as in this case, the p-values are not independent. Further, the advantage of a combined p-value approach (higher CR detection) would presumably be at its largest in the Θ where linkage disequilibrium is very high, such as when h is low, but such Θ have already been established as having very high CR detection power, so the advantage gained from a combined p-value approach would be minimal. On a similar note, we did not attempt to simulate a distribution of selection coefficients across the loci in our simulated genomic regions. Recent studies have raised the question as to whether the majority of heritability for any particular trait is best explained by a small number of mutations of large effect [74] or a large number of small effect mutations [17, 16]. Because there is not a scientific consensus on the question of QTL effect size, and thus selection coefficient, distributions, we chose to avoid making assumptions about selection coefficient distributions, and instead merely simulated a range of selection coefficients and calculated the power to detect CSs at each selection coefficient level. Similarly, we chose not to simulate genomic regions containing multiple CSs because they did not fit the paradigm of this study. The design of this study, in which small genomic regions are simulated, implicitly assumes that CSs are distant enough from each other as to not interact significantly. Were we to relax this assumption, the most appropriate method for simulating multi-CS interactions would be to simulate an entire chromosome and distribute CSs across it. Doing so was outside the scope of this study.

A final model we did not consider is the possibility of plateauing allele frequencies due to diminishing selection pressure as a phenotypic optimum is approached, as hypothesized in Burke et al. (2010) and Burke and Long (2012) (also cf. [143]), based on a model in Chevin and Hospital (2008) and potentially observed in Orozco-Terwengel et al.

(2012). That is, we assumed that immediately following the placement of the selected populations into a novel environment a previously neutral SNP obtains a new fixed positive selection coefficient. An additive CS that follows a plateauing allele frequency trajectory could be more difficult to detect than one in which allele frequencies approach fixation because of the lower total level of divergence expected in a plateauing allele; however, our simulation indicates that there are diminishing returns on power from increased allele frequency divergence over time (Sup. Fig. 8), indicating that plateauing allele frequency trajectories will not severely reduce power. This is evidenced by the fact that, at the Θ where $n = 1000$, $h = 500$, $r = 25$, and $s = 0.05$, a near doubling of mean CS allele frequency over all 500 simulation replicates from 52% at generation 100 to 94% at generation 500 only increased total exact location power from 71% to 76%.

Our simulation indicates that, in spite of their inability to detect CSs of very small effect, E&R studies should be capable of detecting and localizing the majority of CSs of moderate to large effect under conditions that, while more labor-intensive than traditional experimental evolution conditions, are still feasible. The effectiveness of the next generation of E&R experiments will depend on their ability to improve upon the experimental designs of the past by using large, well-replicated, initially diverse populations.

2.7 Materials and methods

2.7.1 The Simulation

We simulated replicated experimental evolution using a two-stage approach. First, we simulated 1,000 replicates of a sample of size 2,000 chromosomes from a Wright-Fisher population using the macs software (version 0.4b, Chen et al. 2009) using the following parameters: `macs 2000 1000000 -t 0.01 -r 0.1 -i 1000 -s $RANDOM`. This command line specifies 1,000 replicate simulations of a sample of 2,000 chromosomes. The locus length is 1 million base pairs mutating at rate $\theta = 4Nu = 0.01$ per site and recombining at rate $\rho = 4N *$ (recombination rate) $= 0.1$ per site, where N is the size of a Wright-Fisher population and u is the mutation rate, per base pair per generation. The mutation parameter was chosen to mimic SNP density in non-African Drosophila melanogaster, and the recombination rate was based on estimates from Chen et al. 2009.

The outputs from macs were used to seed forward-time simulations using the first h haplotypes from a coalescent simulation. The forward-time simulation used here is based on a generic C++ library (Thornton, unpublished) previously used in Thornton, Foran, and Long (2013). The speed of the library compares favorably to existing forward simulations [1, 122], but has the advantage that new models are easily implemented by enabling simple code re-use via the C++ template mechanism. Haplotypes in macs are not sorted, so choosing the first h haplotypes is equivalent to randomly choosing h haplotypes. These h founding haplotypes were replicated r times and then r large “base” populations of n diploids each were generated by sampling with replacement from the initial coalescent simulations. A single site was assigned a positive selection coefficient and experimental evolution was simulated using forward-in-time simulations. The forward in time simulations were carried out with various population sizes (n), numbers of founder haplotypes (h), numbers of replicate populations evolved (r), numbers of generations of experimental evolution (g), and selection coefficients (s). Note that, due to the lack of population structure in these populations, the actual population size should be equal to the effective population size; in a study of real data, the effective population size would be more comparable to the n used here because of non-random mating and population size fluctuations. The SNP under selection, or the causative site (CS), was always the centermost SNP in the region. The selection scheme was codominant with fitnesses 1, $1 + s/2$, and $1 + s$, where s is the selection coefficient on the CS. CSs thus followed the same distribution of initial allele frequencies as all other SNPs in the simulation, consistent with a SNP that is initially neutral, but that is selected upon following a change in environmental conditions. Linkage disequilibrium between SNPs is initially an outcome of the neutral Wright-Fisher sampling process used to generate the h founder haplotypes, and subsequently determined by the details of the forward-in-time simulation. The forward simulations assume no further mutation in the region, and the recombination rate used was 0.025 per diploid per generation assuming that the 1 megabase region is 5% of a “typical” 20 megabase chromosome whose total recombination rate per generation is 0.5.

To find the experimental parameters best suited to E&R CR detection and CS localization, we arranged our simulated genomic regions as one would arrange a set of populations for experimental evolution. In each replicate simulation, we set up an equal number of experimental and control populations (in which the selection coefficient at the CS is equal to zero), all containing individuals with the same genomic region. Haplotypes were generated based on the initial allele frequencies, and individuals carrying these haplotypes were created. A forward-in-time simulation was then used to keep track of the movement of haplotypes over time with recombination and selection applied. Allele frequencies were calculated and recorded at 100, 500, and 1000 generations. Although a true E&R experiment would have allele frequency estimation errors that are a complex function of number of individuals sequenced, library preparation methods, average sequence coverage, and variation in sequencing coverage, we chose to simulate the best-case scenario where all allele frequencies are estimated without error. Thus, our estimates of power are likely somewhat optimistic. We performed simulations that varied in population size (n diploids), number of founder haplotypes ($h =$ twice the number of founding diploids), number of replicate populations (r), and selection coefficient (s) on the CS. In total, 500 replicate simulations were performed under each of 840 possible combinations of experimental parameters (Θ) (Table 1). Combinations in which h was larger than n were not simulated because such populations would presumably closely resemble populations in which h was reduced to the level of n. At each number of generations (g) in which allele frequencies were recorded, a modified t-statistic was calculated on arcsine square-root transformed SNP frequencies using 2, 5, 10, 15, or 25 replicate populations. This empirical Bayesian t-statistic [13] indicates the degree to which the allele frequencies of SNPs in the selected populations have diverged from the same allele frequencies in the control populations. It differs from a standard t-statistic in that it is not infinity in the case where a SNP of interest is fixed in all experimental replicates and lost in all control replicates. The expression for the modified t-statistic is

$$t = \frac{x_{1-?x_2}}{\sqrt{\frac{1-w}{r}\left(v_1 + v_2\right) + \frac{2w}{r}\acute{v}}}$$

where `x1` and `x2` are the mean allele frequencies across all experimental replicates in selected and control treatments, respectively, `v1` and `v2` are the respective variances, `r`is the number of replicates, and `w` $= 0.1$. $\acute{v}$ is the average within treatment variance in allele frequency averaged over all SNPs in the region and both treatments. $\acute{v}$ is then an empirically motivated Bayesian prior on allowable variances in allele frequencies, and has the effect of stabilizing the denominator of the t-statistic. This is especially important in experimental evolution experiments in which a SNP could differentially fix in the experimental versus control replicates purely due to drift alone and thus be associated with a traditional t-statistic of infinity.

p-values were calculated from the modified t-statistic using the pt function in R, using a 2-tailed method (see Sup. Fig. 4 for examples). A 2-tailed t-test was used in order to avoid making a priori assumptions about the nature of the 2 alleles involved at any given locus: since either allele could be beneficial in theory, it is not reasonable to assume that only the allele whose frequency is being tracked could be beneficial. Degrees of freedom were considered to be $2r^{-2}$ (because there are control and experimental treatments, the total number of replicates is twice the number per treatment). The threshold for significance was set independently for each Θ by calculating the false positive rate for every whole-number power of 10 from 10^{-1} to 10^{-14} and choosing the most lenient threshold with an acceptable false positive rate (see results section, "The False Positive Rate"). Power was calculated by finding the fraction out of 500 times that 1) at least one SNP was significantly diverged in a region of interest and 2) a secondary condition was met. These secondary conditions included having the most significant marker (MSM) in the region be the CS (exact location power), having the MSM be within 10kb of the CS (within 10kb power), having the CS be among the top 25 MSMs (top 25 power), and having the CS's LOD (logarithm base 10 of odds) score be within two of the MSM's LOD score (within 2 LOD power). The power to fulfill the first condition without regard for a second condition was termed “CR detection power”. This diversity of methods of SNP localization allowed us to determine which method would be most reliable under any particular set of experimental parameters. The CR detection false positive rate was determined by finding the fraction of cases in which a region with an s of zero contained at least one significant SNP. The distance from the MSM to the CS (MSM-CS distance) and the rank of the CS’s p-value compared to the other SNPs in its region (CS rank) were calculated in every replicate of the 500 replicate simulations per Θ in order to analyze the distribution of significant SNPs across parameter values.

An additional 18,000 replicate genome regions were generated in macs and used to seed forward-in-time simulations with no selection under all of our Θ where $s = 0$. 10,000 replicate neutral simulations were performed for each Θ where $s = 0$. These additional replicates were used to more accurately calculate the false positive CR detection rate. This high level of replication was only required for the calculation of false positive rates because the maximum allowable false positive rate is too small (˜1/2000) to be accurately measured with only 500 replicate experiments.

2.7.2 The Data

This simulation produced 2,205,000 semi-independent experimental results. There are 500 pure replicates of each possible permutation of five distinct experimental parameters’ values (Table 1) – number of replicate populations (r), number of haplotypes in the base population (h), population size (n), selection coefficient at the selected locus (s), and number of generations of selection (g). The resulting data sets are not completely independent because the coalescent simulation used to generate the 1Mb regions used here was only run 500 times (10,000 where $s = 0$), and the resulting 500 (or 10,000) genome regions were re-used for the 500 (or 10,000) replicate experiments for each parameter combination. Further, for any particular combination of n, r, s, and h, the three g values simulated were not entirely independent because the data associated with larger g values was derived from continuing the forward simulations of the smaller g simulations. Values were chosen based on the specifics of the variable: h, n, and g values were chosen based on the levels historically used in experiments of this type. The r values were chosen based on the level of replication commonly used in experimental evolution as well as the level of replication required for high power. Experimental evolution of sexual organisms is usually carried out with 5 or fewer replicate populations due to the difficulty of rearing large numbers of populations, but, especially in genomics, statistical significance is difficult to achieve with low r values because the large number of independent comparisons require a strict significance threshold. s values were chosen based on the minimum selection strength necessary for selection to have an effect sufficiently stronger than genetic drift to produce a measurable change in allele frequencies: when the selection coefficient at a SNP is less than $\sim 1/(2n)$, genetic drift is a more powerful force than selection (Crow and Kimura, 1970, p 425). The s values were thus chosen to cover a range of possible sizes, from an s considerably smaller than $1/(2n)$ to an s larger than $1/(2n)$.

2.7.3 Data Availability

The simulation code and all data and the necessary code to recreate the data are available online at http://www.molpopgen.org/Data, as is a commented copy of the scripts used to calculate p-values, power, and other statistics. Macs is available at http://code.google.com/p/macs/.

2.8 Acknowledgements and funding information

This work was supported by the National Institutes of Health grant R01-RR024862 and R01-GM085251 to ADL and NIH grant R01-GM085183 to KT.

2.9 Tables

	Term	Values used in simulation	Description
r	Number of Replicates	2,5,10,15,25	The number of independent experimental populations that are used in each trial. There are an equal number of control populations.
n	Population Size	100, 250, 500, 1000	The number of diploid individuals that successfully reproduce every generation.
h	Number of Haplotypes	4, 32, 100, 500	The number of haplotypes present in each population at the start of each experiment. A population originally derived from one male and one female would have 4 haplotypes.

g	Number of Generations	100, 500, 1000	The number of generations of selection that both the control populations and the selected populations have undergone before allele frequency calculation.
s	Selection Coefficient	0, 0.0005, 0.005, 0.05, 0.1, 0.2	The strength of selection at the causative locus in a particular genomic region.
Θ	Parameter Combination		The particular set of r, n, h, g, and s used in each set of 500 simulations.
MSM	Most Significant Marker		The SNP that was found to have most significantly diverged in a particular simulation
CS	Causative SNP		The SNP that was selected upon in a particular simulation.
	CR Detection Power		The fraction of studies of a particular Θ that found at least one significantly diverged SNP
	Exact Location Power		The fraction of studies of a particular Θ in which the MSM is the CS.
	Within 10kb Power		The fraction of studies of a particular Θ in which the MSM is within 10kb of the CS.

Top 25 Power	The fraction of studies of a particular Θ in which the CS is one of the 25 most significantly diverged SNPs in the region.
Within 2 LOD Power	The fraction of studies of a particular Θ in which the CS is within a 2 LOD drop of the MSM
Total Power	The fraction of studies of a particular Θ in which the CR is detected and the CS is localized according to one of the CS localization methods above. In other words, CR Detection Power * Localization Power
MSM-CS Distance	The physical distance between the MSM and the CS.
CS Rank	The significance rank of the CS when compared to all other SNPs in the region

Table 2.1: Useful terms

2.10 Chapter 1 Figures

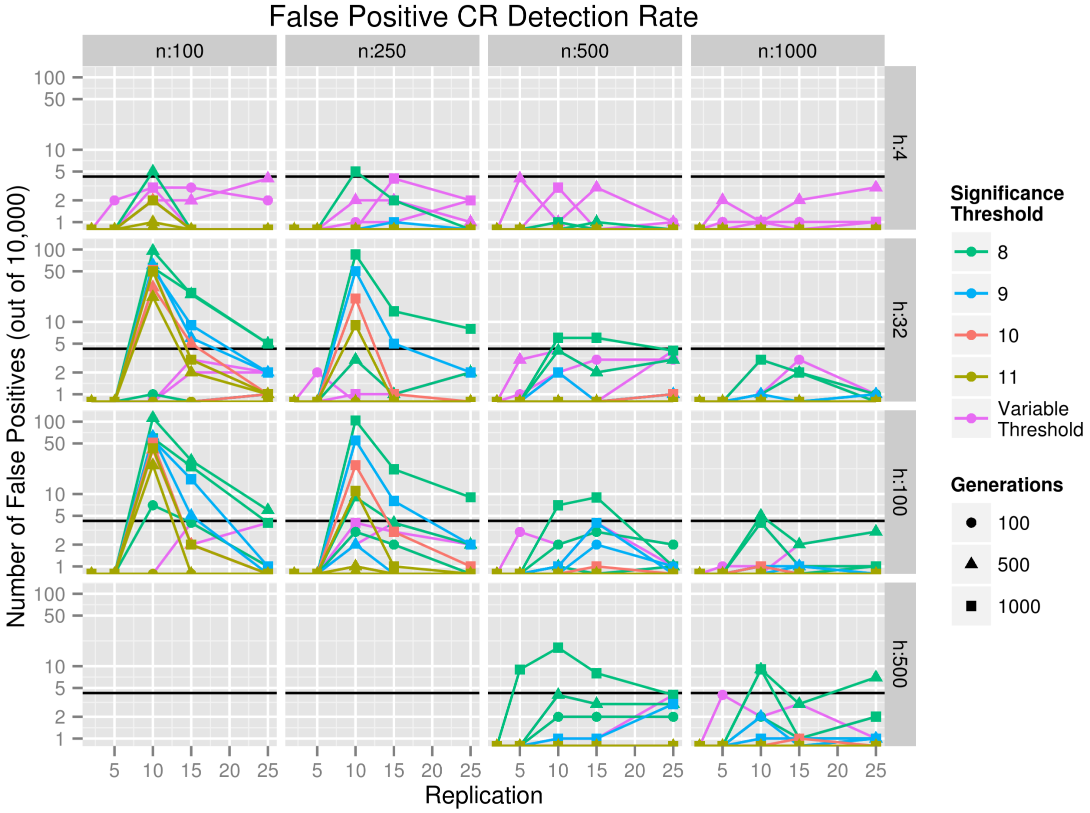


Figure 2.1: The false positive CR detection rate versus replication. This plot depicts the fraction out of 10,000 cases in which a region containing no CS contained at least one significantly diverged SNP for four different per SNP -log10(p-value) thresholds. The black line indicates the maximum allowable false positive rate $(4/10,000)$. n represents population size, while h represents the number of founder haplotypes. The variable significance threshold used in our later power analysis is also included for comparison. When two lines overlap, the line representing a more strict significance threshold is the visible line.

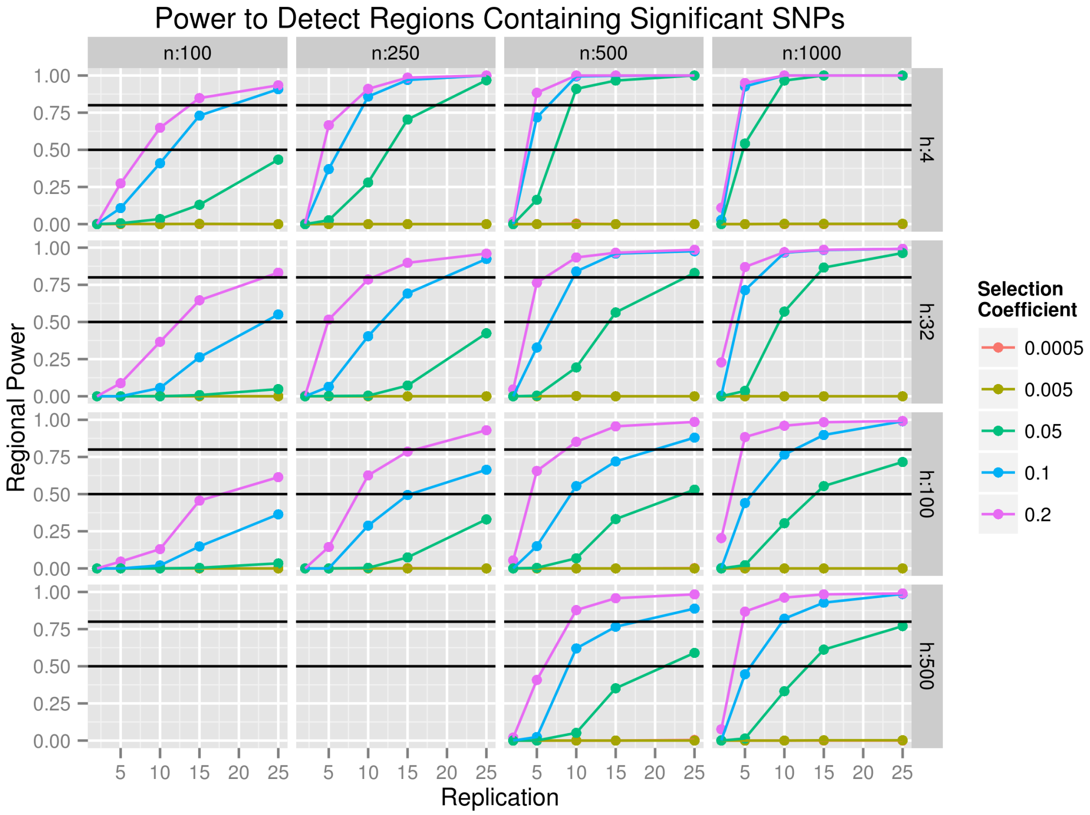


Figure 2.2: CR detection power. This plot depicts the power to detect regions containing one or more significantly diverged SNPs. The Θ in which all of the following are true simultaneously: $r = 10$, $n = 100$, $h = 100$, and $g = 500$ or 1000 would be omitted due to high false positive rates, but only $g = 100$ is shown for ease of viewing. n represents population size. h represents the number of founder haplotypes. The p-value threshold for significance was determined for each Θ by finding the most lenient threshold that sufficiently limited false positives. Each point represents 500 independently replicated sets of populations. All lines that are not visible overlap with $s = 0.005$. The black lines indicate power levels of 50% and 80%.

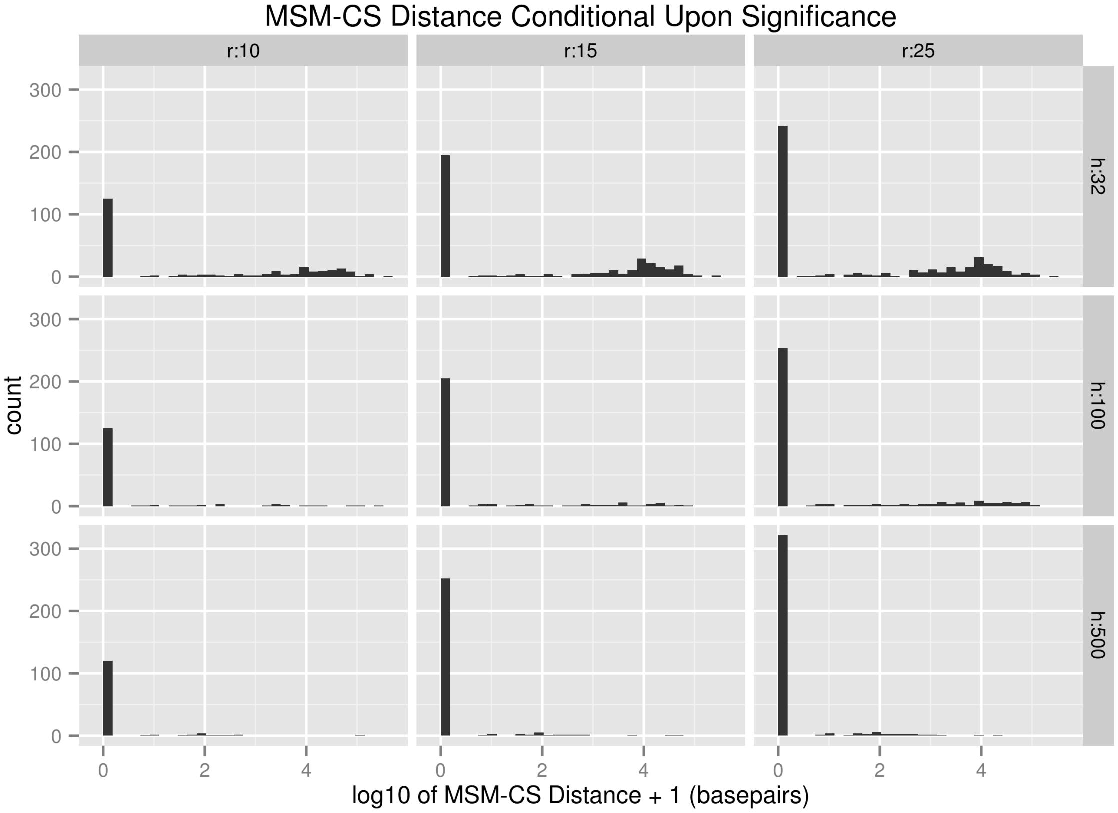


Figure 2.3: A histogram depicting the distribution of the distance from the MSM to the CS (MSM-CS distance) after 500 generations of selection with 500 individuals per population and a selection coefficient at the CS of 0.05 in all cases where the MSM was significant. Variation in population size is not shown because its effects are similar to variation in replication. The MSM-CS distance is shifted by one base pair so that MSM-CS distances of 0 are visible after logarithmic transformation. The count refers to the number of pure replicates out of 500 that fell into a given range. Note the increase in low-MSM-CS-distance hits due to selective sweeps when h is low.

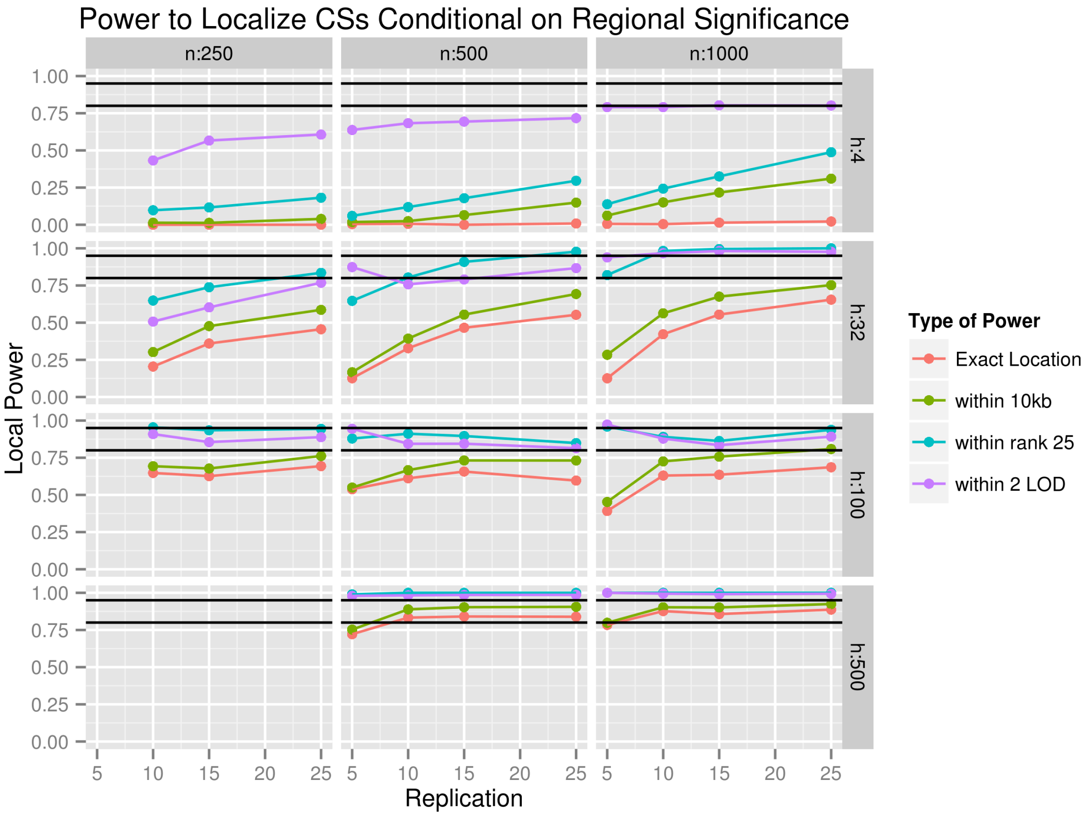


Figure 2.4: Localization power conditional on regional significance. In other words, the fraction of all significant SNP containing regions in which the CS could be either exactly identified or localized to a small number of candidate SNPs. For clarity, only cases where $s = 0.1$ are shown, but similar patterns occur for $s = 0.05$ and $s = 0.2$. This set of plots shows the fraction of experiments that correctly identified the location of the CS out of all experiments in which at least one SNP was significant. Exact location power refers to cases in which the MSM is the CS, top 25 power refers to cases in which the CS is among the 25 most significant SNPs, within 10kb power refers to cases in which the MSM is within 10kb of the CS, and within 2 LOD power refers to cases in which the CS is within 2 LOD of the MSM. The population size is represented by n, while the number of founder haplotypes is represented by h. Non-visible within 10kb power points overlap with top 25 power points. The black lines indicate 80% power and 95% power.

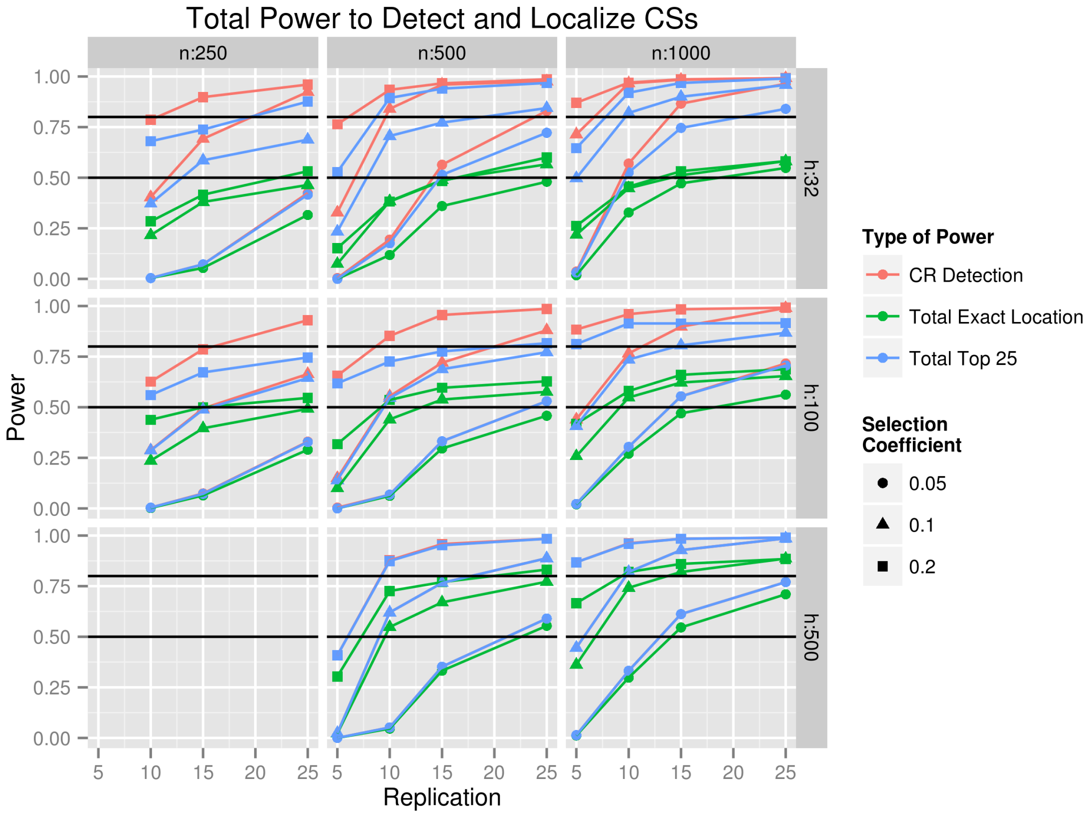


Figure 2.5: Total power to detect and localize CSs. The ability to detect a CS-containing region and either correctly identify the exact location of a CS or decrease the number of candidate loci to a manageable number after 1000 generations with a selection coefficient at the CS of 0.05, 0.1, or 0.2. In other words, the fraction of all simulations in which a region contains a significant SNP and one of two methods of detecting a CS is successful: the MSM is the CS (Total Exact Location Power) or the CS is one of the 25 most significantly diverged SNPs in the region (Total Top 25 Power). Also shown is the CR Detection Power, which is the fraction of regions that contained at least one significant SNP. Other measurements of power are excluded for clarity. By design, all total powers listed here must be lower than the CR detection power. The black lines indicate 50% and 80% power. Where CR detection power is not visible on the plot, it overlaps with total top 25 power.

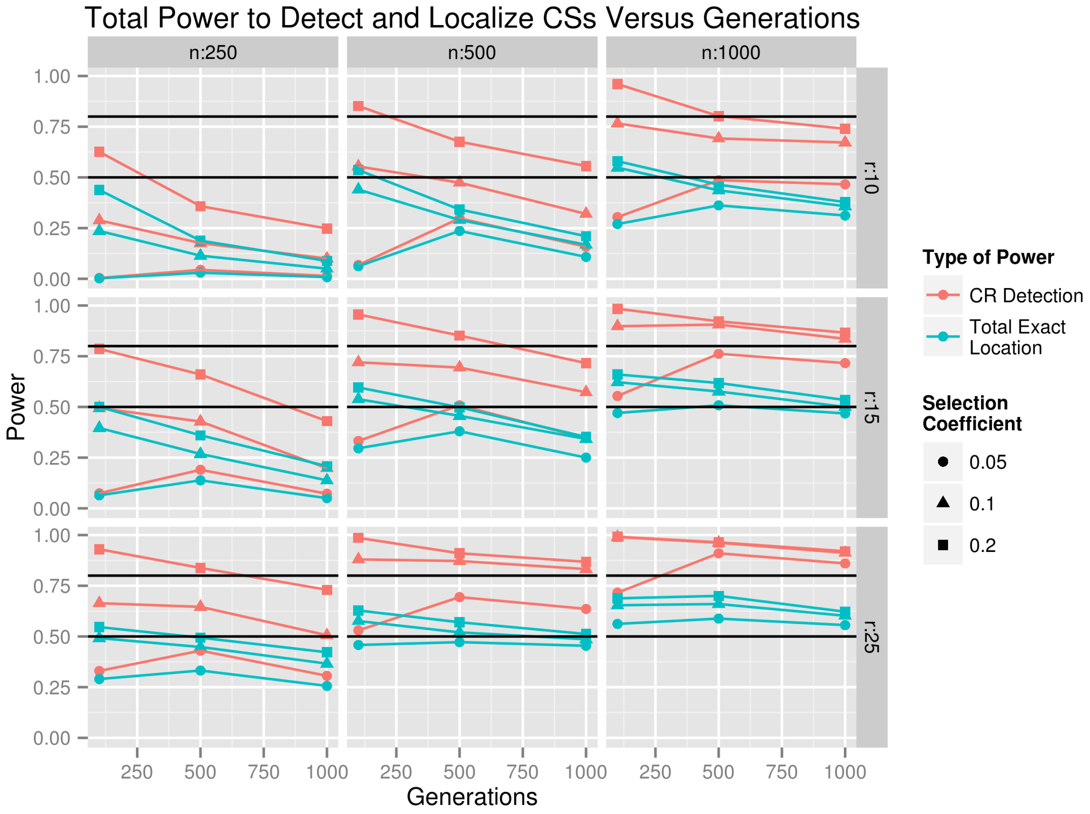


Figure 2.6: The total power to detect and localize SNPs when $s \geq 0.05$ and $h = 100$ versus the number of generations of selection. For simplicity, only CR detection power and total exact location power are shown. Variation in h is not shown because there are no visible interactions between h and g. The black lines indicate 50% and 80% power.

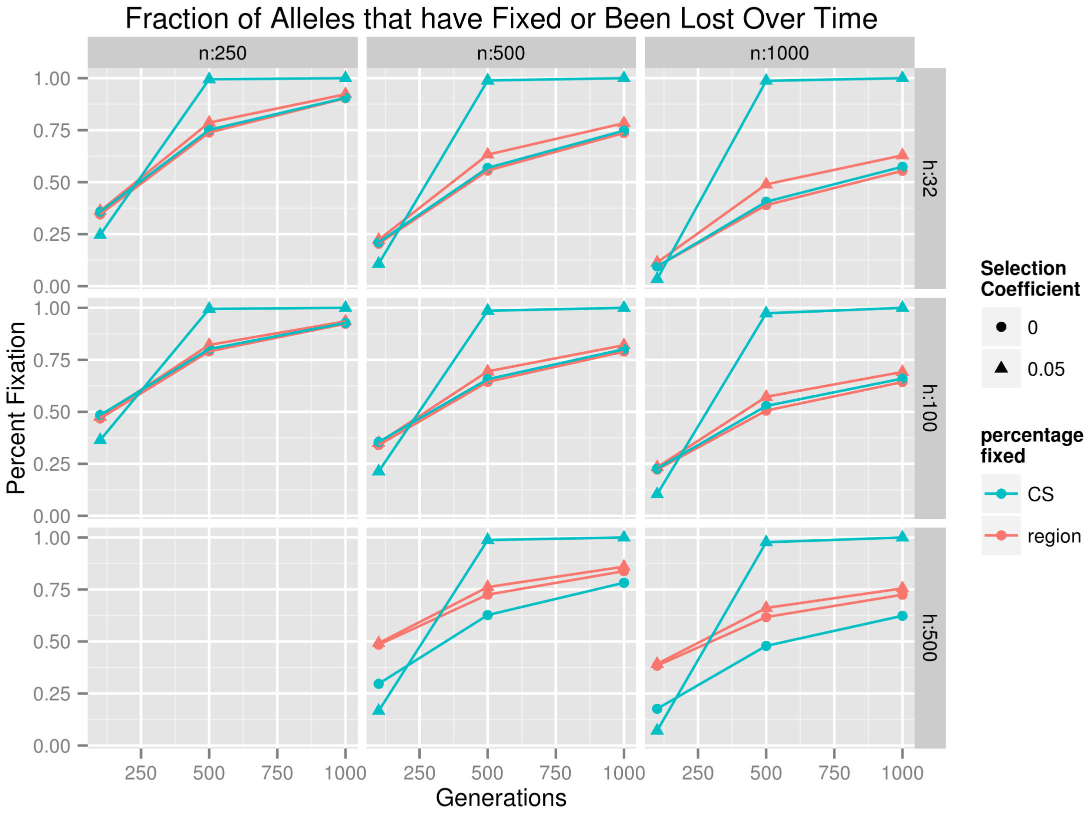


Figure 2.7: The fraction of alleles that have fixed versus number of generations of selection. The blue line indicates the fraction of all CS alleles that reached an allele frequency of 1 or 0, while the red line indicates the fraction of all alleles in the region that reached an allele frequency of 1 or 0. Note that this plot makes use of all available replicates for every Θ. Circles represent $s = 0$, while triangles represent $s = 0.05$. Regions with an s of 0 have no CS, but the fixation frequency of the centermost SNP is included (the blue lines) for comparison with the CS when $s = 0.05$.

Chapter 3

3.1 Article

Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage

Mahul Chakraborty1*, James G. Baldwin-Brown1*, Anthony D. Long1,2, J.J. Emerson1,2†

* These authors contributed equally

† To whom correspondence should be addressed: jje@uci.edu

1 Department of Ecology and Evolutionary Biology, University of California Irvine, Irvine, California, United States of America

2 Center for Complex Biological Systems, University of California Irvine, Irvine, California, United States of America

3.2 Preface

This chapter was originally published in *Nucleic Acids Research* under the title "Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage". It is reprinted here in its original form. Anthony Long and J.J. Emerson advised all aspects of this publication. Mahul Chakraborty performed the whole genome assemblies using PacBio only assemblers, while I performed the whole genome assemblies using hybrid (that is, PacBio plus Illumina data) assemblers. Dr. Chakraborty was the primary author of the *quickmerge* program, but the concept for the program was generated in collaboration amongst the four authors; I was the author of the wrapper program that executes the *quickmerge* pipeline. Dr. Chakraborty and I contributed equally to the writing of the document and creation of the figures.

3.3 Abstract

Genome assemblies that are accurate, complete, and contiguous are essential for identifying important structural and functional elements of genomes and for identifying genetic variation. Nevertheless, most recent genome assemblies remain incomplete and fragmented. While long molecule sequencing promises to deliver more complete genome assemblies with fewer gaps, concerns about error rates, low yields, stringent DNA requirements, and uncertainty about best practices may discourage many investigators from adopting this technology. Here, in conjunction with the platinum standard Drosophila melanogaster reference genome, we analyze recently published long molecule sequencing data to identify what governs completeness and contiguity of genome assemblies. We also present a hybrid meta-assembly approach that achieves

remarkable assembly contiguity for both Drosophila and human assemblies with only modest long molecule sequencing coverage. Our results motivate a set of preliminary best practices for obtaining accurate and contiguous assemblies, a “missing manual” that guides key decisions in building high quality de novo genome assemblies, from DNA isolation to polishing the assembly.

3.4 Introduction

De novo genome assembly is the process of stitching DNA fragments together into contiguous segments (contigs) representing an organism's chromosomes [146]. Until recently, genomes were often assembled using fragments shorter than 1,000 bp. However, such assemblies tend to be highly fragmented when they are generated using sequencing reads shorter than common repeats [146, 125, 24, 12]. Paired end short reads from different sized longer inserts can improve contiguity, but uncertainty of fragment length and the lack of sequence between the insert ends makes resolving many repetitive structures challenging [93]. Longer reads can circumvent this problem, even when such reads exhibit errors rates as high as 20% [93, 100, 124]. Importantly, error-prone reads can be corrected, provided there is sufficient coverage and the errors are approximately uniformly distributed. Single molecule sequencing, like that offered by Pacific Biosciences (PacBio), meets these criteria with reads that are routinely tens of kilobases in length [93, 92, 82, 133]. While PacBio sequences have high error rates (˜15%), errors are nearly uniformly distributed across sequences [93]. With sufficient coverage, these sequences can be used to correct themselves [36]. Assemblies using such correction are referred to as PacBio only assembly [20]. Alternatively, hybrid assembly can be performed using a combination of noisy PacBio long molecules and high quality short reads (e.g. Illumina) [133, 94].

Recently, the value of long molecule sequencing has been definitively demonstrated with the release of several high quality reference-grade genomes assembled from PacBio sequencing data [82, 20, 59]. Indeed, the Drosophila PacBio assembly closed gaps in the reference genome assembly [20], which is often considered the most contiguous metazoan genome assembly. Despite these successes, shepherding a genome project through the process of DNA isolation, sequencing, and assembly is still a challenge, especially for research groups for whom genomes are a means to another goal rather than the goal itself. For example, because high quality genome assembly relies upon long sequencing reads to bridge repetitive genomic regions [100, 97, 25, 126] and high coverage to circumvent read errors [12, 124, 36], the stringent DNA isolation requirements (size, quantity, and purity) for PacBio sequencing [82] intended for genome assembly are higher than those typically employed. Moreover, at present, the low average read quality produced by PacBio sequencing causes coverage requirements to be at least 50-fold [93, 20, 59]. This requirement, combined with the comparatively expensive sequencing, makes striking the right balance between price and assembly quality important. Exacerbating the problem is the fact that rediscovering the optimal approach for a genome project is itself expensive and time consuming. As a consequence of these challenges and uncertainties, many groups may opt out of a long molecule approach, or worse, sink scarce resources into an approach ill-suited for their goals because the consequences of many decisions involved in long molecule sequencing projects have not been synthesized.

In order to optimize a strategy for genome assembly we investigated the consequences of sample preparation (i.e. DNA isolation, quality control, shearing, library loading, etc.), assembly strategies, and properties of the data (i.e. read quality, length, and read filtering). We first evaluate strategies for assembling PacBio reads, and how they perform with differing amounts of sequence coverage. Then, we assess the contribution of read length and read quality to assembly contiguity. We also introduce *quickmerge*, a simple, fast, and general meta-assembler that merges assemblies to generate a more contiguous assembly. Additionally, we describe the protocols, quality-control practices, and size selection strategies that consistently yield high quality DNA reads required for reference grade genome assemblies. Our strategy is flexible enough to yield high quality assemblies using as little as 25X long molecule coverage or as much as >100X.

3.5 Materials and methods

3.5.1 Preparing high quality DNA library for long reads

Obtaining high quality, high molecular weight (HMW) genomic DNA

We used Qiagen's Blood and Cell culture DNA Midi Kit for DNA extraction. As single molecule technologies (PacBio and Oxford Nanopore) do not require any sequence amplification step, a large amount of tissue is required to ensure enough DNA for library preparations that opt for no amplification (as is standard for genome assembly sequencing). For flies, 200 females or 250 males flies is sufficient for optimal yield (40-60µg DNA) from a single anion-exchange column. For other organisms, number of individuals need to be adjusted based on the tissue mass. A good rule of thumb is to keep the total amount of input tissue 100-150mg for optimal yield from each column.

To extract genomic DNA, 0-2 days old flies were starved for two hours, flash frozen in liquid nitrogen, and then ground into fine powder using a mortar and pestle prechilled with liquid nitrogen. The tissue powder was directly transferred into 9.5 ml of buffer G2 premixed with 38µl of RNaseA (100mg/ml) and then 250 µl (0.75AU) of protease (Qiagen) was added to the tissue homogenate. The volume of protease can be increased to 500 µl (1.5AU) to reduce the time of proteolysis. The tissue powder was mixed with the buffer by inverting the tube several times, ensuring that there were no large tissue clumps present in the solution. The homogenate was then incubated at 50[2DA?]C overnight with gentle shaking (with 500µl protease, this incubation time can be reduced to 2 hours or less).

The next day, the sample was taken out of the incubator shaker and centrifuged at 5000xg for 10 minutes at 4[2DA?]C to precipitate the tissue debris. The supernatant was decanted into a fresh 15ml tube. The little remaining particulate debris in the tube was removed with a 1 ml pipette. The sample was then vortexed for 5 seconds to increase the flow rate of the sample inside the column and then poured into the anion-exchange column. The column was washed and the DNA was eluted following the manufacturer's protocol. Genomic DNA was precipitated with 0.7 volumes of isopropanol and resuspended in Tris buffer (pH 8.0). For storage of one week or less, we kept the DNA at 4[2DA?]C to minimize freeze-thaw cycles; for longer storage, we kept the DNA at -20[2DA?]C.

Shearing the DNA

1.5" blunt end needles (Jensen Global, Santa Barbara, CA) were used to shear the DNA. The needle size can be varied to obtain DNA of different length distribution: 24 gauge needles produces a size range of 24-50 kb. To obtain larger fragments, <24 gauge needles need to be used. For the DNA we have sequenced, up to 200µg of high molecular weight raw genomic DNA was sheared using the 24 gauge needle (Fig. 1). Additionally, we have also sheared DNA with 21, 22, and 23 gauge needles to demonstrate the size distribution they generate (supplementary Fig. 1). In brief, the entire DNA solution is drawn into a 1mL Luer syringe and dispensed quickly through the needle. This step is repeated 20 times to obtain the desired distribution of fragment sizes.

Quality Control using FIGE

We verified the size distribution of unsheared and sheared genomic DNA using field inversion gel electrophoresis (FIGE), which allows separation of high molecular weight DNA. The DNA is run on a 1% agarose gel (0.5x TBE) with a pulse field gel ladder (New England Biolabs, Ipswich, MA). The gel is run at 4[2DA?]C overnight in 0.5 x TBE. To avoid temperature or pH gradient buildup, a pump is used to circulate the buffer. The FIGE was run using a BioRad Pulsewave 760 and a standard power supply with the following run conditions:

Initial time A: 0.6s, Final time B: 2.5 s, Ratio: 3, Run time: 8 h, MODE: 10, Initial time A: 2.5s, Final time B: 8s, Ratio: 3, Run time: 8 h, MODE: 11, Voltage: 135 V.

Library preparation

The needle sheared DNA is quantified with Qubit fluorometer (Life Technologies, Grand Island, NY) and NanoDrop (Thermo Scientific,,Wilmington, DE). Following quantification, 20 µg of sheared DNA was optionally run in four lanes of the Blue Pippin size selection instrument (Sage Science, Beverly, MA) using 15-50 kb as the cut-offs for size selection (Fig. 1). This optional size selection step increases final library yield at the cost of requiring more input DNA. This size selected DNA is then used to prepare a SMRTbell template library following PacBio's protocol. A second round of size selection is performed on the SMRTbell template using a 15-50 kb cutoff to remove the smaller fragments generated during the SMRTbell library preparation step (Fig. 1). The second step minimizes the number of DNA fragments less than 15kb subjected to sequencing.

3.5.2 DNA Sequencing

PacBio sequencing was conducted to establish length distributions (D. simulans Fig. 2a) and evaluate the impact of library preparation on quality (Fig. 3), and was performed at the UCI High Throughput Core Facility using DNA isolated using the protocol described above. We note that the D. simulans reads were not used for assemblies reported here – all of our assemblies are constructed with publicly available *D. melanogaster* [82] and Homo sapiens data [133]. We sequenced one SMRTcell of Drosophila genomic DNA with the following conditions to obtain sequences with standard quality and length distribution: 10:1 polymerase to template ratio, 250 pM template concentration, and P6C4 chemistry. The movie time and other conditions were standard for RSII P6C4 chemistry. To demonstrate the tradeoff between yield and quality, we sequenced one SMRTcell each for polymerase:template ratios of 40:1,80:1,100:1 with template concentration held constant at 200pM, and one SMRTcell each with 300pM and 400pM template concentration with the polymerase:template ratio being held constant at 10:1.

3.5.3 PacBio only Assembly

For PacBio sequences, the assembly pipeline is divided into three parts: correction, assembly, and polishing. Correction reduces the error rate in the reads to 0.5-1% [20], and is necessary because reads with a high (˜15%) error rate are extremely difficult to assemble [126]. Correction is facilitated by high PacBio coverage, which allows the error corrector to successfully 'vote out' errors in the PacBio reads. For self correction, we used the *PBcR* pipeline [20] as implemented in *wgs8.3rc1* which, by default, corrects the longest 40X reads. The second step involves assembling the corrected reads into contigs. We used the *Celera* assembler [126], included in the same wgs package, for assembly. A third optional step involves polishing the contigs using *Quiver* and *Pilon* [34, 159], which brings the error rate down to 0.01% or lower. All of the assemblies described in this paper were generated with the same *PBcR* command and spec file (commands and settings, Supplementary materials).

For PacBio only assembly of *D. melanogaster* ISO1 sequences, we used a publicly available PacBio sequence dataset which was generated using the standard P5C3 chemistry. A complete description of this data is available in Kim et al. [82]. We chose the *D. melanogaster* dataset for our experiments and simulations because *D. melanogaster* is widely used in genetics and genomics research and its reference sequence (release 5.57,http://www.fruitfly.org) is one of the best, if not the best, eukaryotic multicellular genome assemblies in terms of assembly contiguity. This is true for both the PacBio generated assembly (21Mb contig N50) [82] and the Sanger assembly (23Mb scaffold N50) of ISO1. The remarkable contiguity of these assemblies becomes more tangible when the theoretical limits of *D. melanogaster* chromosome arms' lengths are considered (20): N50 of both assemblies lie very close to the theoretical maximum N50 (˜28Mb). This high quality assembly serves as a reference for evaluating assemblies presented here.

We evaluated assembly qualities using the standard assembly statistics (average contig size, number of contigs, assembled genome size, N50, etc.) using the *Quast* and *GAGE* [66, 138] packages.

