Application of RAD-seq in Evolutionary Genomics

# of Non-Model Organisms

By

Boryana S. Koseva

Submitted to the graduate degree program in Ecology and Evolutionary Biology and the Graduate Faculty of the University of Kansas in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

____

Chair: John K. Kelly

____

Stuart J. Macdonald

____

Justin P. Blumenstiel

____

Jamie R. Walters

____

Mark T. Holder

Date Defended: 28 April 2017

The dissertation committee for Boryana Koseva certifies that this is the approved version of the following dissertation:

Application of RAD-seq in Evolutionary Genomics of Non-Model Organisms

____

Chair: John K. Kelly

Date Approved: 11 May 2017

Abstract

Next generation sequencing (NGS) technologies are revolutionizing how we study genetics and evolution in the modern world. Data is generated at such a fast pace that scientists are struggling to keep up with the innovations in methodology and analytical tools. Genomes are being sequenced at an unprecedented rate, and scientists in fields that until recently found no use in learning molecular techniques are venturing into the world of high-throughput sequencing.

Almost 10 years ago, a research group developed Restriction-site Associated DNA Sequencing (RAD-seq), a method that targets polymorphisms in close proximity to restriction cut sites in hundreds of samples simultaneously. The beauty of RAD-seq lies in that it is highly customizable and it does not require a reference genome, or intimate prior knowledge of the genetics of the study organism one would like to use. The most exciting part about new RAD-seq methods being developed is that their accessibility has opened the door to many non-model organisms to be used in new areas of research. The overarching theme of my dissertation is the application of RAD-seq data to answer questions in evolutionary, quantitative, or population genetics and genomics using non-model species. A secondary goal is the development of genomic resources for non-model organisms.

In Chapter 1, I studied the genetics of a recent shift from self-incompatibility to self-compatibility in an insular lineage of *Tolpis*, with an aim to identify putative genomic regions responsible for this shift in mating system. To do that, I assembled a draft genome, annotated it, and used RAD-seq data from a mapping population to discover variants. In my second chapter, I focus on a pyralid moth and the genetic basis of male song characters that are attractive to females. For that purpose, I again used RAD-seq data from hundreds of individuals, and, additionally, I assembled and annotated a genome for this non-model organism. In my last chapter, I focused on the bioinformatic challenges associated with RAD-seq data. I explored the question of whether or not using a genome sequence helps in the construction of loci from RAD-seq reads. The evaluation of the last question is on a fairly basic level but it opens up future avenues that I am excited to explore.

Acknowledgements

In looking back on the 6 years I spent working on this dissertation, I know that I didn't do it alone. It really took a village.

First and foremost, I would like to thank my advisors, John Kelly and Stuart Macdonald, for their guidance and support. Their patience, and ability to work with my strengths and around my weaknesses, have allowed me to grow a skillset that will be invaluable in my future professional endeavors. I am especially grateful to have had a great committee of people I admire and look up to. Jamie Walters, Mark Holder, and Justin Blumenstiel allowed me the space to pursue my own interests, and provided me with advice to help in my professional development.

To my undergraduate advisors, Dr. Stephen Hudman and Dr. Chad Montgomery, thank you for instilling in me the love for the scientific method, and teaching me the importance of sharing my knowledge with those willing to learn. I would never have made it to graduate school without you.

To the entities and individuals who provided me with funding, I would never have been able to complete this work without you. Special thanks to the Department of Ecology and Evolutionary Biology (EEB) at the University of Kansas. Aagje Ashe, Dorothy Johanning, and Jaime Keeler work tirelessly to help all EEB graduate students be successful. I hope you know how much I appreciate all you have done for me.

To my fellow Kelly lab members, Patrick, Sarah, Keely, and John, I can't thank you enough for your support and friendship. It has been a great honor and pleasure to have taken this journey with you! I was also fortunate enough to have had an amazing group of colleagues in EEB. Any and all interactions with you all have motivated me in one way or another to keep going even during the difficult times. A few of you have been my unwavering support system the last couple of years, and I couldn't have survived the

storms without you. Jenny Hackett, Alex Erwin, Kaila Colyott, Sally Chang, and honorary EEB members Sophie Loschky and Michael Pope, I appreciate you letting me sleep in your spare bedrooms, feeding me, checking up on me, for the late night conversations, and most of all, for accepting me for who I am.