3.5.4 Hybrid Assembly

PacBio only assembly of high error, long molecule sequences depends upon redundancy between the various low quality reads to 'vote out' errors and identify the true sequence in the sequenced individual. An alternative approach to this problem is to use known high quality sequencing reads to correctly call the bases in the sequence, and then to use PacBio reads to identify the connectivity of the genome. In order to achieve the best possible assembly results, we tested several different hybrid assembly pipelines before choosing *DBG2OLC* (https://arxiv.org/abs/1410.2801, https://sites.google.com/site/dbg2olc/) and *Platanus* [75]. In our early tests, the next highest performing hybrid assembler, a combination of *ECTools* (https://github.com/jgurtowski/ectools) and *Celera*, achieved a highest N50 of 616kb in *Arabidopsis thaliana* using 19 SMRT cells of data; in contrast, using 20 SMRT cells of the same data, the *DBG2OLC* and *Platanus* pipeline produced an N50 of 4.8Mb. We aslo tested the alternative error corrector, *LoRDEC* [137], along with the Celera assembler, but found that the *LoRDEC*-corrected *Celera* assembly of our standard *D. melanogaster* dataset (26X of PacBio data and 67.4X of Illumina data [101]) produced an NG50 of only 109KB. Consequently we adopted *DBG2OLC* as our choice for hybrid assembly. We were not able to exhaustively test all hybrid error correction approaches of PacBio reads followed by overlap assembly and acknowledge that other tools that may operate quite differently (e.g. LSC [9]) could potentially lead to further improvements in the assembly. Using the standard 67.4X of Illumina data discussed above and 26X of PacBio data, we compared *DBG2OLC* runs using three different De Bruijn graph assemblers: *SOAP* [112], *ABySS* [147], and *Platanus*. The NG50s for the three assemblies were, respectively, 2.43Mb, 0.167Mb, and 3.59Mb. Based on this result, we chose to use *Platanus* for the remainder of the assemblies.

We used the pipeline recommended by *DBG2OLC* to perform hybrid assemblies. In this pipeline, we used *Platanus* to perform De Bruijn graph assembly on the Illumina reads. We used 8.36 Gb (67.4X) of Illumina sequence data of the ISO1 *D. melanogaster* inbred line generated by the DPGP project [101] to generate a De Bruijn graph assembly using *Platanus*. We used *DBG2OLC* to align our PacBio reads to the De Bruijn graph assembly to produce a 'backbone', then, according to the *DBG2OLC* standard pipeline, used the backbone to generate the consensus using the programs *BLASR* [32] and *PBDagCon* (https://github.com/PacificBiosciences/pbdagcon). As with the PacBio only assemblies above, we evaluated assembly quality using the *Quast* and *GAGE* packages.

3.5.5 Assembly merging

Hybrid assembly and PacBio assembly were merged using a custom C++ program called *quickmerge* (Fig. 4A, available at https://github.com/mahulchak/quickmerge). The program takes two fasta files (containing contigs from a PacBio only assembly and contigs from a hybrid assembly) as inputs and splices contigs from the two assemblies together to produce an assembly with higher contiguity. As the two assemblies used for merging come from the same genome, gaps in one assembly can be bridged using corresponding sequences from the other assembly The first stage of the assembly merging process involves correctly aligning the corresponding sequences (contigs), which in the second stage are exchanged at the sequence gaps so that the part of the sequence with the gap is replaced with a contiguous sequence from the other assembly. The program *MUMmer* [96] is used to find the correct alignment between the assemblies and assembly merging is handled by *quickmerge*.

First, the program *MUMmer* [96] is used to compute the unique alignments between the contigs from the two assemblies, one of which is used as the reference, or donor, assembly and the other is used as the query, or acceptor, assembly. Distinction between the two assemblies is important because, as described below, the user may choose the more reliable, i.e. with fewer errors, of the two assemblies to bridge gaps in the other assembly. Accurate merging occurs when true correspondance between two sequences is high; conversely, pairing between incorrectly matching regions leads to incorporation of incorrect sequences. Hence, identification of the correct pairing is necessary for error-free sequence merging. Presence of repeats may complicate the situation, but the problem can largely be overcome if the two aligned sequences containing repeats come from the same genome and only the unique best alignments are considered. To obtain the unique best alignment between the reference and the query assembly, spurious matches introduced by gene duplications and repeats are removed using the delta-filter utility (with –r and –q options) of the *MUMmer* package.

Following the repeat filtering step, the alignments are partitioned using a scoring metric called high confidence overlaps (HCO) (Fig. 4B). The program identifies HCOs by dividing the total alignment length between contigs by the length of unaligned but overlapping regions of the alignment partners (Fig. 4B). The metric was chosen under the assumption that the length of the overlapping but unaligned portion between the two sequences relative to the length of the overlapping and aligned parts is high for two unrelated sequences. After the alignment partitioning is done based on a HCO cutoff, only the contig alignments above the HCO cutoff are kept for assembly merging. For fly assemblies, we found that an HCO value of 1.5 was an appropriate default for assembly merges. This cutoff can be increased further, as we did for merging human assemblies. The tradeoff is that increasing HCO cutoff will gradually deplete the pool of matching alignments, thereby leading to a reduction in merging events. Thus, the "HCO" parameter controls merging sensitivity at the cost of increased false positives: the higher the HCO parameter value, the more stringent is the cutoff for HCO selection.

The next step involves searching and ordering the contigs that will be merged. To accomplish that, by default *quickmerge* assigns nodes in the HCO alignment graph with even higher HCO values (>5.0) and reference sequences exceeding a length cutoff (1Mb) as anchor nodes. The high HCO and the length cutoff are used here to ensure that subsequent searches for contigs for merged contig extension do not begin at spurious alignment nodes. Following the assignment of the anchor nodes, a greedy search is initiated on both the left and the right sides (5' and 3' of the reference contig) of the anchor node, in order to find the longest unbroken path through the HCO nodes. In other words, *quickmerge* looks for contigs that connect two adjacent HCO nodes in the graph and this process is continued until no contig can be found to connect two HCO nodes (e.g. a genomic region where both assemblies are broken). For the search, each contig is used only once to connect two HCO nodes, so once a contig from the HCO alignment pool has been used, it is removed from the alignment pool. Query contigs that are completely contained within a reference contig are also removed from the final merged assembly to prevent sequence duplication in the merged assembly.

In the final step, the ordered chain of contigs found in the previous step is joined by swapping portions of the reference assembly into the query assembly in a manner that maximizes retention of sequences from the reference assembly (Fig. 4A). Gap filling within the query assembly occurs as a byproduct of this replacement of sequences; in this way, the process resembles genome editing using homologous recombination.

For coverages of 40X, 53X, 62X, and 77X, merged assemblies were generated using the PacBio only assembly and their corresponding hybrid assemblies. For the 99x and 121x (all reads) SMRTcells datasets, the PacBio only assemblies were merged with the hybrid assembly obtained from the 77X SMRTcells dataset. All hybrid assemblies used for merging were generated without downsampling by read length or quality. The time to merge was limited only by the time required to run *MUMmer*, as *quickmerge* runs in less than 30 seconds on Drosophila-sized genomes, and requires less than 2GB of memory.

3.5.6 Downsampling

We used a number of different downsampling schemes on the *D. melanogaster* data: first, we randomly downsampled the data by drawing a random set of SMRTcells of data from the entire set of 42 SMRTcells; second, from those datasets, we downsampled the longest 50% and 75% of the reads. Next, we downsampled the *D. melanogaster* data to match the read length distributions of PacBio reads from a pilot Drosophila pseudoobscura genome project that was produced using a standard protocol without aggressive size selection (generously made available by Stephen Richards). Finally, we downsampled based on read quality to test the effect of read quality on assembly contiguity. Please see the supplementary text for more details.

3.6 Results

3.6.1 DNA isolation for long reads

As the remainder of the paper will show, read length is an important determinant of genome assembly contiguity. We identified simple and consistent method for isolation of large genomic DNA fragments necessary for PacBio sequencing to achieve long reads. The existing alternative method used for DNA isolation to generate the published PacBio Drosophila assembly involved DNA extraction by CsCl density gradient centrifugation and g-Tube (Covaris, Woburn, MA) based DNA shearing [82]. CsCl gradient centrifugation is a time-consuming method that requires expensive equipment that is not routinely found in most labs. Additionally, g-Tubes are expensive, require specific centrifuges, and are extremely sensitive to both the total mass of DNA input and to its length. We circumvented these problems by using a widely available DNA gravity flow anion exchange column extraction kit in concert with a blunt needle shearing method [63]. Because the DNA fragment size distribution is so important, field inversion gel electrophoresis (FIGE) is an essential quality control step to validate the length distribution of the input DNA (Fig. 1) (see Methods for details). Sequences generated from libraries constructed from this isolation method are comparable to or longer than the published Drosophila PacBio reads [82] (Fig. 2a). The length distribution of the input DNA can potentially be improved further by using wider gauge needles that generate even longer DNA fragments (supplementary Fig. 1).

3.6.2 Long read assembly

PacBio self correction has been used to assemble the *D. melanogaster* reference strain (ISO1) genome so contiguously that most chromosome arms were represented by fewer than 10 contigs [20]. This assembly was generated by using the PBcR pipeline [20] and 121X (15.8 Gb), or 42 SMRTcells' worth, of PacBio long molecule sequences [20]. However, currently, such high coverage may be too expensive for many projects, especially when the genome of the target organism is large. Consequently, we set out to determine how much sequence data is required to obtain assemblies of desired contiguity. We first selected reads from 15, 20, 25, 30, and 35 randomly chosen SMRTcells (40X, 53X, 62X, 77X, and 99X assuming a genome size of 130×106 bp – coverages calculated by dividing total bases of sequence data by total bases in genome) from the 42 SMRTcells of ISO1 PacBio reads [82]. Our sampling method was inclusive and additive: for example, to obtain 20 SMRTcells, we took the 15 previously randomly chosen SMRTcells and then added 5 more randomly selected SMRTcells to it. We then assembled these datasets using the PBcR pipeline. As shown in Fig. 5, the contig NG50 (NG50; G =130×106 bp) continues to improve across the entire range of coverage. At extremely high coverage (121×), the NG50 surges again, approaching the theoretical N50 limit of *D. melanogaster* genome [71]. Notably, despite the extreme contiguity of these sequences, we are still discussing complete contigs, not scaffolds with gaps.

3.6.3 Hybrid assembly

As Fig. 5 makes clear, PacBio only assembly leads to relatively fragmented genomes at lower coverage (Fig. 5), we investigated whether another assembly strategy could perform better with similar amounts of long molecule data. We chose *DBG2OLC* for its speed and its ability to assemble using less than 30X of long molecule coverage (cf. PacBio only methods, which typically require higher coverage [93]). *DBG2OLC* is a hybrid method, which uses both long read data and contigs obtained from a De Bruijn graph assembly. We used contigs from a single Illumina assembly generated using 67.4X of Illumina paired end reads [101]. As shown in Fig. 5, the assembly NG50 increases dramatically as PacBio coverage increased, plateauing near 26X. Beyond this point, NG50 remained relatively constant. Alignment of the test assemblies to the ISO1 reference genome showed that some of the contiguity in the 26X hybrid assembly without downsampling was due to chimeric contigs (ie contigs that possess non-syntenic misjoins), and that these errors are fixed as coverage increases (supplementary Fig. 2-3). Chimeras were also absent when only the longest 50% or 75% of reads from the 26X dataset were used.

To measure the impact of read length on hybrid assembly contiguity, we down-sampled the datasets by discarding the shortest reads such that the resulting datasets contained 50% and 75% of initial total basepairs of data. We then ran the same assembly pipelines using these downsampled datasets and compared to the assemblies constructed from their counterparts that were not downsampled. Our downsampling shows that with high levels of PacBio coverage ($>$ 50x), modest gains in assembly contiguity can be obtained by simply discarding the shortest reads (Fig. 5, green lines). Our hybrid assembly results indicate that improvements in contiguity above 30X are modest, though hybrid assemblies remain more contiguous than PacBio only assemblies up until above 60X coverage. For projects limited by the cost of long molecule sequencing, a hybrid approach using ˜30X PacBio sequence coverage is an attractive target that minimizes sequencing in exchange for modest sacrifices in contiguity that are in any event available only at higher coverages.

3.6.4 Assembly merging

With modest PacBio sequence coverage ($\leq$50X), hybrid assemblies are less fragmented than their self corrected counterparts, but more fragmented than self corrected assemblies generated from higher read coverage (Fig. 5). Despite this, for lower coverage, many contigs exhibit complementary contiguity, as observed in alignments (e.g. Supplementary Fig. 4a) between a PacBio only assembly (53X reads; NG50 1.98 Mb) and a hybrid assembly (longest 30X from 53X reads; NG50 3.2 Mb; not featured in Fig. 5). For example, the longest contig (16.8 Mb) in the PacBio only assembly, which aligns to the chromosome 3R of the reference sequence (Supplementary Fig. 4c), is spanned by 5 contigs in the hybrid assembly (Supplementary Fig. 4b). This complementarity suggests that merging might improve the overall assembly.

We first attempted to merge the hybrid assembly and the PacBio only assembly using the existing meta assembler minimus2 [155], but the program often failed to run to completion when merging a hybrid assembly and a PacBio only assembly, and when it did finish, the run times were measured in days. We therefore developed a program, *quickmerge*, that merges assemblies using the *MUMmer* [96] alignment between the assemblies. Assembly contiguity improved dramatically when we merged the above hybrid and PacBio only assemblies (assembly NG50 9.1 Mb; supplementary Fig. 5); however, assembly contiguity can also be increased with false contig joining. To investigate whether merging leads to false joins or introduces assembly errors at the splice junctions, we investigated the result of merging at base pair resolution for the longest merged contig in the aforementioned assembly.

The longest contig (27.9 Mb) in the merged assembly, which aligns to chromosome arm 3R of the reference sequence (supplementary Fig. 6), was longer than the longest 3R contig in the PacBio assembly based on 42 SMRTcells (25.4Mb) [20] (supplementary Fig. 6). The increased length resulted from closing of gaps present in the published PacBio assembly (supplementary Fig. 6) [20]. All joined contigs map to the chromosome arm 3R in the correct order; we take this as evidence that *quickmerge* does not incorporate spurious sequences or large scale misassemblies Nonetheless, small scale misassemblies could still be introduced at the splice junctions. To check for such errors, we manually inspected a high resolution dot plot between the merged contig and the 3R reference sequence. A total of 18 regions were found where the merged contig differed from the reference sequence (supplementary Table 2). The affected regions ranged from 3bp to 20 kb and involved sequence insertion, deletion, and duplication. All identified misassemblies had a buried Pacbio coverage of 15 or higher, indicating that misassemblies were due not to lack of coverage, but some other factor (for example, repetitive regions of the genome). For buried coverage calculations, reads are mapped to the genome, and only mapped regions supported by 2kb contiguous read coverage on both sides are counted towards buried coverage, ensuring any feature exhibiting buried coverage is strongly supported by the reads overlapping it. That said, such discordance between the merged contigs and the reference could have been carryover from assembly errors from the hybrid and PacBio only assemblies that were used for merging. Indeed, 11 of the 18 errors in the merged contigs came from the PacBio only assembly, whereas the rest came from the hybrid assembly. Additionally, sequences 201bp in length from each of the 29 splice joints (break point is the101th base pair, see Supplementary text) from the aforementioned merged assembly were aligned to the reference sequence. None of the sequences revealed any misassemblies introduced by the merging process. Thus, for this dataset, the *quickmerge* approach splices and merges contigs accurately without introducing any new assembly errors. This indicates that the contiguity of even high coverage PacBio only assemblies can be increased by the addition of inexpensive Illumina reads, and gaps in hybrid assembly can be closed by PacBio only assembly even when the PacBio only assembly quality is suboptimal.

3.6.5 Assessment of assembly quality

We assessed assembly quality using the Quast software package [66] and the quality assessment scripts used in the GAGE study [138]. We confined our assessment to assemblies related to application of the *quickmerge* meta assembler, leaving the assessment of PBcR and *DBG2OLC* assemblies to their respective publications [20]. Quast quantifies assembly contiguity and additionally identifies misassemblies, indels, gaps, and substitutions in an assembly when compared to a known reference. We found that, compared to the *D. melanogaster* reference, all assemblies had relatively few errors, with the primary difference among the assemblies being genome contiguity (NG50). Hybrid assemblies tended to have fewer assembly errors than PacBio only assemblies: the total number of misassemblies and the total number of contigs with misassemblies tended to be higher in PacBio only assemblies compared to hybrid assemblies. Still, PacBio only assemblies tended to have slightly fewer mismatched bases compared to the reference, and slightly fewer small indels. Merged assemblies, being a mix of PacBio only and hybrid assemblies, tended to have intermediate Quast statistics; however, the merged assemblies improved upon the source assemblies in terms of misassemblies and misassembled contigs (Supplementary Fig. 8). Overall, the rate of mismatches was low at an average (across all assemblies) of 47 errors per 100kb (Supplementary Table 1, Supplementary Fig. 8). Mismatches and indels can be further reduced using existing programs, such as Quiver [34]. We used Quiver to polish all non-downsampled hybrid, self, and merged assemblies that used at least 40X of data. After Quiver, the average mismatch rate of the selected assemblies decreased from 24 per 100kb to 15, while the average indel rate decreased from 180 per 100kb to 32 (Supplementary Fig. 9). We also performed post-Quiver polishing on these selected assemblies using Illumina data via the Pilon program [159]. Pilon polishing further reduced the average indel rate per 100kb from 32 to 16 (Supplementary Fig 10).

One concern generated by the pre-polished assemblies was that their N50s were high, but their corrected N50s [138] after accounting for errors were low; however, Quiver and Pilon polishing dramatically improved the corrected N50s of the assemblies, indicating that the low corrected N50 values were due to small local errors that were easily resolved by polishing. The average corrected N50 before polishing was 67kb, while the average corrected N50 after polishing was 530kb. It is evident from the corrected N50s that the first polishing step, Quiver, was responsible for most of the change in corrected N50 (Supplementary Fig. 11). Moreover, Supplementary Fig. 11 shows that, after correcting for misassemblies, polished versions of *quickmerge* are almost always more contiguous than polished versions of the component assemblies.

3.6.6 Size selection and assembly contiguity

Long reads generated by library preparation with aggressive size selection [82] can generate extremely contiguous and accurate de novo assemblies [20]. Unfortunately, some DNA libraries with less stringent size selection produce considerably shorter reads (Fig. 2a). Longer reads are predicted to generate more contiguous genomes [100, 124]. We tested this hypothesis by assembling genomes using randomly sampled whole reads (see Materials and Methods) from the ISO1 dataset to simulate a read length distribution comparable to, but slightly longer than what is typical when size selection is not aggressive. Due to the long read length distribution of the ISO1 dataset relative to the shorter target distribution above, a maximum of 53X of ISO1 data could be sampled.

Consistent with the theoretical prediction that, all else being equal, shorter reads produce more fragmented assemblies [100, 124], reads from the downsampled 53X ISO1 data produced a PacBio only assembly with an NG50 of 1.38 Mb, which is shorter than the NG50 (1.98 Mb) of the assembly from the same amount of ISO1 long read data (Fig. 2c). In addition, nearly all long contigs present in the original 53X assembly are fragmented in the assembly from the shorter reads (Supplementary Fig. 13), although the amount of sequence data (53X) used to build the assemblies is the same.

For hybrid assembly, the shorter dataset also produced significantly less contiguous assemblies, consistent with predictions from theory [124] (Fig. 2b). The NG50 achieved with 26X coverage of the shorter dataset was 1.62Mb, compared to an NG50 of 3.58Mb with the original ISO1 data. This is consistent with the PacBio only result – longer read lengths lead to higher assembly contiguity. Thus, a library preparation procedure that aggressively size selects DNA is crucial in delivering long contigs.

3.6.7 The effects of read quality on assembly

As with reduction in read length, increased read errors are predicted to worsen assembly quality because noisier reads increase the required read length and coverage to attain a high quality assembly [36]. When a PacBio sequencing experiment is pushed for high yield through either high polymerase or template concentration, the data exhibits lower quality scores (Fig. 3). Thus, with equal coverage and read length distribution, reads with higher error rates should result in a more fragmented assembly. To measure this effect, we partitioned the ISO1 PacBio read data into three groups with equal amounts of sequence without changing the read length distribution (see Materials and Methods) (Supplementary Fig. 14). For the first two groups, the data was split in half, with one half comprising the reads from the bottom 50% of phred scores and the other comprising the top 50%. The third dataset was generated by randomly selecting 50% of the reads in the full dataset. We then performed PacBio only and hybrid assemblies with these data.

Low read quality had a particularly dramatic effect on assembly by self correction (Fig. 6): the high quality and randomly sampled reads produced substantially better assemblies (6.23 Mb and 6.15 Mb, respectively) than the assembly made from low quality reads (NG50 146 kb). Hybrid assembly contiguity was far more robust to low quality reads (Fig. 6: NG50 of 3.1Mb for the high quality reads, 2.5Mb for the unfiltered reads, and 2.2Mb for the low quality reads), showing only moderate variation amongst different quality datasets. Throughout this study, we avoided altering the settings from their default states in the various assemblers used in order to do fair comparisons; however, in this case, we chose to also run PBcR in 'sensitive' mode to see if it would improve contiguity when data quality is low. We found assembly contiguity was improved (NG50=4Mb), but was still lower than the assembly generated from unselected reads without the sensitive parameters (NG50=6.23Mb).

3.6.8 Merging of human assemblies of the CHM1 cell line

One challenge in a study of this type is determining whether merging performed on a very different genome, like that of Homo sapiens, would perform as well as on *D. melanogaster*. To do this, we used publicly available sequence data and assemblies for the human hydatidiform mole (CHM1 [133]) to generate a merged assembly for H. sapiens, both to gauge the performance of *quickmerge* on a different species than it was developed on, and to observe its performance on a larger and more repetitive genome (the human genome is ˜3.2Gb, approximately 25X the size of the *D. melanogaster* genome).

Of the available CHM1 data, we chose to re-use the data used in Berlin et al. 2015 [20] (the P5C3 chemistry). We ran our genome assembly pipeline on the 30X longest reads of PacBio data from the 54X in the CHM1 dataset, plus 40.66x of publicly available human CHM1 Illumina data (NCBI accession: PRJNA176729). The hybrid assembly produced an NG50 of 2.4Mb, which is in line with the results observed in Fig. 5. Along with this, we used the PacBio assembly contigs produced by Berlin et al. [20], which had an NG50 of 4.1Mb. We merged the two assemblies with more strict parameters because of the larger genome size: we set HCO to 15, c to 5, and l to 5Mb. Merging the two assemblies produced a final assembly NG50 of 8.85Mb, a substantial improvement upon the PacBio only assembly. This more than doubling of NG50 is in line with our expectations based on the *D. melanogaster* results; all available data indicate that this pipeline improves contiguity for CHM1 to the same extent that it does for the *D. melanogaster* ISO1 strain. We did not polish this assembly with Quiver and Pilon due to computational constraints, but it stands to reason that the gains vis--vis SNP and indel rates would be similar between human and *D. melanogaster*. In order to evaluate misassemblies, we produced a *MUMmer* dnadiff report by comparing the PacBio only, *DBG2OLC*, and merged assemblies to the most recent and highest contiguity CHM1 PacBio only assembly available (GenBank accession number: GCA_001420765.1). The results show that the large increase in contiguity is not a consequence of merging induced misassembly, mirroring the results in Drosophila (Supplementary Fig. 12).

Additionally, we generated *MUMmer* dot plots that indicated that contig orientation and ordering were correct, with the exception of some inversions and translocations that were inherited from the component assemblies (Supplementary Fig. 7). While we attempted to run the Quast and GAGE assessment pipelines on the human assemblies, we found that, in all cases, the programs either crashed or failed to finish successfully in a reasonable time frame.

3.7 Discussion

Genome assembly projects must balance cost against genome contiguity and quality [12]. Self correction and assembly using only long reads clearly produces complete and contiguous genomes (Fig. 5; supplementary Table 1). However, it is often impractical to collect the quantity of PacBio sequence data (>50X) necessary for high quality self correction either because of price or because of scarcity of appropriate biological material, especially when assembling very large genomes. For example, at least 40 µg of high quality genomic DNA is required for us to generate 1.5 µg of PacBio library when we use two rounds of size selection in the library preparation protocol. A 1.5 µg library produces, on average, 15-20 Gb of long DNA molecules. This dramatic loss of DNA during library preparation limits the amount of PacBio data that can be obtained for a given quantity of source tissue. When a project is limited by cost or tissue availability, a hybrid approach using a mix of short and long read sequences is an alternative to self corrected long read sequences.

Our results show that when 67.4X of 100bp paired end Illumina reads is used in combination with 10X –30X of PacBio sequences, reasonably high quality hybrid assemblies can be obtained, with approximately 30X of PacBio sequences yielding the best assembly. In fact, as our results show, a 30X hybrid assembly is less fragmented and higher quality than even a 50X self-corrected assembly (Fig. 5). However, our results also show that with the same long molecule data, PacBio only and hybrid assemblies often assemble complementary regions of the genome. The implication here, that different assemblers are joining complementary contigs, suggesting that future assemblers could generate higher quality assemblies with modest coverage data. The merging of a PacBio only and a hybrid assembly results in a better assembly than either of the two alone (Figure 5, supplementary table 1), regardless of the total amount of long molecule sequences (≥30X) used. Thus, projects for which ≥30X of single molecule sequence can be generated are well-served by collecting an additional 50-100X of Illumina data. These data can then be used to generate both a self-corrected assembly and a hybrid assembly, which can then be merged to obtain an assembly of comparable contiguity to PacBio only assemblies using twice the amount of PacBio data (Fig. 5). This merged assembly approach produced the highest NG50 of any assembly at all coverage levels at which it could be tested, with little or no tradeoff in base accuracy or misassemblies (Supplementary Fig. 8-10).

Nonetheless, it is clear that the tools available for genomic assembly have inherent technical limitations: *DBG2OLC* assembly contiguity asymptotes as PacBio read coverage passes about 30X, and the PBcR pipeline produces the best assembly when the longest reads that make up 40X (of genome size) of data are corrected and only the longest 25X from the corrected sequences are assembled [20]. Indeed, when coverage greater than 25X is used for PacBio only assembly, there is a real loss of assembly quality as coverage increases (data not shown). This may be because an increase in coverage leads to the stochastic accumulation of contradictory reads that cannot be easily reconciled, a limitation of the overlap-layout-consensus (OLC) algorithm used in assembling the long reads [125, 123].

Long read sequencing technologies, such as those offered by PacBio, Oxford Nanopore [58], and Illumina TrueSeq [119] promise to improve the quality of de novo genome assemblies substantially. However, as we have shown using PacBio sequences as an example, not all long read data is equally useful when assembling genomes. We provide empirical validation, perhaps for the first time, of length and quality on assembly contiguity. Additionally, our results provide a novel insight: high throughput short reads can still be useful in improving contiguity of assemblies created with long reads, even when long read coverage is high. In light of our results, we have a compiled a list of best practices for DNA isolation, sequencing, and assembly (Supplementary Fig. 15 and Supplementary Fig. 16). Particularly important for DNA isolation is quality control of read length via pulsed field gel electrophoresis. Regarding assembly, we recommend that researchers obtain between 50x and 100x Illumina sequence. Next, researchers must determine how much long molecule coverage to obtain: between 25x and 35x, or greater than 35x. With coverage below 35X, PacBio only methods often fail to assemble, and produce low contiguity when they do assemble, and thus, we can only confidently recommend a hybrid assembly. Above 35X, we recommend meta assembly of a hybrid and a PacBio only assembly. In this case, we recommend downsampling to the 30X longest PacBio reads when generating the hybrid assembly because hybrid assembly contiguity decreases above this coverage level, but this has not been extensively tested. We show that this approach is effective both in Drosophila and human genomes, which differ in size and extent of repetitive regions.

One challenge in assembly is posed by analyzing data from heterozygous individuals. Heterozygosity is known to make assembly more challenging [93]. All of the data evaluated in this study were produced from either isogenic or highly inbred populations (Drosophila) or from a single haploid cell line (human CHM1). Because there is not a comparable dataset available that was produced using heterozygous individuals, we cannot test the effect of heterozygosity on assembly quality. That said, some assemblers (*Platanus* [75] and Falcon (https://github.com/PacificBiosciences/FALCON)) were designed to produce diploid assemblies from heterozygous sequence data [93]. It stands to reason that substituting Falcon in the place of PBcR in this pipeline could improve assembly quality for highly heterozygous samples, but that claim will require further testing.

The recent rapid development of short read sequencing technology has fostered an explosion of genome sequencing. However, as a result of the cost effectiveness and concomitant popularity of short read technologies, the average quality and contiguity of published genomes has plummeted [4]. Indeed, short read sequences are poorly suited to the task of assembly, especially when compared with long molecule alternatives. While long molecule sequencing has rekindled the promise of high quality reference genomes for any organism, it is currently substantially more expensive than short read alternatives. In order to mitigate uncertainties inherent in adopting this technology, we have outlined the most salient features to consider when planning a genome assembly project. We have recommended effective DNA isolation and preparation practices that result in long reads that take advantage of what the PacBio technology has to offer. We have also provided a guide for assembly that leads to extremely contiguous genomes even when circumstances prevent the collection of large quantities of long molecule sequence data recommended by current methods.

3.8 Funding

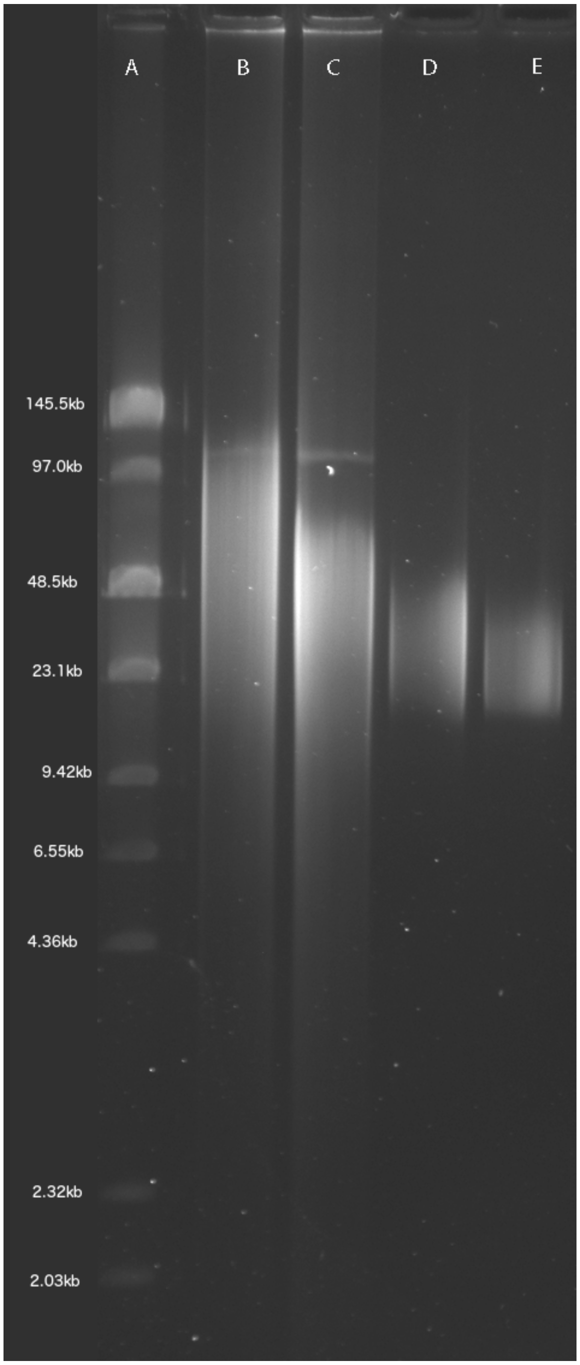
This work was made possible, in part, through access to the Genomic High Throughput Facility Shared Resource of the Cancer Center Support Grant (CA-62203) at the University of California, Irvine and NIH shared instrumentation grants 1S10RR025496-01 and 1S10OD010794-01.

3.9 Acknowledgement

The authors would like to thank Stephen Richards for sharing the length distribution from Drosophila pseudoobscura Pacific Biosciences data and Sergey Koren and Brian Walenz for their assistance with wgs. We would also like to thank Melanie Oakes and Valentina Ciobanu for assistance with sequencing.

3.10 Figure Legends

Figure 3.1: An example of correctly extracted and sheared DNA visualized using field inversion gel electrophoresis. The ladder is the NEB low range PFG marker (no longer produced). The lanes of the gel are as follows: (A) ladder, (B) unsheared DNA, (C) DNA sheared with a 24 gauge needle, (D) sheared DNA size selected with 15-50kb cut-off, (E) SMRTbell template library after 15-50kb size selection. From the gel, it is evident that there is a minimal 'tail' of DNA below ~15kb, the preferred size selection minimum.



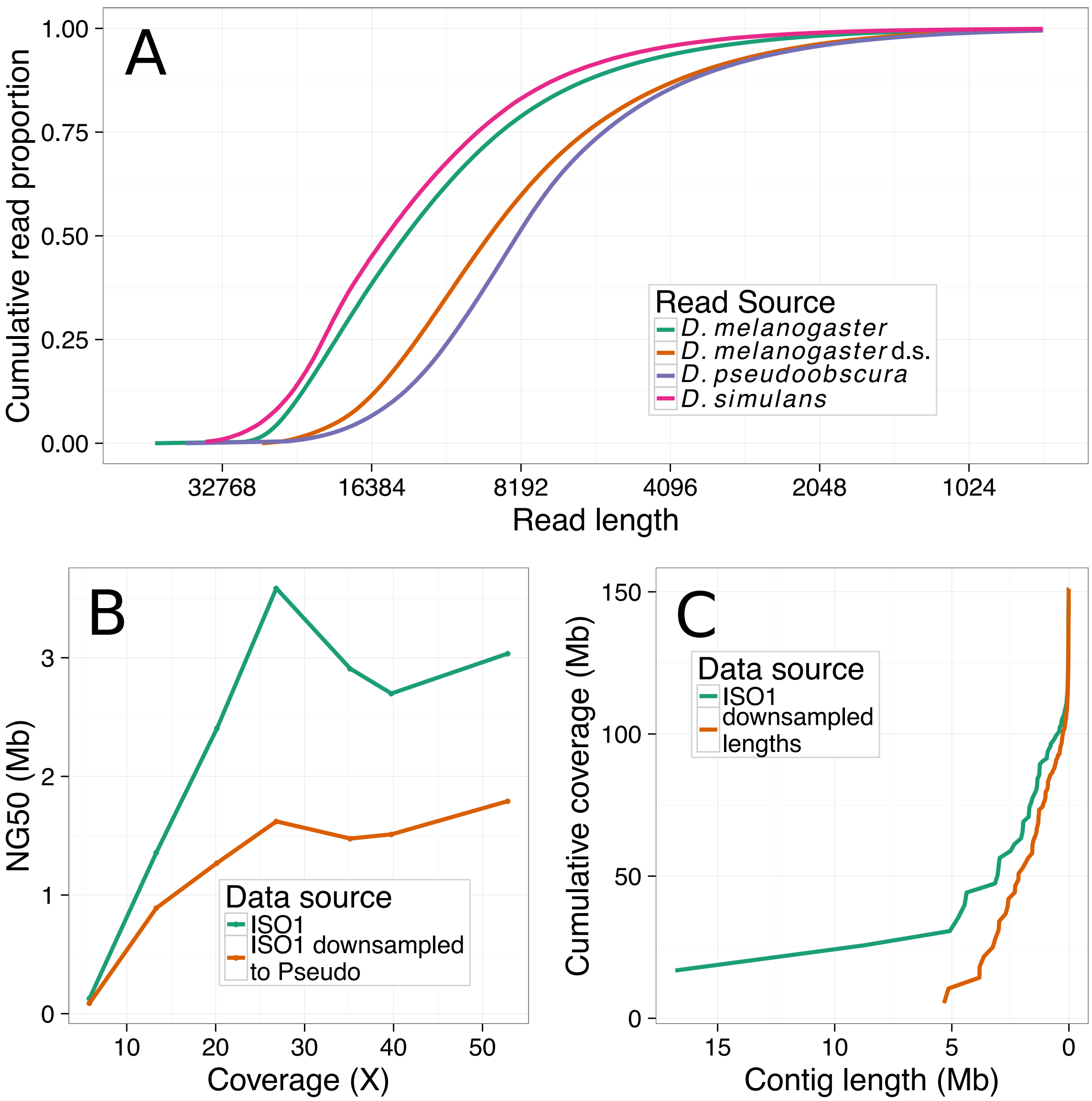
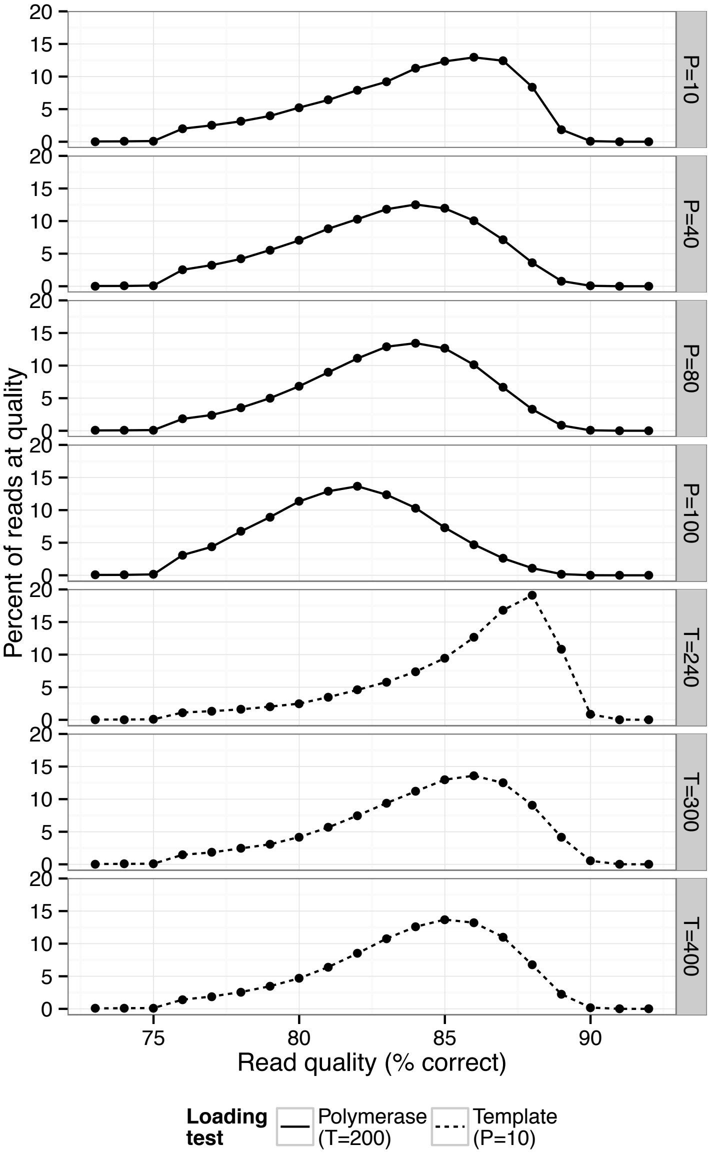


Figure 3.2: (a) The cumulative read length of various data sets, where *D. melanogaster* refers to the original ISO1 data set, *D. pseudoobscura* refers to a publicly available *D. pseudoobscura* dataset with a shorter average read length, D. melanogaster d.s. refers to the *D. melanogaster* data, downsampled to have read lengths resembling the *D. pseudoobscura* dataset, and D. simulans is a D. simulans dataset sequenced using our DNA preparation technique. (b) A plot of NG50 versus coverage of hybrid assemblies, as in Fig. 5. This plot depicts the effect of reduced read length on NG50, while holding read quality and coverage constant. (c) Cumulative contig length distribution of 53X of PacBio only assemblies created with the original ISO1 reads and the ISO1 reads downsampled to resemble Pseudoobscura. Contig lengths in the shorter/downsampled reads assembly are considerably shorter than the contigs in the original reads assembly.

Figure 3.3: The distribution of read quality in sequencing runs performed at the UCI genomics core using our DNA preparation technique. “P” here refers to polymerase loading during sequencing (the proportion of polymerase to template, where 10 would indicate a 10:1 ratio of polymerase to template), while “T” refers to template loading concentration during sequencing (in picomolarity).



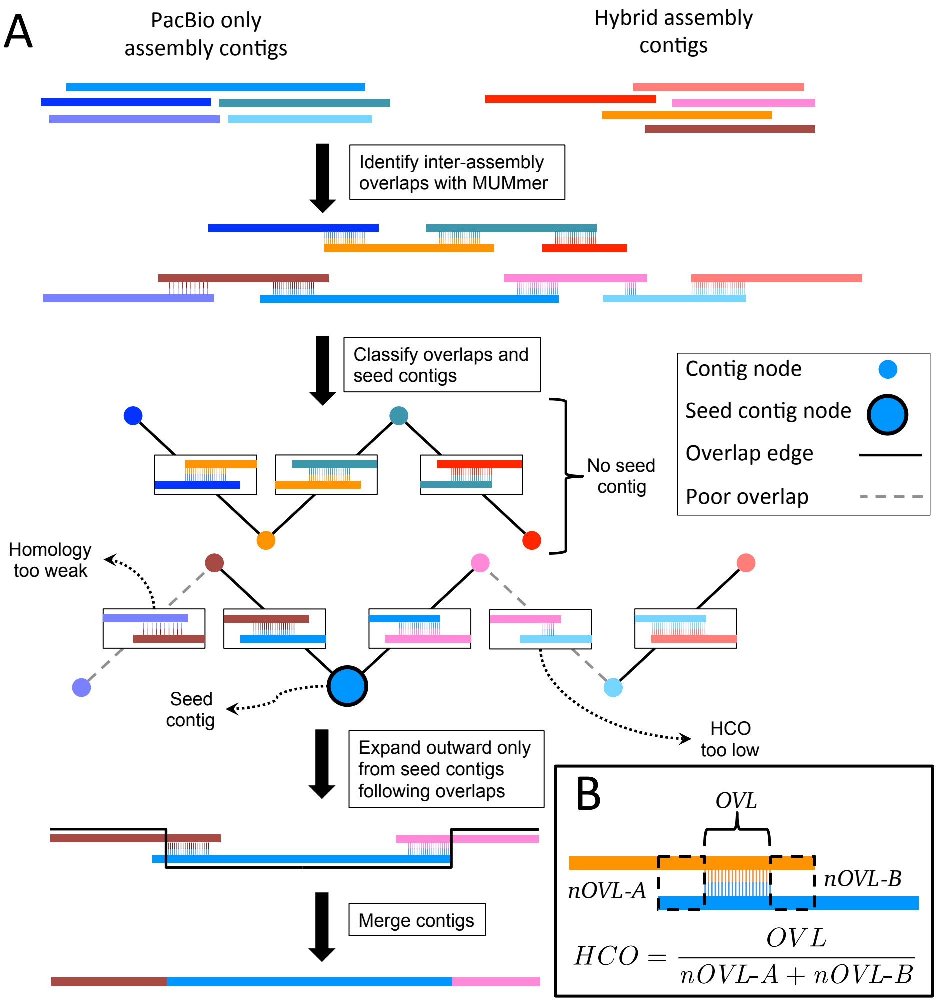

A
PacBio only assembly contigs
Hybrid assembly contigs
Identify inter-assembly overlaps with MUMmer
Classify overlaps and seed contigs
No seed contig
Contig node
Seed contig node
Overlap edge
Poor overlap
Homology too weak
Seed contig
HCO too low
Expand outward only from seed contigs following overlaps
Merge contigs
B
OVL
nOVL-A
nOVL-B
HCO = OVL / (nOVL-A + nOVL-B)

Figure 3.4: A) A diagram representing the algorithm employed by *quickmerge* to improve genome contiguity. (A) *MUMmer* is used to identify overlaps between the two assemblies. High confidence overlaps (HCOs) identified by *MUMmer* will be the primary signal to *quickmerge* that two contigs should be joined. *Quickmerge* clusters contigs according to HCOs. *Quickmerge* identifies seed contigs (contigs in a cluster above a certain size and HCO), and identifies a path that connects it to all other contigs in its cluster by walking from one contig to the next, only stepping to the next contig if the quality of the HCO between the current and next contigs is above the set thresholds. Once the graph connecting available contigs to the seed contig has been constructed, the contigs in the graph are spliced together, with the “Donor” genome’s content preferred over the “acceptor” genome. B) Description of the HCO parameter. HCO represents the ratio between overlapping aligned and overlapping unaligned parts between two contigs.

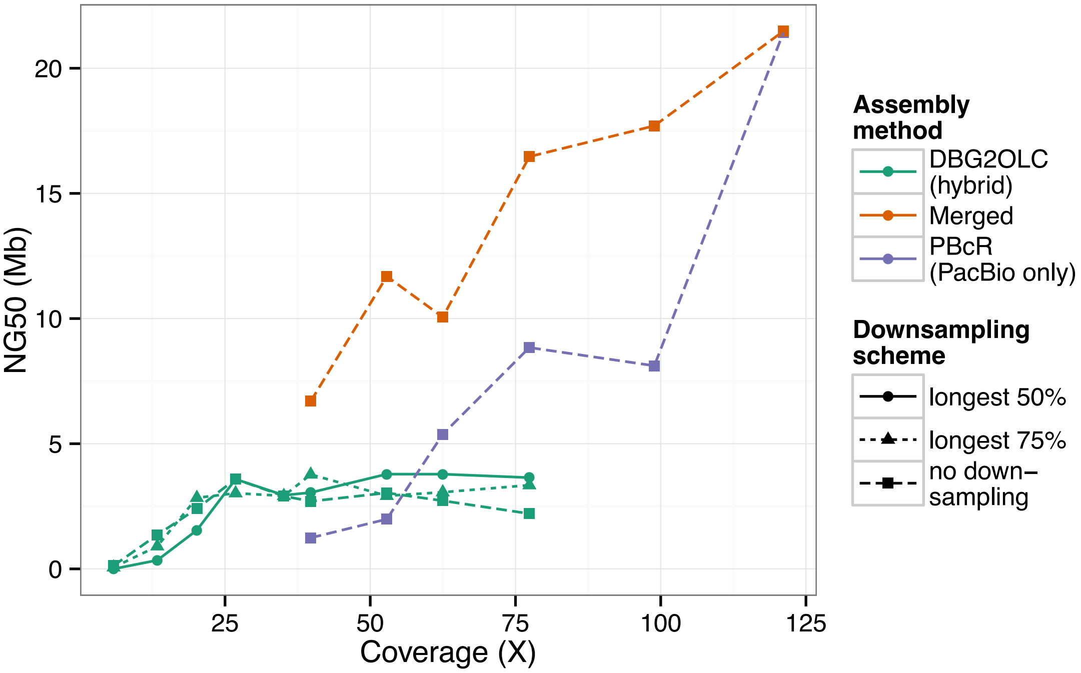


Figure 3.5: The NG50 of *D. melanogaster* assemblies produced using a variety of data sets. NG50 here is the contig size such that at least half of the 130Mb *D. melanogaster* genome (65Mb) is contained in contigs of that size or larger. "longest 50%" and "longest 75%", respectively, refer to datasets in which only the longest 50% or 75% of the available reads have been used. The coverage listed on the x-axis in this case refers to the total amount of available data (before downsampling by length).

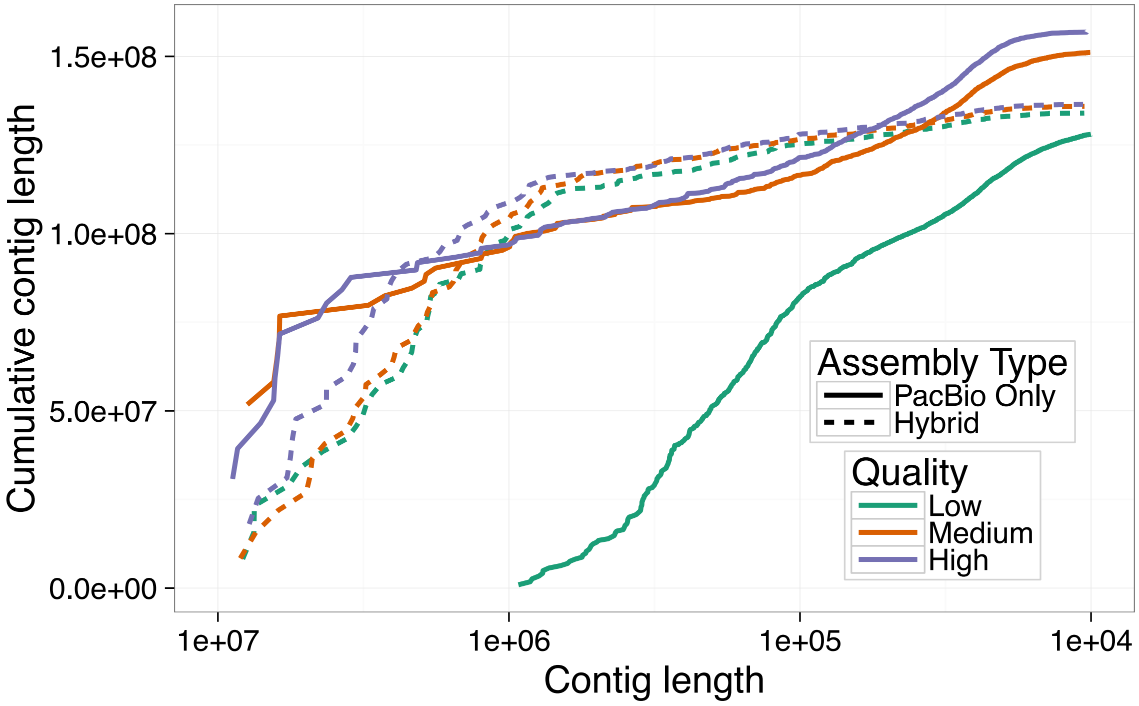


Figure 3.6: As in Figure 2a, a plot of cumulative length distribution. These curves represent the cumulative length distribution of final assemblies using low, medium, and high quality selected reads using either PacBio only assembly or hybrid assembly.

Chapter 4

4.1 Article

Whole genome sequencing of pooled populations reveals signals of differential selection at known genes in the vernal pool clam shrimp *Eulimnadia texana*

Authors:

James G. Baldwin-Brown, Anthony D. Long

4.2 Preface

This chapter is unpublished, but is planned to be submitted in the future under the title "Whole genome sequencing of pooled populations reveals signals of differential selection at known genes in the vernal pool clam shrimp *Eulimnadia texana*" with minor changes. Anthony Long advised all aspects of this publication. Stephen Weeks contributed in the following ways: 1. generation of the JT4(4)5 inbred *E. texana* strain, from which I derived the JT4(4)5-L sequencing strain; 2. collection of all wild populations and ecological data associated with the populations. I was the primary author of this chapter.

4.3 Abstract

Vernal pool clam shrimp (*Eulimnadia texana*) are a promising model system due to their ease of culturing in the lab, short generation time, modest sized genome, and a requirement to produce dessicated resting eggs each generation. Here, we present a genome assembly, annotation, and analysis of pooled population sequencing data for a set of *Eulimnadia texana* clam shrimp populations. We generated a highly contiguous genome assembly using a custom assembly pipeline, 46X of PacBio long read data, and 216X of Illumina short read data; additionally, we annotated the genome using Illumina RNAseq data obtained from adult males or hermaphrodites. 85% of the 120Mb genome is contained in the largest 8 scaffolds of the assembly, the smallest of which is 4.6Mb. Furthermore, the assembly contains 98% of transcripts predicted via RNAseq. Clam shrimp live in small vernal pools in the desert southwest of the USA that differ in many properties. We reasoned that shrimp populations could show local adaptation to the pools in which they live. We then collected 844X of pooled population sequence data from 11 wild *E. texana* populations separated from one another by distances ranging from 0.36 to 253 km, and identified regions of the genome that are putatively locally adapted to the pools they live in as evidenced by extreme allele frequency differences between ponds relative to the bulk of the genome. We looked for excess population subdivision between all the pools, or excess subdivision relative to ecological or physical differences between pools. We identify 12 regions of the genome that are strongly implicated as showing local adaptation and identify genes in these regions that may be responding to selection, including an apparent ortholog of CG10413. Allele frequency at CG10413 differences correlate with latitude, and this gene is predicted to be involved in sodium/potassium/chloride symporter (i.e., active transport of one ion to drive passive transport of another) activity.

4.4 Introduction

The clam shrimp *Eulimnadia texana* has, along with other vernal pool shrimp, been noted for its unique sex determining system [140], its rare (in Metazoa) requirement to reproduce via dessicated diapaused eggs [140], and its unique habitat. This androdioecious [140] species has three common arrangements of sex alleles[140]. Males are always homozygous for the “Z” male allele, while hermaphrodites may be “ZW” or “WW”, with WW females only capable of producing female offspring. Much effort [163] has gone into attempting to identify the *E. texana* sex locus because of this unique arrangement; this, coupled with the fact that close relatives of the species have ordinary male-female sexual dimorphism [164], has led researchers to speculate that the sex locus is recently evolved. The ability of eggs to remain in diapause for years at a time [140] is especially valuable to geneticists because very few macroscopic animals exist in which populations can be archived for long periods without changes occurring in the genetics of the population (genetic drift, loss of linkage disequilibrium, etc.). That said, one of the more remarkable aspects of vernal pool shrimp is the pools in which they live. The naturally limited migration from pool to pool expected of such organisms makes them apparently well suited to the study of populations evolving in relative isolation.

Here, we lay out our attempt to extend genetic research on *E. texana* into the world of whole genome sequence analysis using the latest genomics techniques. We used a combination of short read Illumina [144] and long read PacBio [43] sequencing to generate a high quality draft genome assembly, performed an annotation of genes using RNAseq [160], and used pooled population sequencing [52] to tentatively identify regions and genes that may be under selection in natural populations.

Genome assembly of non-model organisms was financially unrealistic until the advent of high-throughput next generation sequencing. Unfortunately, next generation sequencing methods such as Illumina are limited to short read sequencing, which is not ideal for genome assembly; assemblies produced using Illumina-type short read data tend to have low contiguity [155]. This problem can be overcome by using PacBio[43], Oxford Nanopore [102], or other long read sequencing technologies to supplement or replace Illumina sequencing. A hybrid approach to sequencing and assembly using both short and long reads has been shown to produce highly contiguous assemblies in *Drosophila*-sized genomes[33]. Genome annotation of *de novo* assemblies is routinely performed using RNAseq data [160], and tools for that purpose are already available [151, 62]. We generated a genome assembly for clam shrimp using PacBio data and the genome assembly method developed in Chakraborty et al. 2015 [33], and tools for that purpose are already available [151, 62].

Pooled population sequencing has been in use essentially since the advent of next-generation sequencing [27]. From its first uses, it has not been without controversy – critics have questioned whether the increased cost effectiveness is worth the sampling bias due to uneven coverage [113]; others have argued that the increased sequencing required to account for unevenness of coverage is a minor cost [52]. Although the decision to use pooled population sequencing limits the ability to use haplotype-based inference when founders are unknown, we chose to use pooled population sequencing in this case because it allows for relatively inexpensive estimation of genome-wide allele frequency differences between populations, which was the information we judged most valuable to identify local adaptation. We generated a set of pooled population sequencing runs using populations of individuals found in natural ponds for the purpose of identifying regions of the genome under selection. Two populations were sequenced to approximately 200X coverage, while the remaining populations were sequenced to an average coverage of 48X. We set strict minimum thresholds on coverage in order to reduce the effects of variable coverage. Here, we present the first genome wide estimates of site frequency spectra and population differentiation statistics, as well as Θ, ρ, and other basic population genetics statistics for *E. texana.*

We used pooled population sequencing to estimate F_{ST} [168], *Bayenv2*'s [68] $X^T X$ and Bayes factors, and *LFMM*'s [50] z-values in order to identify regions of the genome that have differentiated due to local adaptation; additionally, we used composite likelihood ratios from *SweeD* [132] to identify site-frequency-spectrum evidence of recent selection on these populations. Various methods ([50, 68, 168, 128, 158]) have been proposed for identifying signals of selection in wild populations. Haplotype focused methods [158] and site frequency spectrum methods [128] rely upon the availability of individual genotypes. There have been numerous attempts to detect population differentiation by identifying allele frequency differences localized to specific regions of the genome. The most venerable of these methods is F_{ST}, an estimator of the probability of identity by descent of individuals within a population. The statistic can be calculated by estimating the variation in allele frequency both within and between populations in order to identify whether structure exists between populations. A higher F_{ST} value than expected (more population structure than expected based on the genome-wide F_{ST} estimate) is an indicator that a force outside of genetic drift and migration is acting upon variation in an area of the genome [2]. Numerous statistics analogous to F_{ST} have been developed, each with advantages and disadvantages. Recently, several different statistics, including *Bayenv2*'s Bayes factors [68] and *LFMM*'s z-values [50], have been developed that used Bayesian statistical methods to identify the likelihood or probability that a given polymorphism's pattern of allele frequencies across subpopulations is explained by the shared ancestry of the populations. These statistics have an advantage over raw F_{ST} in that they account for existing relationships between populations. Although there is no perfect method for detecting selection in wild populations, several studies [109] indicate that *Bayenv2* and *LFMM* are especially powerful in this type of analysis. We used both *Bayenv2* and *LFMM* to analyze this data, and compare the results of the two methods, as well as standard F_{ST}. Through *Bayenv2*'s statistics, we identified 13 loci that appear to be locally adapted in the pools that we surveyed, and identified correlations between these loci and various environmental variables, including collection year, geographic location, pool dimensions, pH, and others.

4.5 Methods

4.5.1 Shrimp collection and rearing

Clam shrimp populations were sampled from New Mexico and Arizona as previously described [166]. We acquired 11 soil samples, each from a different natural clam shrimp pool, to grow shrimp for sequencing (Figure 4.1, Supplementary data table 1); additionally, we sequenced one laboratory population (EE) that is directly descended from the WAL wild population, but has been reared in the lab for six generations. We hydrated the soil samples, then collected 100 individuals from each population on day 10 of their life cycles. These particular clam shrimp populations were chosen because ecological data were already available for these sites (Sup. Table A.5). Clam shrimp populations were reared in 50 × 30 × 8 cm disposable aluminum foil catering trays (Catering Essentials, full size steam table pan). In each pan, we mixed 500mL of soil with 6L of water purified via reverse osmosis. 0.3 grams of aquarium salt (API aquarium salt, Mars Fishcare North America, inc.) were added to each tray to ensure that necessary nutrients were available to the shrimp. Trays were checked daily for non-clam shrimp, especially the carnivorous *Triops longicaudatus*, and all non-clam shrimp were immediately removed from trays. We identified the following non-clam shrimp: *Triops longicaudatus*, *Daphnia pulex*, and an unknown species of *Anostraca* fairy shrimp.

4.5.2 Library preparation and sequencing

Illumina library for assembly

DNA for Illumina sequencing was extracted from 50 inbred monozygotic hermaphrodites from the JT4(4)5-L strain. We performed the Illumina Truseq library preparation protocol. We chose this method over Nextera library preparation for the library for genome assembly for two reasons: first, Nextera library preparation has been shown to produce a bias in coverage that can cause problems during genome assembly[99]; second, the Covaris shearing used in the Truseq protocol allowed us to control the fragment length of the DNA to produce sequencing reads that could be joined into read pairs, or 'pontigs'. In order to produce an average fragment length of 150bp, we used the following Covaris shearing settings: 60 seconds × 6 at 10% duty cycle. 5 intensity, 200 cycles per burst. We size selected the final library on an agarose gel to get the desired 150bp read length. We ran one lane of paired-end 100bp Illumina sequencing on an Illumina HiSeq 2500, producing 124.9Gb of sequence data.

Pacbio library for assembly

We followed the general protocol outlined in [33] to generate the PacBio library used here. We homogenized 265 inbred monozygotic hermaphrodites from the JT4(4)5-L strain in liquid nitrogen using a mortar and pestle. We then extracted DNA using the Qiagen Blood and Cell culture DNA Midi Kit (Qiagen, Valencia, CA, USA). We made two modifications to the protocol: first, we incubated the tissue powder in the mixture of G2 buffer, RNaseA, and protease for 18 hours, rather than the 2 hours listed in the protocol; second, we doubled the RNaseA added from 19ul up to 38ul, and halved the protease added from 500ul to 250ul. We made these changes based on the presence of RNA in earlier attempts to use this kit. After gDNA extraction, we sheared the gDNA using a 1.5-inch, 24-gauge blunt tipped needle for 20 strokes. We visualized both the original gDNA and the sheared DNA using field inversion gel electrophoresis as in Chakraborty and Baldwin-Brown et al. 2016 [33]. We size selected the DNA using a 15kb-50kb cutoff using the BluePippin gel electrophoresis platform (Sage Science, Beverly, MA). We prepared the sequencing library using 5ug of this product, then size selected again using a 15kb-50kb cutoff using the BluePippin gel electrophoresis platform. This produced a total of 0.149nmol of library. We sequenced this library using 10 SMRTcells on the PacBio RS II sequencer, producing 6.7Gb of sequence data and a read length N50 of 15.2kb.

Illumina libraries from wild populations

These libraries were produced using the Nextera Library Preparation Kit. We collected 100 random individuals from each population and pooled the individuals from each population to make each of the 12 libraries (one library per population). 13 cycles of PCR were used during the Nextera protocol, except in the case of the LTER and Tank 011 populations, where 15 cycles of PCR were used due to low yield. Each library was barcoded (Table A.2). Equal aliquots of each library were pooled, and the pooled samples were size selected on a Pippin (Sage Science, Beverly, MA) size selection instrument. The pooled libraries were sequenced over four runs of paired end 100bp Illumina sequencing, producing a total of 127Gb of data, or 844× of coverage. Full coverage statistics for each library are included in Supplementary Table A.3.

RNA sequencing

Individuals for RNA sequencing were derived from the WAL wild population. Adult males and hermaphrodites were sequenced separately. RNA extraction was performed using Trizol [35]. We cleaned the RNA using RNeasy Mini columns (74104, Qiagen) following the manufacturer's protocols, then used this RNA to generate Illumina TruSeq RNAseq libraries according to the standard Illumina protocol. The male and hermaphrodite libraries were sequenced using 1 lane each of paired end 100 bp Illumina sequencing. We generated 23Gb of sequence data for males and 23Gb of sequence data for hermaphrodites.

4.5.3 k-mer counting

We generated k-mers using Jellyfish, v. 1.1.6 [118]. We counted all 25-mers in the joined, but uncorrected, pontigs, then identified a local maximum coverage of 76×, then computed the genome size using the following formula:

$$\text{Genome size} = \frac{T \times \frac{(L-M)}{L}}{C}$$

Where $T = 15.7Gb$ = total basepairs of pontig data, $L = 112.7$ = mean read length, $M = 24$ = mer length − 1, and $C = 76$ = coverage (cf. [98]). This produced a genome size estimate of 144Mb. We used this genome size estimate throughout this work.

4.5.4 Genome Assembly

Hybrid assembly

Genome assembly was performed according to the protocol established in [33]. We first generated pontigs from the PE100 reads obtained from the 150bp insert library by assembling individual read pairs. There is some evidence (cf. read joining with a third read in [56]) that such long, contiguous, error-free reads are slightly better for genome assembly than trimmed paired reads. We generated pontigs using the fq-join function in ea-utils [8], then used *Quake* [79] to error correct the pontigs. We then assembled the corrected pontigs using *Platanus*[75], a De Bruijn graph assembler, with its default settings. This produced an assembly with an N50 of 5.2kb. We input this assembly, plus the raw PacBio reads, into *DBG2OLC* [171]. In order to identify the input dataset that would produce the highest contiguity assembly, we generated a set of hybrid assemblies using a range of quality cutoffs – we tested every whole numbered quality cutoff from 82% to 92%, and, in keeping with [33], downsampled each dataset down to the longest 30×. The 85% cutoff produced the highest N50 of 1.92Mb and an assembly size of 120Mb. All N50s are summarized in table A.1.

PacBio-only assembly

We used *Celera* 8.2, release candidate 3[126], to generate the PacBio-only assembly, using the specfile listed in the supplementary materials (Supplementary Text A.3.1). The assembly had an N50 of 3.4Mb, and a genome size of 126Mb.

Assembly merging

We used *Quiver* [34] to correct both the hybrid assembly and the PacBio assembly, then performed merging using *quickmerge* [33]. We used the following command line settings:

```
python merge_wrapper.py -pre merged_quivered_shrimp_assemblies
-hco 5.0 -c 1.5
/path/to/quivered/hybrid /path/to/quivered/pbonly
```

Here, `-hco` refers to the stringency with which seed high confidence overlaps are filtered, and `-c` refers to the stringency with which other HCOs are filtered. After merging, we corrected the resultant assembly using *Quiver* again. In keeping with the *Quiver* standard practices, we ran *Quiver* on this assembly one more time, then quantified differences between the assemblies using *MUMmer*[96]. We noted a decrease in the number of SNPs and indels identified between the final two *Quiver* runs, so we took the final quivered assembly as our final assembly.

4.5.5 Annotation

We used *Trinity*[62] and *Augustus*[151] to generate the annotation for the genome assembly. We ran *Trinity* three times: once for the male data, once for the hermaphrodite data, and once for the combination of both males and hermaphrodites. We used a custom script to convert *Augustus* data into a generic gff3 file, and another custom script to identify 4-fold degenerate sites based on the same annotation. We used *BLAST* [5] to align the entire *Drosophila melanogaster* proteome against the *Augustus*-generated shrimp CDS and vice-versa. Mutual best hits with an e-value below 10^{-5} were considered significant. We tentatively assert that these genes are correctly annotated, and that they are orthologous or paralogous to genes in *D. melanogaster*.

4.5.6 Differential expression analysis

We identified differences in expression between males and hermaphrodites using *Tophat* [154] and the *DESeq* package [110] (Fig. 4.2). *Tophat* was used for transcript counting, while *DESeq* was used for differential expression analysis. Because we did not have replicated RNAseq data, we used the 'blind' method to estimate dispersion using the following R code:

```
cds <- estimateDispersions(cds,method='blind',sharingMode=c("fit-only"))
```

We then identified differences between the base means of the 'male' and 'herm' groups using the modified binomial test featured in *DESeq*, using the following R code:

```
res=nbinomTest(cds,"herm","male")
```

4.5.7 Identification of repetitive regions

We used *Repeatmasker* (Smit, AFA, Hubley, R & Green, P. *RepeatMasker Open-4.0.* 2013-2015 <`http://www.repeatmasker.org`>.) and *Repeatmodeler* (Smit, AFA, Hubley, R. *RepeatModeler Open-1.0.* 2008-2015 <`http://www.repeatmasker.org`>.) to identify repetitive regions of the genome.

4.5.8 Comparison of wild populations

Data preparation

Our pipeline for cleaning Illumina sequencing data, aligning to the reference, and calling SNPs was as follows: deduplicate data using *Picard tools* (`https://sourceforge.net/projects/picard/`), align using *BWA* [108], and call SNPs using *GATK* [120]. After SNP calling, we censored SNPs by coverage using the following protocol: after merging the WAL and EE populations, remove all SNPs that have a mapped coverage of less than 10 or more than 200 in any population (in the two deeply sequenced samples, the 200 cutoff was applied to the coverages after random downsampling of reads to match the less well covered populations), and remove all SNPs that, in any population, have a coverage more than 3 standard deviations from the population's mean coverage. We performed this censoring separately for each of the three population comparisons examined in section 4.6.6. This removed a variable number of SNPs from the population depending on the coverages in each comparison, leaving a total of 1.4 million SNPs for further analysis in the full 11-population comparison. Command line options for *Picard tools*, *BWA*, and *GATK* are included in Supplementary Texts A.3.1, A.3.1, and A.3.1.

Calculation of population genetics statistics

Our simple, genome wide Θ was calculated by:

$$\Theta = \frac{\text{SNPs per base}}{\sum_{i=1}^{n-1} \frac{1}{i}}$$

Where Θ is the frequency of polymorphisms and n is the sample size (approximately the average coverage of the genome). Here, Watterson's estimator is used to estimate Θ

[161]. We calculated the genome-wide average Θ per basepair using the entire dataset independently for each of our populations, then averaged the result to produce our reported Θ value. Note that Watterson's estimator will be inherently biased in cases where SNP ascertainment is imperfect: rare alleles may be underrepresented if SNP detection has strict standards, but may be overestimated if sequencing errors are called as SNPs. We accounted for this error by fitting the neutral site frequency spectrum expected by chance [51] to all SNPs with a minor allele frequency above 0.1, then using that projected allele frequency spectrum to identify the expected number of SNPs. Fu 1995 notes that the expected minor allele count, for coverage n, is equal to $\phi\Theta$, where Θ is an estimator of $4N_e\mu$, and

$$\phi = \frac{1}{1+\delta_{i,n-i}}\left(\frac{1}{i}+\frac{1}{n-i}\right)$$

For each population, we computed the expected ϕ based on the average coverage of that population, then used that to calculate the proportion of SNPs expected for each allele count class i using $p(i) = \phi$. From there, we computed the fraction of the total distribution contained in the SNPs with a frequency above 0.1, and divided our empirical SNP count by this value to arrive at a projected number of SNPs. We then used the first equation to calculate the projected Θ per basepair based on this estimate of the SNP count.

We calculated ρ by first estimating the short-distance linkage disequilibrium using *LDx* [48]. We then estimated ρ by modeling decay of linkage disequilibrium (r^2) with distance in basepairs using a non-linear model, as in [117] (Sup. text A.3.1). See results for more detail on ρ.

Fourfold Degenerate Sites

We generated a custom Python script for identifying fourfold degenerate sites for use in *Bayenv2*. This script identified sites based on the codon contents of the CDS in all *Augustus*-identified candidate genes. Naturally, candidate sites would not be considered fourfold degenerate if even one transcript disagreed with that assessment. Fourfold degenerate sites were used for *Bayenv2*'s covariance matrix generation step because fourfold degenerate sites have been shown to be under selection less often than any other class of genomic site [170].

Identifying differentiation

We calculated F_{ST} via the Weir and Cockerham method [168] using a custom *R* script. We found pairwise F_{ST} for each pair of wild populations, then reported the mean at each locus. We also calculated Bayes factors using *Bayenv2* [68] both for population differentiation and ecological factor correlation. We did not use *Bayenv2*'s option to incorporate pooled sequencing variation into the Bayes factors because we could not get the program to finish when using that setting (Supplementary Text A.3.1). In the case of population differentiation, we did not use the pooled sequence option because it hasn't been correctly formulated[68].

Sliding window calculations

We performed windowed analyses by averaging values across SNPs with all statistics except *LFMM*'s z-values, which were combined in windows using the Fisher-Stouffer method, as detailed in the manual for *LFMM* [50]. We note that the Fisher-Stouffer method assumes independence of the values of the combined statistics, which is not necessarily true in this case due to linkage, so the window averaging should be taken with a grain of salt. In any case, we largely disregard *LFMM*'s results due to coverage as discussed below, and window averaged statistics are not used in any of the values reported here, but are merely used in plotting where indicated. We used 25-SNP windows in all cases except *LFMM*, where 100-SNP windows were used. We chose all of these based on visual examination of the statistics – these window sizes appeared to reduce noise while making peaks more visible.

Testing for overlap between selection detection methods

We used the hypergeometric distribution in R to test for overlap between detection methods. We first divided the SNPs in the genome into 100,000 equal-sized bins, then took the top 0.1% of 'hits' from each method. If a bin contained even a single such SNP, the bin was considered a detected region. We assumed that this corresponded to a hypergeometric distribution, where hits from one method correspond to one of the colors of balls in the urn, and the other method's hits correspond to the draws from the urn. Significance is achieved if more of the 'hit' balls are drawn than expected, or if more of the drawn balls are 'hit' balls than expected. We used the following R command for this test:

```
phyper(overlap - 1, xhitcount, totalregions - xhitcount,
yhitcount, lower.tail = FALSE)
```

where `xhitcount` is the number of hits via method 1, `yhitcount` is the number of hits via method 2, `overlap` is the number of regions that are hits for both methods, and `totalregions` is the total number of blocks that could be hits.

4.6 Results

4.6.1 Inbred shrimp populations sampled for genome assembly

An inbred population of clam shrimp, here referred to by its numerical title JT4(4)5, was derived from the JT4 wild population and used for Illumina sequencing for the genome assembly. This population was generated by collecting a set of JT4 monogenic hermaphrodites and raising them in the laboratory for 6 generations [162]. Because monogenic hermaphrodites cannot interbreed and can only produce hermaphroditic offspring, this population was the exclusive product of selfing for 6 generations. Although diversity may between individuals exists in this population, each individual is highly homozygous. Thus, we generated an isohermaphrodite line from one individual (JT4(4)5-L) and used it for sequencing. We sequenced a pool of 262 inbred isohermaphrodite individuals to produce the sequencing data for genome assembly.

4.6.2 Genome assembly

We assembled the genome using both the hybrid approach suggested by DBG2OLC [171] and the PacBio-only approach used in PBcR [20], then merged the two assemblies using *quickmerge* [33] to produce the final assembly. The genome assembled into 112 contigs totaling 120Mb. These contigs had an N50 of 18Mb. A plot of cumulative coverage versus contig length (Sup. Fig. A.30) demonstrates that a substantial portion (85%) of the genome is contained in only a few (8) contigs, the largest of which is 41Mb in length and the smallest of which is 4.6Mb. This level of contiguity is a dramatic improvement for vernal pool research: the highest quality vernal pool species currently assembled is *Daphnia pulex*, with a genome size of 227Mb and a scaffold N50 of 470kb (from wfleabase, [37]). Note that scaffold N50 differs from contig N50 in that scaffolds are inferred by joining contigs with gaps, while contigs are gapless; thus, the difference between the assemblies is more dramatic than the numbers seem to indicate.

The fact that the estimated genome size is 144Mb, and the final assembly size is 120Mb, indicates that some portions of the genome were not assembled. This is ordinary in genome assembly, as highly repetitive heterochromatin regions tend to be impossible to assemble with current technology. For instance, the *Drosophila melanogaster* genome is estimated to consist of 175Mb [45], yet the *D. melanogaster* assembled genome is only 143Mb [39].

Two lines of evidence lead us to have confidence in this genome: the quality of other genomes produced under the same circumstances, and empirical evidence of the quality of this assembly. The genome assembly pipeline used in Chakraborty 2016 [33] has been thoroughly evaluated under a variety of genome size and coverage circumstances, and the genome size and coverage of these test assemblies match very closely to the genome size and coverage of our *E. texana* assembly. In particular, the Chakraborty 2016 assembly that used 39X of coverage to assemble a 140Mb genome and was corrected using *Quiver* had an assembly N50 of 6.69Mb, only 3194 misassemblies, and 12.25 mismatched bases per 100kb. Empirical evidence of the quality of a never-before-assembled genome is difficult to acquire, but we can report on the fraction of the *Trinity*-assembled [62] RNAseq-derived transcripts that are present within the final assembly. We find that, if we use transcripts assembled entirely from RNA from hermaphrodites of the reference strain JT4(4)5-L, 98.9% of the transcripts align with above 92% identity, according to *BLAT* [80]. Interestingly, using the entire RNAseq dataset, which contained both the hermaphrodites from the reference strain and males from the WAL strain, produced 95.5% successful alignment, which opens the possibility that some genes are present only in males (unfortunately, they may also be strain-specific rather than male-specific, with no simple way to differentiate those possibilities beyond carrying out a second assembly from males).

Repeatmasker identified 624 SINEs, 16,044 LINEs, 2302 LTRs, 24817 DNA elements, and 88928 unclassified elements, together making up 26.4% of the genome. This contrasts with the relatively low rate of repetitive elements in *D. melanogaster*, at 3.9% [76]. That said, a large portion of this repetitive sequence is 'unclassified'; if we remove the unclassified repeats from the count, only 9.8% of the genome consists of interspersed repeats. Other (non-interspersed) repeats make up 5.1% of the genome.

4.6.3 Annotation and differential expression

We collected one lane of RNAseq data from 25 male clam shrimp from the WAL wild population, and another lane from 25 inbred monozygotic females from the JT4(4)5-L population (the reference population used for the assembly). We used a combination of *Trinity* [62] and *Augustus* [151] to generate an annotation based on our collected RNAseq data. We did three runs of trinity – one run using only the males, one run using only the hermaphrodites, and one run using both together. The combined run produced 85,721 transcripts, while the male and hermaphrodite runs produced 77,257 and 55,845 transcripts, respectively. We ran *Augustus* using the combined run to generate gene predictions for *E. texana*. This generated a total of 17,667 genes and 23,965 transcripts. Of these genes, 5,438 were found to be mutual best hits with known *D. melanogaster* genes.

Of the 17,667 genes, 486 are identified as being differentially expressed between males and hermaphrodites (the Benjamini-Hochberg-Yekutieli [19] adjusted p-value below 0.05) (Fig. 4.2). 40 of these genes are amongst the genes with *D. melanogaster* orthologs. Gene ontology enrichment analysis with *GOrilla* [42] indicates an enrichment of the following GO terms based on the rank order of significance of differential expression (GO terms with a Benjamini-Hochberg corrected P-value below 0.05 are listed): structural contituent of cuticle, chitin binding, structural constituent of chitin-based larval cuticle, structural constituent of chitin-based cuticle, carboxypeptidase activity, chitin deacetylase activity, and association with the condensin complex, extracellular region, and DNA packaging complex (Sup. Table A.5). Hermaphrodites have both testes and ovaries, while males have only testes; additionally, hermaphrodites typically store up to several hundred large, eggs in their carapace prior to ovipositioning [165]. These two large phenotypic differences between males and females are likely to drive many of the observed expression differences.

4.6.4 Sex locus localization

The quality of the clam shrimp genome assembly allows for a more thorough examination of the sex locus in *E. texana.* Previous analyses of allozymes and microsatellites [163] seem to indicate the presence of a sex determining locus that is linked to several markers, with at least three markers so tightly linked that they can be used to genotype the sex locus status of individuals (ZZ vs. ZW vs. WW) with relatively high accuracy. In one study [162], researchers used the *Fum*, *Idh-1*, and *Idh-2* allozymes as markers to identify the sex locus genotype of eggs from several populations. 24 egg banks from a variety of populations, including three (WAL, JD1, and JT4) used in this study, were typed as monogenic using these allozymes; 18 of these were actually monogenic. In another study [163], these allozymes, plus a set of microsatellites, were examined for evidence of linkage to the sex determining locus. Of the loci studied, four were found to be associated with the sex determining locus, and these four (the allozyme *Fum* and microsatellites *CS8*, *CS11*, and *CS15*) were used to build a linkage map for the sex chromosome. We used *BLAST* [5] to align the sequences of these known, linked markers to the *E. texana* genome in order to identify the approximate location of the sex locus (Sup. Fig. A.38). We found that many of the microsatellite loci, including the ones linked to the sex locus, aligned to the 41Mb largest chromosome in the genome in the following order, with other microsatellites being found elsewhere in the genome:

CS19 —CS8 —CS11 —CS15 —CS12

Additionally, of the allozymes that we aligned to the genome (minus Idh-1, for which we could not find a sequence), the two that are believed to be linked to the sex locus (*Idh-2* and *Fum*) were located on a 1Mb contig that we postulate is likely part of the same chromosome as the 41Mb contig, given that they both contain sex locus linked markers.

These results do not fully agree with the results of Weeks 2010 [163]. Their mapping study produced the following order for the mapped markers:

Fum —CS11 —CS15 —CS8

This ordering of markers, specifically markers *CS8*, *CS11*, and *CS15*, is incompatible with our *BLAST* results, indicating either a problem with mapping or a problem with our assembly. While we are relatively confident in the quality of our assembly for the reasons stated above, there is one reason to think that the mapping in this case could be incorrect. Weeks 2010 found a very high rate of recombination between the four mapped loci when looking at males (in fact, approximately 50%; the map pattern of recombination was not significantly different from a lack of linkage between the loci). In contrast, in hermaphrodites, adjacent markers were separated by a very small number of recombinants (3% crossing over rate, 170 individuals —approximately 5 crossover events). We posit that recombination does not occur in amphigenic hermaphrodites, or occurs very seldom, and that much of the inference of marker order may actually be due to genotyping errors. This would also explain the relatively large span of the genome covered by these apparently tightly linked markers – 22.8Mb.

An additional complication here is that the genome assembly was produced using data from WW hermaphrodites. Thus, the male version of the sex determining locus is not expected to be present in the genome assembly. This may make detection of the sex determining locus more difficult, depending on the divergence of the 'Z' and 'W' versions of the sex locus. If the two are highly diverged, they may not align to each other; on the other hand, if they are not highly diverged, they may align to each other, but show a signal of increased polymorphism. Future studies may benefit by assembling the existing whole genome data of the pooled populations to attempt to isolate such a contig.

4.6.5 Population Genetic Statistics

We aligned the Illumina data from the inbred JT4(4)5-L line (the one used for assembly) to the reference genome and produced a SNP rate of 0.00018 per bp, indicating, as expected, a very low SNP rate within the inbreed strain we sequenced. Further investigation revealed that there were no SNPs in this population where the minor allele was the result of more than one sequencing read. Thus, while it is possible that heterozygosity exists in the genome, it is likely that the majority of the SNPs observed in the inbred line are actually the result of sequencing errors. We find no evidence of the existence of blocks of high heterozygosity, as would be expected if there were a segregating polymorphic haplotype in the inbred line.

In addution to the JT4(4)5-L inbred population, we also sequenced pools of individuals from a set of natural clam shrimp populations. Populations were sampled from 11 wild pools (Supplementary data table 1). In addition, a population derived from the WAL population was kept in the laboratory for 6 generations and sequenced. This lab population was maintained at a minimum of 250 individuals per generation. All sequenced populations were sequenced as pools of 100 individuals. We collected pooled population sequencing data from these 12 populations (11 natural populations, and 1 lab population, EE, descended from the WAL natural population) (Fig. 4.1), calculated allele frequencies at each SNP using *GATK* [120], and used these allele frequencies to compare the populations. We first compared the allele frequencies of the entire genome across the populations to identify relationships that exist amongst the populations, then attempted to identify genomic regions that show a greater degree of differentiation than expected based on the existing relationships. A simple hierarchical clustering tree indicated that many of the populations were quite similar to each other (Fig. 4.1): the populations EE and WAL, being directly related by only 6 generations of laboratory maintenance, should be quite similar; many of the natural populations appear to be as closely related to each other as EE and WAL are. This similarity seems to go against some conventional wisdom in the vernal pool world, where vernal pool shrimp are believed to have difficulty crossing from one pool to another, but also increases the difficulty of identifying genomic regions of high differentiation.

We initially calculated Watterson's theta [161] to be 0.0017 per bp, or about 1 SNP per 100 basepairs. This estimate is in the same order of magnitude as *Drosophila melanogaster* [148] and many other invertebrates. This estimate does not take into account the problem of re-sequencing individuals: in pooled sequencing of a finite sample of 2k individuals, some of the 2k alleles will be sequenced multiple times. Treating all reads as being from different individuals will thus underestimate theta. *Popoolation* [88] solves this problem, among others, by assuming that the number of copies of the 2k allele sample is a hypergeometric distribution, and corrects for this effect; however, *popoolation* does not take into account another problem that is likely more problematic: in order to avoid mistacenly calling sequencing errors as SNPs, practitioners typically censor SNPs lower than some arbitrary frequency (or count). Of course, if coverage is much less than 2k it is totally reasonable to expect true positive SNPs to only be observed a single time in the sample. This mis-estimation of Θ due to ascertainment bias is problematic because, under, neutrality we expect low allele frequency SNPs to be more common throughout the genome than intermediate frequency SNPs.

Supplementary Figure A.31 plots the observed minor allele frequency spectrum and the expected allele frequency spectrum under neutrality for a Θ that matches the frequency distribution for SNPs at a frequency greater than 10% (see the methods for a description of this correction). Although we calculated Θ independently for each of our 12 sequenced populations, supplementary figure A.31 displays the result produced if all alleles from all populations are aggregated for ease of viewing. The results in each population are qualitatively the same. As the figure shows, the expected neutral allele frequency spectrum contains a large number of rare alleles that are not correctly identified as SNPs by our SNP calling pipeline.

We calculated Θ separately for each of the populations in order to generate unbiased estimates. Pooling alleles from all populations to calculate one meta-population Θ value is appealing, but allows the site frequency spectrum to be dominated by populations that are sequenced more deeply, and does not account for relationships that among the populations.

A look at the allele frequency spectrum (Sup. Fig A.31) indicated a lack of rare alleles, most likely due to ascertainment bias where rare SNPs are written off as errors, reducing the SNP density. We accounted for this by re-computing Θ using only high frequency SNPs and inferring the existence of a number of rare SNPs by assuming wright-fisher neutrality [51]. This method, which we hold to be the most accurate, produced a Θ per basepair of .00387 if Θ is calculated independently for each population and then averaged. We use this 0.00387 value throughout the following equations.

We note that this value of Θ is fairly typical for invertebrates, and is close to the Θ observed in *Drosophila melanogaster* of 0.0053 [6].

Under population genetic theory, the expected value of theta is equal to $4N_e\mu$, assuming a mutation rate (μ) of 2.8×10^{-9} (the *Drosophila melanogaster* mutation rate per site per generation, from [78]), we estimate the effective population size of the clam shrimp to be $\Theta/(4 \times \mu) = 3.45 \times 10^5$.

We used the result of *LDx* to estimate average linkage disequilibrium at various distances up to about 400bp, then identified the recombination rate based on decay of linkage disequilibrium according to equations in [117]. We estimated the population average "rho" ρ per basepair, the population-adjusted recombination rate per basepair, to be 0.0036, which would make ρ for the entire genome equal to 436,000. We can estimate the recombination rate per basepair per mutation by dividing ρ by Θ. This produces $\rho/\Theta = (4N_e r/bp)/(4N_e\mu/bp) = r/\mu$, where r is the total map distance (genome wide recombination rate), bp is the number of bases in the genome, and μ is the genome-wide, i.e. not per-base, mutation rate. We did this independently for each population and averaged. We get an average r/μ value of 0.94 and an r value of 0.33. Because we do not know the number of chromosomes that make up the genome, and whether or not recombination happens in all individuals, it is difficult to know if this is an accurate estimate of recombination, but assuming that the estimate is correct, it is quite low.

Note that this may or may not reflect that the putative sex chromosome, or the entire genome, may not recombine in amphigenic hermaphrodites, as noted in section 4.6.4. A straightforward reading of the genome-wide apparent map distance of 0.33 seems to imply very limited recombination in the population, which would be consistent with a lack of recombination in amphigenic hermaphrodites, which are all but guaranteed to be the most common sex in a population. The value of 33 centimorgans of recombination is used in supplementary figure A.38 to scale the physical and linkage maps of the putative sex chromosome. It is interesting to note that the total length of the mapped region is approximately similar between the physical map and the scaled hermaphrodite linkage map, but very differenti between the physical map and the scaled male linkage map. This may indicate that the hermaphrodite recombination rate is indeed low and does indeed dramatically influence the average recombination rate across entire populations. That said, because the linkage disequilibrium was estimated using short reads, LD was estimated out to only 450bp; thus, if LD does not follow the decay rate predicted in [70], long distance LD may not be well estimated.

In performing these calculations, we found a SNP rate of 0.0165 for the wild JT4 (i.e., non-inbred) line. We note that, as mentioned above, the inbred JT4 line had a SNP density of 0.00018, which is a reduction in heterozygosity of 91X, and, in fact, we expect that most of the polymorphisms observed in our sequencing data are due to sequencing errors, rather than actual polymorphisms. The standard model of inbreeding predicts a reduction in heterozygosity of only 64X, so all indications are that inbreeding did effectively remove heterozygosity from the inbred JT4(4)5-L shrimp line.

4.6.6 Genome-wide Selection Detection

Pairwise population differentiation comparisons

We began our comparison of the sequenced natural populations by comparing the WAL and EE populations. Because the EE population is a direct descendant (6 generations in the laboratory at ≥ 250 individuals per generation) of the WAL population, there are several reasons that a pairwise comparison of WAL and EE are of interest: first, if minimal differences between WAL and EE exist, then they can be combined to increase coverage of WAL in the 11-population analysis; second, if there are substantial differences between them, there may be evidence of selection due to domestication in the EE population (that is, selection due to being reared in laboratory conditions); third, the level of differentiation between the WAL and EE populations, which have a known history, could inform inferences about the history and relatedness of the wild populations. We computed F_{ST} for this pair of populations. In order to test if the WAL and EE populations were sufficiently similar to each other that they could be pooled, we compared the F_{ST} distribution generated from the WAL vs. EE comparison to the theoretical F_{ST} expected. F_{ST} is expected to be exponentially distributed with a lambda that can be calculated from the empirical F_{ST} distribution [44]. We generated the correct exponential distribution, then compared it to the observed data using a quantile-quantile (Q-Q) plot (Sup. Fig. A.34). In a Q-Q plot, all data points from two datasets are put into rank order and plotted against each other. This test is often used as a way to see if two datasets have similar distributions. One common practice is to plot the distribution of an experimentally determined statistic against a theoretical or simulated distribution that represents the null hypothesis for the statistic in question. If the two distributions match exactly, the line formed by these points will lie along the $x = y$ 1:1 line. If the distributions have the same shape, but one is linearly scaled compared to the other, then we expect to see a straight line that is off of the 1:1 line. We would expect to see this if the populations are differentiated from each other due to random genetic drift. If there are differences in the distributions, the points will not form a straight line, but a more complex shape, with "hockey stick" shaped changes (that is, sudden increases) in slope at the tails being the most common disturbance. A "hockey stick" in one of the tails indicates that the distributions do not match each other, and is consistent with a small number of data points in one dataset being drawn from a different distribution than the rest of the dataset. We would expect to see this if there is a signal of differentiation due to selection at a small number of loci. In this case, we plotted the distribution produced by our F_{ST} test against the theoretical distribution of an F_{ST} statistic, which has been demonstrated to be exponential with a lambda that can be calculated from the empirical distribution [44]. Q-Q plots of F_{ST} (Sup. Fig. A.34, panel A) indicate that, while there may be a small bias toward higher allele frequency differences between WAL and EE compared to the theoretical expectation, the datasets are approximately Q-Q-linear, assuming that F_{ST} is indeed exponentially distributed [44]. Panel B of supplementary figure A.34 represents the same Q-Q plot arrangement as panel A, but uses a 25-SNP-windowed average of F_{ST} rather than raw F_{ST}, with lambda being calculated from the windowed dataset for the purpose of producing the exponential distribution for Q-Q plotting. In a trend that we observe throughout our Q-Q plotting, we find that this plot closely resembles the un-windowed “A” plot, but has a much lower slope. F_{ST}, like many of the statistics we examine here, is skewed toward a long tail in the positive direction, so it is not surprising that the average of a set of F_{ST} values is lower than the raw values (more extreme values are averaged away). We decided to pool the WAL and EE populations based on this result. That is, for the duration of this study, we treat the two population samples as being randomly drawn from a larger sample, and generate our allele frequency statistics using read data from both populations. This is advantageous because it allows us to take advantage of more of the available data and make up for somewhat low coverage of the WAL population.

We next examined the two high coverage lines, LTER and Tank011. Because these populations are more deeply sequenced than other populations, it is possible that they will have reduced allele frequency noise and, thus, a higher power to detect differences in allele frequencies at locally adapted sites. We generated Q-Q plots of F_{ST} in this pair of lines according to the same scheme used to compare WAL and EE (Sup. Fig. A.35). As demonstrated in panel A of the figure, the Q-Q line is not on the 1:1 line, but it is quite linear. This seems to indicate that there is differentiation between the populations due to random genetic drift, but there is no evidence of local adaptation in these lines because there is no “hockey stick” to the Q-Q line. There is a dramatic decrease in the slope of the line as F_{ST} approaches 0.5, but that appears to be due to F_{ST} approaching its maximum value, and does not seem informative. As an aside, the difference in the scale of this Q-Q plot versus the WAL vs. EE one should be expected because there should be more isolation due to distance between these populations (Fig. 4.1). That said, a look at the 25-SNP-windowed “B” panel seems to tell a slightly different story. There, in the same manner as with the WAL and EE, the overall slope is lower than in the “A” panel. On top of this, we see what appears to be the “hockey stick” shape that we expect to see if there is differentiation amongst the populations. That said, the power for F_{ST} to detect differences between populations seems to be weak here if windowing is necessary in order to find a signal of local adaptation. We will later show that F_{ST} seems to do a poor job of identifying local adaptation even when all 11 populations are taken into account.

Site frequency spectrum derived inference of selection

From here, we moved on to analyzing all 11 populations together. We began by performing a scan for selection based on site frequency spectra using *SweeD*. *SweeD* identifies variation in the site frequency spectrum (specifically, a lack of rare alleles) and takes this as evidence of a selective sweep in the recent past. We ran SweeD both for each individual population and for all populations together, but owing to the similarity between the results, and the fact that we believe the results are not highly informative, here we present only the full 11-population result. We plotted *SweeD*'s CLR (composite likelihood ratio) and alpha (significance) statistics both as Q-Q (Fig. 4.3) plots and as Manhattan plots (Sup. Fig. A.32). The Q-Q plots (plotted against a uniform distribution) seem quite promising, with both plots showing a strong "hockey stick" shape at the upper tail, which should be indicative of a signal of local adaptation. *SweeD* superficially seemed to discover numerous regions that had heightened signals of selection, but because we know that the site frequency spectrum is not accurately represented by the SNPs that we have identified (see the section on population genetics for more detail), *SweeD* should produce many false positive calls under these conditions. Since rare SNPs are difficult to distinguish from errors, many rare SNPs are never called, leading to an allele frequency spectrum that is skewed toward common alleles. Because *SweeD* identifies selection by finding regions of the genome whose site frequency spectra deviate dramatically from neutrality, we should expect (and do indeed find – Sup Fig. A.32) that a large portion of the genome appears be under selection. Thus, the results of *SweeD* should be treated with skepticism in this case, and perhaps in other cases where pooled sequencing data has been used to ascertain SNPs.

Population differentiation in all 11 wild populations

We next carried out an analysis to identify population differentiation using all 11 natural populations. We chose to use F_{ST} and *Bayenv2*'s X^TX statistic to scan for differentiation. For all of the 11-population analyses to follow (except for F_{ST} which, as above, is still compared to the exponential distribution), we used a set of simulated read count values as the control values both for Q-Q plotting and for setting significance thresholds (we consistently use a genome-wide significance threshold of 0.05). We generated this neutral distribution using simulation machinery from Gautier 2015 [54]. This machinery uses a covariance matrix produced by *Bayenv2*, plus information on sampling and coverage, to generate the distribution of allele frequencies that we would expect if the sequenced populations are evolving neutrally and are related in the way that the covariance matrix describes. We then calculated our statistics on both the simulated and actual allele count values. We first examined Q-Q plots comparing these simulated and actual values in order to determine whether or not a signal of local adaptation was present. Q-Q plots of experimental F_{ST} vs. the exponential distribution (Fig. 4.3) did not demonstrate any evidence of population differentiation. Indeed, the F_{ST} distribution here seems to be poorly approximated by the exponential distribution, as the Q-Q plot both of single SNPs and 25-SNP windows is highly nonlinear, and gradually curves in the direction of depressed significance. Regardless of the reasons for this, it appears that there is no strong signal of local adaptation visible through F_{ST} analysis. On the other hand, $X^T X$ seemed to display a signal of selection (Fig. 4.3). The Q-Q plots of both single SNPs and 25-SNP windows seem to be essentially linear with the simulated control statistic (with a slight depression in the 25-SNP window, in keeping with our theme), with an uptick in the slope at the upper tail. The increased power in $X^T X$ compared to F_{ST} is perhaps unsurprising if we take into account the level of relatedness amongst the populations. A simple UPGMA tree comparing the populations based on genome-wide allele frequencies (Fig. 4.1) indicates that several of the populations are very similar to each other. In fact, knowing that "EE" is directly descended from "WAL" and separated by only 6 generations of laboratory breeding, the UPGMA tree makes it evident that some of the populations are nearly identical in terms of allele frequencies. Thus, statistics such as $X^T X$ that take into account the relationships between the populations should perform better when attempting to identify differentiated loci.

We then proceeded to generate Manhattan plots of F_{ST} and $X^T X$, and used the same simulated data to set significance thresholds, with the genome wide false positive rate set to 0.05 (Fig. 4.4. F_{ST} (panel E) stands out here as having no peaks that indicate local adaptation. As discussed above, this may be unsurprising in the case of a set of populations that are closely related to each other in complex ways (Fig. 4.1). Thus, we will not discuss F_{ST} in great detail. On the other hand, 13 large peaks, plus a handful of smaller peaks that we will not discuss in detail, are immediately evident in the Manhattan plot for $X^T X$ (panel A). These peaks extend well above our significance threshold, and consist of, in most cases, tens to hundreds of SNPs in regions small enough to contain one to three gene candidates, as discussed in more detail below. The 25-SNP sliding window average of $X^T X$ (panel B) seems to produce a result broadly consistent with the single SNP Manhattan plot, with some peaks becoming more pronounced (i.e., peak 4). We note that there is nothing special about the 25-SNP window size, as it was chosen arbitrarily based on the ability to resolve the peaks clearly. All numbers reported throughout this document (the number of significant SNPs, etc.) rely upon the un-windowed statistics.

Although it is tempting to use the simulation-derived significance threshold as the primary indicator of the differentiation of a SNP, there is an undeniable trend in the data wherein a small number of loci have clusters of highly differentiated SNPs. These are detailed in figures 4.4 and 4.5, as well as in table 4.1, and seem to be more dramatic when examined with windowed statistics (Fig. 4.4). We performed a manual analysis of these sites which, while not statistically rigorous, provides some insight into the function of these sites and may explain their differentiation in some cases. The resolution of these peaks is quite narrow, with peak widths in the range of 1kb to 15kb, and most peaks containing one to three genes. Thus, this may be classified as a relatively high resolution test for selection, which is likely capable of identifying individual genes under selection in many cases.

Seeing as these loci stand out quite dramatically in our analysis, the genes that underlie them may have important effects on the fitness of clam shrimp in various environments. We identified the probable identities of genes in these regions that did not have mutual best BLAST hits in *Drosophila melanogaster* by taking the most significant BLAST hit for each gene (identified using *blastp* against the *D. melanogaster nr* protein database) and assigning that putative identity to the gene of interest. We compiled a table (Table 4.2) of these genes and identified their exact locations relative to the peaks nearby to them (Fig. 4.5). We found several genes with well documented functions, including *Rumpelstiltskin*, *okra*, *Cp1*, *SNS*, *Dscam2*, *pyridoxal kinase*, *Ublcp1*, and many more.

Manhattan plots of this and other statistics are presented in figure 4.4. In addition, A total of 501 SNPs were found to be significant according to the significance threshold we set for the $X^T X$ statistic. 198 well-annotated genes were within 5kb of these SNPs; *GOrilla* indicates that genes related to external visual stimuli are overrepresented in this set (Sup. Table A.6).

Associations of population differentiation and environmental variables

In addition to genomic data, we collected 24 environmental and biological variables relating to the pools in which the wild clam shrimp were found (Supplementary data table 1, Sup. table A.4). We generated Bayes factors and *LFMM* z-values for all of these variables, but acknowledge that many of the variables tested are not likely to influence the fitnesses of individuals in these populations. Additionally, some of the values may be measured inaccurately, or may have a large number of missing data points. Thus, in the following section, many of our measured variables are excluded from thorough discussion. Some variables deserve special description. Date is a measure of the day of collection of the soil. Percent males refers to the fraction of individuals that were male in hydrated samples. Surface area and volume were calculated based on measurements taken on-site at these pools. *Streptocephalus mackeni* and *Thamnocephalus platyurus* refer to the presence or absence of these species of *Anostraca* fairy shrimp, and “Fairy shrimp” refers to all fairy shrimp where the species was unknown. Variables that produced no strong signals of differentiation, such as the presence or absence, or the counts, of *Triops* tadpole shrimp, are left out of this discussion (Sup. table A.4). We note that, because of the relatively small number of assayed populations, many of the environmental variables measured here are highly correlated. Thus, it will be difficult to distinguish between the effects of certain environmental variables, and certain environmental variables that are perfectly correlated cannot be distinguished at all (Sup. Fig. A.37)

We used *Bayenv2* to generate Bayes factors at each SNP, and for each ecological variable. Bayes factors differ from X^TX and other measures of population differentiation in that they are only elevated when the allele frequencies at a SNP are, broadly put, correlated with the environmental variable in question. For each ecological variable, we generated Bayes factors [57] at each polymorphic site in the genome to compare two hypotheses: either that the observed allele frequencies are due to ancestry alone, or that they are due to a combination of ancestry and selection that is correlated with an environmental variable of interest. Our Bayes factors were elevated if the "selection" hypothesis was more likely than the "ancestry alone" hypothesis. We began our analysis of the Bayes factors by subjecting them to the same Q-Q plotting test that we used with X^TX, wherein we plot the Bayes factor values against the distribution of Bayes factors generated by running *Bayenv2* on our simulated neutral dataset. We present here Q-Q plots for a single Bayes factor (association with pool surface area, Figure 4.3, panel C), because other Bayes factors had nearly identical Q-Q plots. We find that the simulated and empirical results are highly linear, and have a strong "hockey stick" shape in the upper tail, indicating a set of SNPs that may be a signal of population differentiation above what is expected by neutrality. We find that, while the line is highly linear, it does not sit on the 1:1 line as we would expect. We have no intuitive explanation for that in this case. The Q-Q plot of the 25-SNP windowed Bayes factor is qualitatively identical to the single SNP plot. Across the 24 environmental variables, we found 1,663 SNPs associated with one or more environmental variables, and 645 of our annotated genes were associated with at least one significant SNP. The Bayes factors thus seem to have picked up more significant hits than X^TX, although some of that may be due to multiple testing. We show a select set of Bayes factors Manhattan plots in figure 4.6. Although there were a variety of random SNPs that were marginally significant, all of the major peaks that were present in these Bayes factor analyses were already detected in the X^TX analysis. This seems to imply that the power to detect population differentiation is as high as or higher than the power to associate SNPs with environmental variables, which speaks to the power of these statistical techniques. A sample Bayes factor plot is included in figure 4.4 because of the striking similarity between the Bayes factor results and the X^TX results. In one sense, this agreement between X^TX and the Bayes factors is not surprising because the Bayes factors used by *Bayenv2* are derived, in part, from X^TX values that are computed during the Bayes factor calculations, with the distinction that X^TX does not take environmental variables into account, but merely indicates divergence from the model based on known relationships between populations. On the other hand, studies have historically had much higher power when comparing allele frequencies to environment, rather than merely to each other, with some selected loci being identifiable only when examined in the context of correlation to environment [21]. This seems to be a demonstration that modern statistical techniques, combined with whole-genome SNP discovery and analysis, have a much higher power to detect differentially selected sites without knowledge of the ecology of the organisms in question. We indicate which peaks are present in which Bayes factor analyses in figure 4.7. We note that several of the environmental variables seem to share the same pattern of significant peaks. In particular, the date of collection seems to match the percent of males, presence of *Streptocephalus* seems to match with presence of *Thamnocephalus*, and volume seems to match with surface area. A heat map of correlations can be found in supplementary figure A.37, and a table of correlation coefficients is available in supplementary data table 2. The correlation coefficients between the three pairs listed here are, respectively, -0.32, 1 (they are identical), and 0.99 (they are nearly identical). This seems to indicate that the correlation of the environmental variables explains the similarity of the peaks, which is unsurprising in a system where only 11 populations have been surveyed. Thus, it may be difficult to precisely identify the environmental variable explained by any given trait. We also discuss an alternative to Bayes factors, *LFMM*'s z values, but only in limited detail, because *LFMM* does not take into account coverage, and so seems to produce inflated values in regions where coverage is low and allele frequency estimates are inaccurate. Again, we begin by generating Q-Q plots of all *LFMM* p-values, though for the sake of brevity, we only report one sample plot here, as all others were similarly shaped. Our Q-Q plot of the p-values produced by *LFMM* for the environmental variable of pool surface area (Fig. 4.3, panel A) against the same statistic calculated from our simulated neutral dataset was highly linear and somewhat above the 1:1 line, with no "hockey stick" shape to it. This indicates differentiation amongst the populations that may exceed somewhat what is predicted in the neutral case, but with no signal of extreme differentiation at specific loci. The 25-SNP windowed Q-Q plot was superficially simlar, but with a lower slope, in keeping with many of the other Q-Q plots generated here. Genome wide analysis of our data with *LFMM* produced conclusions superficially similar to the Bayes factor results, which may be somewhat surprising, given that the Q-Q plot did not indicate the presence of strongly differentiated loci: a relatively small number of loci had visibly large numbers of strongly significant SNPs adjacent to each other (Sup. Figs. A.33,A.36). That said, while a few *LFMM* hits seem to correspond to Bayes factor hits, many hits are unique to one of the two methods. Where there is disagreement, we believe *Bayenv2* is a more reliable indicator of the presence of local adaptation. *Bayenv2* incorporates count data into its significance calculations, while *LFMM* uses only allele frequencies. Comparison to coverage indicates that many of *LFMM*'s strongest hits are in areas of low sequencing coverage. This is unsurprising, as inaccurate estimates of allele frequencies should produce allele frequencies that are not in agreement with the existing relationships between the populations (Fig. 4.5). For this reason, we do not here report the LFMM peaks except in supplementary figures (Sup. Figs. A.36,A.33).

We searched for trends in the genes underlying the highly significant loci with regard to Bayes factor associations with environmental variables. In spite of the problem of correlations noted above, many of these loci produced significant Bayes factor signals only in relation to specific environmental variables. In particular, the surface area of the vernal pools was associated with numerous loci. Other factors that have strong Bayes factor hits include latitude, the percent of males in the population (notably, the one major locus hit by this factor occurs on contig 1, which is believed to contain the sex determining locus), and the presence or absence of various species of fairy shrimp (Figs. 4.4, 4.6, 4.5; Sup. Table 4.2). Specific regions of the genome appear to be correlated with multiple environmental variables. Region 3 (the region aside from regions 12 and 13 that was clearly significant in the largest number of comparisons) displays a strong peak when compared with the date of collection, latitude, the fraction of males in the sample, the presence or absence of *Triops* tadpole shrimp, the surface area of the ponds, the volume of the ponds, and more. Region 4 correlated with longitude, presence/absence of *Streptocephalus* fairy shrimp, presence/absence of *Thamnocephalus* fairy shrimp, and pH, among other, smaller peaks. It is difficult, therefore, to conclusively say that any one locus corresponds with one environmental variable. That said, some gene functions do seem to suggest a relationship between genotype and phenotype —for example, CG10413, a gene in region 11, is believed to have sodium/potassium/chloride symporter activity; one might speculate this influences salinity tolerance in these shrimp, though we do not have salinity data to confirm this. Most strikingly, regions 12 and 13 display elevated Bayes factors for every environmental variable we have measured. Because these loci occur in a small contig that is more likely to contain repetitive content, they should be viewed with some skepticism, but the putative genes identified at these regions (Cp1, CG7627, CG4562, multidrug resistance-like protein 1, octopamine β-1 Receptor), which relate to various nervous system functions and wound healing, are likely worthy of further study. The precise cause of the significance of regions 12 and 13 is unknown, but one possibility that would explain the repeated significance for all environmental variables could be that one population has extreme allele frequencies at this locus.

4.7 Discussion

4.7.1 On non-model organisms and genome assembly

One of the long standing assumptions in genomics is that high quality whole-genome genetic analysis is not possible with non-model organisms because of the lack of genetics resources available for such systems, such as genome assemblies, annotations, and accurate estimates of population genetics parameters. Here, we demonstrate that the generation of a genomic resource for a new model organism is not necessarily difficult or costly. Modern sequencing techniques (i.e., PacBio) allow for de novo genome assembly on a budget on the order of $10K USD. Pooled population sequencing allows for the measurement of essential population genetics statistics in a reasonably large number of populations for a similar cost. Genome annotation with RNAseq is now more reasonable than it has ever been. This combination of factors makes genomics in non-model systems an attractive target for evolutionary biologists. Admittedly, this study was performed on an organism with a relatively small genome of about 150Mb, but the cost of sequencing the whole genome is merely linear with genome size, and RNA sequencing costs vary little between organisms.

We present here a *de novo* whole genome assembly for *E. texana* with an N50 of 18Mb. We hope that this genome will be a useful resource for the vernal pool research community, and that clam shrimp will be a useful model organism in the future. Sequencing of wild populations revealed approximately 1.5 million SNPs that can be used as markers in future studies. Additionally, we present a draft annotation of the genome that allows for accurate identification of genic, intergenic, etc. regions, as well as homology-based comparisons with genes in other species.

4.7.2 Sex-driven differential expression

We used the *DEseq* package in *R* to identify genes that were differentially expressed in males and hermaphrodites. We may expect our power to detect differential expression to be low in this case because of the lack of samples: we have only two samples, one of which is a pooled set of males, and the other of which is a pooled set of hermaphrodites. In addition, hermaphrodites share more in common with males than females do: both males and hermaphrodites must have testes and produce sperm. Thus, we expect that differences between the sexes are most likely to be associated with egg production and laying. Our test for outliers produced a set of 486 genes that are differentially expressed, 40 of which could be matched to *D. melanogaster* genes with mutual best hit *BLAST*. Interestingly, the GO terms in this group of genes were enriched for terms relating to chitin structure in *D. melanogaster*. While it is difficult to know the function of these genes in vernal pool shrimp, it is notable that there are significant differences in body shape between males and hermaphrodites that make hermaphrodites differentiable from males. Specifically, a hermaphrodite has a 'hump' on its dorsal region that provides a space for the brood pouch that holds eggs before laying. Further studies might compare structural morphology of hermaphrodites after gene knockouts to that of males and wild type hermaphrodites.

4.7.3 The sex locus

Much effort has gone into identifying the structure of the sex locus in individuals with recently-derived sex chromosomes. *E. texana* is androdioecious, but is believed [164] to be descended from a dioecious ancestor that was ancestral to the entire *Eulimnadia* clade. Prior evidence has indicated that the sex determining locus may be autosomal, and that known markers may have incomplete linkage with the sex determining locus. We identified a single contig that contained all but one of the sex-linked markers, indicating that this contig is likely the sex determining chromosome. That said, the markers were spread across the entire 42-Mb contig, and the order of the markers differed from the order predicted by linkage mapping. Thus, we were unable to identify a small region of the chromosome associated with sex determination. One explanation for this discrepancy may be that the inferred linkage between known linked markers near the sex determining locus is due to a lack of recombination in the heterogametic sex, as is the case in *Drosophila melanogaster* [106] and other organisms. This is supported somewhat by the low rate of recombination in amphigenic *E. texana* hermaphrodites inferred in previous studies [163], though this is unconfirmed.

We observed that there was a lower rate of successful mapping of RNAseq-derived transcripts to the genome when male-specific transcripts were included in the transcripts to map. Based on this, we posit that there are transcripts present in males that are too distinct to map to the hermaphrodite derived genome assembly. This suggests one of three possibilities: first, there may be a genomic region that only occurs in males, which is thus absent from our current assembly; second, that there is a region present in both male and hermaphrodite versions of the genome, but that is so far diverged that the male version fails to map to the assembly; third, that there are differences among the sequenced strains that account for the difference, because the male individuals and the hermaphrodites were sampled from different strains. A further study could elucidate which of these is the cause of the difference by generating a whole genome assembly of the male genome.

4.7.4 Wild populations and selection

We used F_{ST}, *Bayenv2*'s X^TX and Bayes factors, and *LFMM*'s z-values to identify signals of selection in these populations. We largely disregarded F_{ST} in this case because there were clear relationships between the populations that made F_{ST} poorly suited to identifying selection. Additionally, we largely disregarded the results of *LFMM* because it fails to take into account coverage when computing significance, and many of *LFMM*'s peaks appear to occur in areas of suspiciously low coverage, where estimates of allele frequencies are inaccurate. Therefore, we relied largely on the results of *Bayenv2*'s X^TX and Bayes factor statistics when dissecting this data.

Conventional wisdom indicates that vernal pool organisms are not capable of a great deal of migration under most circumstances —indeed, Bohonak (1998) [22] indicates that a geographic distance of only a few hundred meters should be sufficient for a high degree of differentiation of populations in the *Anostracoda* (fairy shrimp). The inability of vernal pool shrimp to escape the pools in which they are born seems to prohibit migration between distinct pools. In fact, we find here (Figs. 4.1, 4.1) that there is a great deal of migration between pools, both across short and long geographic distances, with shorter distances leading to increased migration. Mean pairwise F_{ST} is 0.038 across these samples. The source of this ability to migrate, whether it be animal tracking, wind dispersal, periodic flooding, or some other mechanism, should be the subject of further study.

We identify several genomic loci that appear to be under selection, as well as several variables in the environment that appear to be correlated with these selected loci. Of note are two regions that appear to be subject to selection that are both related to RNA-to-protein translation, including *CG10306*, which is expected to be involved in regulation of translation initiation (Flybase Curators, personal communication to Flybase), *La*, which is experimentally validated as binding to rRNA primary transcript [172], and *rumpelstiltskin*, which is experimentally validated as binding to the $3'$ UTR of mRNAs. It is not clear what would drive protein translation machinery to be under differential selection in different pools, though we note that not all genes in our assembly have orthologs in *D. melanogaster*, so it is possible that unannotated genes, or even undiscovered genes, could drive these signals of local adaptation.

Correlation with environmental variables seems to indicate that certain variables have a larger effect on allele frequencies than others. For example, there are a number of loci whose allele frequencies are strongly correlated with surface area of the pool in which the shrimp reside. Pool surface area could be influential for a number of reasons, including the persistence of water over longer periods (though the surface area to volume ratio is not strongly correlated with pool differentiation), the presence of predatory shrimp (there may be a relationship between pool size and presence of predatory shrimp), consistency of food availability, mate choice, etc. That said, there is not an obvious pattern in the genes associated with surface area. Surface area correlates with latitude, the other most strongly influential environmental variable, with r=0.36, suggesting conflation of latitude and surface area in our analysis; that said, many other aspects of the ecology of these pools have higher correlations with latitude, and not all have obvious signals of differentiation. Overall, although a large number of SNPs were significantly differentiated according to our comparison to simulated data, a small number of loci showed visibly strong signals reminiscent of selective sweeps, while other significant loci were often individual SNPs with no evidence of allele frequency change in the surrounding SNPs. These few, strongly differentiated loci certainly seem to be worthy of further study. The clam shrimp ortholog of CG10413 in region 11, for example, is predicted to have sodium/potassium/chloride symporter activity: it has long been believed [135] that sodium/potassium pumps and chloride pumps with passive sodium diffusion are important for regulating osmotic stress in vernal pool shrimp. We have no data on salinity in these pools, but it is tempting to speculate that salinity is correlated with the significant region 9 hits, especially date and latitude.

4.7.5 The future

We identified a relatively small number of candidate genes that appear to be associated with differentiation of these populations. Genetic studies, perhaps using CRISPR-Cas [139] or gene knockouts/knockdowns, could reveal much about the effect of these genes on phenotype, especially if wild type and mutant alleles could be swapped in an individual. The approach employed in this study acts as an effective template for establishing a non-model organism as a viable system for genomic study. The observation that somewhat reproductively isolated populations living in different ecological or physical settings can show strong geographical isolation in isolated genomic regions suggests a powerful paradigm for identifying the genes contributing to adaptation in the wild. We hope that future studies will gain insight into the genetics of never-before-sequenced species using the methodology of high quality genome assembly and whole genome short read sequencing of natural populations.

4.8 Data Availability

All sequencing data is available at the NCBI All data will be made available at the NCBI Sequencing Read Archive under the Bioproject "PRJNA352082". Additional files are available at the following URL:
`http://www.wfitch.bio.uci.edu/~tdlong/PapersRawData/BaldwinShrimp.tar.gz`.
Additionally, all scripts used for analysis will be made available at the following GitHub page: `https://github.com/jgbaldwinbrown/jgbutils`

4.9 Tables

Locus index	Contig	Range	Width	Maximum X^TX
1	1	7960415:7961229	814	52.6524
2	1	11807825:11808199	374	47.6688
3	1	18205259:18209499	4240	70.4348
4	1	28626427:28634469	8042	75.102
5	1	41068868:41081593	12725	71.9472
6	3	5849714:5855825	6111	65.5048
7	4	2920923:2921932	1009	64.5828
8	4	8446194:8447185	991	53.1936
9	5	2113388:2117449	4061	59.1548
10	5	3903600:3908963	5363	61.8584
11	6	1318370:1320640	2270	96.1406
12	14	369627:420280	50653	89.70358
13	14	791617:817899	26282	67.7304

Table 4.1: Major significant sites according to the 11-way X^TX population differentiation analysis.

Site	Name	Function	Citation
1	*SNS* (sticks and stones)	Actin filament related. Absence of body wall muscles and presence of unfused myoblasts in mutants. An IgSF.	[23, 41]

2	*Dscam1*, *Dscam2*	An IgSF. Overexpression in fetal brain leads to down syndrome. Every *D. melanogaster* neuron has a unique Dscam1 isoform mix.	[141]
2	mini chromosome maintenance 2	Helicase. *MCM* is a polymer of *MCM2* through *MCM7*.	[105].
3	no significant hits		
3	*CG10306*	regulation of translational initiation (predicted by Flybase)	
3	*CG4049*	ATP binding, DNA repair (predicted by Flybase)	
4	chondroitin synthase-like protein (*CG9220*)	synthesizes chondroitin sulfate, a glycosaminoglycan expressed on most cell surfaces. Regulates many processes.	[87]
4	*rumpelstiltskin*	mRNA 3-UTR binding [72], anterior/posterior axis specification in embryo [72], intracellular mRNA localization, mitotic nuclear division, pole cell development, pole plasm oskar mRNA localization, segmentation	[72, 149, 40]

4	*okra*	Helicase. Homologous DNA repair. Meiotic recombination. Mutant females sterile. Oogenesis, response to ionizing radiation, double-stranded break repair, dorsal appendage formation, chromatin remodeling.	inferred from various mutants. [142, 3, 90, 167, 111, 60]
4	*CG43370*	cilium assembly	[10]
5	kinesin-like protein 67a	Localization of mitochondria in undifferentiated cells. *E. coli* moves mitochondria toward *KLP67A*. Causes movement along microtubules (+ direction). Drives disassembly of microtubule arrays.	[53, 134, 61]
5	*pyridoxal kinase*	Enzyme that generates pyridoxal-5-phosphate (Vit. B6).	[121]
5	*CG5514*	Non associative learning, synaptic growth at neuromuscular junction.	[46]
5	*Ublcp1*	26S proteasome phosphatase. Regulates nuclear proteasome activity.	[65]
6	*CG8500*	Mutations are viable and fertile [18]; small GTPase mediated signal transduction	GTPase inferred from similarity with mouse *Rap1a* (Flybase curators 2008)

6	no significant hits		
7	no significant hits		
7	no significant hits		
8	*LqfR-L*	cell proliferation; oogenesis; positive regulation of Wnt signaling pathway; sensory perception of pain	inferred from mutant. [104, 107, 103, 127]
11	*CG10413*	Predicted to have sodium/potassium/chloride symporter activity.	See Flybase.
11	*branchless*	Drosophilas only known FGF (fibroblast growth factor). Influences branching morphogenesis in trachea, lungs. Receptor for baculovirus FGF in *Lepidoptera*.	[55, 77]
12	*CG7627*	wound healing	inferred from mutant phenotype, [30]
12	*CG7627*, *CG4562*, multidrug resistance-like protein 1	wound healing (*CG7627*), methotrexate resistance in malpighian tubules	inferred from mutant [30].

12	Octopamine Beta1 Receptor	octopamine receptor activity, negative regulation of synaptic growth at the neuromuscular junction	inferred from mutant phenotype. [91]
13	*Cp1*	Cp1 mutants do not exhibit retinal degeneration when exposed to six days of constant light.	[85]
13	*CG4847*	no significant hits	

Table 4.2: A table indicating the putative functions of orthologs of genes under the major $X^T X$ peaks.

4.10 Figures

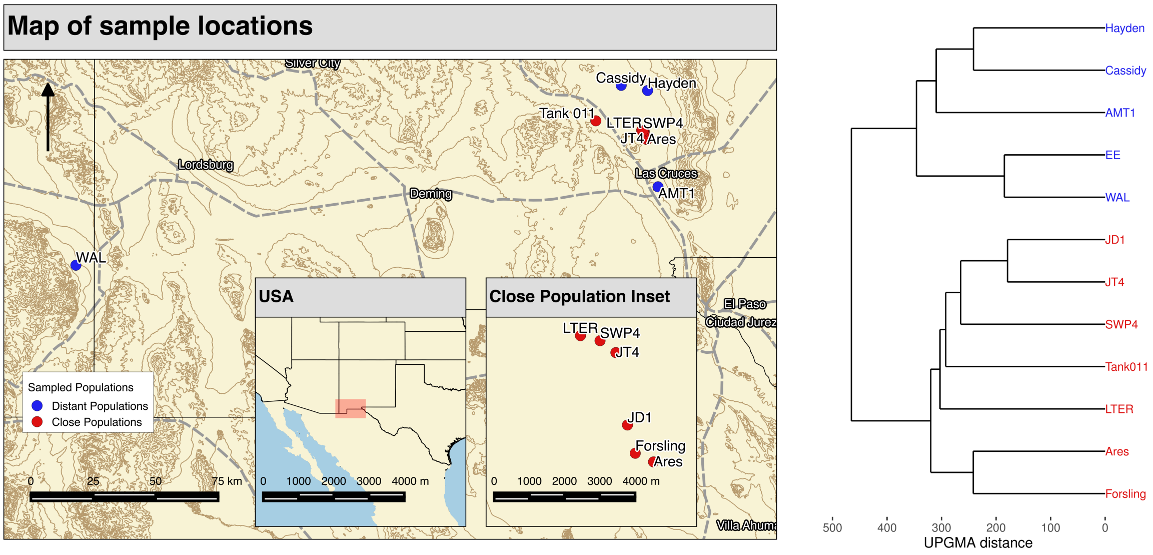


Figure 4.1: A map of the sampling locations for all populations sequenced in this experiment, and UPGMA tree corresponding to the relationships of the populations based on genome-wide allele frequency similarity. Colors correspond between the map and the tree. All populations were taken as soil samples from field sites in New Mexico and Arizona. Note that the “EE” strain is descended from the WAL population.

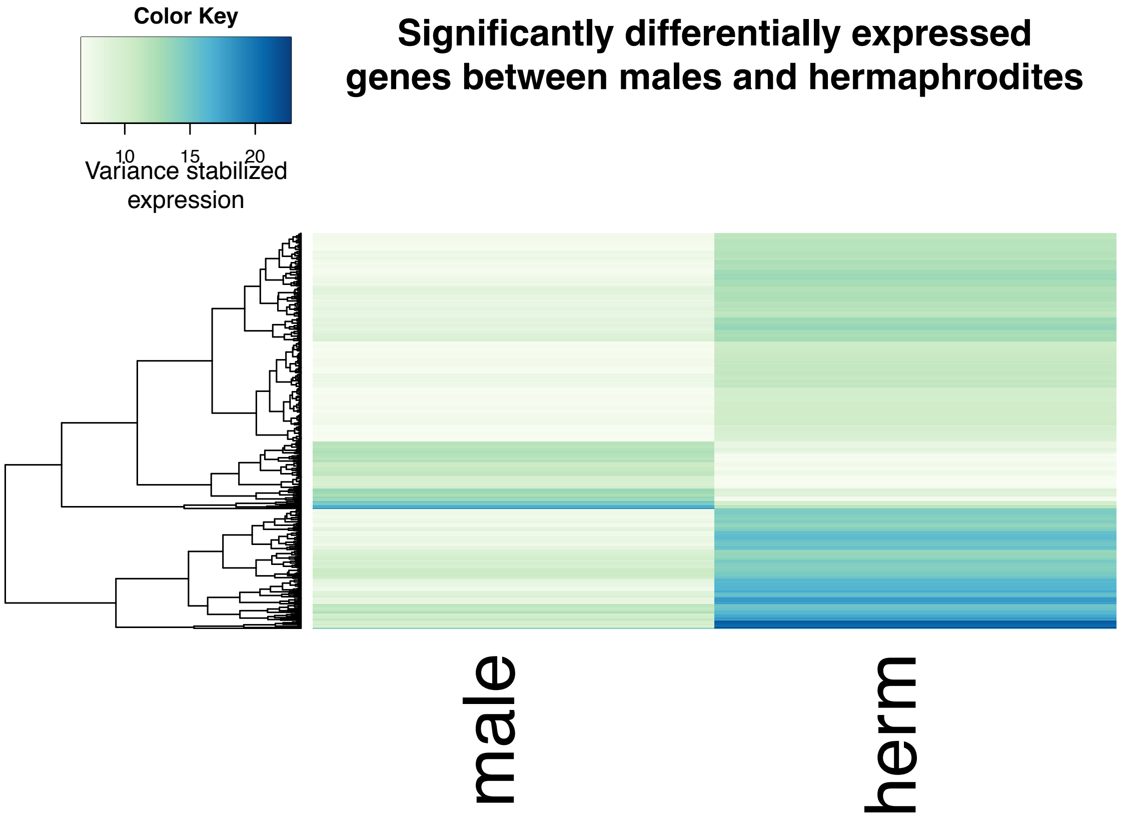
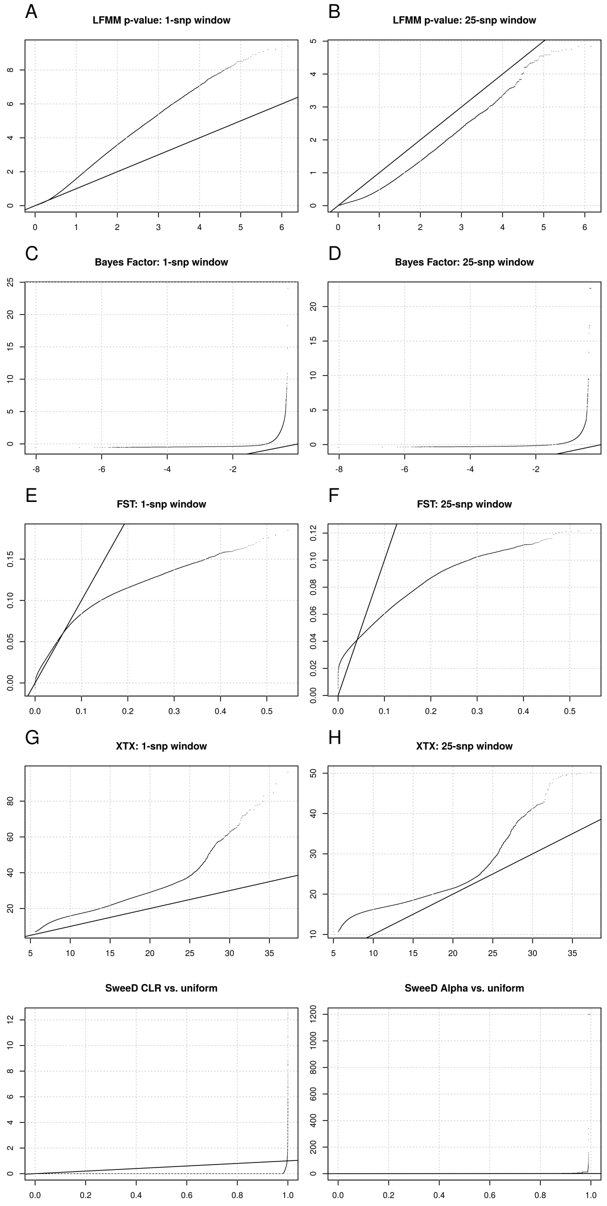


Figure 4.2: A heat map of expression in genes that are significantly differentially expressed (adjusted $p = 0.05$). Note the small portion of genes that have nearly 0 expression in males, and high expression in hermaphrodites.

Figure 4.3: Quantile-quantile plots of five different statistics used to identify regions under selection in the clam shrimp genome. Left hand plots are single SNP statistics, while right hand plots are 25-SNP windows. Plots are as follows: A,B: *LFMM* p-values for the trait of pool surface area versus a uniform distribution; C,D: *Bayenv2* Bayes factors for the same trait, log-log plotted against Bayes factors for the same trait computed from our simulated neutral dataset; E,F: mean pairwise F_{ST} versus an exponential distribution with $\lambda = 1/\bar{F_{ST}}$; G,H: *Bayenv2* X^TX values versus X^TX values calculated from our neutral simulation; I: *SweeD* CLR values vs. a uniform distribution; J: *SweeD* alpha values vs. a uniform distribution.



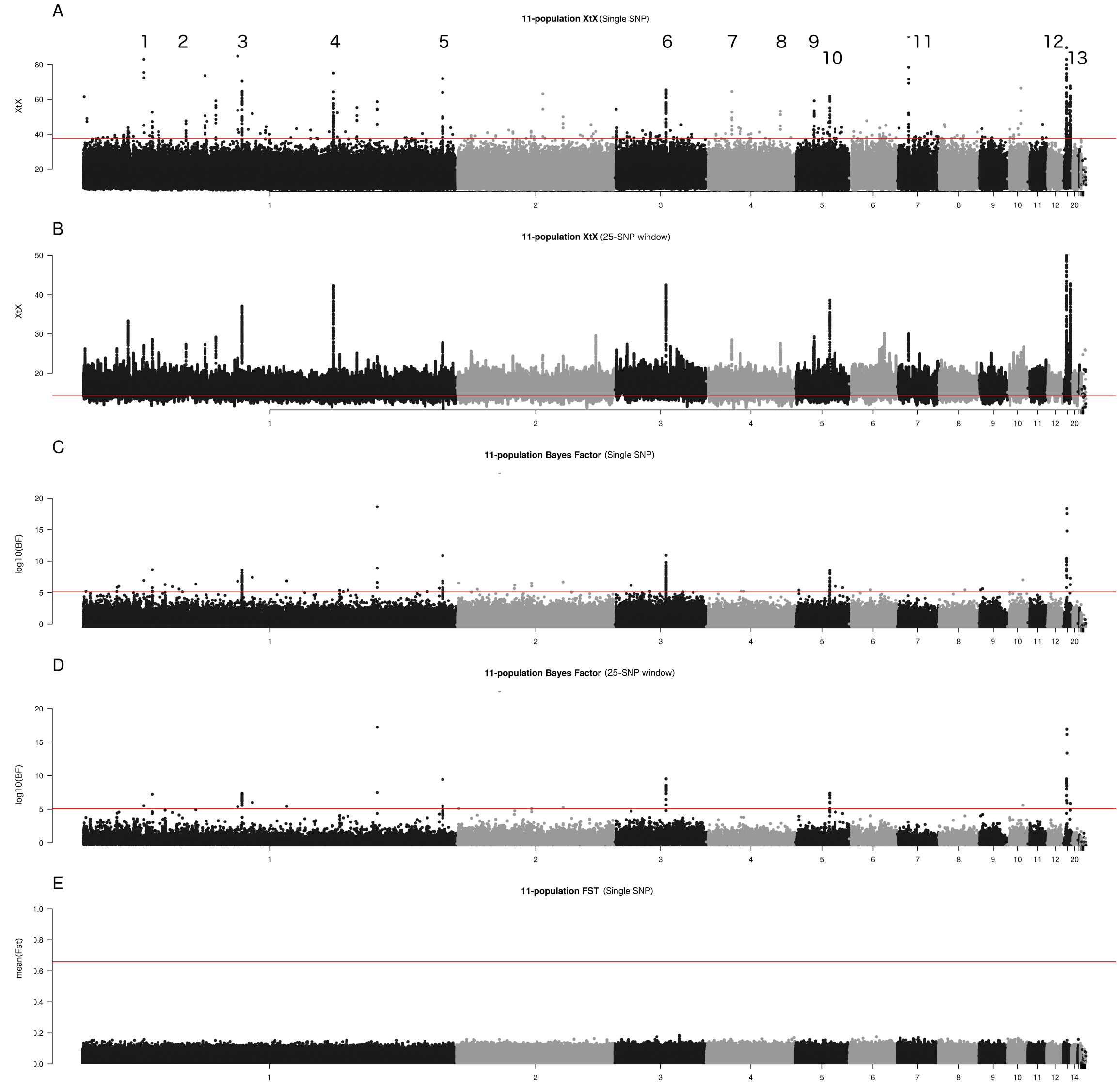


Figure 4.4: Manhattan plots of all statistics of population differentiation. Plots are as follows: A,B: *Bayenv2* $X^T X$ values (B is a 25-SNP windowed statistic); C,D: *Bayenv2* Bayes factors associating pool surface area with allele frequency differences (D is a 25-SNP window); E: mean pairwise F_{ST}.

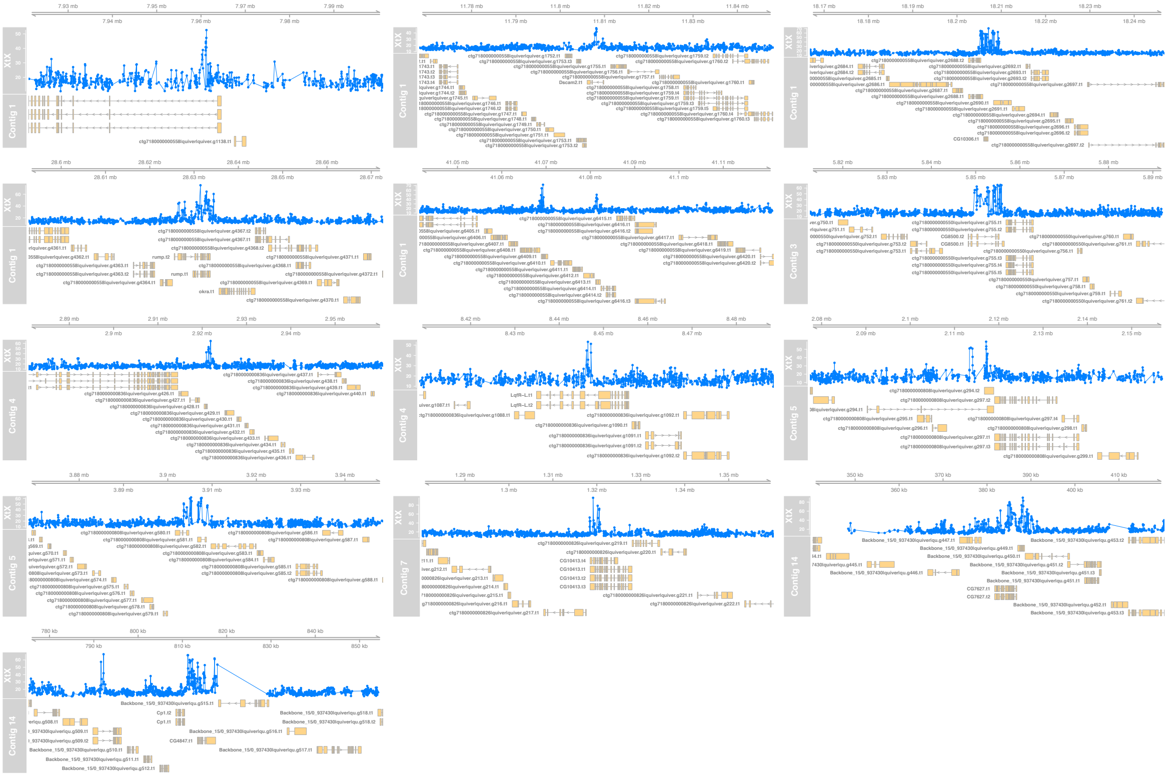


Figure 4.5: A set of Manhattan plots showing X^TX values for a select set of loci with high X^TX values, as indicated in figure 4.4. The order of these plots corresponds to the numbered loci as depicted in figure 4.4.

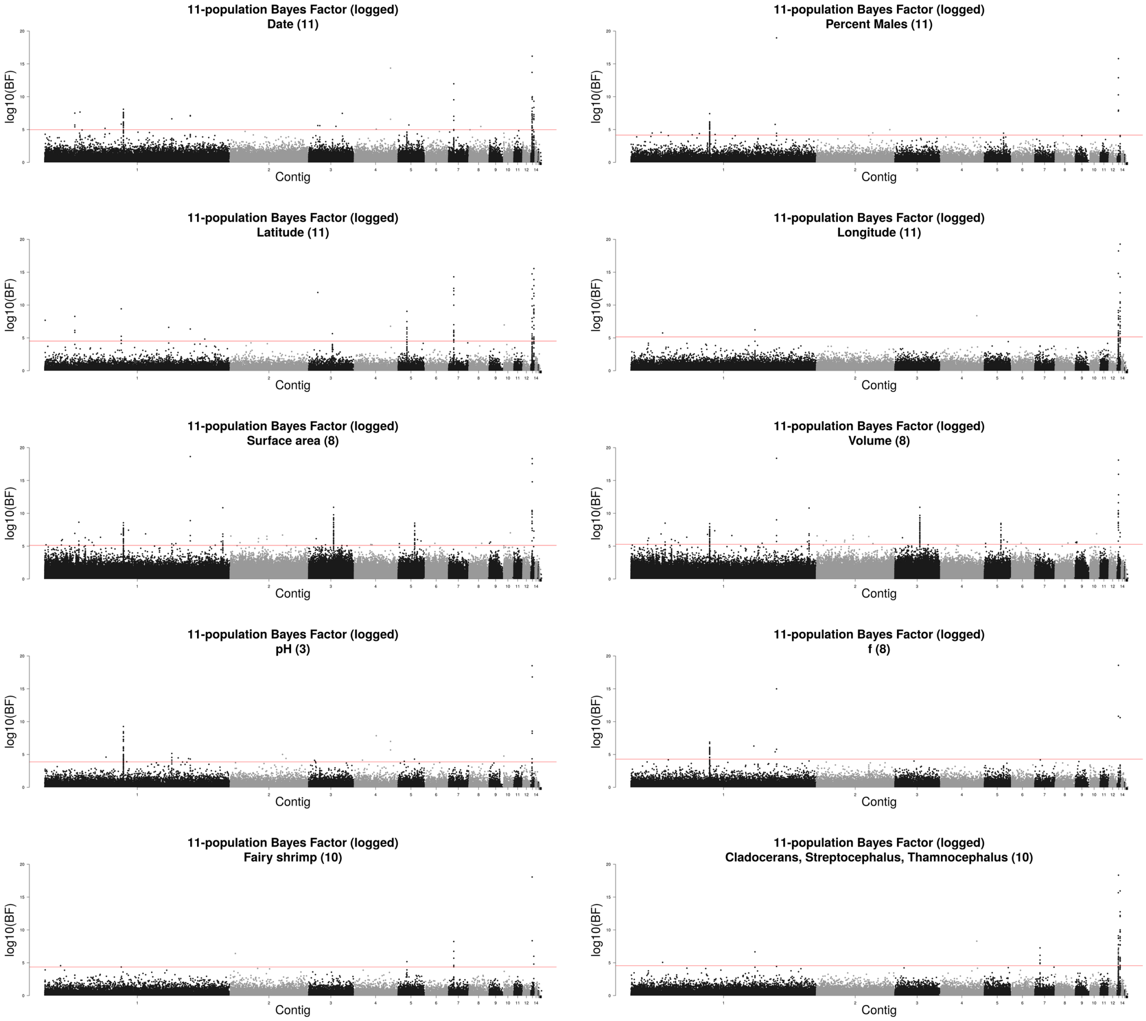


Figure 4.6: Un-windowed Manhattan plots of Bayes' factors across the entire genome, for all examined environmental variable. The environmental variable is printed below each plot. Here, Ap, He, and f refer, respectively, to average number of allele per polymorphic locus, expected number of heterozygotes, and inbreeding coefficient, as presented in [166].

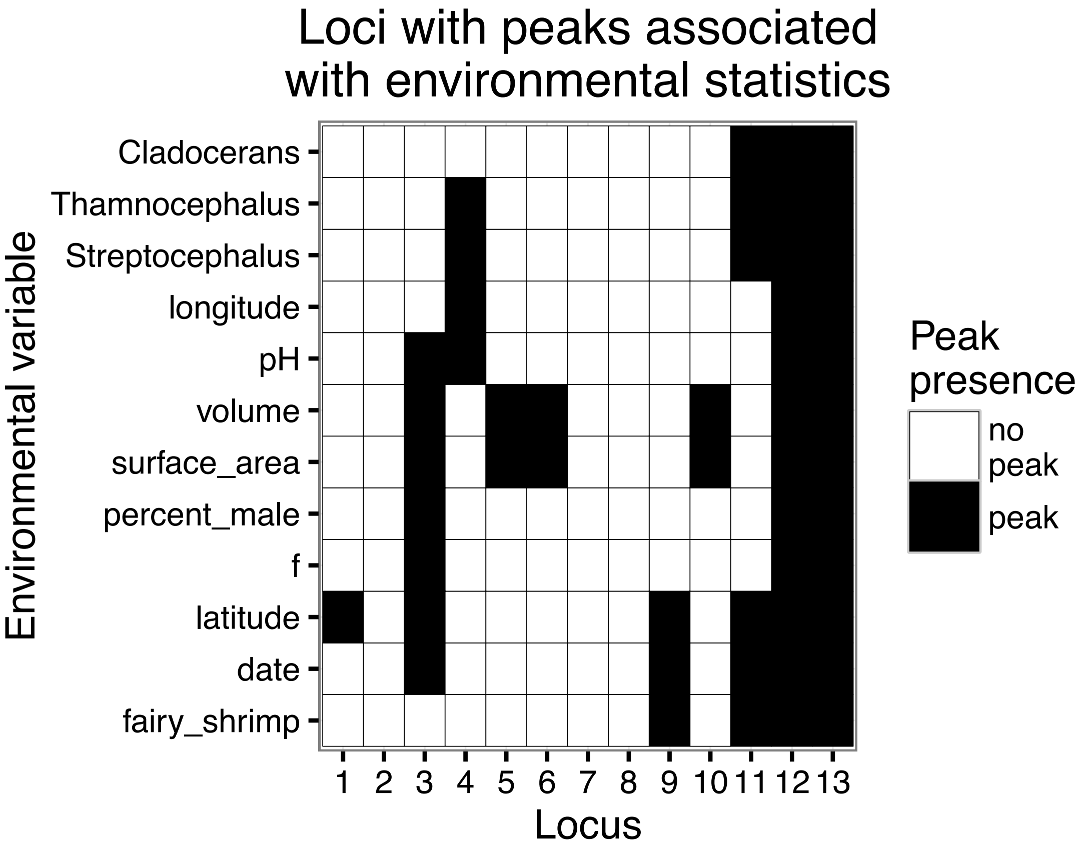


Figure 4.7: A matrix associating loci of interest with environmental variables. The same loci and variables are represented here as is figure 4.6. A black square indicates the presence of a peak at that locus, when associations are tested using Bayes factors for that variable. A white square indicates no peak.

Chapter 5

5.1 Chapter description

Experimental evolution toward salinity resistance in the clam shrimp *Eulimnadia texana*

5.2 Preface

This chapter is not planned to be published outside of this dissertation document. The experimental evolution and all experiments associated with it were carried out by me. Soil samples containing clam shrimp were graciously provided by Stephen Weeks. The improved simulation machinery used herein was written by Kevin Thornton and run and analyzed by me.

5.3 Abstract

Highly replicated experimental evolution of a macroscopic, non-model organism is a novel experimental paradigm that allows for precise detection of loci underlying quantitative traits. I describe here a set of replicate populations of the clam shrimp *Eulimnadia texana* that have been experimentally evolved for approximately 10 generations. I demonstrate that there are repeated phenotypic differences between the experimental populations, which are under selection for salinity tolerance, and the control populations, which are not. Finally, I lay out a set of genomic experiments that could be performed to identify the genetic loci underlying salinity tolerance in the clam shrimp using these populations.

5.4 Introduction

Experimental evolution and resequencing of populations has been proposed as a method for identifying genomic regions that govern a trait under selection. [14]. In an idealized case where selection is applied precisely to a single quantitative trait of interest, this would allow for the identification of genes that influence that trait; in reality, it is often likely that multiple traits are (knowingly or not) under selection. Thus, it is in some ways analogous to QTL mapping and GWAS, other methods for understanding quantitative traits, but not totally so. In terms of its ability to detect loci of interest (whether or not they are QTL underlying the expected trait), there is some evidence that, under certain conditions [14], experimental evolution and resequencing may have both high power to detect loci of importance, and high resolution to accurately localize them. This is in contrast to other QTL-related methodologies: traditional QTL mapping has resulted in high power to explain the heritability of a trait of interest, but low resolution [115]; on the other hand, genome-wide association studies (GWAS) have produced high resolution, but low power [116]. Unfortunately, several evidences [14, 89, 81] indicate that the power to detect selected loci via experimental evolution and resequencing (E&R) is limited except under conditions of large population sizes and high replication. This lack of power is due to the prevalence of genetic drift in relatively small populations of the type generally used in experimental evolution of macroscopic organisms. Although experimental evolution of microorganisms can and has been performed with very high population sizes and/or levels of replication [136, 15, 86, 152, 131], E&R studies in macroorganisms have historically been limited by replication. Here, I detail a set of experimentally evolved populations of the clam shrimp *Eulimnadia texana*, which I have carried through between 8 and 13 generations of evolution (depending on the population). I show that there is an effect of selection for salinity tolerance on the phenotypes of the shrimp, and describe the procedure that would allow for detection of the loci underlying salinity tolerance in *E. texana*. I also describe the rearing methods used to maintain a high volume of large populations of clam shrimp in parallel, which is the first instance of such rearing of which I am aware.

5.5 Methods and Results

5.5.1 Clam shrimp collection and rearing

Clam shrimp populations have been sampled in New Mexico and Arizona in previous experiments [166]. This series of experiments was performed entirely with individuals derived from the WAL population (collected in 2005) (Figure 5.1). We generated the initial population to be used for experimental evolution by simultaneously hydrating a number of soil samples, then collecting all 265 individuals from the initial sample in a single rearing tray. We then allowed these shrimp to have sex and lay eggs together, maintaining a simple population structure. We carried this population through six generations of laboratory survival in order to "acclimate" the shrimp to the laboratory environment. Populations were check each generation to ensure a consistent population size above 250 adults. I performed this check after at least 7 days (time to adulthood is approximately 7 days [165]). In this check, I performed a cursory visual examination of the shrimp and a statistical estimate of the population size. In the visual examination, I determined if a large portion of the population consisted of egg carrying hermaphrodites (egg carrying is, naturally, an indicator of reproductive maturity). I performed the estimate of population size using a sampling without replacement methodology [129]. Soil containing these 6th-generation eggs was used as the progenitor stock of all 32 experimentally evolved lines.

Clam shrimp populations were reared in 50 × 30 × 8 cm disposable aluminum foil catering trays (Catering Essentials, full size steam table pan). In each pan, we mixed 500mL of soil with 6L of water purified via reverse osmosis. 0.3 grams of aquarium salt (API aquarium salt, Mars Fishcare North America, inc.) were added to each tray to ensure that necessary nutrients were available to the shrimp. Trays were checked daily for non-clam shrimp, especially the carnivorous *Triops longicaudatus*, and all non-clam shrimp were immediately removed from trays. We identified the following non-clam shrimp: *Triops longicaudatus*, *Daphnia pulex*, and an unknown species of *Anostraca* fairy shrimp. Trays of shrimp were raised in shelving units containing standard "cool white" fluorescent bulbs (T12 bulb, 40 watt, 1220mm). The top edges of the shrimp trays were within 6cm of the fluorescent lights, and there were 3 trays per light. In order to facilitate hatching [165], lights were switched on upon hydration of the shrimp, and were switched off 48 hours after hydration in order to prevent algal growth. This is in keeping with [165], which indicates that most shrimp hatch within the first 48 hours of hydration. I allowed each generation of shrimp to grow unimpeded for one week (14 days) with a constant water level; after one week, water was allowed to dry naturally, and tanks were completely dry within 21 days of hydration. I maintained a temperature of 28°C in the shrimp rearing room: first, for the health of the shrimp (Stephen Weeks, personal communication), and second, to increase the rate of evaporation of the tanks.

5.5.2 Serial dilution of saline water

I tested the effects of water salinity on clam shrimp survivorship through a serial dilution. I added various quantities of aquarium salt to the shrimp's water upon hydration, then tracked survivorship for 5 days to identify the death rate of the shrimp. I double checked salinity using a standard salinity meter, calibrated against known standard salt concentrations. I estimated the 5-day and 7-day survivorship rates at intervals of .5 ppt NaCl, and found that approximately 75% of individuals died before adulthood in the case of a salinity of 1.2 ppt. I used this salinity as the "experimental" salinity throughout the experiment and used a very low salinity of 0.16 ppt for the "control" salinity.

5.5.3 Experimental evolution setup

Results from Baldwin-Brown et al. 2014 [14] indicated the importance of replication and population size on the power to detect loci under selection in E&R experiments.

The number of circulating haplotypes, likewise, influences the power to resolve these loci accurately. I endeavored to match the quantitative results of this study to the best of my abilities. I seeded 32 populations with soil from the master population described above (raised for 6 generations in the laboratory). I visually inspected each tank after 1 week (approximately the time to adulthood of the clam shrimp [165]); any populations that had a suspicious population size (that is, not obviously in the range of several hundred individuals) was restarted with a fresh soil sample from the previous generation of that population. Population sizes were sporadically checked using capture-without-replacement methods [129] in order to establish that visual inspections of the shrimp populations were not drastically underestimating population sizes. In such a test, a fixed amount of sampling effort is repeatedly used to remove individuals without replacement, with the number of individuals removed counted each time. The distribution of the counts establishes a curve that allows statistical estimation of the number of individuals in a population in cases where the population is too large to manually count. My sampling method was to pass a net through the shrimp-containing tank in five circles over the course of five seconds. As long as this is done consistently, relatively accurate estimation is possible [129]. Although the control populations proceeded relatively smoothly, the experimental populations were subject to more necessary re-runs of low-population generations: the control populations reached an average of 9.05 generations in 422 total hydrations of soil, while the experimental populations reached an average of 5.6 generations in 430 total hydrations (Table 5.1). This may be due to the salinity stress on experimental populations.

5.5.4 Predicted experimental design power

Baldwin-Brown et al. 2014 [14] indicates that, with population sizes in excess of 500, a quantity of founding haplotypes greater than 100, and replication of 15 populations, power to detect the exact location of a causative SNP is just over 50%. That said, Baldwin-Brown et al. 2014 never simulated less than 100 generations, and while generations were found to have a small effect on causative site localization, that may not remain true when generation times are much, much less than 100. The linear model from that paper, which provides a rough estimate of the power to localize causative polymorphisms down to single-SNP resolution, indicates, under very conservative conditions of only 10 generations of selection, 10 populations, 300 individuals per population, 500 founding haplotypes, and a selection coefficient of 0.05 at the hypothetical selected locus, produces a power of 31%. Increasing the number of populations to 14 (a more realistic number, given the number of populations at close to 10 generations) produces a power of 41%.

Still, using the linear model outside of the range in which the original simulations were done leaves some questions open. We generated a more realistic set of simulated data (unpublished) using an updated version of the same simulation machinery from Baldwin-Brown et al. 2014 [14] to test this. This version of the simulation machinery simulates an entire chromosome with multiple causative sites Rather than using a fixed selection coefficient as in the original study, we generated a phenotypic value for each individual as an additive trait summed up from the effect sizes of its causative sites. A gaussian curve was placed around an arbitrarily determined optimum phenotype, and fitness for each individual was determined by calculating the value of this gaussian function for the distance of this individual's phenotype from the optimum phenotype. The options were as follows:

```
expevol_region_qtrait_additive_onerep --selected -i
    ↪ stone_final_table_minor_fixed.bin -I indexfile_n1000_d0
    ↪ .1.txt -o outfile_n1000_d0.1.bin --replicate $REP -S
    ↪ $SEED1 -S $SEED2 -m 100 -N 1000 -t 10 --heritability 0.5
    ↪  -d 0.1 -O 5 -C 0.2 -V 10 --flies --weight 0.1
```

where “m” is the number of funding chromosomes, “N” is the population size per generation, “t” is the number of generations, “heritability” is the heritability of the trait, “d” is the density of causative sites per centimorgan (0.1 corresponds to an average of 5 causative sites per chromosome), “O” indicates the phenotypic optimum value, “C” indicates the fraction of the genome occupied by the chromosome of interest, and V indicates the ratio of the gaussian fitness function’s variance to the variance of the phenotypic variation (phenotypic standard deviation is equal to 1). The recombination map used here is from the *Drosophila melanogaster* X chromosome from release 5 of the *D. melanogaster* reference genome [31], and the starting allele frequencies are derived from those of the DGRP [114] X chromosome. The number of causative sites was poisson distributed, and sites were placed randomly throughout the genome. We generated 100 simulated replicates of a conservative set of experimental values (100 initial haplotypes, a population size of 500, 10 generations of selection, 10 replicate populations, 5 causative sites across the chromosome). After correcting for false positives by setting a cyber-t threshold of 10^{-7} (this reduced the number of false positive SNPs in our control simulations, where no sites were selected, to 0), we found that 44% of causative sites were significant, and 99% of causative sites were within 1Mb of a significant polymorphism. This seems to indicate that the power to detect regions containing a causative site is quite high, but should not be taken as evidence that localization will be easy or possible. As a 1Mb region might contain numerous genes and polymorphisms, the power to detect that such a region is under selection may not be of great value.

We now consider the ability of a study using this experimental design to localize the causative sites precisely. We note one indicator that exact localization of causative sites is low in this case: the number of significant sites with our significance threshold is 103,000. Raising the significance threshold to reduce the number of these pseudo false positive sites to less than 10 required an increase of the significance threshold to 10^{-20}, which reduced the power to detect causative sites to 0.006. Further analysis could determine the precise resolution under these conditions. I speculate that the small number of generations of selection reduces the amount of recombination available during selection, preventing the independent segregation of polymorphisms and causing a small number of large haplotype blocks to be selected. Thus, while the power to detect causative sites at all is likely quite high in these lines, there are good reasons to think that localization of the causative sites here may be low.

5.5.5 Detecting adaptation to salinity

We performed a 2-way nested ANOVA to identify adaptation to salinity in the experimental populations [7]. We identified the two healthiest stocks from the experimental and control populations (EEX11-7 and EEC8-7, respectively) and grew small populations of individuals in cups under both high salt (1.2 ppt) and control salt (0.16ppt) conditions. This formed a set of 4 treatments (88 cups; 22 cups for each treatment; treatments: high salt shrimp in high salt environment, low salt shrimp in low salt environment, high salt shrimp in low salt environment, low salt shrimp in high salt environment, Figure 5.2). We measured the body length of each shrimp in each cup (length being a correlate of fitness [165]) and found that there were several factors with effects on body size: the strain of shrimp and the salt environment were both (unsurprisingly) found to have effects, but most importantly, there was an interaction effect indicating that, compared to the control shrimp, the experimental shrimp had a much higher fitness in the salt environment compared to the control environment (Table 5.1; Fig. 5.3. This indicated that the experimental shrimp were adapted to the salt environment, though there was no evidence that the control shrimp were better adapted to the control environment. We performed this ANOVA in r using the following command: `aov(popbodysize   (strain*treat) + Error(popnum/(strain*treat)), data=alldata)`

5.5.6 organization of samples

We organized all samples systematically to simplify future use of this system. All samples are stored in plastic zip-sealing bags, inside of cardboard boxes. Each box is labeled with its contents (one box with wild samples, one box with the first generation of experimental evolution lines, etc.). The ancestral stock for the experimental evolution is in a set of bags labeled “Big population 7”, with generations 1 through 6 each stored separately. Each set of soil that was ever hydrated is separately stored in a plastic bag. “Experimental” lines (those that have been subjected to high salinity) are labeled with the following scheme: EEXna-g, where “n” is the (numerical) stock ID, “g” is the generation number, and “a” is a letter corresponding to the instance of the particular combination of stock and generation. For instance, EEX3b-7 would refer to the seventh generation of the third experimental line. The “b” indicates that this is the second attempt to produce this population from the previous (EEX3-6) population.

The control (freshwater) lines are labeled similarly, but with “EEC” in the place of “EEX”.

5.6 Discussion

One of the crucial difficulties in experimental evolution has been identifying a model organism with complex traits that can easily be used for experimental evolution and resequencing. Here, we show an example of experimental evolution performed on the clam shrimp *Eulimnadia texana* using extremely high replication to generate a set of populations that will allow for high power to detect regions under selection under a regime of salinity stress. We have documented the techniques used to maintain large numbers of clam shrimp at a low cost, and have shown that adaptation toward resistance of high salinity is occurring in the clam shrimp.

Based on the results of [14] and follow up simulations, this population should be fairly well suited to resequencing and detection of genomic sites of selection via simple statistics. A survey of population allele frequencies, carried out via pooled population sequencing or sequencing of individuals, should produce data well suited to identification of loci under selection. Thus, a fairly straightforward study could likely identify some of the major loci underlying salinity tolerance in the shrimp, though it is possible that more precise localization will not be possible without more generations of selection. It is likely that, using allele frequency estimates of these populations, at least some causative sites would produce a detectable signal, but that the signal would be broad enough that precise localization was not possible. Additionally, because of the existence of an archive of eggs from each generation of evolution, follow-up studies could examine the trajectories of alleles, both selected and not, in these shrimp lines.

We demonstrate here that *E. texana* has a fairly low salinity tolerance in the wild, and that adaptation to higher salinity is possible; however, we do not know what physiological mechanism underlies the difference between these shrimp lines. Evidence exists that osmotic regulation in vernal pool shrimp occurs through specialized neck organs that putatively allow passive transport of sodium and chloride through an outer membrane, and control osmotic levels through active transport of sodium and potassium through a membrane that faces the interior of the organism [135]. One might speculate that this is the region most likely to be physiologically differentiated between the experimental and control shrimp lines. Alternatively, osmotic potential differences are the signal that drive the hatching of newly hydrated shrimp [26]. It is possible that selection in these lines is occurring at this stage of development, or any other. Follow up studies could benefit from direct physiological observation of osmotic potential across the membranes of these neck organs in the various evolved lines, or observation of hatching rates of eggs in solutions of varying salinity.

5.7 Listings

Listing 5.1: ANOVA sum-of-squares table

```
Error: popnum
        Df Sum Sq Mean Sq
strain7  1  4.134   4.134

Error: popnum:strain7
        Df Sum Sq Mean Sq
```

```
strain7  1  177.8   177.8

Error: popnum:treat7
        Df Sum Sq Mean Sq
strain7  1  34.91   34.91

Error: popnum:strain7:treat7
        Df Sum Sq Mean Sq
strain7  1  58.08   58.08

Error: Within
                 Df Sum Sq Mean Sq F value  Pr(>F)
strain7           1    2.0   2.008   2.961 0.08547 .
treat7            1    5.3   5.280   7.785 0.00532 **
strain7:treat7    1    2.6   2.577   3.800 0.05141 .
Residuals      1884 1277.8   0.678
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

5.8 Tables

Stock ID	Control	Experimental
1	9	7
2	5	4
3	4	4

4	11	8
5	8	7
6	13	8
7	NA	NA
8	13	7
9	14	9
10	13	4
11	NA	9
12	10	6
13	12	4
14	9	NA
15	11	6
16	9	8
17	10	9
18	12	9
19*	NA	4
20*	NA	4
21*	NA	5
22*	NA	4
23*	NA	3

Table 5.1: A table showing the number of successful generations of experimental evolution of clam shrimp populations. IDs with an asterisk are those started late into the experiment, after recognizing the slow progress of the salt challenged lines and the possibility of attrition of lines over time.

5.9 Figures

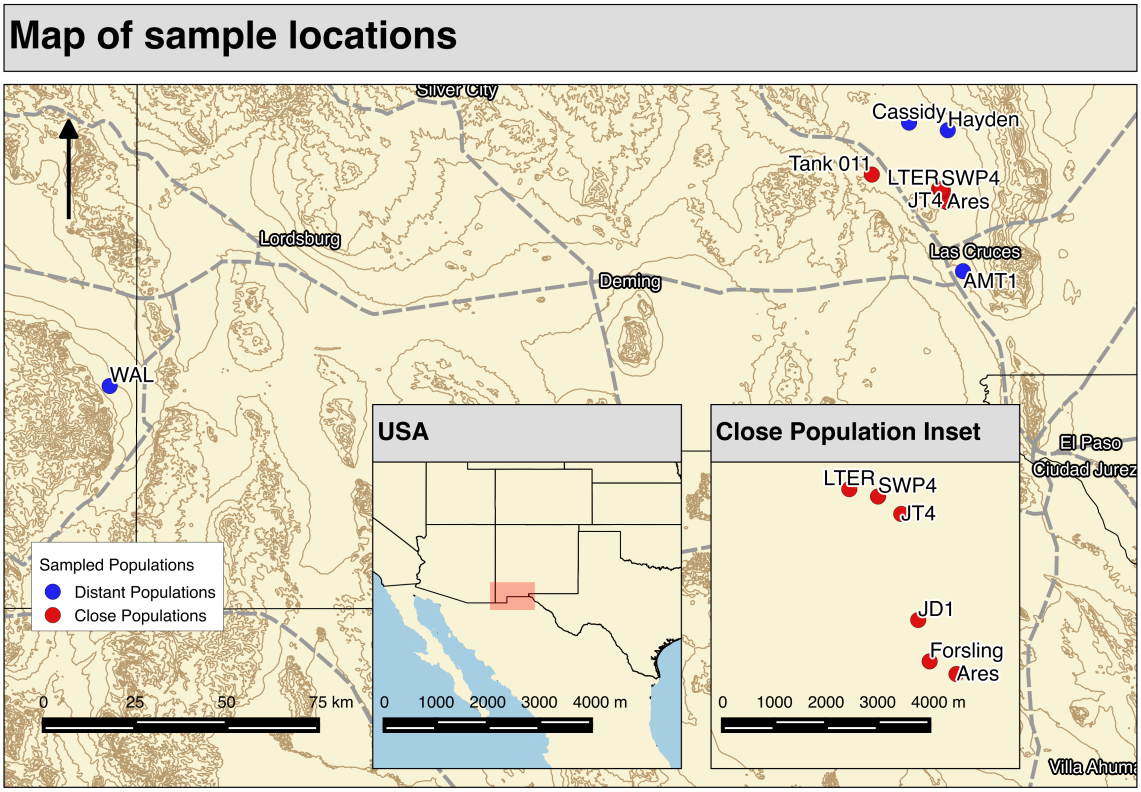


Figure 5.1: A map of the sampling locations for all populations sequenced in this experiment. Note that lab ancestor strain is descended from the WAL population. All populations were taken directly from field sites in New Mexico and Arizona.

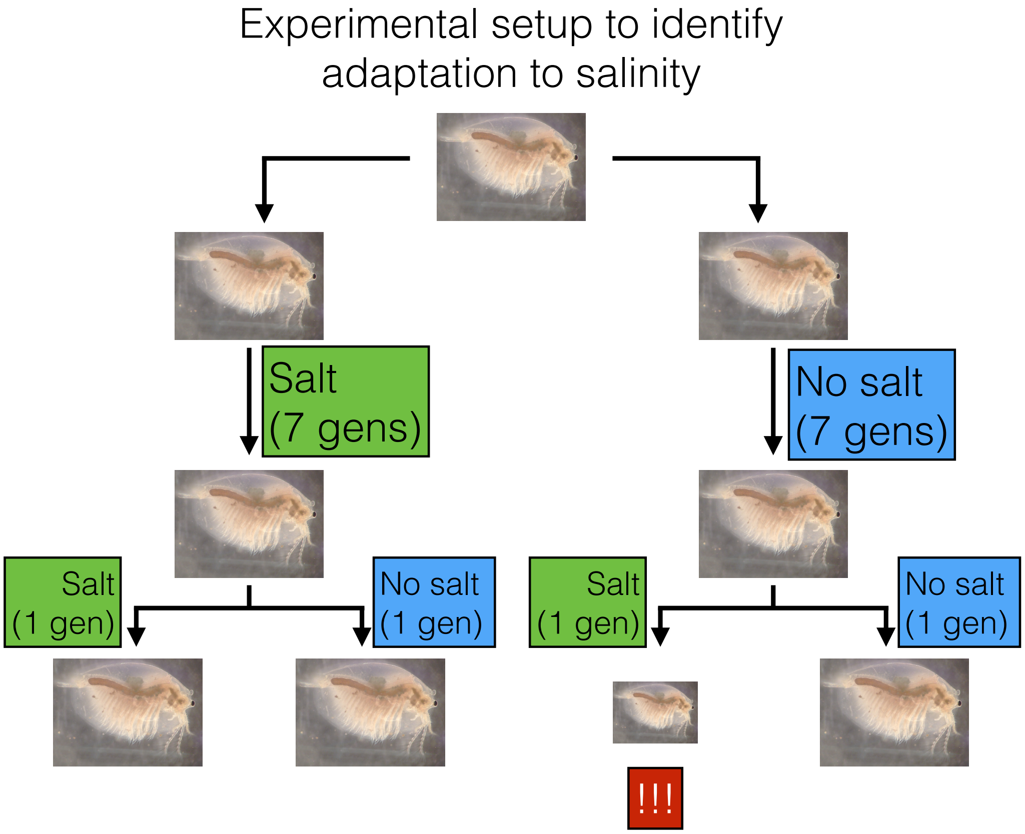


Figure 5.2: A diagram depicting the arrangement of the clam shrimp populations used to identify adaptation to a high salinity environment. Each shrimp picture represents a population of shrimp. The initial population of clam shrimp was divided into subpopulations, which were raised independently for 6 generation, one in salt and one in non-salt. Offspring of these adapted lines were split into 88 groups. 22 "salt" shrimp populations were raised in salt water (SS), 22 "salt" populations were raised in fresh water (SF), 22 "fresh" populations were raised in salt water (FS), and 22 "fresh" populations were raised in fresh water (FF). We performed an ANOVA that showed that the ratio of FF body size to FS body size was significantly larger than the ratio of SF body size to SS body size, because fresh water shrimp raised in salt water are not adapted to the high salinity, and therefore grow poorly.

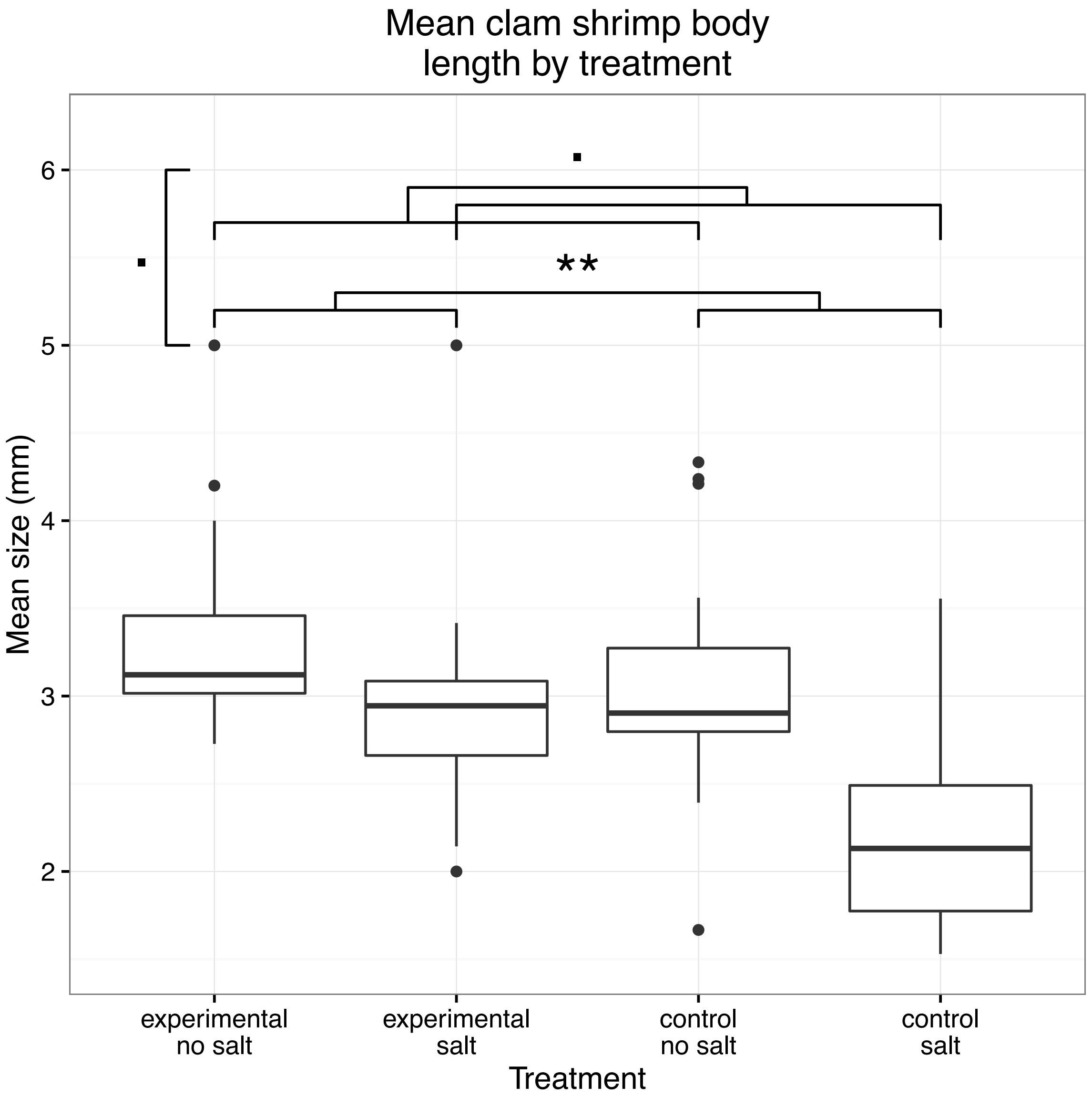


Figure 5.3: A box-and-whisker plot indicating the distribution of body length by treatment, where “experimental” refers to shrimp adapted to salt water, “control” refers to shrimp adapted to fresh water, and “salt” and “no salt” respectively refer to the absence or presence of high salt in the growing medium of the cups. Values used to generate these figures are the mean values of each cup of shrimp in order to minimize the effect of any individual cup, as some cups contained more individuals than others. The tree-like diagram indicates the significance of different comparisons. Namely, the comparison of salt vs. non-salt treatments is highly significant, and the other two comparisons (salt strain vs non-salt strain and the interaction of the other two comparisons) are significant, but not highly so.

Chapter 6

6.1 Chapter description

Conclusion

6.2 Using population differentiation to identify selection

In the previous chapters, I have demonstrated the value and limitations of population differentiation in identifying and understanding selection where it occurs differentially between populations. I demonstrated via simulation that the power to detect loci under selection using statistical tests of allele frequencies in a set of replicated experimentally evolved populations was only high under a narrow set of experimental parameters that have infrequently been met in experimental evolution of macroscopic species. I showed that it was possible to use analyses of allele frequency to identify differentiated loci in the genomes of highly related populations of individuals through the example of a set of natural *Eulimnadia texana* vernal pool clam shrimp populations, where we found a set of loci that we strongly suspect to be under selection in these populations. In service of this *E. texana* analysis, I also developed a new pipeline for genome assembly and compared it to existing assembly methods in an extensive power analysis, then demonstrated its effectiveness on a non-model organism via assembly of the *E. texana* genome. Finally, I developed a set of experimentally evolved populations of *E. texana* that is highly replicated and, matches the levels of replication and population size laid out in my power analysis of experimental evolution, though the number of generations of selection is currently well below the simulated number. I evolved the populations in the laboratory for approximately 10 generations and demonstrated statistically that the experimental populations were phenotypically differentiated from the control populations after only 7 generations of selection.

Many of these projects suggest further studies. My power analysis of experimental evolution indicates a setup for experimental evolution that is very likely to result in highly precise localization of polymorphisms underlying quantitative traits. Since the publication of that first chapter, several other studies [89, 81] have been published that broadly agree with these results, but more sophisticated models of quantitative traits could be applied to such analyses in the future, possibly taking into account the relationship between genotype, phenotype, and fitness, or incorporating optimum phenotype [169] fitness models into the simulations. Additionally, these studies suggest the practical application of experimental evolution and resequencing of macroscopic species, including my experimentally evolved set of *E. texana* populations.

My analysis of genome assembly techniques indicates a gap in current techniques that is waiting to be filled. The central result of the paper, that different genome assembly techniques fail to assemble the genome for different reasons and at different locations in the genome, suggests that it should be possible to write an assembler that can produce considerably higher quality genome assemblies using the same data that is available now. Additionally, our genome assembly merging tool, *quickmerge*, should allow for higher contiguity genome assemblies under a range of circumstances, using only existing data and tools. I hope that this tool will be used by future researchers to improve genome contiguity.

Finally, my examination of natural populations of *E. texana* suggests three further avenues of research. First, our results with the clam shrimp suggest there is much to learn about the ecology and evolution of vernal pool shrimp through genome sequencing and analysis. This small-scale study successfully identified a set of genomic loci that appear to be under selection based on only 11 natural populations, most of which are geographically and genetically very similar to each other. A larger scale experiment, with more thorough measurement of ecological variables and a wider sampling of natural populations, is very likely to identify more loci under selection, and to be able to precisely identify the ecological variables associated with each locus. Second, the loci that we identified present several tempting candidate genes whose effects on phenotype have not been thoroughly examined in *E. texana*, but which appear to be under selection. Gene knockout or swapping studies, performed with CRISPR [139] or some other gene manipulation technology, could reveal a great deal about the traits under selection in these shrimp. Third, this study provides a template for the exploration of population, quantitative, and ecological genetics of non-model organisms using whole-genome sequencing and assembly to identify and compare polymorphism in natural populations. This template can be applied, at a relatively low cost, to any organism that can be reliably sequenced, and we hope that this study design will inform other researchers working on non-model organisms in the future.

Bibliography

[1] A. J. Aberer and A. Stamatakis. Rapid forward-in-time simulation at the chromosome and genome level. *BMC Bioinformatics*, 14(1):216, July 2013.

[2] J. M. Akey, G. Zhang, K. Zhang, L. Jin, and M. D. Shriver. Interrogating a High-Density SNP Map for Signatures of Natural Selection. *Genome Research*, 12(12):1805–1814, Dec. 2002.

[3] V. Alexiadis, A. Lusser, and J. T. Kadonaga. A Conserved N-terminal Motif in Rad54 Is Important for Chromatin Remodeling and Homologous Strand Pairing. *Journal of Biological Chemistry*, 279(26):27824–27829, June 2004.

[4] C. Alkan, B. P. Coe, and E. E. Eichler. Genome structural variation discovery and genotyping. *Nature reviews Genetics*, 12(5):363–76, 2011.

[5] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic local alignment search tool. *Journal of Molecular Biology*, 215(3):403–410, Oct. 1990.

[6] P. Andolfatto and M. Przeworski. Regions of Lower Crossing Over Harbor More Rare Variants in African Populations of Drosophila melanogaster. *Genetics*, 158(2):657–665, June 2001.

[7] F. J. Anscombe. The Validity of Comparative Experiments. *Journal of the Royal Statistical Society. Series A (General)*, 111(3):181–211, 1948.

[8] E. Aronesty. Comparison of Sequencing Utility Programs. *The Open Bioinformatics Journal*, 7(1):1–8, Jan. 2013.

[9] K. F. Au, J. G. Underwood, L. Lee, and W. H. Wong. Improving pacbio long read accuracy by short read alignment. *PLoS One*, 7(10):e46679, 2012.

[10] T. Avidor-Reiss, A. M. Maer, E. Koundakjian, A. Polyanovsky, T. Keil, S. Subramaniam, and C. S. Zuker. Decoding Cilia Function: Defining Specialized Genes Required for Compartmentalized Cilia Biogenesis. *Cell*, 117(4):527–539, May 2004.

[11] D. L. Aylor, W. Valdar, W. Foulds-Mathes, R. J. Buus, R. A. Verdugo, R. S. Baric, M. T. Ferris, J. A. Frelinger, M. Heise, M. B. Frieman, L. E. Gralinski, T. A. Bell, J. D. Didion, K. Hua, D. L. Nehrenberg, C. L. Powell, J. Steigerwalt, Y. Xie, S. N. P. Kelada, F. S. Collins, I. V. Yang, D. A. Schwartz, L. A. Branstetter, E. J. Chesler, D. R. Miller, J. Spence, E. Y. Liu, L. McMillan, A. Sarkar, J. Wang, W. Wang, Q. Zhang, K. W. Broman, R. Korstanje, C. Durrant, R. Mott, F. A. Iraqi, D. Pomp, D. Threadgill, F. P.-M. d. Villena, and G. A. Churchill. Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Research*, 21(8):1213–1222, Aug. 2011.

[12] M. Baker. De novo genome assembly: what every biologist should know. *Nature Methods*, 9(4):333–337, 2012.

[13] P. Baldi and A. D. Long. A Bayesian framework for the analysis of microarray expression data: regularized t -test and statistical inferences of gene changes. *Bioinformatics*, 17(6):509–519, June 2001.

[14] J. G. Baldwin-Brown, A. D. Long, and K. R. Thornton. The Power to Detect Quantitative Trait Loci Using Resequenced, Experimentally Evolved Populations of Diploid, Sexual Organisms. *Molecular Biology and Evolution*, 31(4):1040–1055, Apr. 2014.

[15] J. E. Barrick, D. S. Yu, S. H. Yoon, H. Jeong, T. K. Oh, D. Schneider, R. E. Lenski, and J. F. Kim. Genome evolution and adaptation in a long-term experiment with Escherichia coli. *Nature*, 461(7268):1243–1247, Oct. 2009.

[16] N. H. Barton and P. D. Keightley. Understanding quantitative genetic variation. *Nature Reviews Genetics*, 3(1):11–21, Jan. 2002.

[17] N. H. Barton and M. Turelli. Evolutionary Quantitative Genetics: How Little Do We Know? *Annual Review of Genetics*, 23(1):337–370, 1989.

[18] H. J. Bellen, R. W. Levis, G. Liao, Y. He, J. W. Carlson, G. Tsang, M. Evans-Holm, P. R. Hiesinger, K. L. Schulze, G. M. Rubin, R. A. Hoskins, and A. C. Spradling. The BDGP Gene Disruption Project. *Genetics*, 167(2):761–781, June 2004.

[19] Y. Benjamini and D. Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[20] K. Berlin, S. Koren, C. S. Chin, J. P. Drake, J. M. Landolin, and A. M. Phillippy. Assembling large genomes with single-molecule sequencing and locality-sensitive hashing. *Nat Biotechnol*, 33(6):623–30, 2015.

[21] A. Berry and M. Kreitman. Molecular analysis of an allozyme cline: alcohol dehydrogenase in Drosophila melanogaster on the east coast of North America. *Genetics*, 134(3):869–893, July 1993.

[22] A. J. Bohonak. Genetic population structure of the fairy shrimp Branchinecta coloradensis (Anostraca) in the Rocky Mountains of Colorado. *Canadian Journal of Zoology*, 76(11):2049–2057, Nov. 1998.

[23] B. A. Bour, M. Chakravarti, J. M. West, and S. M. Abmayr. Drosophila SNS, a member of the immunoglobulin superfamily that is essential for myoblast fusion. *Genes & Development*, 14(12):1498–1511, June 2000.

[24] K. R. Bradnam, J. N. Fass, A. Alexandrov, P. Baranay, M. Bechner, I. Birol, S. Boisvert, J. A. Chapman, G. Chapuis, R. Chikhi, H. Chitsaz, W.-C. Chou, J. Corbeil, C. Del Fabbro, T. R. Docking, R. Durbin, D. Earl, S. Emrich, P. Fedotov, N. A. Fonseca, G. Ganapathy, R. A. Gibbs, S. Gnerre, E. Godzaridis, S. Goldstein, M. Haimel, G. Hall, D. Haussler, J. B. Hiatt, I. Y. Ho, J. Howard, M. Hunt, S. D. Jackman, D. B. Jaffe, E. D. Jarvis, H. Jiang, S. Kazakov, P. J. Kersey, J. O. Kitzman, J. R. Knight, S. Koren, T.-W. Lam, D. Lavenier, F. Laviolette, Y. Li, Z. Li, B. Liu, Y. Liu, R. Luo, I. Maccallum, M. D. Macmanes, N. Maillet, S. Melnikov, D. Naquin, Z. Ning, T. D. Otto, B. Paten, O. S. Paulo, A. M. Phillippy, F. Pina-Martins, M. Place, D. Przybylski, X. Qin, C. Qu, F. J. Ribeiro, S. Richards, D. S. Rokhsar, J. G. Ruby, S. Scalabrin, M. C. Schatz, D. C. Schwartz, A. Sergushichev, T. Sharpe, T. I. Shaw, J. Shendure, Y. Shi, J. T. Simpson, H. Song, F. Tsarev, F. Vezzi, R. Vicedomini, B. M. Vieira, J. Wang, K. C. Worley, S. Yin, S.-M. Yiu, J. Yuan, G. Zhang, H. Zhang, S. Zhou, and I. F. Korf. Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species. *GigaScience*, 2(1):10, 2013.

[25] G. Bresler, M. Bresler, and D. Tse. Optimal assembly for high throughput shotgun sequencing. *BMC Bioinformatics*, 14 Suppl 5:S18, 2013.

[26] L. R. Brown and L. H. Carpelan. Egg Hatching and Life History of a Fairy Shrimp Branchinecta Mackini Dexter (Crustacea: Anostraca) in a Mohave Desert Playa (Rabbit Dry Lake). *Ecology*, 52(1):41–54, Jan. 1971.

[27] M. K. Burke, J. P. Dunham, P. Shahrestani, K. R. Thornton, M. R. Rose, and A. D. Long. Genome-wide analysis of a long-term evolution experiment with Drosophila. *Nature*, 467(7315):587–590, Sept. 2010.

[28] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, J. A. Todd, P. Donnelly, J. C. Barrett, P. R. Burton, D. Davison, P. Donnelly, D. Easton, D. Evans, H.-T. Leung, J. L. Marchini, A. P. Morris, C. C. A. Spencer, M. D. Tobin, L. R. Cardon, D. G. Clayton, A. P. Attwood, J. P. Boorman,

B. Cant, U. Everson, J. M. Hussey, J. D. Jolley, A. S. Knight, K. Koch, E. Meech, S. Nutland, C. V. Prowse, H. E. Stevens, N. C. Taylor, G. R. Walters, N. M. Walker, N. A. Watkins, T. Winzer, J. A. Todd, W. H. Ouwehand, R. W. Jones, W. L. McArdle, S. M. Ring, D. P. Strachan, M. Pembrey, G. Breen, D. St Clair, S. Caesar, K. Gordon-Smith, L. Jones, C. Fraser, E. K. Green, D. Grozeva, M. L. Hamshere, P. A. Holmans, I. R. Jones, G. Kirov, V. Moskvina, I. Nikolov, M. C. O'Donovan, M. J. Owen, N. Craddock, D. A. Collier, A. Elkin, A. Farmer, R. Williamson, P. McGuffin, A. H. Young, I. N. Ferrier, S. G. Ball, A. J. Balmforth, J. H. Barrett, D. T. Bishop, M. M. Iles, A. Maqbool, N. Yuldasheva, A. S. Hall, P. S. Braund, P. R. Burton, R. J. Dixon, M. Mangino, S. Stevens, M. D. Tobin, J. R. Thompson, N. J. Samani, F. Bredin, M. Tremelling, M. Parkes, H. Drummond, C. W. Lees, E. R. Nimmo, J. Satsangi, S. A. Fisher, A. Forbes, C. M. Lewis, C. M. Onnie, N. J. Prescott, J. Sanderson, C. G. Mathew, J. Barbour, M. K. Mohiuddin, C. E. Todhunter, J. C. Mansfield, T. Ahmad, F. R. Cummings, D. P. Jewell, J. Webster, M. J. Brown, D. G. Clayton, G. M. Lathrop, J. Connell, A. Dominiczak, N. J. Samani, C. A. B. Marcano, B. Burke, R. Dobson, J. Gungadoo, K. L. Lee, P. B. Munroe, S. J. Newhouse, A. Onipinla, C. Wallace, M. Xue, M. Caulfield, M. Farrall, A. Barton, , T. B. i. R. G. Genomics (BRAGGS), I. N. Bruce, H. Donovan, S. Eyre, P. D. Gilbert, S. L. Hider, A. M. Hinks, S. L. John, C. Potter, A. J. Silman, D. P. M. Symmons, W. Thomson, J. Worthington, D. G. Clayton, D. B. Dunger, S. Nutland, H. E. Stevens, N. M. Walker, B. Widmer, J. A. Todd, T. M. Frayling, R. M. Freathy, H. Lango, J. R. B. Perry, B. M. Shields, M. N. Weedon, A. T. Hattersley, G. A. Hitman, M. Walker, K. S. Elliott, C. J. Groves, C. M. Lindgren, N. W. Rayner, N. J. Timpson, E. Zeggini, M. I. McCarthy, M. Newport, G. Sirugo, E. Lyons, F. Vannberg, A. V. S. Hill, L. A. Bradbury, C. Farrar, J. J. Pointon, P. Wordsworth, M. A. Brown, J. A. Franklyn, J. M. Heward, M. J. Simmonds, S. C. L. Gough, S. Seal, B. C. Susceptibility Collaboration (UK), M. R. Stratton, N. Rahman, M. Ban, A. Goris, S. J. Sawcer, A. Compston, D. Conway, M. Jallow, M. Newport, G. Sirugo, K. A. Rockett, D. P. Kwiatkowski, S. J. Bumpstead, A. Chaney, K. Downes, M. J. R. Ghori, R. Gwilliam, S. E. Hunt, M. Inouye, A. Keniry, E. King, R. McGinnis, S. Potter, R. Ravindrarajah, P. Whittaker, C. Widden, D. Withers, P. Deloukas, H.-T. Leung, S. Nutland, H. E. Stevens, N. M. Walker, J. A. Todd, D. Easton, D. G. Clayton, P. R. Burton, M. D. Tobin, J. C. Barrett, D. Evans, A. P. Morris, L. R. Cardon, N. J. Cardin, D. Davison, T. Ferreira, J. Pereira-Gale, I. B. Hallgrimsdttir, B. N. Howie, J. L. Marchini, C. C. A. Spencer, Z. Su, Y. Y. Teo, D. Vukcevic, P. Donnelly, D. Bentley, M. A. Brown, L. R. Cardon, M. Caulfield, D. G. Clayton, A. Compston, N. Craddock, P. Deloukas, P. Donnelly, M. Farrall, S. C. L. Gough, A. S. Hall, A. T. Hattersley, A. V. S. Hill, D. P. Kwiatkowski, C. G. Mathew, M. I. McCarthy, W. H. Ouwehand, M. Parkes, M. Pembrey, N. Rahman, N. J. Samani, M. R. Stratton, J. A. Todd, and J. Worthington. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, June 2007.

[29] P. R. Burton, D. G. Clayton, L. R. Cardon, N. Craddock, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, N. J. Samani, J. A. Todd, P. Donnelly (Chair), J. C. Barrett, P. R. Burton, D. Davison, P. Donnelly, D. Easton, D. M. Evans, H.-T. Leung, J. L. Marchini, A. P. Morris, C. C. Spencer, M. D. Tobin, L. R. Cardon, D. G. Clayton, A. P. Attwood, J. P. Boorman, B. Cant, U. Everson, J. M. Hussey, J. D. Jolley, A. S. Knight, K. Koch, E. Meech, S. Nutland, C. V. Prowse, H. E. Stevens, N. C. Taylor, G. R. Walters, N. M. Walker, N. A. Watkins, T. Winzer, J. A. Todd, W. H. Ouwehand, R. W. Jones, W. L. McArdle, S. M. Ring, D. P. Strachan, M. Pembrey, G. Breen, D. S. Clair, S. Caesar, K. Gordon-Smith, L. Jones, C. Fraser, E. K. Green, D. Grozeva, M. L. Hamshere, P. A. Holmans, I. R. Jones, G. Kirov, V. Moskivina, I. Nikolov, M. C. O'Donovan, M. J. Owen, N. Craddock, D. A. Collier, A. Elkin, A. Farmer, R. Williamson, P. McGuffin, A. H. Young, I. N. Ferrier, S. G. Ball, A. J. Balmforth, J. H. Barrett, T. D. Bishop, M. M. Iles, A. Maqbool, N. Yuldasheva, A. S. Hall, P. S. Braund, P. R. Burton, R. J. Dixon, M. Mangino, S. Stevens, M. D. Tobin, J. R. Thompson, N. J. Samani, F. Bredin, M. Tremelling, M. Parkes, H. Drummond, C. W. Lees, E. R. Nimmo, J. Satsangi, S. A. Fisher, A. Forbes, C. M. Lewis, C. M. Onnie, N. J. Prescott, J. Sanderson, C. G. Matthew, J. Barbour, M. K. Mohiuddin, C. E. Todhunter, J. C. Mansfield, T. Ahmad, F. R. Cummings, D. P. Jewell, J. Webster, M. J. Brown, D. G. Clayton, M. G. Lathrop, J. Connell, A. Dominiczak, N. J. Samani, C. A. B. Marcano, B. Burke, R. Dobson, J. Gungadoo, K. L. Lee, P. B. Munroe, S. J. Newhouse, A. Onipinla, C. Wallace, M. Xue, M. Caulfield, M. Farrall, A. Barton, I. N. Bruce, H. Donovan, S. Eyre, P. D. Gilbert, S. L. Hilder, A. M. Hinks, S. L. John, C. Potter, A. J. Silman, D. P. Symmons, W. Thomson, J. Worthington, D. G. Clayton, D. B. Dunger, S. Nutland, H. E. Stevens, N. M. Walker, B. Widmer, J. A. Todd, T. M. Frayling, R. M. Freathy, H. Lango, J. R. B. Perry, B. M. Shields, M. N. Weedon, A. T. Hattersley, G. A. Hitman, M. Walker, K. S. Elliott, C. J. Groves, C. M. Lindgren, N. W. Rayner, N. J. Timpson, E. Zeggini, M. I. McCarthy, M. Newport, G. Sirugo, E. Lyons, F. Vannberg, A. V. S. Hill, L. A. Bradbury, C. Farrar, J. J. Pointon, P. Wordsworth, M. A. Brown, J. A. Franklyn, J. M. Heward, M. J. Simmonds, S. C. Gough, S. Seal, M. R. Stratton, N. Rahman, M. Ban, A. Goris, S. J. Sawcer, A. Compston, D. Conway, M. Jallow, M. Newport, G. Sirugo, K. A. Rockett, D. P. Kwiatkowski, S. J. Bumpstead, A. Chaney, K. Downes, M. J. Ghori, R. Gwilliam, S. E. Hunt, M. Inouye, A. Keniry, E. King, R. McGinnis, S. Potter, R. Ravindrarajah, P. Whittaker, C. Widden, D. Withers, P. Deloukas, H.-T. Leung, S. Nutland, H. E. Stevens, N. M. Walker, J. A. Todd, D. Easton, D. G. Clayton, P. R. Burton, M. D. Tobin, J. C. Barrett, D. M. Evans, A. P. Morris, L. R. Cardon, N. J. Cardin, D. Davison, T. Ferreira, J. Pereira-Gale, I. B. Hallgrimsdttir, B. N. Howie, J. L. Marchini, C. C. Spencer, Z. Su, Y. Y. Teo, D. Vukcevic, P. Donnelly, D. Bentley, M. A. Brown, L. R. Cardon, M. Caulfield,

D. G. Clayton, A. Compston, N. Craddock, P. Deloukas, P. Donnelly, M. Farrall, S. C. Gough, A. S. Hall, A. T. Hattersley, A. V. S. Hill, D. P. Kwiatkowski, C. G. Matthew, M. I. McCarthy, W. H. Ouwehand, M. Parkes, M. Pembrey, N. Rahman, N. J. Samani, M. R. Stratton, J. A. Todd, J. Worthington, S. L. Mitchell, P. R. Newby, O. J. Brand, J. Carr-Smith, S. H. S. Pearce, S. C. L. Gough, R. McGinnis, A. Keniry, P. Deloukas, J. D. Reveille, X. Zhou, L. A. Bradbury, A.-M. Sims, A. Dowling, J. Taylor, T. Doan, L. R. Cardon, J. C. Davis, J. J. Pointon, L. Savage, M. M. Ward, T. L. Learch, M. H. Weisman, P. Wordsworth, and M. A. Brown. Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nature Genetics*, 39(11):1329–1337, Oct. 2007.

[30] I. Campos, J. A. Geiger, A. C. Santos, V. Carlos, and A. Jacinto. Genetic Screen in Drosophila melanogaster Uncovers a Novel Set of Genes Required for Embryonic Epithelial Repair. *Genetics*, 184(1):129–140, Jan. 2010.

[31] S. E. Celniker and G. M. Rubin. The Drosophila Melanogaster Genome. *Annual Review of Genomics and Human Genetics*, 4(1):89–117, 2003.

[32] M. J. Chaisson and G. Tesler. Mapping single molecule sequencing reads using basic local alignment with successive refinement (blasr): application and theory. *Bmc Bioinformatics*, 13, 2012.

[33] M. Chakraborty, J. G. Baldwin-Brown, A. D. Long, and J. J. Emerson. Contiguous and accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Research*, page gkw654, July 2016.

[34] C. S. Chin, D. H. Alexander, P. Marks, A. A. Klammer, J. Drake, C. Heiner, A. Clum, A. Copeland, J. Huddleston, E. E. Eichler, S. W. Turner, and J. Korlach. Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nat Methods*, 10(6):563–9, 2013.

[35] P. Chomczynski and N. Sacchi. Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Analytical Biochemistry*, 162(1):156–159, Apr. 1987.

[36] G. A. Churchill and M. S. Waterman. The accuracy of dna sequences: estimating sequence quality. *Genomics*, 14(1):89–98, 1992.

[37] J. K. Colbourne, V. R. Singan, and D. G. Gilbert. wFleaBase: the Daphnia genome database. *BMC Bioinformatics*, 6:45, 2005.

[38] N. Craddock, M. E. Hurles, N. Cardin, R. D. Pearson, V. Plagnol, S. Robson, D. Vukcevic, C. Barnes, D. F. Conrad, E. Giannoulatou, C. Holmes, J. L. Marchini, K. Stirrups, M. D. Tobin, L. V. Wain, C. Yau, J. Aerts, T. Ahmad, T. Daniel Andrews, H. Arbury, A. Attwood, A. Auton, S. G. Ball, A. J.

Balmforth, J. C. Barrett, I. Barroso, A. Barton, A. J. Bennett, S. Bhaskar, K. Blaszczyk, J. Bowes, O. J. Brand, P. S. Braund, F. Bredin, G. Breen, M. J. Brown, I. N. Bruce, J. Bull, O. S. Burren, J. Burton, J. Byrnes, S. Caesar, C. M. Clee, A. J. Coffey, J. M. C. Connell, J. D. Cooper, A. F. Dominiczak, K. Downes, H. E. Drummond, D. Dudakia, A. Dunham, B. Ebbs, D. Eccles, S. Edkins, C. Edwards, A. Elliot, P. Emery, D. M. Evans, G. Evans, S. Eyre, A. Farmer, I. Nicol Ferrier, L. Feuk, T. Fitzgerald, E. Flynn, A. Forbes, L. Forty, J. A. Franklyn, R. M. Freathy, P. Gibbs, P. Gilbert, O. Gokumen, K. Gordon-Smith, E. Gray, E. Green, C. J. Groves, D. Grozeva, R. Gwilliam, A. Hall, N. Hammond, M. Hardy, P. Harrison, N. Hassanali, H. Hebaishi, S. Hines, A. Hinks, G. A. Hitman, L. Hocking, E. Howard, P. Howard, J. M. M. Howson, D. Hughes, S. Hunt, J. D. Isaacs, M. Jain, D. P. Jewell, T. Johnson, J. D. Jolley, I. R. Jones, L. A. Jones, G. Kirov, C. F. Langford, H. Lango-Allen, G. Mark Lathrop, J. Lee, K. L. Lee, C. Lees, K. Lewis, C. M. Lindgren, M. Maisuria-Armer, J. Maller, J. Mansfield, P. Martin, D. C. O. Massey, W. L. McArdle, P. McGuffin, K. E. McLay, A. Mentzer, M. L. Mimmack, A. E. Morgan, A. P. Morris, C. Mowat, S. Myers, W. Newman, E. R. Nimmo, M. C. ODonovan, A. Onipinla, I. Onyiah, N. R. Ovington, M. J. Owen, K. Palin, K. Parnell, D. Pernet, J. R. B. Perry, A. Phillips, D. Pinto, N. J. Prescott, I. Prokopenko, M. A. Quail, S. Rafelt, N. W. Rayner, R. Redon, D. M. Reid, A. Renwick, S. M. Ring, N. Robertson, E. Russell, D. St Clair, J. G. Sambrook, J. D. Sanderson, H. Schuilenburg, C. E. Scott, R. Scott, S. Seal, S. Shaw-Hawkins, B. M. Shields, M. J. Simmonds, D. J. Smyth, E. Somaskantharajah, K. Spanova, S. Steer, J. Stephens, H. E. Stevens, M. A. Stone, Z. Su, D. P. M. Symmons, J. R. Thompson, W. Thomson, M. E. Travers, C. Turnbull, A. Valsesia, M. Walker, N. M. Walker, C. Wallace, M. Warren-Perry, N. A. Watkins, J. Webster, M. N. Weedon, A. G. Wilson, M. Woodburn, B. P. Wordsworth, A. H. Young, E. Zeggini, N. P. Carter, T. M. Frayling, C. Lee, G. McVean, P. B. Munroe, A. Palotie, S. J. Sawcer, S. W. Scherer, D. P. Strachan, C. Tyler-Smith, M. A. Brown, P. R. Burton, M. J. Caulfield, A. Compston, M. Farrall, S. C. L. Gough, A. S. Hall, A. T. Hattersley, A. V. S. Hill, C. G. Mathew, M. Pembrey, J. Satsangi, M. R. Stratton, J. Worthington, P. Deloukas, A. Duncanson, D. P. Kwiatkowski, M. I. McCarthy, W. H. Ouwehand, M. Parkes, N. Rahman, J. A. Todd, N. J. Samani, and P. Donnelly. Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–720, Apr. 2010.

[39] G. dosSantos, A. J. Schroeder, J. L. Goodman, V. B. Strelets, M. A. Crosby, J. Thurmond, D. B. Emmert, W. M. Gelbart, and t. F. Consortium. FlyBase: introduction of the Drosophila melanogaster Release 6 reference genome assembly and large-scale migration of genome annotations. *Nucleic Acids Research*, 43(D1):D690–D697, Jan. 2015.

[40] D. Ducat, S.-i. Kawaguchi, H. Liu, J. R. Yates, and Y. Zheng. Regulation of Mi-

crotubule Assembly and Organization in Mitosis by the AAA+ ATPase Pontin. *Molecular Biology of the Cell*, 19(7):3097–3110, July 2008.

[41] K. Dumstrei, C. Nassif, G. Abboud, A. Aryai, A. Aryai, and V. Hartenstein. EGFR signaling is required for the differentiation and maintenance of neural progenitors along the dorsal midline of the Drosophila embryonic head. *Development*, 125(17):3417–3426, Sept. 1998.

[42] E. Eden, R. Navon, I. Steinfeld, D. Lipson, and Z. Yakhini. GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. *BMC Bioinformatics*, 10:48, 2009.

[43] J. Eid, A. Fehr, J. Gray, K. Luong, J. Lyle, G. Otto, P. Peluso, D. Rank, P. Baybayan, B. Bettman, A. Bibillo, K. Bjornson, B. Chaudhuri, F. Christians, R. Cicero, S. Clark, R. Dalal, A. deWinter, J. Dixon, M. Foquet, A. Gaertner, P. Hardenbol, C. Heiner, K. Hester, D. Holden, G. Kearns, X. Kong, R. Kuse, Y. Lacroix, S. Lin, P. Lundquist, C. Ma, P. Marks, M. Maxham, D. Murphy, I. Park, T. Pham, M. Phillips, J. Roy, R. Sebra, G. Shen, J. Sorenson, A. Tomaney, K. Travers, M. Trulson, J. Vieceli, J. Wegener, D. Wu, A. Yang, D. Zaccarin, P. Zhao, F. Zhong, J. Korlach, and S. Turner. Real-Time DNA Sequencing from Single Polymerase Molecules. *Science*, 323(5910):133–138, Jan. 2009.

[44] E. Elhaik. Empirical Distributions of F ST from Large-Scale Human Polymorphism Data. *PLOS ONE*, 7(11):e49837, Nov. 2012.

[45] L. L. Ellis, W. Huang, A. M. Quinn, A. Ahuja, B. Alfrejd, F. E. Gomez, C. E. Hjelmen, K. L. Moore, T. F. C. Mackay, J. S. Johnston, and A. M. Tarone. Intrapopulation Genome Size Variation in D. melanogaster Reflects Life History Variation and Plasticity. *PLOS Genetics*, 10(7):e1004522, July 2014.

[46] S. Esmaeeli-Nieh, M. Fenckova, I. M. Porter, M. M. Motazacker, B. Nijhof, A. Castells-Nobau, Z. Asztalos, R. Weimann, F. Behjati, A. Tzschach, U. Felbor, H. Scherthan, S. M. Sayfati, H. H. Ropers, K. Kahrizi, H. Najmabadi, J. R. Swedlow, A. Schenck, and A. W. Kuss. BOD1 Is Required for Cognitive Function in Humans and Drosophila. *PLOS Genet*, 12(5):e1006022, May 2016.

[47] S. N. Ethier and T. Nagylaki. Diffusion Approximations of Markov Chains with Two Time Scales and Applications to Population Genetics. *Advances in Applied Probability*, 12(1):14–49, Mar. 1980. ArticleType: research-article / Full publication date: Mar., 1980 / Copyright 1980 Applied Probability Trust.

[48] A. F. Feder, D. A. Petrov, and A. O. Bergland. LDx: Estimation of Linkage Disequilibrium from High-Throughput Pooled Resequencing Data. *PLoS ONE*, 7(11):e48588, Nov. 2012.

[49] R. A. Fisher. The Distribution of Gene Ratios for Rare Mutations. *Proceedings of the Royal Society of Edinburgh*, 1930. Reproduced with permission of the Royal Society of Edinburgh.

[50] E. Frichot, S. D. Schoville, G. Bouchard, and O. Franois. Testing for Associations between Loci and Environmental Gradients Using Latent Factor Mixed Models. *Molecular Biology and Evolution*, 30(7):1687–1699, July 2013.

[51] Y. X. Fu. Statistical Properties of Segregating Sites. *Theoretical Population Biology*, 48(2):172–197, Oct. 1995.

[52] A. Futschik and C. Schltterer. The Next Generation of Molecular Markers From Massively Parallel Sequencing of Pooled DNA Samples. *Genetics*, 186(1):207–218, Sept. 2010.

[53] R. Gandhi, S. Bonaccorsi, D. Wentworth, S. Doxsey, M. Gatti, and A. Pereira. The Drosophila Kinesin-like Protein KLP67a Is Essential for Mitotic and Male Meiotic Spindle Assembly. *Molecular Biology of the Cell*, 15(1):121–131, Jan. 2004.

[54] M. Gautier. Genome-Wide Scan for Adaptive Divergence and Association with Population-Specific Covariates. *Genetics*, 201(4):1555–1579, Dec. 2015.

[55] A. Ghabrial, . Stefan Luschnig, . Mark M. Metzstein, and a. M. A. Krasnow. Branching Morphogenesis of the Drosophila Tracheal System. *Annual Review of Cell and Developmental Biology*, 19(1):623–647, 2003.

[56] S. Gnerre, I. MacCallum, D. Przybylski, F. J. Ribeiro, J. N. Burton, B. J. Walker, T. Sharpe, G. Hall, T. P. Shea, S. Sykes, A. M. Berlin, D. Aird, M. Costello, R. Daza, L. Williams, R. Nicol, A. Gnirke, C. Nusbaum, E. S. Lander, and D. B. Jaffe. High-quality draft assemblies of mammalian genomes from massively parallel sequence data. *Proceedings of the National Academy of Sciences*, 108(4):1513–1518, Jan. 2011.

[57] S. N. Goodman. Toward Evidence-Based Medical Statistics. 2: The Bayes Factor. *Annals of Internal Medicine*, 130(12):1005–1013, June 1999.

[58] S. Goodwin, J. Gurtowski, S. Ethe-Sayers, P. Deshpande, M. C. Schatz, and W. R. McCombie. Oxford nanopore sequencing, hybrid error correction, and de novo assembly of a eukaryotic genome. *Genome research*, 25(11):1750–6, 2015.

[59] D. Gordon, J. Huddleston, M. J. Chaisson, C. M. Hill, Z. N. Kronenberg, K. M. Munson, M. Malig, A. Raja, I. Fiddes, L. W. Hillier, C. Dunn, C. Baker, J. Armstrong, M. Diekhans, B. Paten, J. Shendure, R. K. Wilson, D. Haussler, C. S. Chin, and E. E. Eichler. Long-read sequence assembly of the gorilla genome. *Science*, 352(6281):aae0344, 2016.

[60] M. M. Gorski, J. C. J. Eeken, A. W. M. d. Jong, I. Klink, M. Loos, R. J. Romeijn, B. L. v. Veen, L. H. Mullenders, W. Ferro, and A. Pastink. The Drosophila melanogaster DNA Ligase IV Gene Plays a Crucial Role in the Repair of Radiation-Induced DNA Double-Strand Breaks and Acts Synergistically With Rad54. *Genetics*, 165(4):1929–1941, Dec. 2003.

[61] G. Goshima and R. D. Vale. Cell Cycle-dependent Dynamics and Regulation of Mitotic Kinesins in Drosophila S2 Cells. *Molecular Biology of the Cell*, 16(8):3896–3907, Aug. 2005.

[62] M. G. Grabherr, B. J. Haas, M. Yassour, J. Z. Levin, D. A. Thompson, I. Amit, X. Adiconis, L. Fan, R. Raychowdhury, Q. Zeng, Z. Chen, E. Mauceli, N. Hacohen, A. Gnirke, N. Rhind, F. di Palma, B. W. Birren, C. Nusbaum, K. Lindblad-Toh, N. Friedman, and A. Regev. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nature biotechnology*, 29(7):644–652, May 2011.

[63] C. A. Graham and A. J. Hill. Introduction to dna sequencing. *Methods in molecular biology (Clifton, N J)*, 167:1–12, 2001.

[64] S. J. Gratz, A. M. Cummings, J. N. Nguyen, D. C. Hamm, L. K. Donohue, M. M. Harrison, J. Wildonger, and K. M. O'Connor-Giles. Genome Engineering of Drosophila with the CRISPR RNA-Guided Cas9 Nuclease. *Genetics*, 194(4):1029–1035, Aug. 2013.

[65] X. Guo, J. L. Engel, J. Xiao, V. S. Tagliabracci, X. Wang, L. Huang, and J. E. Dixon. UBLCP1 is a 26s proteasome phosphatase that regulates nuclear proteasome activity. *Proceedings of the National Academy of Sciences*, 108(46):18649–18654, Nov. 2011.

[66] A. Gurevich, V. Saveliev, N. Vyahhi, and G. Tesler. Quast: quality assessment tool for genome assemblies. *Bioinformatics*, 29(8):1072–5, 2013.

[67] R. N. Gutenkunst, R. D. Hernandez, S. H. Williamson, and C. D. Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLoS Genet*, 5(10):e1000695, Oct. 2009.

[68] T. Gnther and G. Coop. Robust Identification of Local Adaptation from Allele Frequencies. *Genetics*, 195(1):205–220, Sept. 2013.

[69] W. G. Hill, M. E. Goddard, and P. M. Visscher. Data and Theory Point to Mainly Additive Genetic Variance for Complex Traits. *PLoS Genet*, 4(2):e1000008, Feb. 2008.

[70] W. G. Hill and B. S. Weir. Variances and covariances of squared linkage disequilibria in finite populations. *Theoretical Population Biology*, 33(1):54–78, Feb. 1988.

[71] R. A. Hoskins, C. D. Smith, J. W. Carlson, A. B. Carvalho, A. Halpern, J. S. Kaminker, C. Kennedy, C. J. Mungall, B. A. Sullivan, G. G. Sutton, J. C. Yasuhara, B. T. Wakimoto, E. W. Myers, S. E. Celniker, G. M. Rubin, and G. H. Karpen. Heterochromatic sequences in a drosophila whole-genome shotgun assembly. *Genome Biol*, 3(12):RESEARCH0085, 2002.

[72] R. A. Jain and E. R. Gavis. The Drosophila hnRNP M homolog Rumpelstiltskin regulates nanos mRNA localization. *Development*, 135(5):973–982, Mar. 2008.

[73] A. M. Johansson, M. E. Pettersson, P. B. Siegel, and O. Carlborg. Genome-wide effects of long-term divergent selection. *PLoS genetics*, 6(11):e1001188, Nov. 2010.

[74] T. Johnson and N. Barton. Theoretical models of selection and mutation on quantitative traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1459):1411–1425, July 2005.

[75] R. Kajitani, K. Toshimoto, H. Noguchi, A. Toyoda, Y. Ogura, M. Okuno, M. Yabana, M. Harada, E. Nagayasu, H. Maruyama, Y. Kohara, A. Fujiyama, T. Hayashi, and T. Itoh. Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads. *Genome Res*, 24(8):1384–95, 2014.

[76] J. S. Kaminker, C. M. Bergman, B. Kronmiller, J. Carlson, R. Svirskas, S. Patel, E. Frise, D. A. Wheeler, S. E. Lewis, G. M. Rubin, M. Ashburner, and S. E. Celniker. The transposable elements of the Drosophila melanogaster euchromatin: a genomics perspective. *Genome Biology*, 3(12):research0084.1–84.2, 2002.

[77] S. Katsuma, T. Daimon, K. Mita, and T. Shimada. Lepidopteran Ortholog of Drosophila Breathless Is a Receptor for the Baculovirus Fibroblast Growth Factor. *Journal of Virology*, 80(11):5474–5481, June 2006.

[78] P. D. Keightley, R. W. Ness, D. L. Halligan, and P. R. Haddrill. Estimation of the Spontaneous Mutation Rate per Nucleotide Site in a Drosophila melanogaster Full-Sib Family. *Genetics*, 196(1):313–320, Jan. 2014.

[79] D. R. Kelley, M. C. Schatz, and S. L. Salzberg. Quake: quality-aware detection and correction of sequencing errors. *Genome Biology*, 11:R116, 2010.

[80] W. J. Kent. BLATThe BLAST-Like Alignment Tool. *Genome Research*, 12(4):656–664, Apr. 2002.

[81] D. Kessner and J. Novembre. Power analysis of artificial selection experiments using efficient whole genome simulation of quantitative traits. *Genetics*, 199(4):991–1005, Apr. 2015.

[82] K. E. Kim, P. Peluso, P. Babayan, P. J. Yeadon, C. Yu, W. W. Fisher, C. S. Chin, N. A. Rapicavoli, D. R. Rank, J. Li, D. E. Catcheside, S. E. Celniker, A. M. Phillippy, C. M. Bergman, and J. M. Landolin. Long-read, whole-genome shotgun sequence data for five model organisms. *Sci Data*, 1:140045, 2014.

[83] M. Kimura. Diffusion Models in Population Genetics. *Journal of Applied Probability*, 1(2):177–232, Dec. 1964. ArticleType: research-article / Full publication date: Dec., 1964 / Copyright 1964 Applied Probability Trust.

[84] E. G. King, C. M. Merkes, C. L. McNeil, S. R. Hoofer, S. Sen, K. W. Broman, A. D. Long, and S. J. Macdonald. Genetic dissection of a model complex trait using the Drosophila Synthetic Population Resource. *Genome Research*, 22(8):1558–1566, Aug. 2012.

[85] R. D. Kinser and P. J. Dolph. Cathepsin proteases mediate photoreceptor cell degeneration in Drosophila. *Neurobiology of Disease*, 46(3):655–662, June 2012.

[86] T. Kishimoto, L. Iijima, M. Tatsumi, N. Ono, A. Oyake, T. Hashimoto, M. Matsuo, M. Okubo, S. Suzuki, K. Mori, A. Kashiwagi, C. Furusawa, B.-W. Ying, and T. Yomo. Transition from Positive to Neutral in Mutation Fixation along with Continuing Rising Fitness in Thermal Adaptive Evolution. *PLoS Genet*, 6(10):e1001164, Oct. 2010.

[87] H. Kitagawa, T. Uyama, and K. Sugahara. Molecular Cloning and Expression of a Human Chondroitin Synthase. *Journal of Biological Chemistry*, 276(42):38721–38726, Oct. 2001.

[88] R. Kofler, P. Orozco-terWengel, N. D. Maio, R. V. Pandey, V. Nolte, A. Futschik, C. Kosiol, and C. Schltterer. PoPoolation: A Toolbox for Population Genetic Analysis of Next Generation Sequencing Data from Pooled Individuals. *PLOS ONE*, 6(1):e15925, Jan. 2011.

[89] R. Kofler and C. Schltterer. A guide for the design of evolve and resequencing studies. *Molecular Biology and Evolution*, 31(2):474–483, Feb. 2014.

[90] R. Kooistra, K. Vreeken, J. B. Zonneveld, A. d. Jong, J. C. Eeken, C. J. Osgood, J. M. Buerstedde, P. H. Lohman, and A. Pastink. The Drosophila melanogaster RAD54 homolog, DmRAD54, is involved in the repair of radiation damage and recombination. *Molecular and Cellular Biology*, 17(10):6097–6104, Oct. 1997.

[91] A. C. Koon and V. Budnik. Inhibitory Control of Synaptic and Behavioral Plasticity by Octopaminergic Signaling. *The Journal of Neuroscience*, 32(18):6312–6322, May 2012.

[92] S. Koren, G. P. Harhay, T. P. Smith, J. L. Bono, D. M. Harhay, S. D. McVey, D. Radune, N. H. Bergman, and A. M. Phillippy. Reducing assembly complexity

of microbial genomes with single-molecule sequencing. *Genome Biol*, 14(9):R101, 2013.

[93] S. Koren and A. M. Phillippy. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology*, 23:110–120, 2015.

[94] S. Koren, M. C. Schatz, B. P. Walenz, J. Martin, J. T. Howard, G. Ganapathy, Z. Wang, D. A. Rasko, W. R. McCombie, E. D. Jarvis, and M. P. Adam. Hybrid error correction and de novo assembly of single-molecule sequencing reads. *Nat Biotechnol*, 30(7):693–700, 2012.

[95] P. X. Kover, W. Valdar, J. Trakalo, N. Scarcelli, I. M. Ehrenreich, M. D. Purugganan, C. Durrant, and R. Mott. A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana. *PLoS Genet*, 5(7):e1000551, July 2009.

[96] S. Kurtz, A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu, and S. L. Salzberg. Versatile and open software for comparing large genomes. *Genome Biology*, 5(2), 2004.

[97] K.-K. Lam, A. Khalak, and D. Tse. Near-optimal assembly for shotgun sequencing with noisy reads. *Bmc Bioinformatics*, 15, 2014.

[98] S. Lamichhaney, G. Fan, F. Widemo, U. Gunnarsson, D. S. Thalmann, M. P. Hoeppner, S. Kerje, U. Gustafson, C. Shi, H. Zhang, W. Chen, X. Liang, L. Huang, J. Wang, E. Liang, Q. Wu, S. M.-Y. Lee, X. Xu, J. Hglund, X. Liu, and L. Andersson. Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nature Genetics*, 48(1):84–88, Jan. 2016.

[99] J. H. Lan, Y. Yin, E. F. Reed, K. Moua, K. Thomas, and Q. Zhang. Impact of three Illumina library construction methods on GC bias and HLA genotype calling. *Human Immunology*, 76(23):166–175, Mar. 2015.

[100] E. S. Lander and M. S. Waterman. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics*, 2(3):231–9, 1988.

[101] C. H. Langley, K. Stevens, C. Cardeno, Y. C. Lee, D. R. Schrider, J. E. Pool, S. A. Langley, C. Suarez, R. B. Corbett-Detig, B. Kolaczkowski, S. Fang, P. M. Nista, A. K. Holloway, A. D. Kern, C. N. Dewey, Y. S. Song, M. W. Hahn, and D. J. Begun. Genomic variation in natural populations of drosophila melanogaster. *Genetics*, 192(2):533–98, 2012.

[102] T. Laver, J. Harrison, P. A. ONeill, K. Moore, A. Farbos, K. Paszkiewicz, and D. J. Studholme. Assessing the performance of the Oxford Nanopore Technologies MinION. *Biomolecular Detection and Quantification*, 3:1–8, Mar. 2015.

[103] J.-H. Lee and J. A. Fischer. Drosophila Tel2 Is Expressed as a Translational Fusion with EpsinR and Is a Regulator of Wingless Signaling. *PLOS ONE*, 7(9):e46357, Sept. 2012.

[104] J.-H. Lee, E. Overstreet, E. Fitch, S. Fleenor, and J. A. Fischer. Drosophila liquid facets-Related encodes Golgi epsin and is an essential gene required for cell proliferation, growth, and patterning. *Developmental Biology*, 331(1):1–13, July 2009.

[105] M. Lei. The MCM Complex: Its Role in DNA Replication and Implications for Cancer Therapy. *Current Cancer Drug Targets*, 5(5):365–380, Aug. 2005.

[106] T. Lenormand. The Evolution of Sex Dimorphism in Recombination. *Genetics*, 163(2):811–822, Feb. 2003.

[107] P. A. Leventis, T. R. D. Sylva, N. Rajwans, S. Wasiak, P. S. McPherson, and G. L. Boulianne. Liquid facets-Related (lqfR) Is Required for Egg Chamber Morphogenesis during Drosophila Oogenesis. *PLOS ONE*, 6(10):e25466, Oct. 2011.

[108] H. Li and R. Durbin. Fast and accurate short read alignment with BurrowsWheeler transform. *Bioinformatics*, 25(14):1754–1760, July 2009.

[109] K. E. Lotterhos and M. C. Whitlock. Evaluation of demographic history and neutral parameterization on the performance of FST outlier tests. *Molecular Ecology*, 23(9):2178–2192, May 2014.

[110] M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, 15(12):1–21, Dec. 2014.

[111] W.-J. Lu, J. Chapo, I. Roig, and J. M. Abrams. Meiotic Recombination Provokes Functional Activation of the p53 Regulatory Network. *Science*, 328(5983):1278–1281, June 2010.

[112] R. Luo, B. Liu, Y. Xie, Z. Li, W. Huang, J. Yuan, G. He, Y. Chen, Q. Pan, Y. Liu, J. Tang, G. Wu, H. Zhang, Y. Shi, Y. Liu, C. Yu, B. Wang, Y. Lu, C. Han, D. W. Cheung, S. M. Yiu, S. Peng, Z. Xiaoqian, G. Liu, X. Liao, Y. Li, H. Yang, J. Wang, T. W. Lam, and J. Wang. Soapdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, 1(1):18, 2012.

[113] M. Lynch, D. Bost, S. Wilson, T. Maruki, and S. Harrison. Population-Genetic Inference from Pooled-Sequencing Data. *Genome Biology and Evolution*, 6(5):1210–1218, May 2014.

[114] T. F. C. Mackay, 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. Barrn, 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. Rmia, 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. The Drosophila melanogaster Genetic Reference Panel. *Nature*, 482(7384):173–178, Feb. 2012.

[115] T. F. C. Mackay, E. A. Stone, and J. F. Ayroles. The genetics of quantitative traits: challenges and prospects. *Nature Reviews Genetics*, 10(8):565–577, Aug. 2009.

[116] T. A. Manolio, F. S. Collins, N. J. Cox, D. B. Goldstein, L. A. Hindorff, D. J. Hunter, M. I. McCarthy, E. M. Ramos, L. R. Cardon, A. Chakravarti, J. H. Cho, A. E. Guttmacher, A. Kong, L. Kruglyak, E. Mardis, C. N. Rotimi, M. Slatkin, D. Valle, A. S. Whittemore, M. Boehnke, A. G. Clark, E. E. Eichler, G. Gibson, J. L. Haines, T. F. C. Mackay, S. A. McCarroll, and P. M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, Oct. 2009.

[117] F. Marroni, S. Pinosio, G. Zaina, F. Fogolari, N. Felice, F. Cattonaro, and M. Morgante. Nucleotide diversity and linkage disequilibrium in Populus nigra cinnamyl alcohol dehydrogenase (CAD4) gene. *Tree Genetics & Genomes*, 7(5):1011–1023, Apr. 2011.

[118] G. Marais and C. Kingsford. A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics*, 27(6):764–770, Mar. 2011.

[119] R. C. McCoy, R. W. Taylor, T. A. Blauwkamp, J. L. Kelley, M. Kertesz, D. Pushkarev, D. A. Petrov, and A.-S. Fiston-Lavier. Illumina truseq synthetic long-reads empower de novo assembly and resolve complex, highly-repetitive transposable elements. *PloS one*, 9(9):e106689, 2014.

[120] A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, and M. A. DePristo. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, 20(9):1297–1303, Sept. 2010.

[121] N. T. Meisler, L. M. Nutter, and J. W. Thanassi. Vitamin B6 Metabolism in Liver and Liver-derived Tumors. *Cancer Research*, 42(9):3538–3543, Sept. 1982.

[122] P. W. Messer. SLiM: Simulating Evolution with Selection and Linkage. *Genetics*, 194(4):1037–1039, Aug. 2013.

[123] J. R. Miller, S. Koren, and G. Sutton. Assembly algorithms for next-generation sequencing data. *Genomics*, 95(6):315–27, 2010.

[124] A. Motahari, K. Ramchandran, D. Tse, and N. Ma. Optimal dna shotgun sequencing: Noisy reads are as good as noiseless reads. *2013 Ieee International Symposium on Information Theory Proceedings (Isit)*, pages 1640–1644, 2013.

[125] E. W. Myers. Toward simplifying and accurately formulating fragment assembly. *Journal of computational biology : a journal of computational molecular cell biology*, 2(2):275–90, 1995.

[126] E. W. Myers, G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams, and J. C. Venter. A whole-genome assembly of drosophila. *Science*, 287(5461):2196–204, 2000.

[127] G. G. Neely, A. Hess, M. Costigan, A. C. Keene, S. Goulas, M. Langeslag, R. S. Griffin, I. Belfer, F. Dai, S. B. Smith, L. Diatchenko, V. Gupta, C.-p. Xia, S. Amann, S. Kreitz, C. Heindl-Erdmann, S. Wolz, C. V. Ly, S. Arora, R. Sarangi, D. Dan, M. Novatchkova, M. Rosenzweig, D. G. Gibson, D. Truong, D. Schramek, T. Zoranovic, S. J. F. Cronin, B. Angjeli, K. Brune, G. Dietzl, W. Maixner, A. Meixner, W. Thomas, J. A. Pospisilik, M. Alenius, M. Kress, S. Subramaniam, P. A. Garrity, H. J. Bellen, C. J. Woolf, and J. M. Penninger. A Genome-wide Drosophila Screen for Heat Nociception Identifies 23 as an Evolutionarily Conserved Pain Gene. *Cell*, 143(4):628–638, Nov. 2010.

[128] R. Nielsen, S. Williamson, Y. Kim, M. J. Hubisz, A. G. Clark, and C. Bustamante. Genomic scans for selective sweeps using SNP data. *Genome Research*, 15(11):1566–1575, Nov. 2005.

[129] D. H. Ogle. *FSA: Fisheries Stock Analysis*, 2016. R package version 0.8.10.

[130] P. Orozco-Terwengel, M. Kapun, V. Nolte, R. Kofler, T. Flatt, and C. Schltterer. Adaptation of Drosophila to a novel laboratory environment reveals temporally heterogeneous trajectories of selected alleles. *Molecular ecology*, 21(20):4931–4941, Oct. 2012.

[131] L. Parts, F. A. Cubillos, J. Warringer, K. Jain, F. Salinas, S. J. Bumpstead, M. Molin, A. Zia, J. T. Simpson, M. A. Quail, A. Moses, E. J. Louis, R. Durbin, and G. Liti. Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Research*, 21(7):1131–1138, July 2011.

[132] P. Pavlidis, D. ivkovi, A. Stamatakis, and N. Alachiotis. SweeD: Likelihood-Based Detection of Selective Sweeps in Thousands of Genomes. *Molecular Biology and Evolution*, page mst112, June 2013.

[133] M. Pendleton, R. Sebra, A. W. Pang, A. Ummat, O. Franzen, T. Rausch, A. M. Stutz, W. Stedman, T. Anantharaman, A. Hastie, H. Dai, M. H. Fritz, H. Cao, A. Cohain, G. Deikus, R. E. Durrett, S. C. Blanchard, R. Altman, C. S. Chin, Y. Guo, E. E. Paxinos, J. O. Korbel, R. B. Darnell, W. R. McCombie, P. Y. Kwok, C. E. Mason, E. E. Schadt, and A. Bashir. Assembly and diploid architecture of an individual human genome via single-molecule technologies. *Nat Methods*, 12(8):780–6, 2015.

[134] A. J. Pereira, B. Dalby, R. J. Stewart, S. J. Doxsey, and L. S. B. Goldstein. Mitochondrial Association of a Plus EndDirected Microtubule Motor Expressed during Mitosis in Drosophila. *The Journal of Cell Biology*, 136(5):1081–1090, Mar. 1997.

[135] W. T. W. Potts and C. T. Durning. Physiological evolution in the branchiopods. *Comparative Biochemistry and Physiology Part B: Comparative Biochemistry*, 67(3):475–484, Jan. 1980.

[136] M. M. Riehle, A. F. Bennett, and A. D. Long. Genetic architecture of thermal adaptation in Escherichia coli. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2):525–530, Jan. 2001.

[137] L. Salmela and E. Rivals. Lordec: accurate and efficient long read error correction. *Bioinformatics*, 30(24):3506–3514, 2014.

[138] S. L. Salzberg, A. M. Phillippy, A. Zimin, D. Puiu, T. Magoc, S. Koren, T. J. Treangen, M. C. Schatz, A. L. Delcher, M. Roberts, G. Marcais, M. Pop, and J. A. Yorke. Gage: A critical evaluation of genome assemblies and assembly algorithms. *Genome Res*, 22(3):557–67, 2012.

[139] J. D. Sander and J. K. Joung. CRISPR-Cas systems for editing, regulating and targeting genomes. *Nature Biotechnology*, 32(4):347–355, Apr. 2014.

[140] C. Sassaman and S. C. Weeks. The Genetic Mechanism of Sex Determination in the Conchostracan Shrimp Eulimnadia texana. *The American Naturalist*, 141(2):314–328, Feb. 1993.

[141] D. Schmucker, J. C. Clemens, H. Shu, C. A. Worby, J. Xiao, M. Muda, J. E. Dixon, and S. L. Zipursky. Drosophila Dscam Is an Axon Guidance Receptor Exhibiting Extraordinary Molecular Diversity. *Cell*, 101(6):671–684, June 2000.

[142] T. Schpbach and E. Wieschaus. Female sterile mutations on the second chromosome of Drosophila melanogaster. II. Mutations blocking oogenesis or altering egg morphology. *Genetics*, 129(4):1119–1136, Dec. 1991.

[143] D. Sellis, B. J. Callahan, D. A. Petrov, and P. W. Messer. Heterozygote advantage as a natural consequence of adaptation in diploids. *Proceedings of the National Academy of Sciences of the United States of America*, 108(51):20666–20671, Dec. 2011.

[144] R. Shen, J.-B. Fan, D. Campbell, W. Chang, J. Chen, D. Doucet, J. Yeakley, M. Bibikova, E. Wickham Garcia, C. McBride, F. Steemers, F. Garcia, B. G. Kermani, K. Gunderson, and A. Oliphant. High-throughput SNP genotyping on universal bead arrays. *Mutation Research*, 573(1-2):70–82, June 2005.

[145] H. Signer-Hasler, C. Flury, B. Haase, D. Burger, H. Simianer, T. Leeb, and S. Rieder. A Genome-Wide Association Study Reveals Loci Influencing Height and Other Conformation Traits in Horses. *PLoS ONE*, 7(5):e37282, May 2012.

[146] J. T. Simpson and M. Pop. The theory and practice of genome sequence assembly. *Annu Rev Genomics Hum Genet*, 16:153–72, 2015.

[147] J. T. Simpson, K. Wong, S. D. Jackman, J. E. Schein, S. J. Jones, and I. Birol. Abyss: a parallel assembler for short read sequence data. *Genome Res*, 19(6):1117–23, 2009.

[148] N. D. Singh, J. D. Jensen, A. G. Clark, and C. F. Aquadro. Inferences of Demography and Selection in an African Population of Drosophila melanogaster. *Genetics*, 193(1):215–228, Jan. 2013.

[149] K. S. Sinsimer, R. A. Jain, S. Chatterjee, and E. R. Gavis. A late phase of germ plasm accumulation during Drosophila oogenesis requires Lost and Rumpelstiltskin. *Development*, 138(16):3431–3440, Aug. 2011.

[150] C. C. A. Spencer, Z. Su, P. Donnelly, and J. Marchini. Designing Genome-Wide Association Studies: Sample Size, Power, Imputation, and the Choice of Genotyping Chip. *PLoS Genet*, 5(5):e1000477, May 2009.

[151] M. Stanke and S. Waack. Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics*, 19(suppl 2):ii215–ii225, Sept. 2003.

[152] O. Tenaillon, A. Rodrguez-Verdugo, R. L. Gaut, P. McDonald, A. F. Bennett, A. D. Long, and B. S. Gaut. The molecular diversity of adaptive convergence. *Science (New York, N.Y.)*, 335(6067):457–461, Jan. 2012.

[153] H. Teotnio, I. M. Chelo, M. Bradi, M. R. Rose, and A. D. Long. Experimental evolution reveals natural selection on standing genetic variation. *Nature genetics*, 41(2):251–257, Feb. 2009.

[154] C. Trapnell, L. Pachter, and S. L. Salzberg. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics*, 25(9):1105–1111, May 2009.

[155] T. J. Treangen and S. L. Salzberg. Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics*, 13(1):36–46, Nov. 2011.

[156] T. L. Turner and P. M. Miller. Investigating Natural Variation in Drosophila Courtship Song by the Evolve and Resequence Approach. *Genetics*, 191(2):633–642, June 2012.

[157] T. L. Turner, A. D. Stewart, A. T. Fields, W. R. Rice, and A. M. Tarone. Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. *PLoS genetics*, 7(3):e1001336, Mar. 2011.

[158] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A Map of Recent Positive Selection in the Human Genome. *PLOS Biol*, 4(3):e72, Mar. 2006.

[159] B. J. Walker, T. Abeel, T. Shea, M. Priest, A. Abouelliel, S. Sakthikumar, C. A. Cuomo, Q. Zeng, J. Wortman, S. K. Young, and A. M. Earl. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*, 9(11):e112963, 2014.

[160] Z. Wang, M. Gerstein, and M. Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, Jan. 2009.

[161] G. A. Watterson. On the number of segregating sites in genetical models without recombination. *Theoretical Population Biology*, 7(2):256–276, Apr. 1975.

[162] S. C. Weeks. Levels of inbreeding depression over seven generations of selfing in the androdioecious clam shrimp, Eulimnadia texana. *Journal of evolutionary biology*, 17(3):475–484, May 2004.

[163] S. C. Weeks, C. Benvenuto, T. F. Sanderson, and R. J. Duff. Sex chromosome evolution in the clam shrimp, Eulimnadia texana. *Journal of Evolutionary Biology*, 23(5):1100–1106, 2010.

[164] S. C. Weeks, E. G. Chapman, D. C. Rogers, D. M. Senyo, and W. R. Hoeh. Evolutionary transitions among dioecy, androdioecy and hermaphroditism in limnadiid clam shrimp (Branchiopoda: Spinicaudata). *Journal of Evolutionary Biology*, 22(9):1781–1799, 2009.

[165] S. C. Weeks, V. Marcus, and S. Alvarez. Notes on the life history of the clam shrimp, Eulimnadia texana. *Hydrobiologia*, 359(1-3):191–197, Dec. 1997.

[166] S. C. Weeks and N. Zucker. Rates of inbreeding in the androdioecious clam shrimp Eulimnadia texana. *Canadian Journal of Zoology*, 77(9):1402–1408, Nov. 1999.

[167] D. S. Wei and Y. S. Rong. A Genetic Screen For DNA Double-Strand Break Repair Mutations in Drosophila. *Genetics*, 177(1):63–77, Sept. 2007.

[168] B. S. Weir and C. C. Cockerham. Estimating F-Statistics for the Analysis of Population Structure. *Evolution*, 38(6):1358–1370, Nov. 1984.

[169] S. Wright. *The role of mutation, inbreeding, crossbreeding, and selection in evolution.*, volume 1. 1932.

[170] Z. Yang and J. P. Bielawski. Statistical methods for detecting molecular adaptation. *Trends in Ecology & Evolution*, 15(12):496–503, Dec. 2000.

[171] C. Ye, C. M. Hill, S. Wu, J. Ruan, and Z. S. Ma. DBG2olc: Efficient Assembly of Large Genomes Using Long Erroneous Reads of the Third Generation Sequencing Technologies. *Scientific Reports*, 6:31900, Aug. 2016.

[172] C. J. Yoo and S. L. Wolin. La proteins from Drosophila melanogaster and Saccharomyces cerevisiae: a yeast homolog of the La autoantigen is dispensable for growth. *Molecular and Cellular Biology*, 14(8):5412–5424, Aug. 1994.

Appendix A

Supplementary information for all chapters

A.1 Chapter 1 supplementary figures

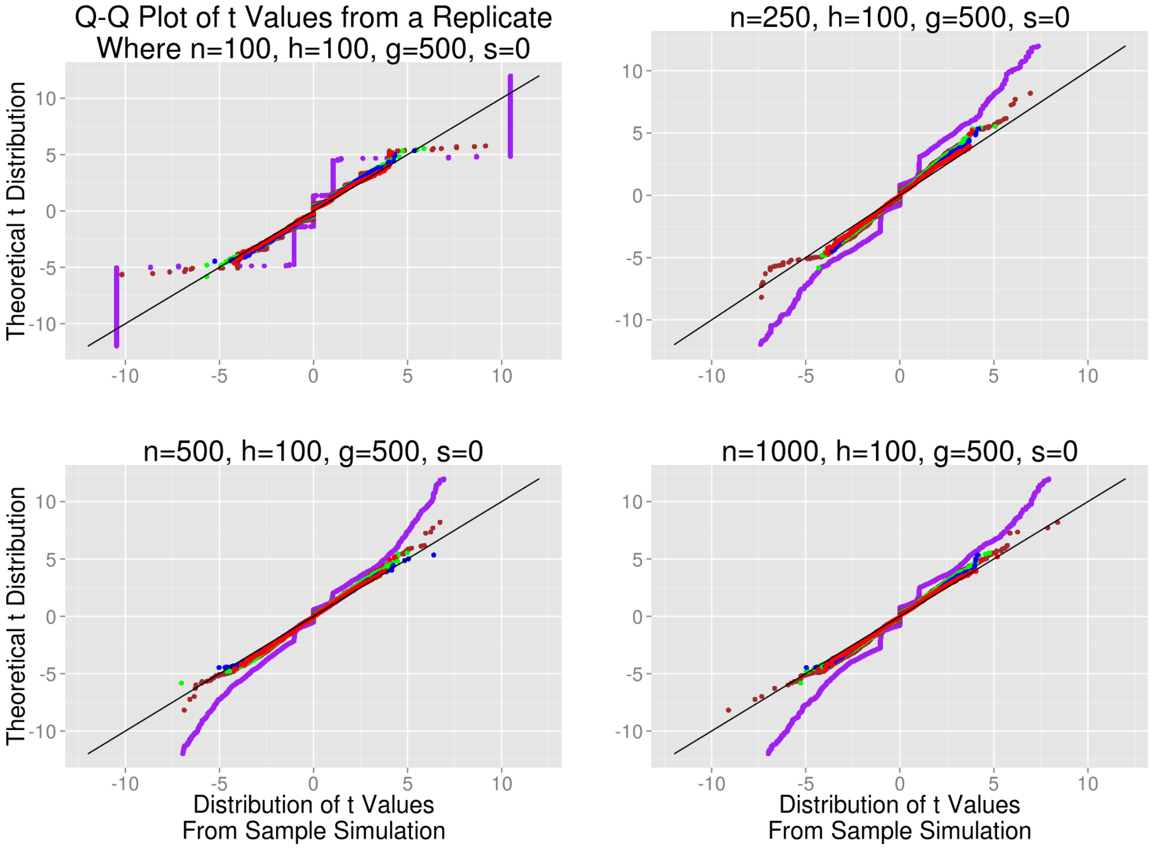


Figure A.1: This figure depicts a set of Q-Q plots illustrating the distribution of the t statistic used in this study versus a theoretical t distribution. In each case, the t values were calculated from a single, randomly chosen replicate simulation. In all cases, $s = 0$ so that the plots will illustrate the effect of drift alone on the generated t values. Purple, brown, green, blue and red points correspond, respectively, to the cases where $r = 2$, $r = 5$, $r = 10$, $r = 15$, and $r = 25$.

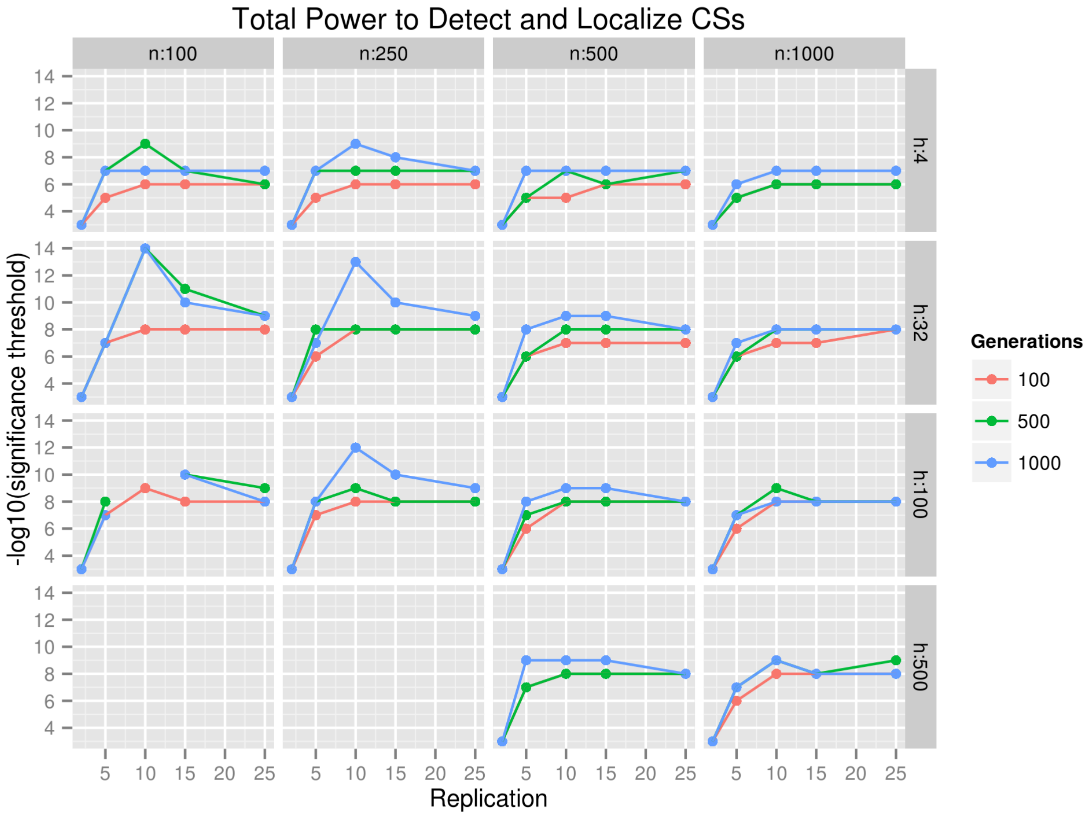


Figure A.2: The log10(p) value chosen as the significance threshold for every simulated parameter combination. Because false positive rates are independent of s, s is not shown. Parameter combinations that produced unacceptably high false positive rates even when the significance threshold was $p10 - 14$ are depicted as missing points.

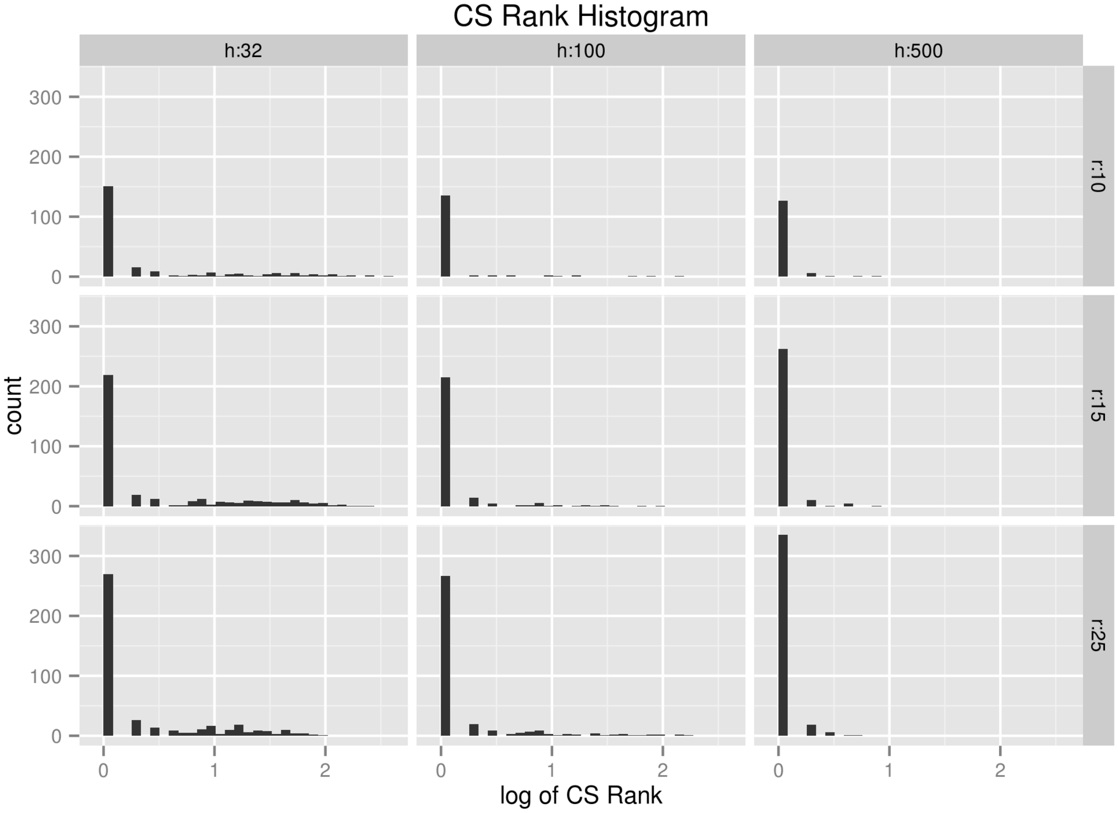


Figure A.3: A histogram depicting the distribution of the rank order position of the CS when all SNPs in a region are ordered from most significant to least significant after 500 generations of selection with 500 individuals per population and a selection coefficient at the CS of 0.05 in all cases where the MSM was significant. Variation in population size is not shown because its effects are similar to variation in replication. The count refers to the number of pure replicates out of 500 that fell into a given range. Note the increase in low-CS-rank hits due to selective sweeps when haplotype number is low.

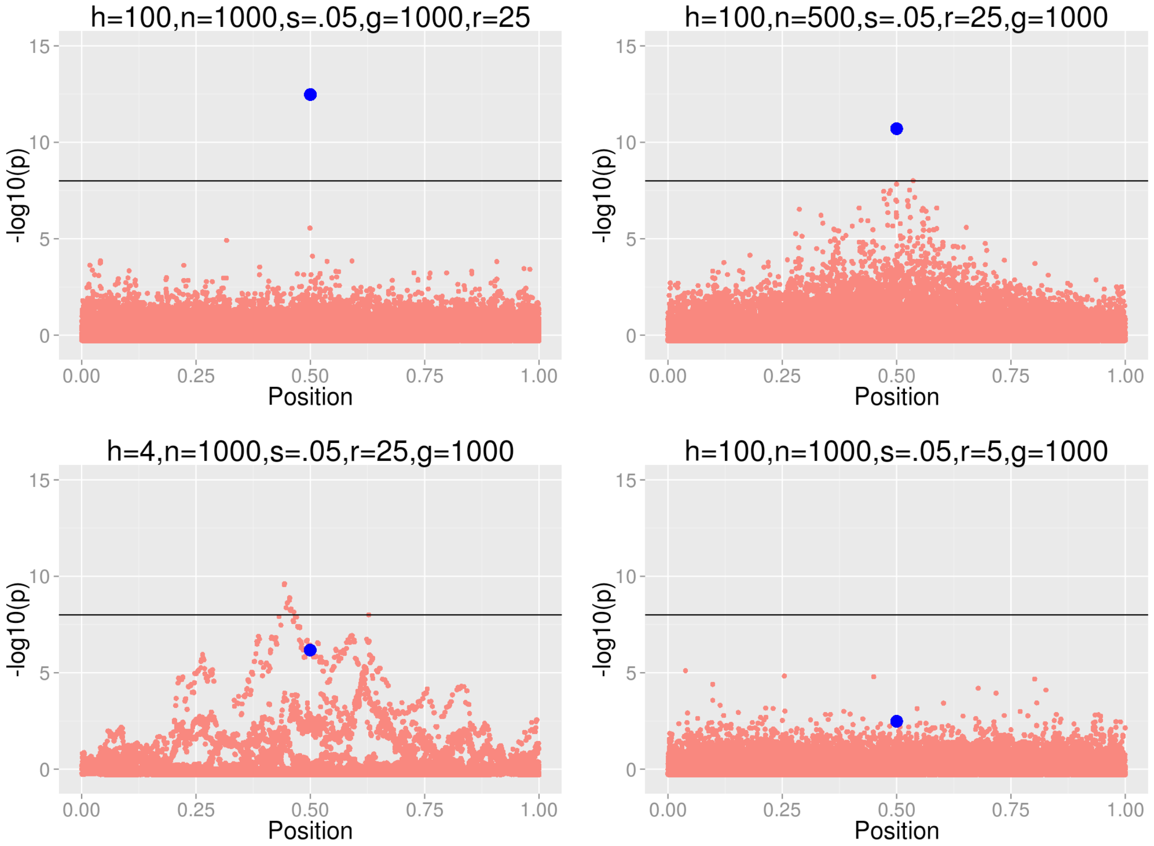


Figure A.4: Sample plots of significance across genome regions under various . The blue point is the CS, while the red points represent all other SNPs. Top left is a plot of relatively ideal conditions, where the CS is the MSM and no selective sweeps are evident. Top right is a plot showing diminished n. Bottom left shows the blocks of linkage disequilibrium found when h is low. Bottom right shows the inability to detect the causative SNP when r is low.

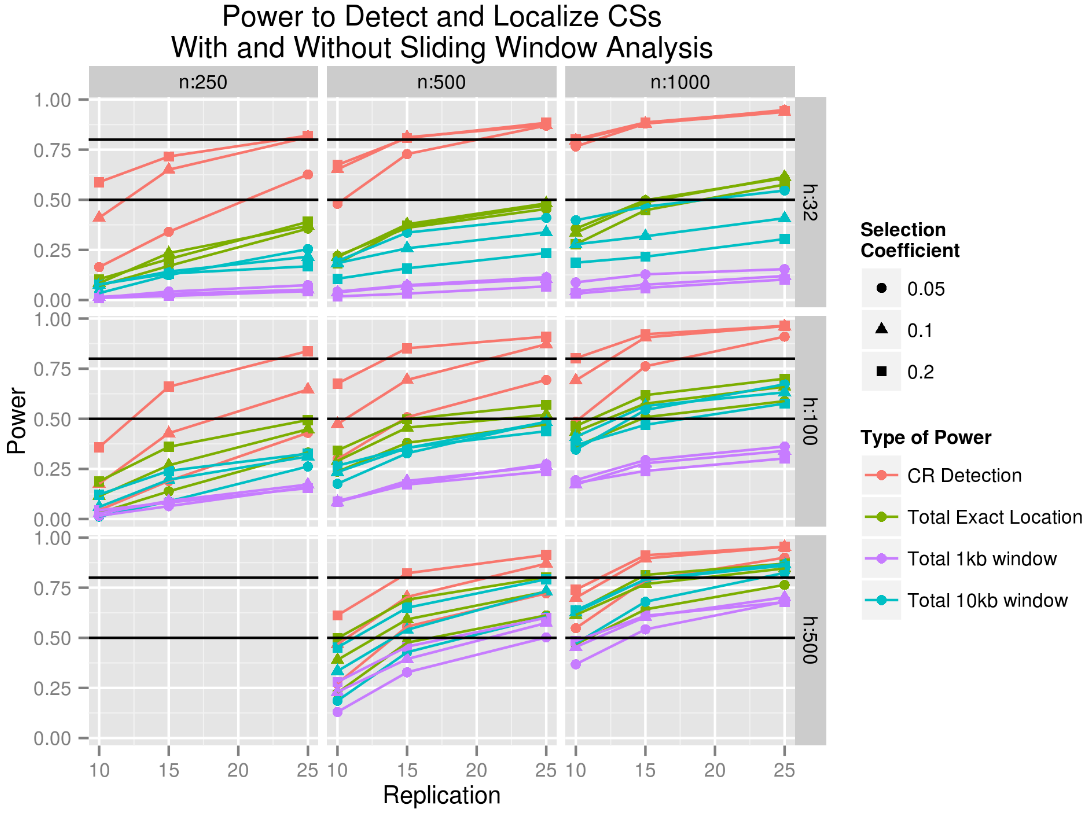


Figure A.5: The ability (total power) to detect a CS-containing region and either correctly identify the exact location of a CS or decrease the number of candidate loci to a manageable number after 1000 generations with a selection coefficient at the CS of 0.05 or higher. In other words, the fraction of all simulations in which a region is a region contains a significant SNP and one of three methods of detecting a CS is successful: the MSM is the CS (Total Exact Location Power), the CS is included in the most significant 1kb window in the region (Total 1kb window power) or the CS is included in the most significant 10kb window in the region (Total 10kb window power). Also shown is the CR Detection Power, which is the fraction of regions that contained at least one significant SNP. By design, all total powers listed here must be lower than the CR detection power. In every instance where at least one SNP was significant, the window with the largest sum of log10(p) values was found. If the most significant window in the region contained the CS, the CS was considered correctly detected. The black lines indicate 50% power and 80% power.

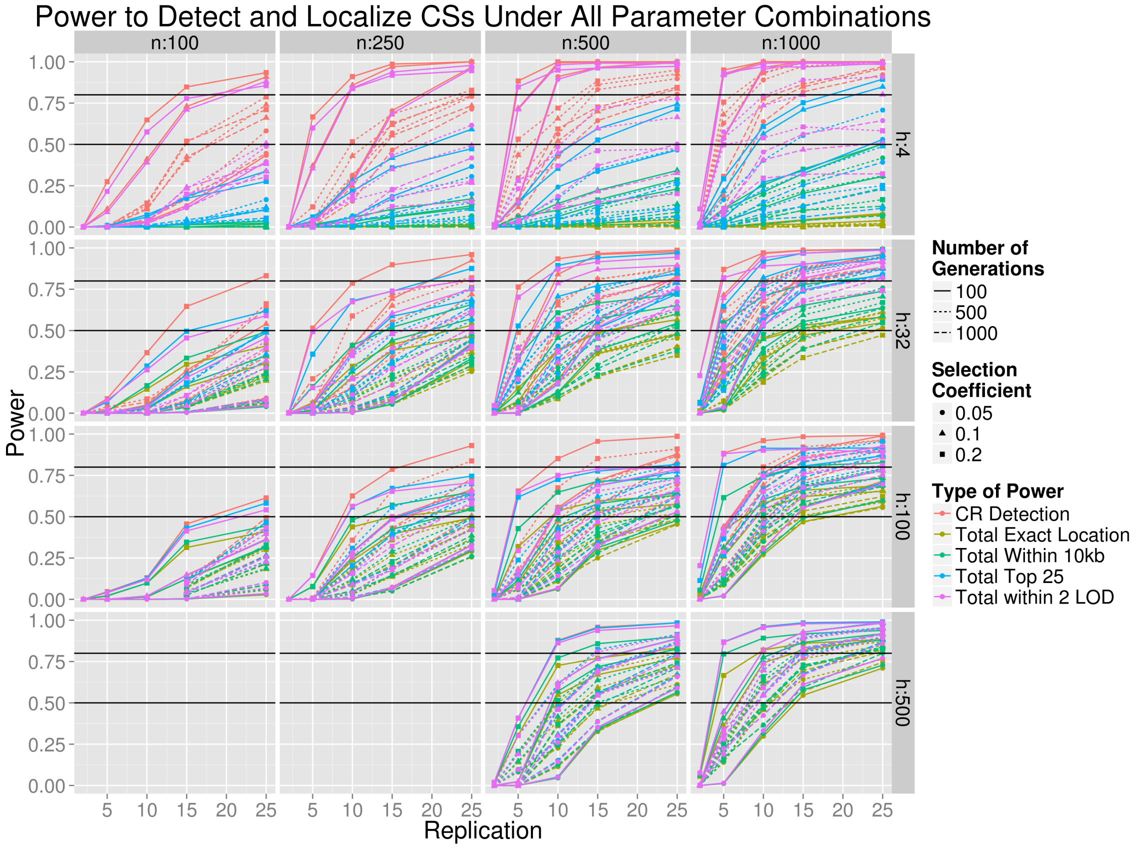


Figure A.6: The ability (total power) to detect a CS-containing region and either correctly identify the exact location of a CS or decrease the number of candidate loci to a manageable number under all . In other words, the fraction of all simulations in which a region is a region contains a significant SNP and one of three methods of detecting a CS is successful: the MSM is the CS (Total Exact Location Power), the CS is within 10kb of the MSM (Total Within 10kb Power), the CS is one of the 25 most significantly diverged SNPs in the region (Total Top 25 Power), or the CS is within 2 LOD of the MSM (Total within 2 LOD power). Also shown is the CR Detection Power, which is the fraction of regions that contained at least one significant SNP. The black lines indicate 50% power and 80% power.

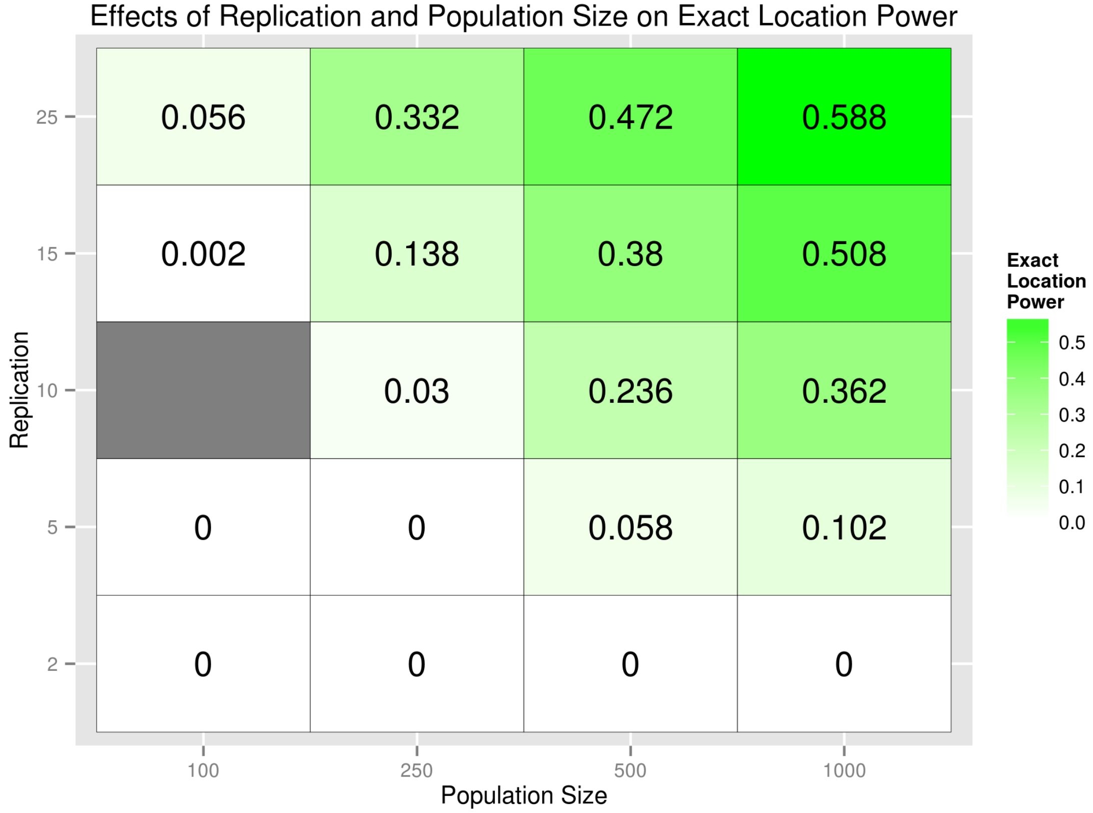


Figure A.7: Heat map depicting exact location power rate at various levels of n and h. $s = 0.05$, $h = 100$, $g = 500$. The power where $n = 100$ and $r = 10$ is missing because none of our tested significance thresholds was strict enough to sufficiently limit false positives in that .

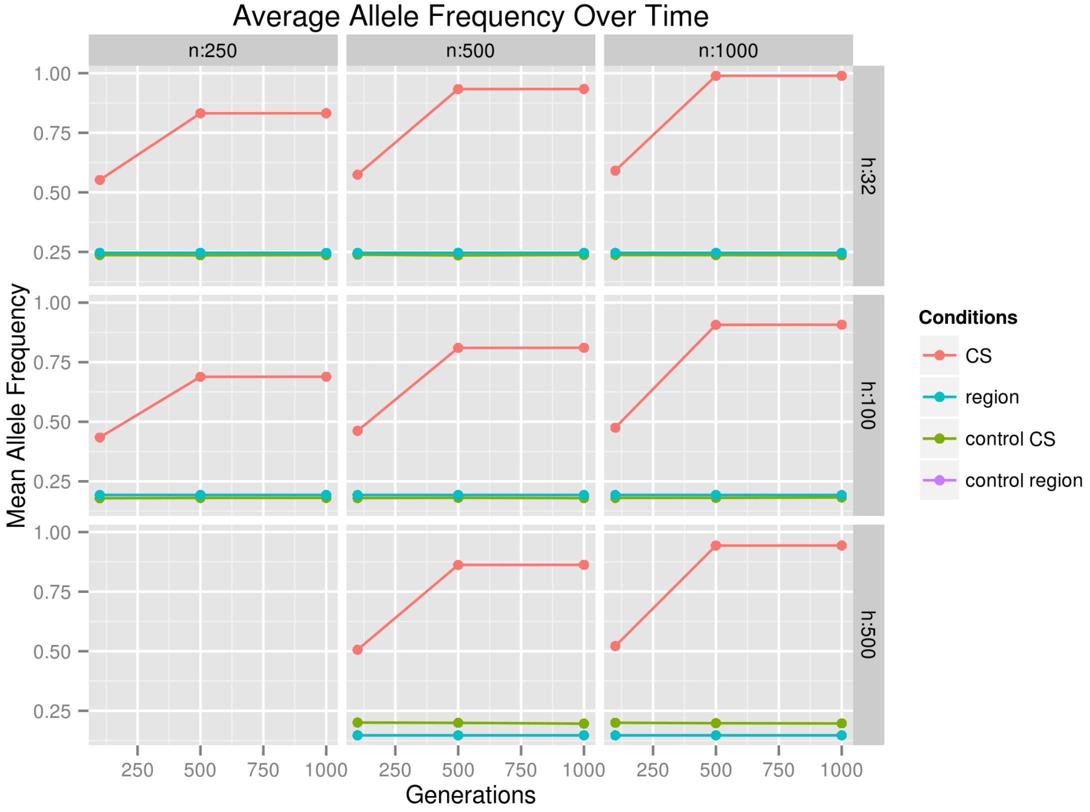


Figure A.8: Average allele frequency versus number of generations of selection. $r = 25$, $s = 0.05$. The red line indicates the average allele frequency of CS alleles, while the blue line indicates the average allele frequency for every SNP across regions. The green and purple lines correspond to these same values, but in the non-selected control populations. Note that this plot makes use of all available replicates for every .

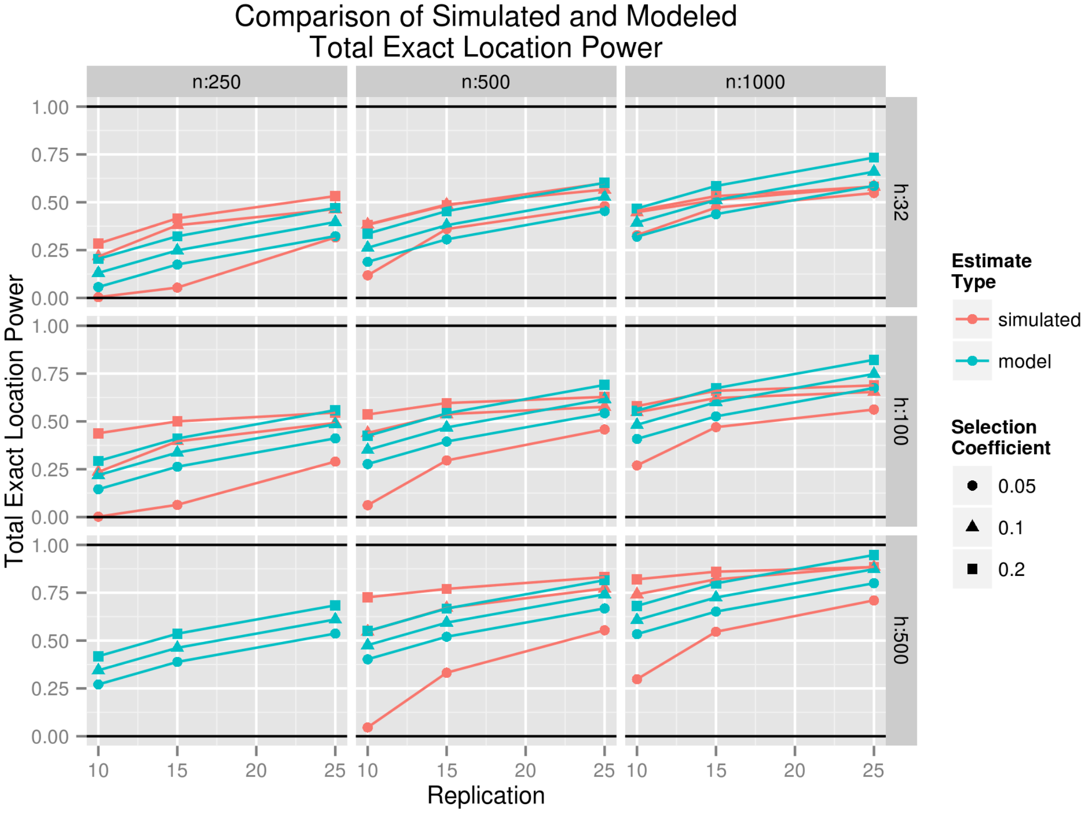


Figure A.9: A comparison of the total exact location power estimates generated by our simulations and the power estimate generated by our linear model. Only the space in which the model is relatively accurate is shown. Simulated estimate type refers to the total exact location power generated by our simulation. Model estimate type refers to the total exact location power from our mathematical model. Variation in g is not shown for clarity.

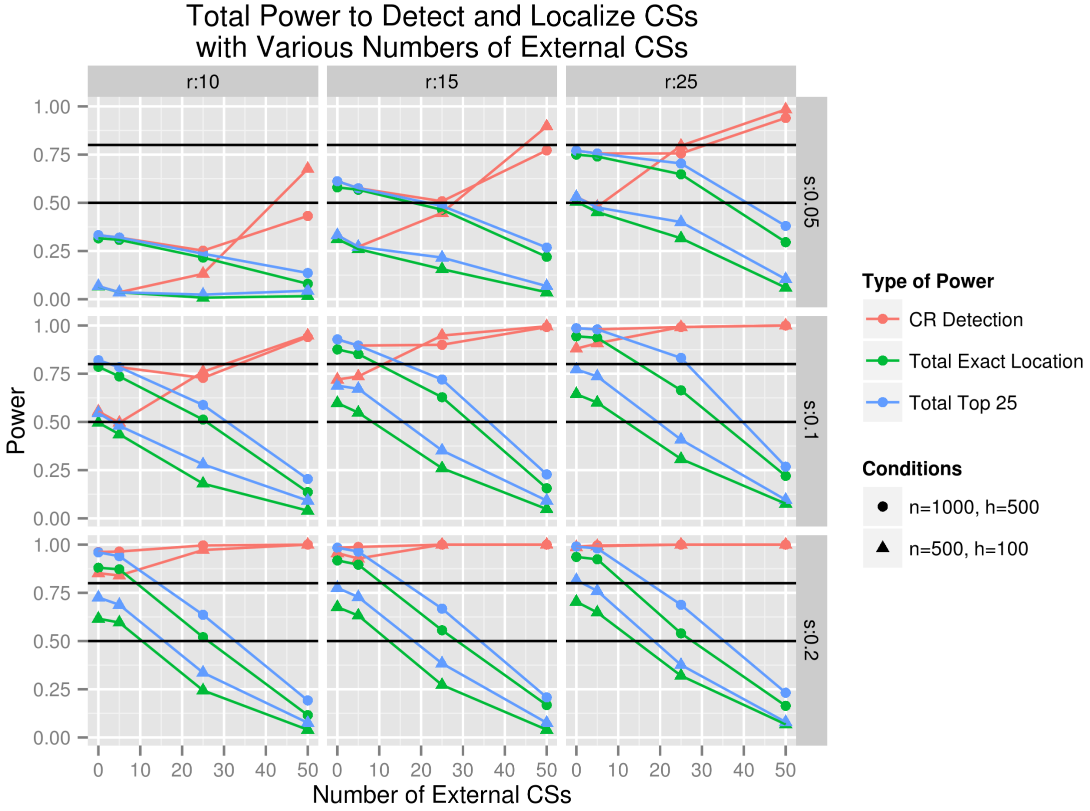


Figure A.10: A comparison of CR detection power and total power when the number of external CSs is varied. External CSs here refer to CSs that are outside of the 1Mb region surrounding the focal CS. These external CSs are randomly distributed throughout a 20Mb region and have the same selection coefficient as the focal CS.

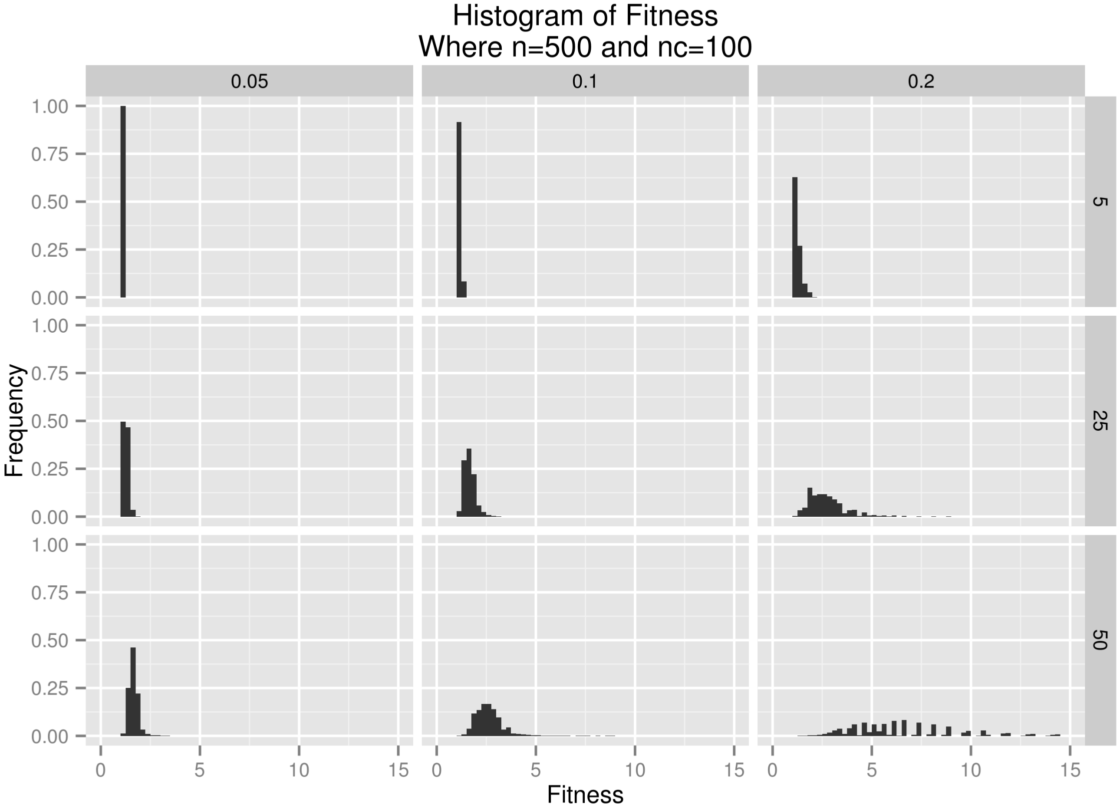


Figure A.11: This histogram illustrates the distribution of fitnesses across all simulated populations at a given . All fitnesses shown here are relative to the base (mutation free) fitness, which was assigned a relative fitness of 1.

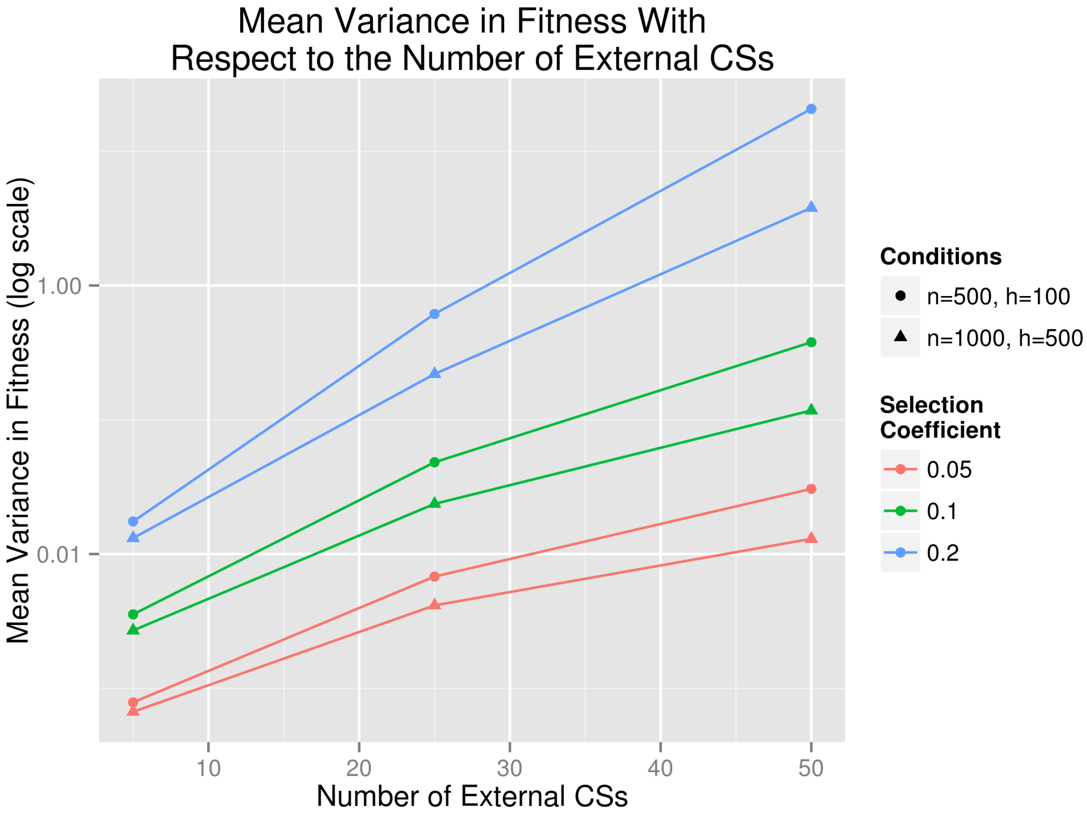


Figure A.12: This plot demonstrates the change in variance in fitness caused by the increase in the number of external CSs in a simulated chromosome. Each point represents the mean of the variances from 250 independent simulations. Note that, as these values were computed at the beginning of the forward simulation, number of generations is not taken into account.

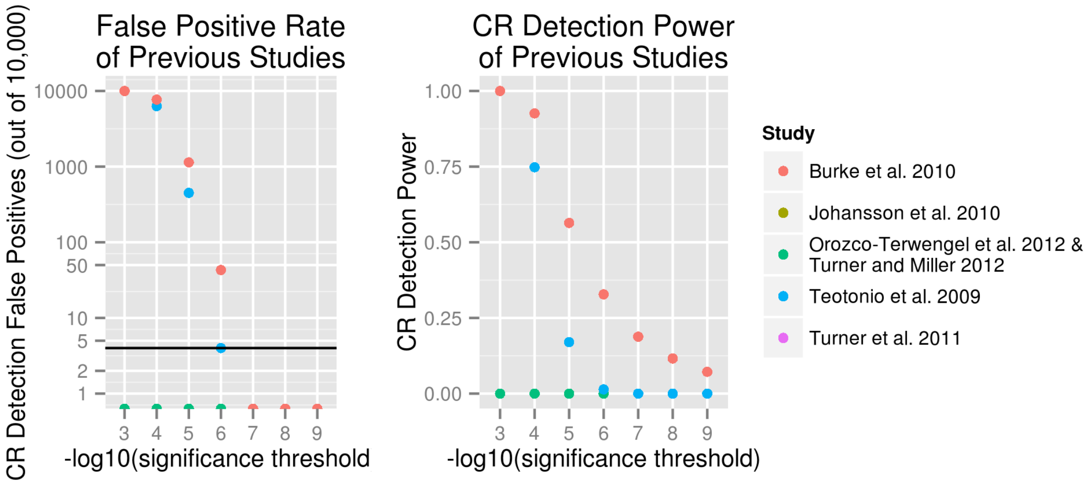


Figure A.13: CR detection power and false positive rate versus significance threshold. The that most closely correspond to the experimental parameters used in existing E&R experiments are depicted. Burke et al. 2010: $h = 500$, $n = 1000$, $r = 5$, $g = 500$. Johansson et al. 2010: $h = 32$, $n = 100$, $r = 2$, $g = 100$. Orozco-Terwengel et al. 2012 and Turner and Miller 2012: $h = 100$, $n = 1000$, $r = 2$, $g = 100$. Turner et al. 2011: $h = 100$, $n = 250$, $r = 2$, $g = 500$. When a parameter value was unknown, the value that provided the highest power was chosen. Only $s = 0.05$ is shown. Any points that are not visible are overlapping at $y = 0$.

A.2 Chapter 2 supplementary figures

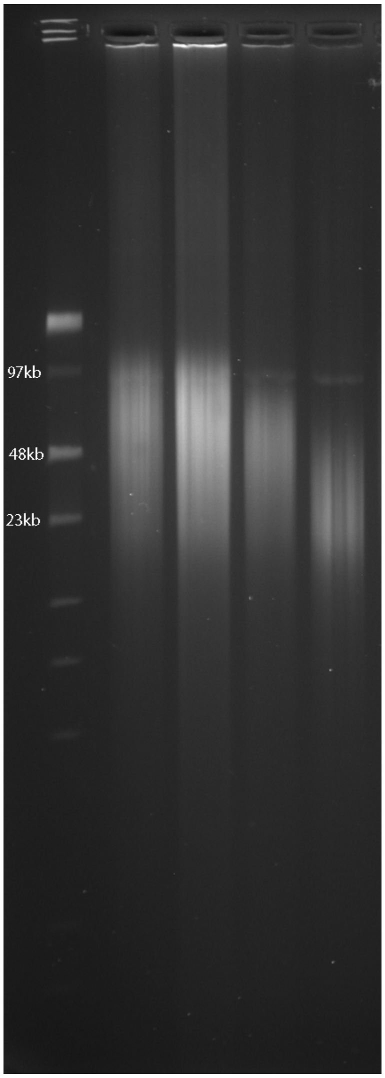


Figure A.14: A FIGE gel showing the size distribution of sheared genomic DNA fragments generated using different sized needles. From left to right: lane 1 ladder, lane 2 DNA sheared with 21 gauge needle, lane 3 DNA sheared with 22 gauge needle, lane 4 DNA sheared with 23 gauge needle, lane 5 DNA sheared with 24 gauge needle.

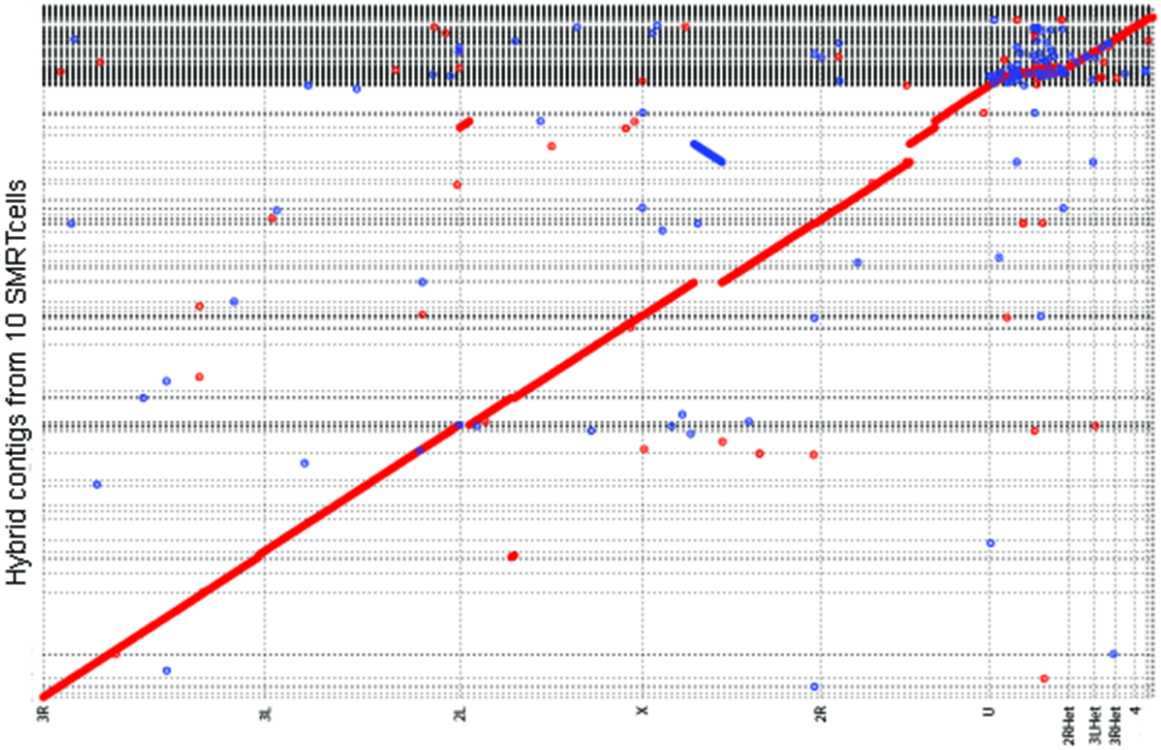


Figure A.15: An alignment plot between the 10 SMRT cell hybrid assembly and the *D. melanogaster* reference genome version 5. Red lines indicate correctly oriented contigs, while blue lines indicate inversions. A single major inversion/translocation is visible on the X chromosome. This plot was generated using *MUMmer* version 3.23 using the 'fat' and 'filter' plotting options.

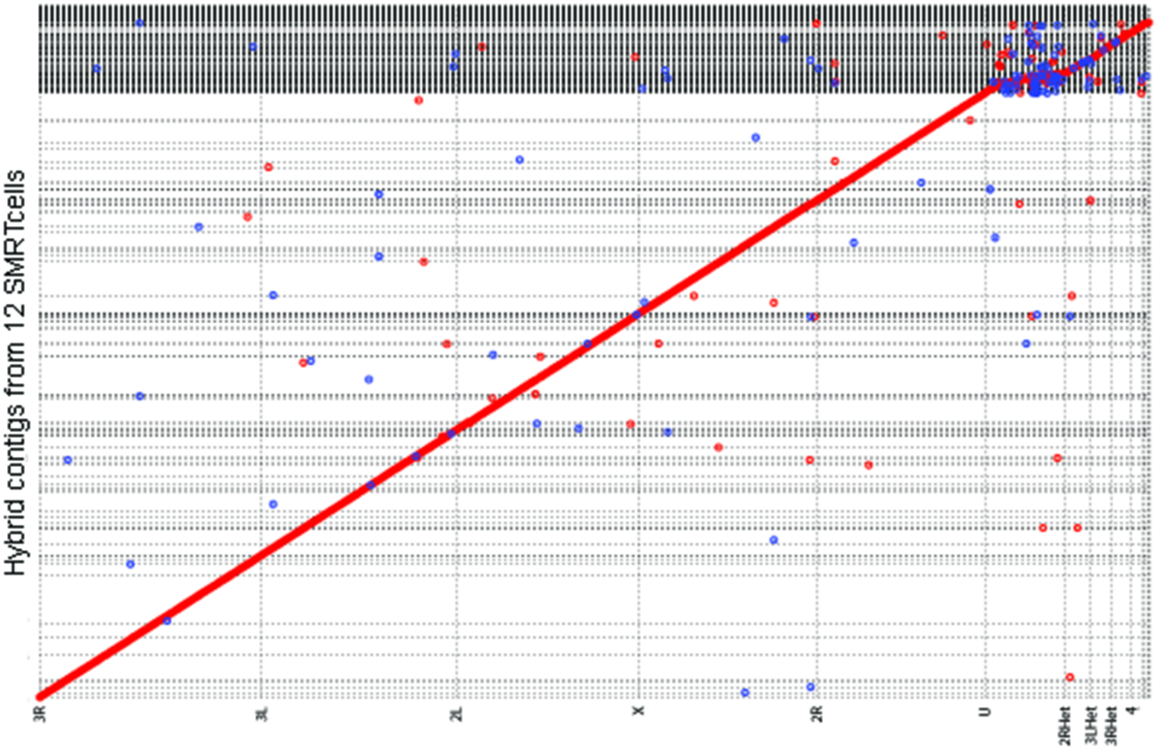


Figure A.16: An alignment plot between the 10 SMRT cell hybrid assembly and the *D. melanogaster* reference genome version 5. Red lines indicate correctly oriented contigs, while blue lines indicate inversions. A single major inversion/translocation is visible on the X chromosome. This plot was generated using *MUMmer* version 3.23 using the 'fat' and 'filter' plotting options.

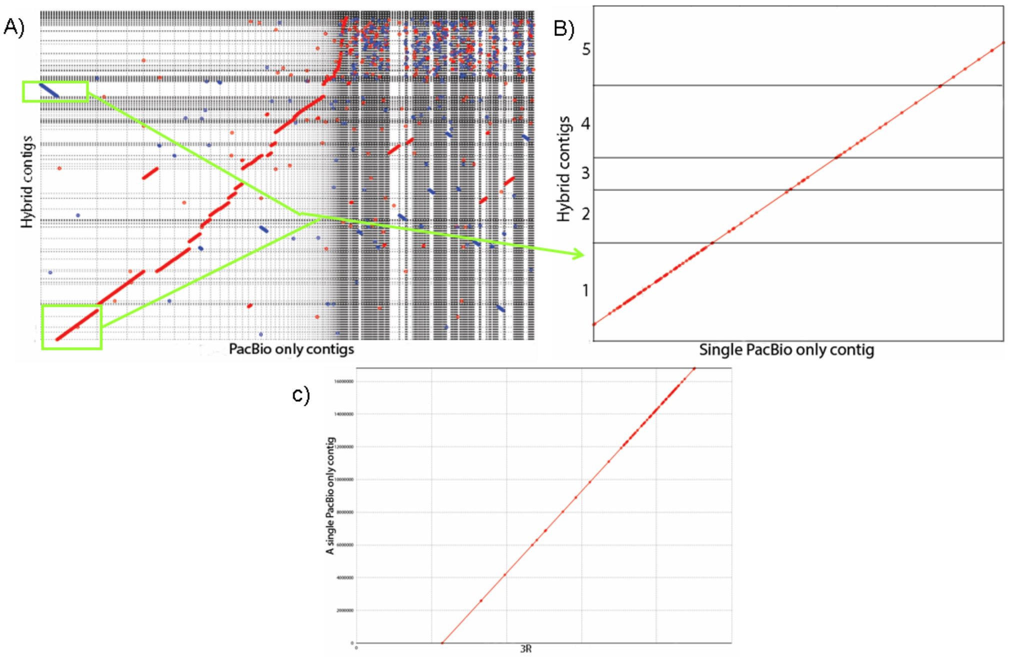


Figure A.17: a) A mummer plot depicting the alignment of the 20 SMRT cell PacBio only assembly of *D. melanogaster* to the hybrid assembly of the 50% longest reads from the same data. This plot demonstrates that large contigs produced by the PacBio only assembly are not necessarily contiguous in the complementary hybrid assembly, and vice versa. This indicates that a meta-assembly of the hybrid and PacBio only assemblies should produce a higher NG50 than either individual assembly. (b) is a zoomed in portion of (a) illustrating a PacBio only contig that contains five hybrid contigs. (c) The PacBio only contig in (b) aligns to the reference chromosome 3R.

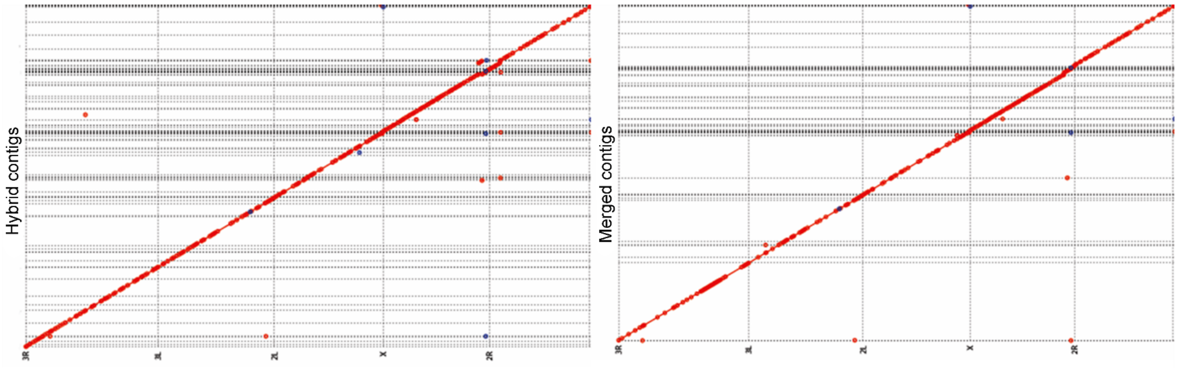


Figure A.18: A side by side comparison of the alignment of a hybrid assembly to the reference versus a merged assembly to the reference. These assemblies are both produced using the same 20 SMRTcells of data, and conform to each other closely. It is evident that the merged assembly has increased contiguity without increased misassembly.

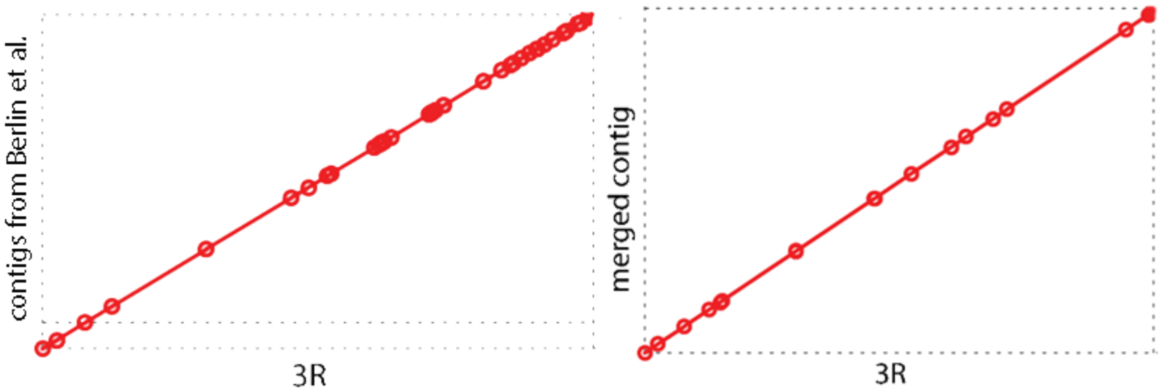


Figure A.19: A comparison of a portion of the assembly from Berlin et al. [20] (121× PacBio reads) to our merged assembly (52× PacBio reads), demonstrating increased contiguity when merging hybrid and PacBio only assemblies as compared to PacBio only alone.

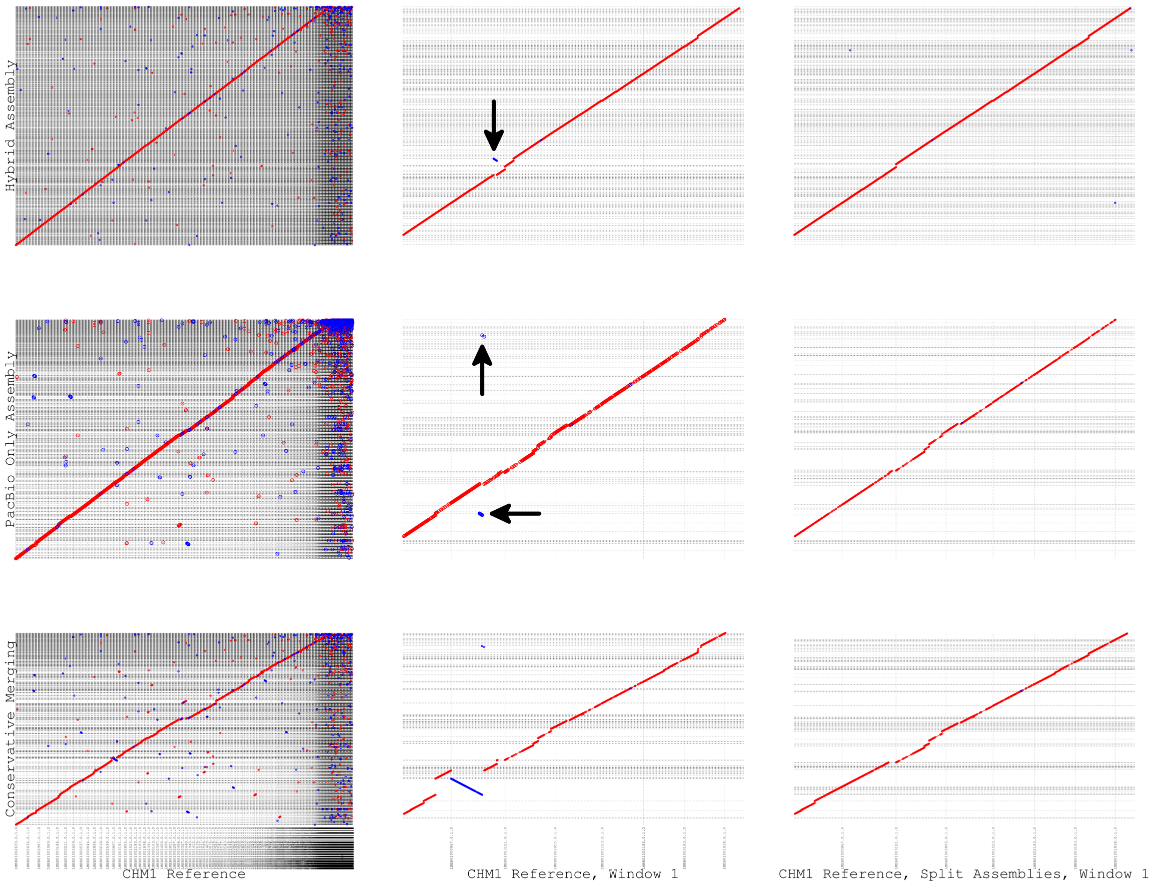


Figure A.20: Mummer dotplots between the human hybrid assembly (top row), PacBio only assembly (second row from the top), and merged assembly (bottom row) versus an extremely contiguous human genome assembly from NCBI (GenBank assembly accession GCA_001420765.1). The reference contigs are on the X-axis and the contigs from the assemblies reported here are on the Y-axis. From left to right: the first column represents the dotplots of the entire assemblies. The second column shows a magnified view of the dotplot where the most conspicuous misassembly is present in the merged assembly and the same regions in the PacBio and hybrid assemblies. The third column shows the same view as the second column after deliberately splitting contigs (indicated by the arrows) in the component assemblies that contained inversions. The contigs were split at the inversion breakpoints (at positions 2618934 and 2099497 in the PacBio only contigs utg7180000013520 and utg7180000000047, respectively, and position 3292143 in the hybrid contig Backbone_94), and the component assemblies were merged, producing a final merged assembly that did not contain the inversions present in column two. This demonstrates that at least this set of misassemblies in the merged assembly was due to misassemblies in the hybrid and PacBio assemblies and was not introduced by the merging process. The merged assembly was produced using quickmerge parameters $hco = 15$, $c = 5$, $l = 5000000$.

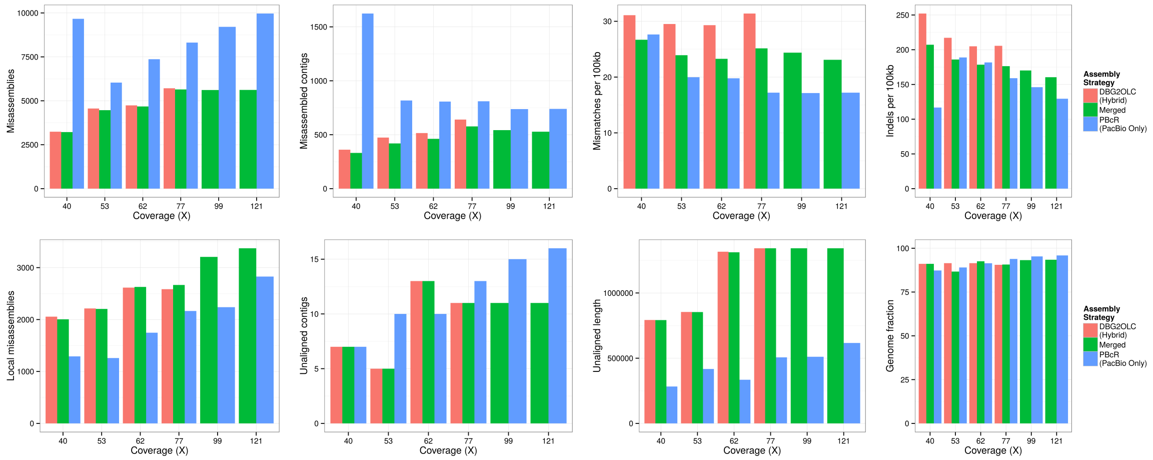


Figure A.21: Summary of unpolished assembly quality metrics from *Quast*. All hybrid and PacBio only assemblies are based on non-downsampled reads. The merged assemblies based on up to 77× data are generated by merging the hybrid and the PacBio only assemblies made using the same amount of PacBio reads. For PacBio only assemblies made with PacBio reads $>$ 77× coverage, the hybrid assembly based on 77× PacBio reads were used for assembly merging.

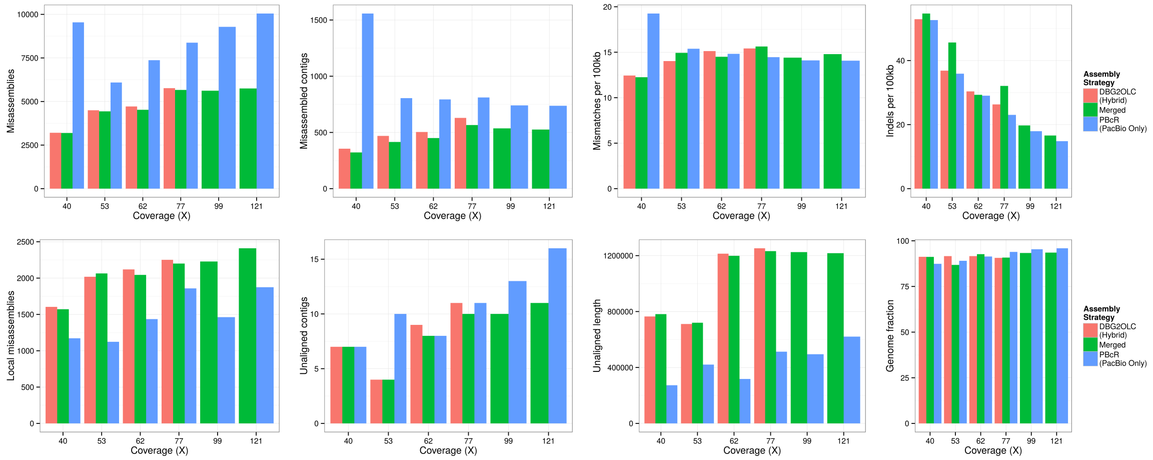


Figure A.22: Summary of *Quiver*-polished assembly quality metrics from quast. All assemblies are same as in Supplementary Fig. A.21. As evidenced here, polishing by *Quiver* improved all assemblies.

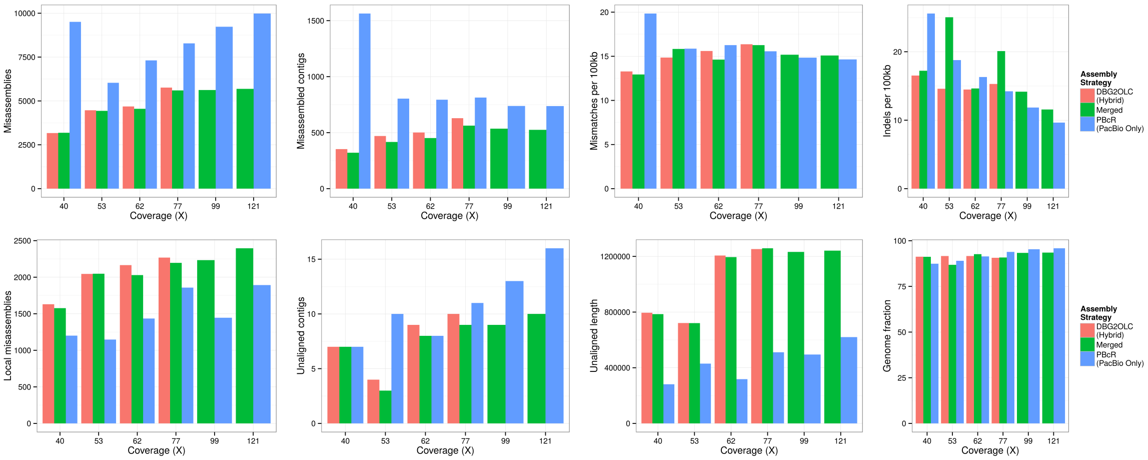


Figure A.23: Following Supplementary Fig. A.22, quality metrics after both *Quiver* and *pilon* polishing.

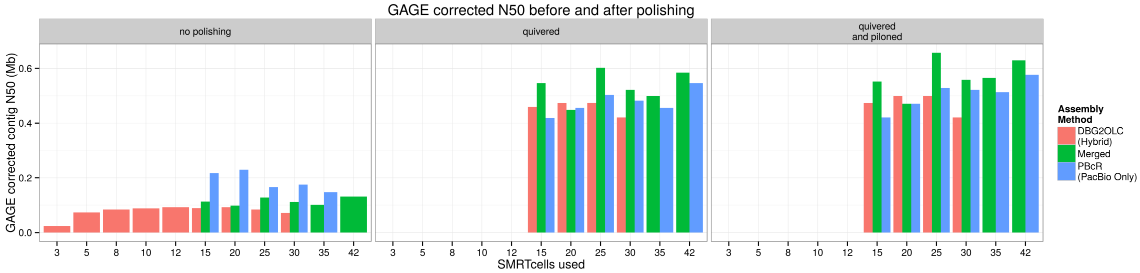


Figure A.24: Summary of *GAGE* adjusted N50 before polishing, and with polishing either by *Quiver* only or by both *Quiver* and *Pilon*.

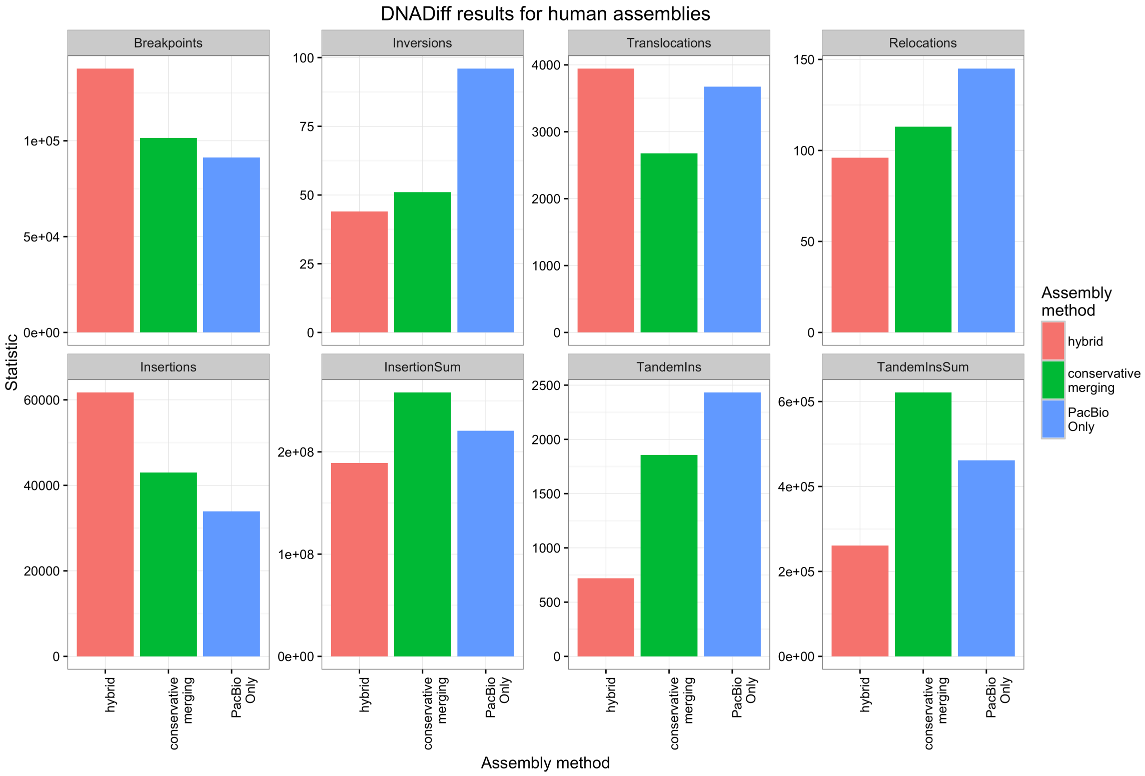


Figure A.25: *DNAdiff* results for hybrid, PacBio, and merged human assemblies. The conservative merged assembly is from merging the hybrid and the PacBio assembly using the parameters: $hco = 15$,$c = 5.0$,$l = 5000000$.

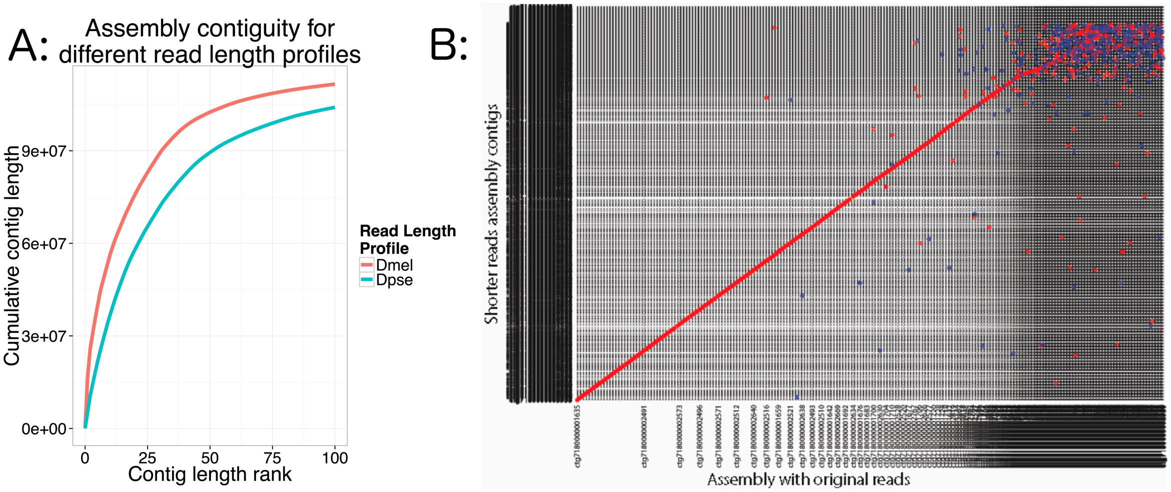


Figure A.26: A) Cumulative contig length distribution of a PacBio only assembly produced using the ISO1 long reads2 and a PacBio only assembly using the same reads downsampled to resemble the length of the shorter *D. pseudoobscura* data. B) A mummer alignment dot plot illustrating the difference in contiguity between PacBio only assemblies produced using the same assemblies as in A.

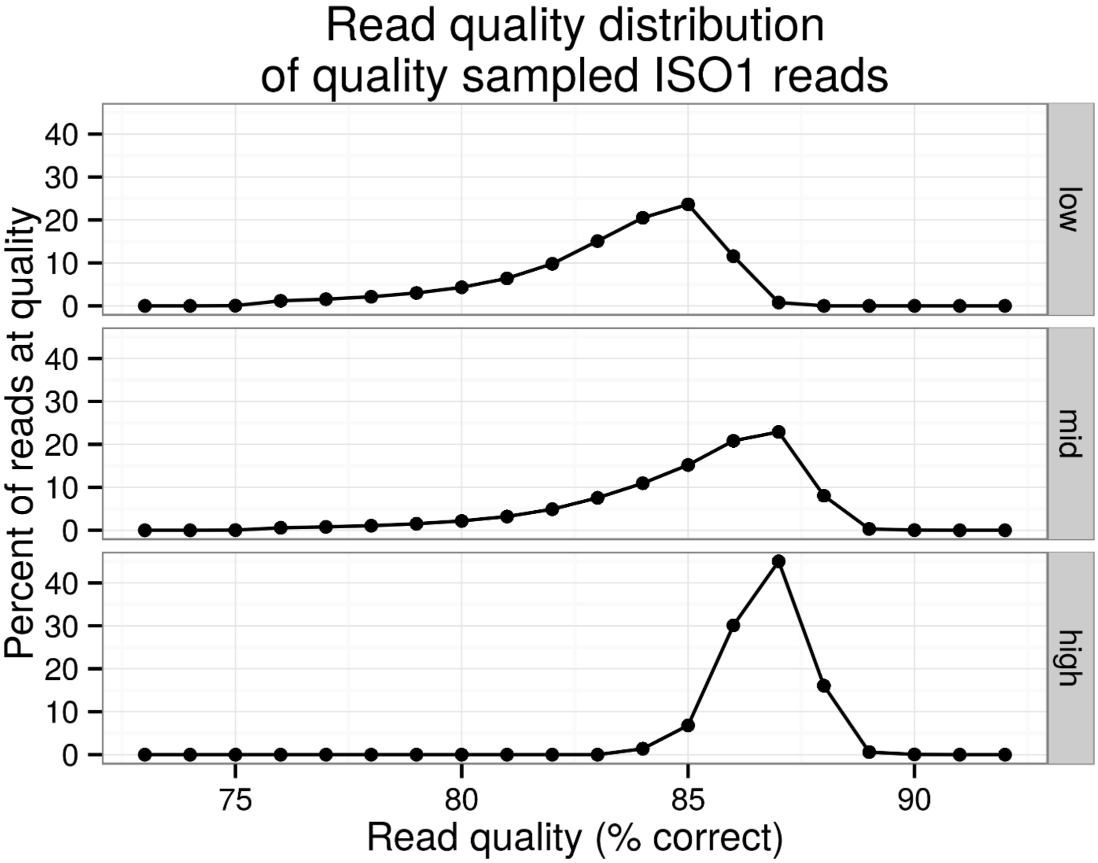


Figure A.27: Read quality distributions of the ISO1 dataset after downsampling to produce low (50% lowest quality reads), medium (a random 50% of reads), and high (50% highest quality reads) quality distributions.

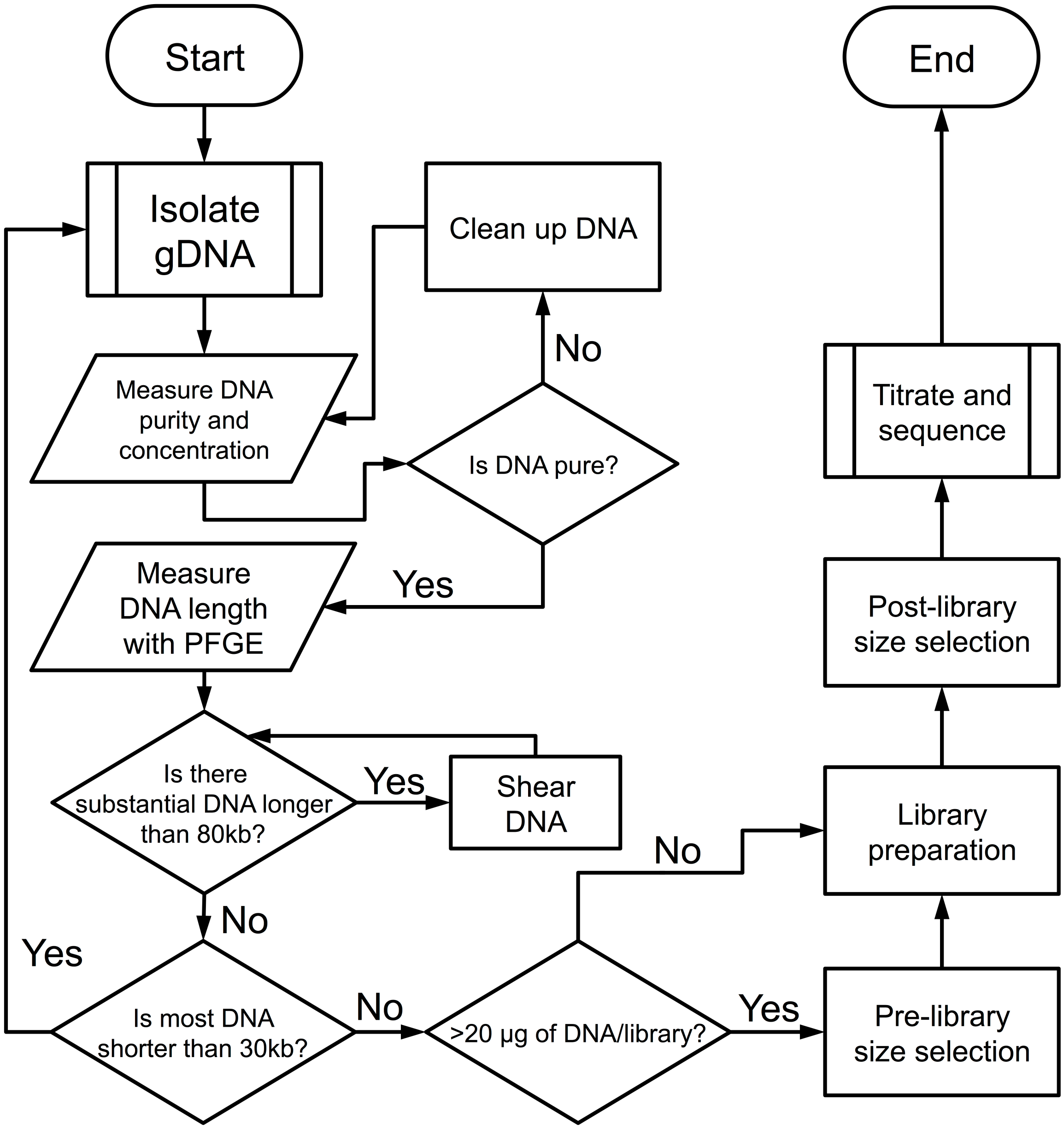

Start
Isolate gDNA
Clean up DNA
Measure DNA purity and concentration
No
Is DNA pure?
Yes
Measure DNA length with PFGE
Is there substantial DNA longer than 80kb?
Yes
Shear DNA
No
Yes
Is most DNA shorter than 30kb?
No
>20 µg of DNA/library?
No
Yes
Pre-library size selection
Library preparation
Post-library size selection
Titrate and sequence
End

Figure A.28: Flow chart showing the key steps involved in generating long reads for optimal genome assembly.

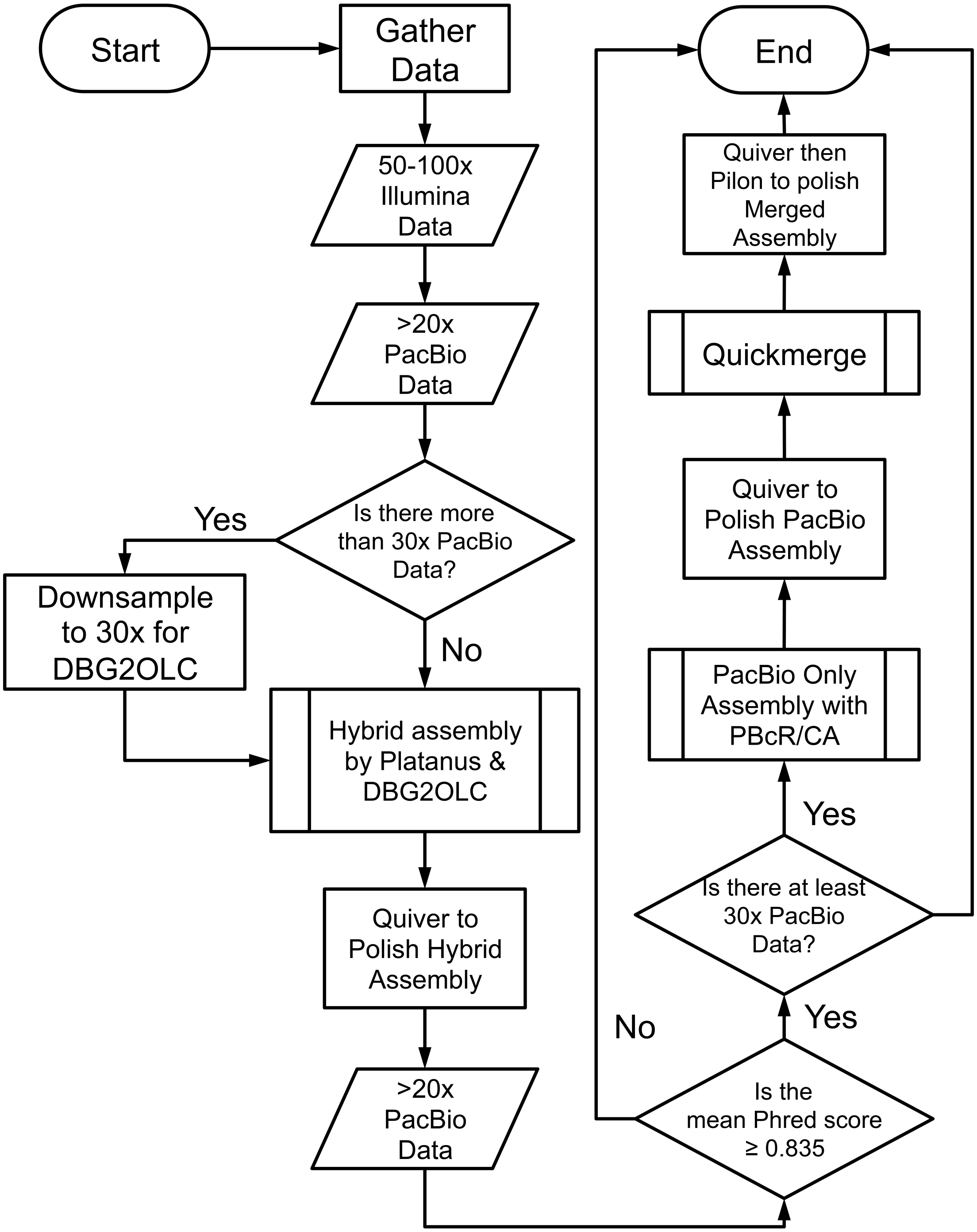

Start
Gather Data
50-100x Illumina Data
>20x PacBio Data
Is there more than 30x PacBio Data?
Yes
Downsample to 30x for DBG2OLC
No
Hybrid assembly by Platanus & DBG2OLC
Quiver to Polish Hybrid Assembly
>20x PacBio Data
Is the mean Phred score ≥ 0.835
No
Yes
Is there at least 30x PacBio Data?
No
Yes
PacBio Only Assembly with PBcR/CA
Quiver to Polish PacBio Assembly
Quickmerge
Quiver then Pilon to polish Merged Assembly
End

Figure A.29: Flow chart showing the key steps involved in assembling the long reads into a highly contiguous reference grade assembly.

A.3 Chapter 3 supplement

A.3.1 Chapter 3 supplementary texts

Bayenv settings

Bayenv was run in two modes: one with population differentiation statistics and non-parametric statistics switched on, and one with pooled population analysis turned on. The command line options for these, respectively, are as follows:

```
./bayenv2 -i $f -m $MATFILE -e $ENVFILE -n 24 -p $POPNUM -k
    ↪ 100000 -t -f -X -c -r $RSEED
./bayenv2 -i $f -m $POOLMATFILE -e $tempenv -s $samplefile -n
    ↪ 1 -p $POPNUM -k $ITNUM -t -x -r $RSEED
```

PBcR settings

PBcR, from the Celera WGS version 8.3, release candidate 1, was used. We used the following command line options:

```
PBcR -sensitive -libraryname shrimp_pacbio_round2-3-4 -s
    ↪ pacbio.spec -fastq_linebreak shrimp_pb_data.fastq
```

Where the file pacbio.spec contained the following:

```
asmOvlErrorRate=0.1
```

```
asmUtgErrorRate=0.1
asmCnsErrorRate=0.1
asmCgwErrorRate=0.1
asmOBT=1
asmObtErrorRate=0.08
asmObtErrorLimit=4.5
utgGraphErrorRate=0.05
utgMergeErrorRate=0.05
ovlHashBits=24
ovlHashLoad=0.8
utgMergeErrorLimit=5.25
useGrid=1
scriptOnGrid=1
ovlCorrOnGrid=1
frgCorrOnGrid=1
ovlMemory=128
ovlStoreMemory=128000
threads=32
ovlConcurrency=1
cnsConcurrency=32
merylThreads=32
merylMemory=128000
frgCorrThreads = 16
frgCorrBatchSize = 100000
ovlCorrBatchSize = 100000}
```

DBG2OLC settings

We ran DBG2OLC with the following command line options:

```
./DBG2OLC_Linux k 17 KmerCovTh 2 MinOverlap 20 AdaptiveTh
    ↪ 0.002 LD1 0 MinLen 200 Contigs linebreak
    ↪ platanus_contigs.fa RemoveChimera 1 f pacbio_data.fa
```

quickmerge settings

We ran quickmerge with the following command line options:

```
python merge_wrapper.py -pre merged_quivered_shrimp_assemblies
    ↪  -hco 5.0 -c 1.5 -l 1000000 linebreak hybrid_assembly.
    ↪ fasta .self_assembly.fasta
```

BWA settings

BWA and *samtools* command line settings for aligning and filtering reads were as follows:

```
bwa aln -t ${CORES}  $REFPATH $FDATAPATH > ${OUTPATH}.F.sai
bwa aln -t ${CORES}  $REFPATH $RDATAPATH > ${OUTPATH}.R.sai
bwa sampe ${REFPATH} ${OUTPATH}.F.sai ${OUTPATH}.R.sai
    ↪ $FDATAPATH $RDATAPATH | samtools view -q 20 -bS - |
    ↪ samtools sort - data/bam/$PREFIX
```

picard-tools settings

picard-tools command line settings for deduplication were as follows:

```
java -jar picard.jar MarkDuplicates INPUT=${prefix}.bam OUTPUT
    ↪ =${prefix}.dedup.bam METRICS_FILE=${prefix}.dedup.
    ↪ metrics.txt REMOVE_DUPLICATES=true}
```

GATK settings

GATK command line settings for calling SNPs were as follows:

```
java -d64 -Xmx128g -jar GenomeAnalysisTK.jar -T
    ↪ UnifiedGenotyper -nt ${CORES} -R ${REFPATH} -I merged-
    ↪ realigned-deduped.bam -gt_mode DISCOVERY -
    ↪ stand_call_conf 30 -stand_emit_conf 10 -o rawSNPS-Q30_v2
    ↪ .vcf
java -d64 -Xmx128g -jar GenomeAnalysisTK.jar -T
    ↪ VariantAnnotator -nt ${CORES} -R ${REFPATH} -I merged-
    ↪ realigned-deduped.bam -G StandardAnnotation -V:variant,
    ↪ VCF rawSNPS-Q30_v2.vcf -XA SnpEff -o rawSNPS-Q30-
    ↪ annotated_v2.vcf
java -d64 -Xmx128g -jar GenomeAnalysisTK.jar -T
    ↪ UnifiedGenotyper -nt ${CORES} -R ${REFPATH} -I merged-
    ↪ realigned-deduped.bam -gt_mode DISCOVERY -glm INDEL -
    ↪ stand_call_conf 30 -stand_emit_conf 10 -o inDels-Q30_v2.
    ↪ vcf
```

```
java -d64 -Xmx20g -jar GenomeAnalysisTK.jar -T
  ↪ VariantFiltration -R ${REFPATH} -V rawSNPS-Q30-
  ↪ annotated_v2.vcf --mask inDels-Q30_v2.vcf --
  ↪ maskExtension 5 --maskName InDel --clusterWindowSize 10
  ↪ --filterExpression "MQ0 >= 4 && ((MQ0 / (1.0 * DP)) >
  ↪ 0.1)" --filterName "BadValidation" --filterExpression "
  ↪ QUAL < 30.0" --filterName "LowQual" --filterExpression "
  ↪ QD < 5.0" --filterName "LowVQCBD" --filterExpression "FS
  ↪  > 60" --filterName "FisherStrand" -o Q30-SNPs_v2.vcf
cat Q30-SNPs_v2.vcf | grep 'PASS|textasciicircum#' > only-PASS
  ↪ -Q30-SNPs_v2.vcf
java -d64 -Xmx20g -jar GenomeAnalysisTK.jar -T
  ↪ VariantFiltration -R ${REFPATH} -V inDels-Q30_v2.vcf --
  ↪ clusterWindowSize 10 --filterExpression "MQ0 >= 4 && ((
  ↪ MQ0 / (1.0 * DP)) > 0.1)" --filterName "BadValidation"
  ↪ --filterExpression "QUAL < 30.0" --filterName "LowQual"
  ↪ --filterExpression "QD < 5.0" --filterName "LowVQCBD" --
  ↪ filterExpression "FS > 60" --filterName "FisherStrand" -
  ↪ o Q30-INDEL_v2.vcf
cat Q30-INDEL_v2.vcf | grep 'PASS|textasciicircum#' > only-
  ↪ PASS-Q30-INDEL_v2.vcf
```

R **code to estimate ρ from** ***LDx*** **data**

R code for estimating ρ using r^2 estimates at various distances is as follows:

```
ld_info <- read.table(”r2_data.txt”)
ld_info$bp <- 1:nrow(ld_info)
colnames(ld_info)[1] <- ”r2”
distance<- ld_info$bp
LD.data<-ld_info$r2
n<-844
HW.st<-c(C=0.1)
HW.nonlinear<-nls(LD.data~((10+C*distance) / ((2+C*distance)
    ↪ *(11+C*distance))) * (1+((3+C*distance)*(12+12*C*
    ↪ distance+(C*distance)^2)) / (n*(2+C*distance)*(11+C*
    ↪ distance))), start=HW.st, control=nls.control(maxiter
    ↪ =100))
tt<-summary(HW.nonlinear)
new.rho<-tt$parameters[1]
```

A.3.2 Chapter 3 supplementary tables

cutoff	n50	contigs	length	largest
82	1420491	253	120342282	4450974
83	1450029	251	120059725	4448668
84	1424862	230	119347677	10883472
85	1926101	210	118647318	10769207
86	1886292	183	116223020	7955201
87	219999	864	106905299	1281479
88	34294	1080	29263120	130428
89	9895	20	156585	21076
90	14014	3	22150	14014
91	12650	3	21171	12650
92	12644	3	21161	12644

Table A.1: Assembly statistics for hybrid genome assemblies using various quality thresholds.

population	oligo used	adapter sequence	index sequence
All	hb501	5’-AATGATACGGCGACCACCGA GATCTACACTAGATCGCTCGTCGGCAGCGTC	TAGATCGC
Cassidy	hb701	5’-CAAGCAGAAGACGGCATACG AGATTCAAGTGGTCTCGTGGGCTCGG	CACTTGA
WAL	hb702	5’-CAAGCAGAAGACGGCATACG AGATATTCCGGGTCTCGTGGGCTCGG	CCGGAAT
Hayden	hb703	5’-CAAGCAGAAGACGGCATACG AGATCGGTCTAGTCTCGTGGGCTCGG	TAGACCG
JT4	hb704	5’-CAAGCAGAAGACGGCATACG AGATGAGATACGTCTCGTGGGCTCGG	GTATCTC
Forsling	hb705	5’-CAAGCAGAAGACGGCATACG AGATCTATAGCGTCTCGTGGGCTCGG	GCTATAG
ARES	hb706	5’-CAAGCAGAAGACGGCATACG AGATGACGGAAGTCTCGTGGGCTCGG	TTCCGTC
LTER	hb707	5’-CAAGCAGAAGACGGCATACG AGATGCACTCTGTCTCGTGGGCTCGG	AGAGTGC
AMT1	hb708	5’-CAAGCAGAAGACGGCATACG AGATAAGAACGGTCTCGTGGGCTCGG	CGTTCTT
SWP4	hb709	5’-CAAGCAGAAGACGGCATACG AGATCGTCGAAGTCTCGTGGGCTCGG	TTCGACG
JD1	hb710	5’-CAAGCAGAAGACGGCATACG AGATTGCTGGTGTCTCGTGGGCTCGG	ACCAGCA

Tank 011	hb711	5’-CAAGCAGAAGACGGCATACG GGAATGT AGATACATTCCGTCTCGTGGGCTCGG
EE	hb712	5’-CAAGCAGAAGACGGCATACG CGATGGA AGATTCCATCGGTCTCGTGGGCTCGG

Table A.2: Oligos used in Nextera library construction for the wild populations.

Number	Name	read (million)	sequence (Mb)	coverage
1	cassidy	82	8209	54
2	wal	30	3000	19
3	hayden	62	6173	41
4	jt4	45	4542	30
5	forsling	49	4877	32
6	ares	56	5568	37
7	lter	271	27108	180
8	amt1	108	10753	71
9	swp4	111	11141	74
10	jd1	49	4881	32
11	tank011	348	34836	232
12	ancestor	64	6351	42

Table A.3: Coverage of sequenced populations.

Number	Name	Description
1	Date	The date of collection

2	Males	The total number of males in hydrated samples
3	Hermaphrodites	The total number of hermaphrodites in hydrated samples
4	Percent Males	The fraction of observed individuals that were male
5	Latitude	Latitude of collection site
6	Longitude	Longitude of collection site
7	Elevation	Elevation of collection site
8	Surface area	Surface area of collection site pool
9	Depth	Depth of collection site pool
10	Volume	Volume of collection site pool
11	S/V ratio	Ratio of surface area to volume in collection site pool
12	pH	pH of tank after hydrating soil
13	Ap	Average number of alleles per polymorphic allozyme locus
14	He	Expected number of heterozygotes based on allozyme polymorphism
15	f	Inbreeding coefficient, based on allozyme data
16	Fairy shrimp	Presence or absence of unclassified fairy shrimp in hydrated soil
17	Cladocerans	Presence or absence of cladocerans in hydrated soil

18	*Leptestheria*	Presence or absence of *Leptestheria* clam shrimp in hydrated soil
19	*Triops longicaudatus*	Presence or absence of *Triops longicaudatus* tadpole shrimp in hydrated soil
20	Tadpole	Presence or absence of unclassified tadpole shrimp in hydrated soil
21	*Streptocephalus mackeni*	Presence or absence of *Streptocephalus mackeni* fairy shrimp in hydrated soil
22	*Thamnocephalus platyurus*	Presence or absence of *Thamnocephalus platyurus* fairy shrimp in hydrated soil
23	*Eocyzicus*	Presence or absence of *Eocyzicus* clam shrimp in hydrated soil
24	Tadpole shrimp count	Number of tadpole shrimp observed in hydrated soil

Table A.4: A key of abbreviations for the measured environmental variables.

GO type	GO term	Description	P-value	FDR q-value
Function	GO:0042302	structural constituent of cuticle	2.44E-06	5.32E-03
Function	GO:0008061	chitin binding	3.1E-06	3.38E-03
Function	GO:0008010	structural constituent of chitin-based larval cuticle	5.31E-06	3.87E-03
Function	GO:0005214	structural constituent of chitin-based cuticle	2.34E-05	1.28E-02
Function	GO:0004180	carboxypeptidase activity	2.52E-05	1.1E-02

Function	GO:0004099	chitin deacetylase activity	8.21E-05	2.99E-02
Function	GO:0016490	structural constituent of peritrophic membrane	1.93E-04	6.03E-02
Function	GO:0070026	nitric oxide binding	4.34E-04	1.19E-01
Function	GO:0070025	carbon monoxide binding	4.34E-04	1.05E-01
Function	GO:0019826	oxygen sensor activity	4.34E-04	9.49E-02
Function	GO:0030594	neurotransmitter receptor activity	8.06E-04	1.6E-01
Function	GO:0008094	DNA-dependent ATPase activity	9.34E-04	1.7E-01
Component	GO:0000796	condensin complex	4.7E-06	4.54E-03
Component	GO:0005576	extracellular region	1.77E-05	8.55E-03
Component	GO:0044815	DNA packaging complex	2.36E-05	7.6E-03
Component	GO:0008074	”guanylate cyclase complex	soluble”	4.34E-04
Component	GO:0044421	extracellular region part	9.8E-04	1.89E-01
Process	GO:0006022	aminoglycan metabolic process	3.79E-05	2.12E-01
Process	GO:1903046	meiotic cell cycle process	4.48E-05	1.25E-01
Process	GO:0006030	chitin metabolic process	5.61E-05	1.05E-01
Process	GO:1901071	glucosamine-containing compound metabolic process	1.08E-04	1.51E-01
Process	GO:0006040	amino sugar metabolic process	1.32E-04	1.47E-01
Process	GO:1903047	mitotic cell cycle process	1.98E-04	1.84E-01

Process	GO:0040003	chitin-based cuticle development	2.63E-04	2.1E-01
Process	GO:0007366	periodic partitioning by pair rule gene	3.96E-04	2.76E-01
Process	GO:0006721	terpenoid metabolic process	5.05E-04	3.13E-01
Process	GO:0007512	adult heart development	6.52E-04	3.64E-01
Process	GO:0007376	cephalic furrow formation	6.52E-04	3.31E-01
Process	GO:0042335	cuticle development	9.47E-04	4.4E-01

Table A.5: GO terms for differentially expressed genes between males and hermaphrodites.

GO type	GO term	Description	p-value	FDR q-value
Function	GO:0051606	detection of stimulus	6.50E-05	3.63E-01
Function	GO:0009582	detection of abiotic stimulus	1.46E-04	4.07E-01
Function	GO:0009581	detection of external stimulus	1.46E-04	2.71E-01
Function	GO:0007602	phototransduction	1.80E-04	2.52E-01
Function	GO:0009583	detection of light stimulus	4.68E-04	5.23E-01

Table A.6: GO terms for genes with differential allele frequencies according to $X^T X$.

A.3.3 Chapter 3 supplementary figures

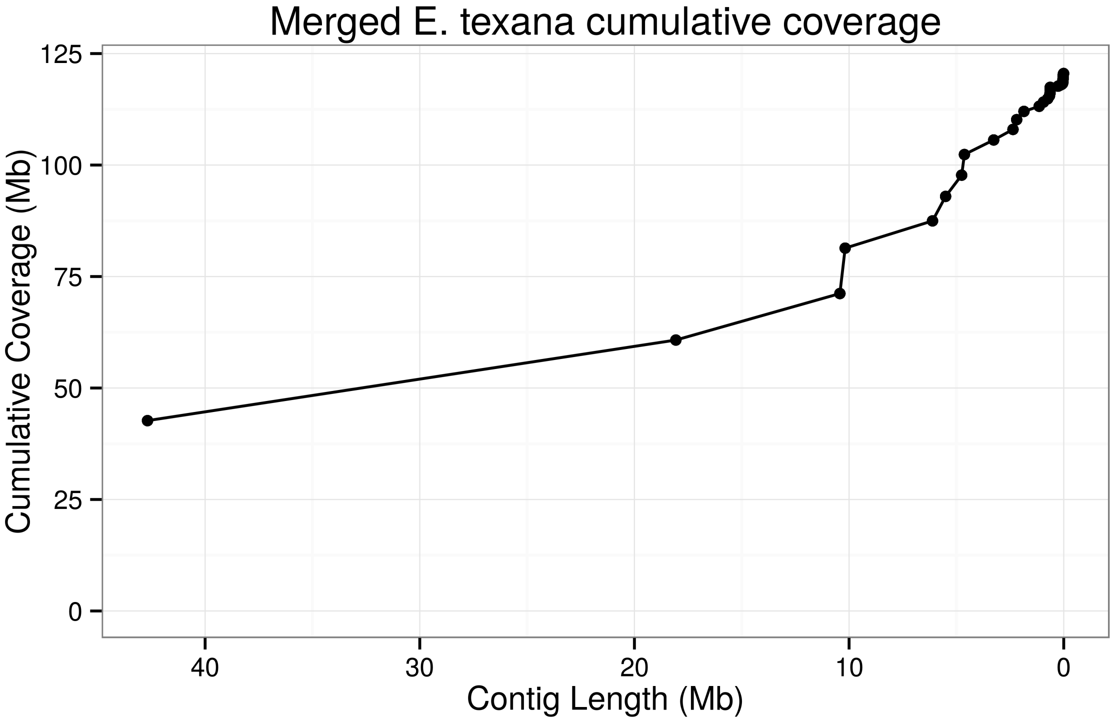


Figure A.30: a plot of cumulative genome coverage of the *E. texana* genome assembly by contig. As the plot progresses from left to right, the contig lengths are added to the cumulative coverage in order from largest to smallest. A high quality assembly should achieve a high cumulative coverage with a small number of contigs.

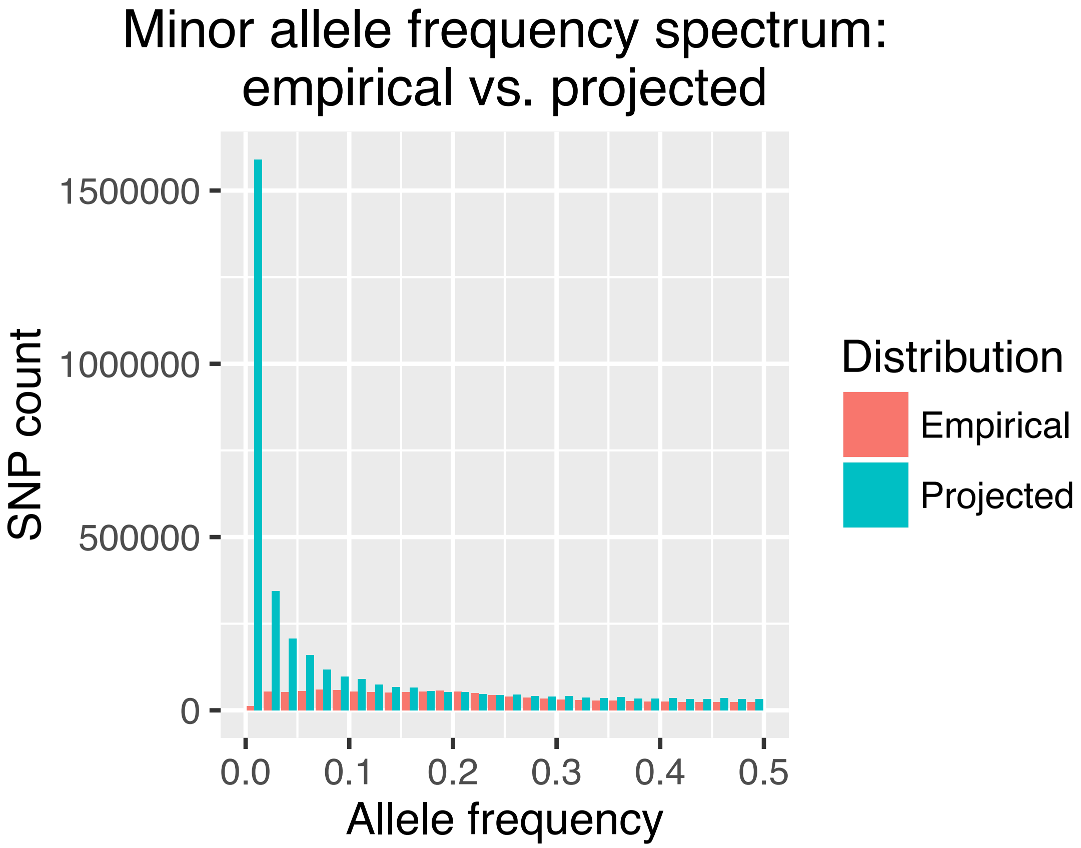


Figure A.31: An allele frequency histogram of all SNPs that passed coverage censoring, both with and without the projection of rare SNPs. Unsurprisingly, extremely low frequency SNPs are rare because of coverage limitations. The overall site frequency spectrum approximately matches the site frequency spectrum expected by chance, except when comparing low frequency SNPs.

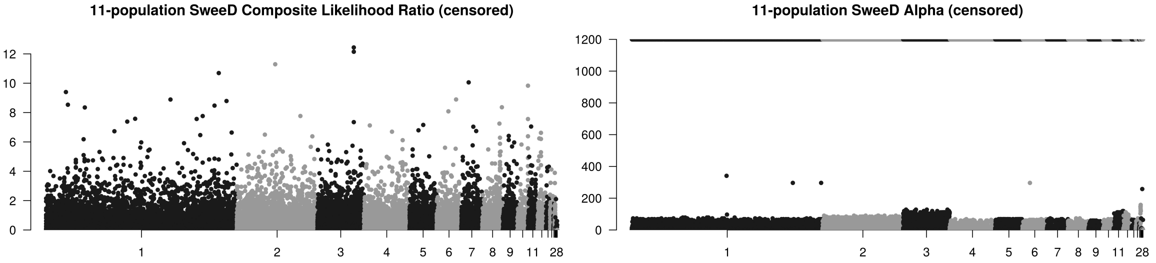


Figure A.32: Manhattan plot of *SweeD* CLR (composite likelihood ratio) values and alpha values. The CLR values (left) represent the ratio of the likelihood of non-neutrality vs. neutrality, while the alpha values (right) represent the $-\log_{10}(p)$ probability that the region is neutral, based on its site frequency spectrum. It is evident that a large portion of the genome appears to be non-neutral, but that is likely due to a lack of rare alleles in the sample because of SNP ascertainment bias.

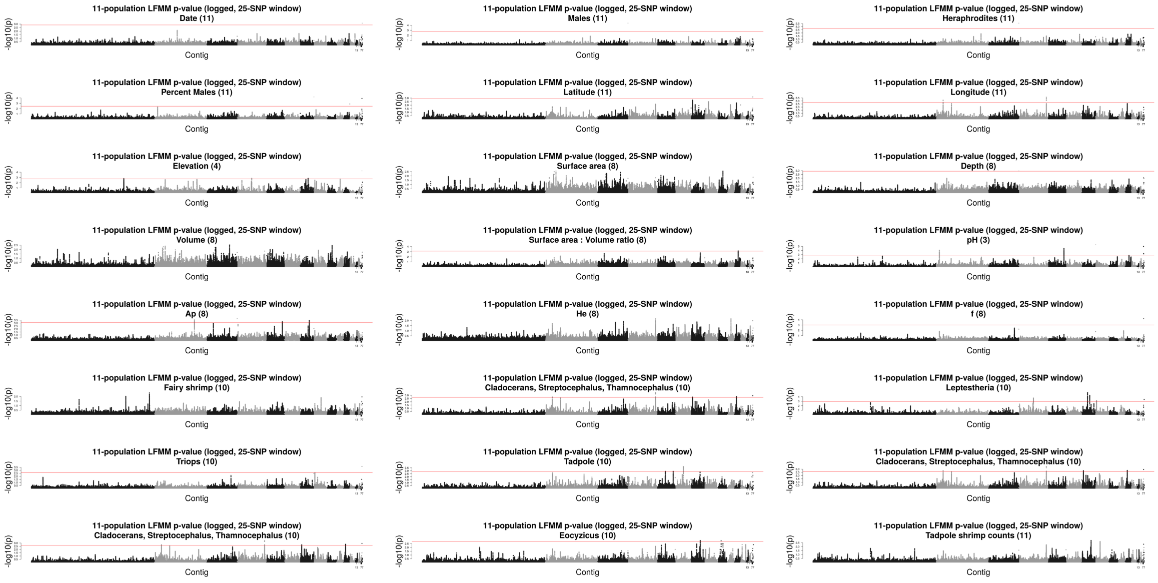


Figure A.33: Manhattan plots of $-log_{10}(p)$-values from LFMM, in 100-SNP windows, across the entire genome, for all examined environmental variable. The environmental variable is printed below each plot.

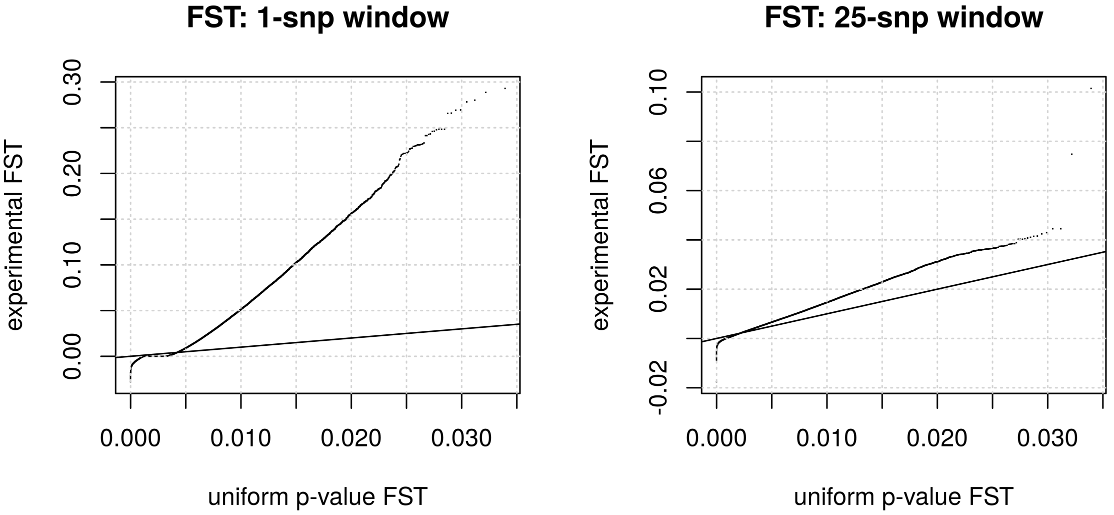


Figure A.34: A quantile-quantile plot of the pairwise F_{ST} of the WAL population and the EE population versus the FST expected via an exponential distribution with $\lambda = 1/\bar{F_{ST}}$. Left: single-SNP values; Right: 25-SNP averages.

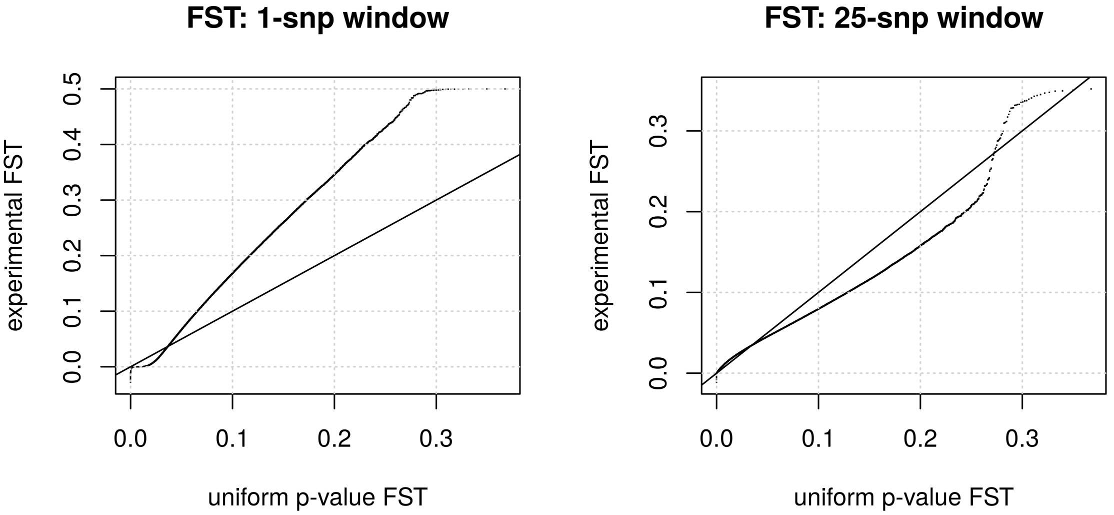


Figure A.35: A quantile-quantile plot of the pairwise F_{ST} of the LTER population and the Tank011 population versus the FST expected via an exponential distribution with $\lambda = 1/\bar{F_{ST}}$. Left: single-SNP values; Right: 25-SNP averages.

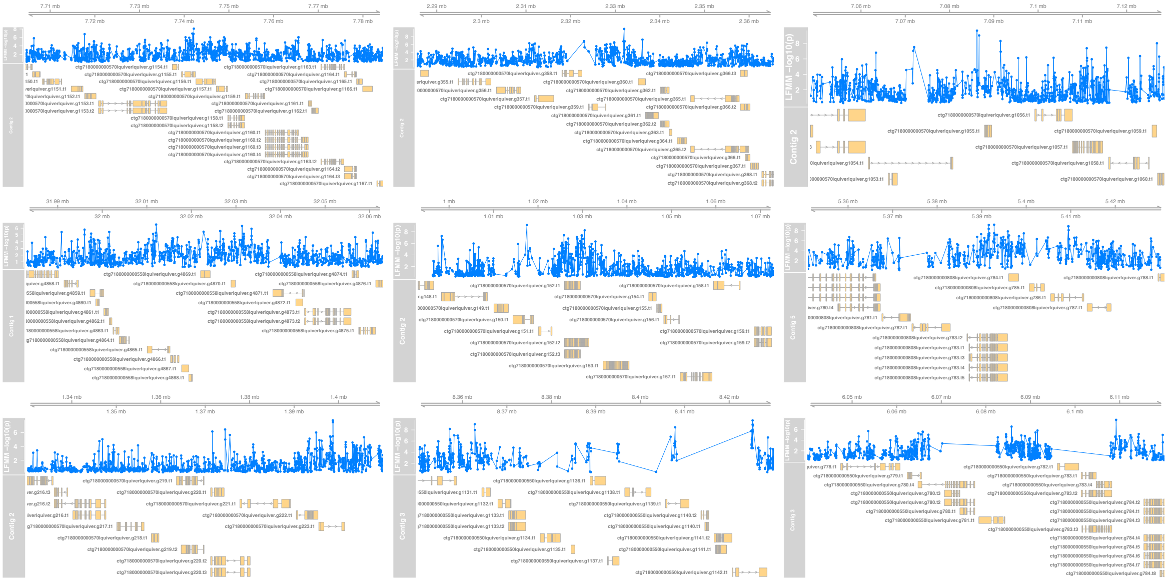


Figure A.36: A set of manhattan plots showing *LFMM* z-values for a select set of loci with high z-values. Note that regions with few data points tend to be areas of low coverage.

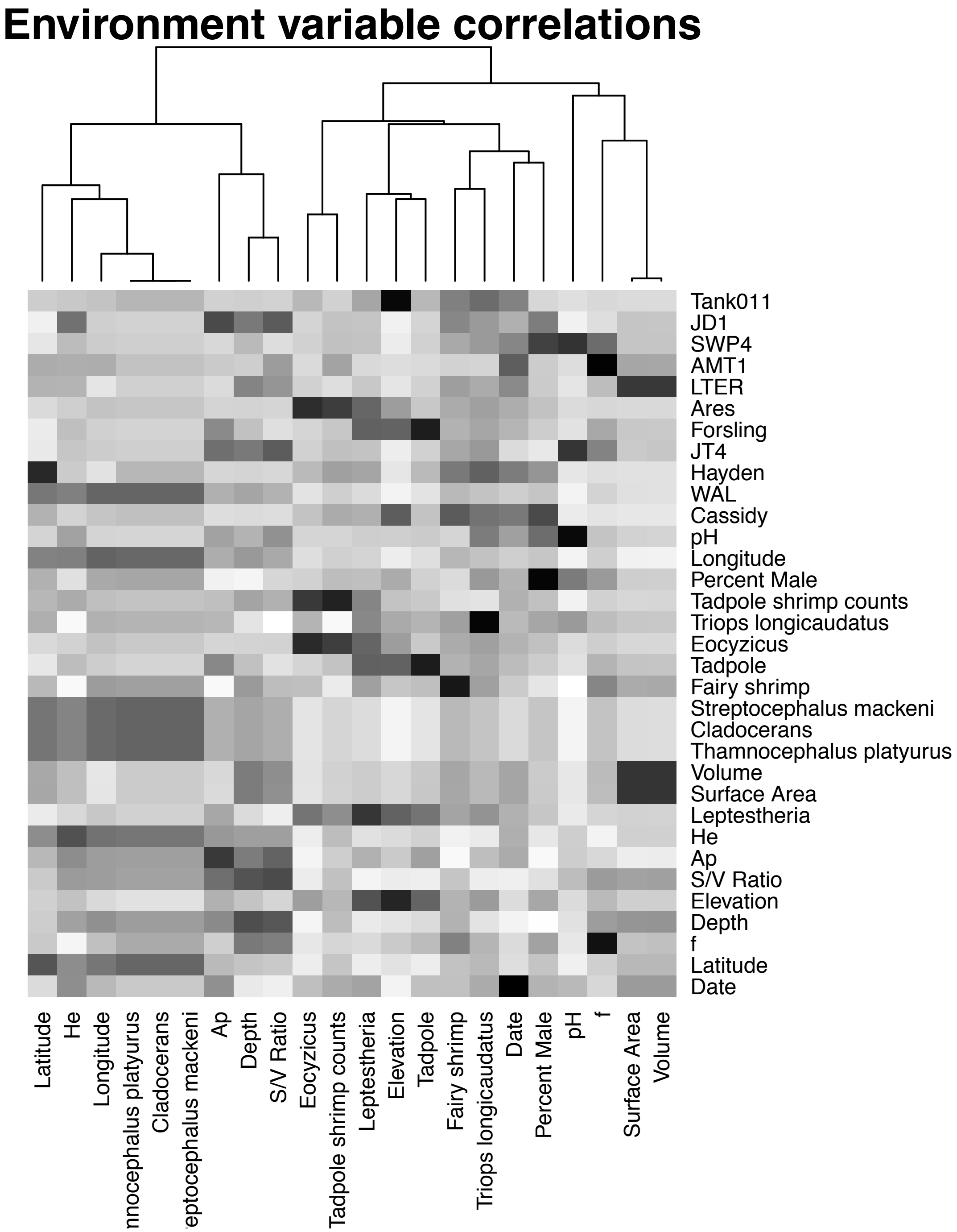


Figure A.37: A heatmap depicting the correlation coefficients between the measured environmental variables, plus, dummy variables indicating the various populations. Black indicates a high level of correlation, while white indicates a lack of correlation.

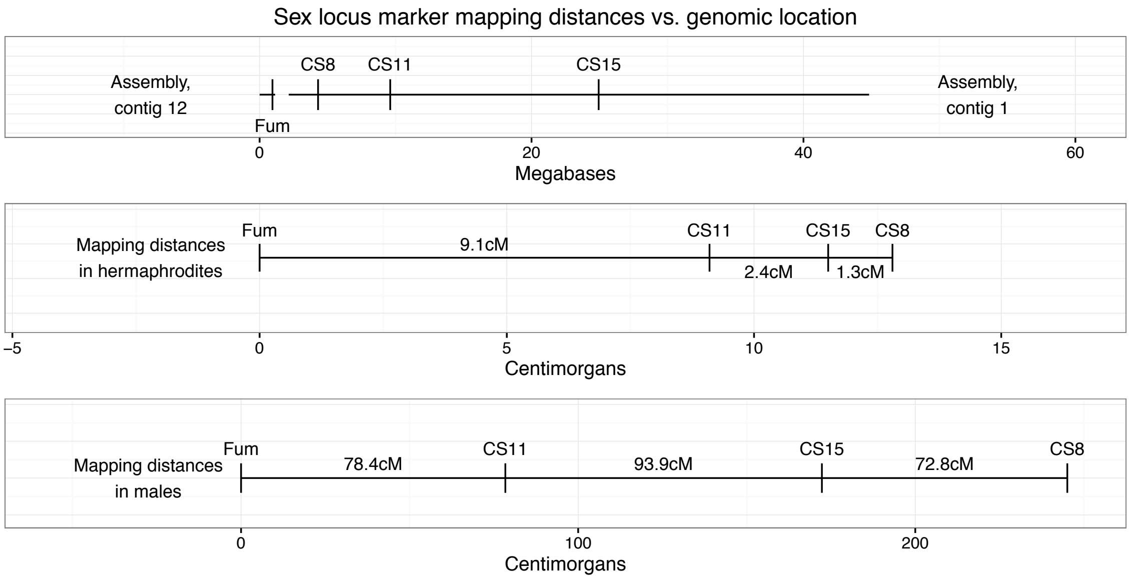


Figure A.38: A set of line diagrams indicating the mapping of a set of markers onto the putative sex chromosome using a variety of methods. The methods are as follow, from top to bottom: 1. *BLAST* alignment results of the markers onto the genome assembly contigs that we hypothesize make up the sex chromosome. 2. A map of the sex chromosome using linkage mapping information from a hermaphrodite cross, as performed in Weeks 2010. 3. A map of the sex chromosome using linkage mapping information from a male cross, as performed in Weeks 2010. The lengths of the hermaphrodite linkage map here is to the assembly contig lengths according to the genome-wide recombination rate calculated in this paper. Note that the total mapping distance in hermaphrodites resembles the total physical distance in the assembly, but the total mapping distance in males is much longer.