Finally, I have to thank my family. My mother, Irina, wasn't really sure why I wanted to go to graduate school but has supported me without questions, even when she was thousands of miles away. When she visited, she spoiled me to no end with care and love. My sister, Nena, has always reminded me that I can do anything I set my mind to even when I am contemplating giving up. She has been not just my sister but also my best friend and confidante. My love, Alec, has showed patience, generosity, and endless love to me during the final stretch, and has helped me be a better version of myself. Mom, Nena, and Alec, for all that, and much more, I thank you and love you.

You all are the best village anyone could ever ask for!

Table of Contents

Abstract ... **iii**

Acknowledgements ... **iv**

Table of Contents ... **vi**

List of Figures ... **ix**

List of Tables ... **xi**

Chapter 1 The Breakdown of Self-Incompatibility in an Insular Lineage of *Tolpis* (Asteraceae) ... **1**

Abstract ... **2**

Introduction ... **3**

Materials and Methods ... **5**

Overview ... 5

Genomic Data ... 6

Draft Genome Assembly ... 6

Annotation ... 7

Mapping Population and Genotyping ... 7

Linkage Map Construction ... 9

QTL Mapping ... 10

Identification of Putative Genes in the QTL region ... 10

Results ... **11**

Phenotype Distributions ... 11

Genome Assembly and Annotation ... 11

Linkage Map ... 11

QTL Mapping Analysis ... 11

Candidate genes ... 12

Discussion ... **12**

Sporophytic Self Incompatibility (SSI) ... 12

The Genetics of the Transition to Selfing ... 14

Pseudo Self-Compatibility and Baker's Law ... 15

Figures ... **18**

Tables ... **22**

Chapter 2 *De novo* Genome Assembly and QTL Mapping of Male Song Characters in *Achroia grisella*... 26

Abstract..27

Introduction..28

Methods...30

Overview..30

Genomic Data..31

Genome Assembly and Evaluation..32

Transcriptomic Data...33

Genome Annotation...34

Mapping Populations and Phenotypes...34

Library Construction..35

Marker Discovery...36

Genotyping Recombinants...36

Genetic Map Construction...37

QTL Mapping...38

Mapping Marker Sequences to the Genome Assembly......................................38

Identification of the Z chromosome..39

Results...39

Phenotypes...39

Genetic Map Construction..40

QTL mapping..42

Genome Assembly and Annotation..44

Discussion..44

Developmental Traits..44

Male Song Characters...45

Choices Have Consequences...47

Conclusion..48

Figures...50

Tables...57

Chapter 3 Draft Genome Assemblies in Non-Model Organisms: Do they facilitate or hinder the construction of genomic loci from RAD-seq data?.. 69

Abstract...70

Introduction **71**

Methods **76**

Study Organisms 76

Mapping populations and Genomic Library Preparation 77

Marker Discovery 78

Evaluation 79

Results **81**

Discussion **82**

Effect of mapping RAD tags to a reference genome 82

Effects of assembly errors on population genetic estimates 84

Conclusion 87

Tables **89**

Literature Cited **91**

List of Figures

Figure 1.1. Distribution of F_2 phenotype: (a) percent self-seed set, (b) log of percent self-seed set, and (c) absolute frequency of SC and SI individuals. 18

Figure 1.2. *Tolpis coronopifolia* (n=9) linkage map containing 9 linkage groups and a total of 2,125 markers from 92 individuals. 19

Figure 1.3. A QTL analysis revealed a single QTL around 3.5cM on LG2. (a) LOD scores for the 9 LGs using standard interval mapping of percent self-seed set. The dotted line is the genome-wide threshold established by a permutation test (N=1000) at α=0.01. (b) LOD scores from 3 analyses for LG2, which contains the QTL peak. The turquoise line is from standard interval mapping of percent self-seed, the violet is from standard interval mapping of log-transformed percent self-seed set, and the pink line is based on a QTL analysis using a binary model for the SI/SC scoring of the phenotype. The dotted line is the genome-wide threshold established by a permutation test (N=1000) at α=0.01, and is approximately the same for all analyses. 20

Figure 1.4. The effect plot for the peak marker at the QTL associated with SC in *T. coronopifolia*; phenotype average (+ symbols indicate the span of the standard error). A single copy of the SI allele (*R*) appears to be enough to inhibit self-fertilization (strong dominance effect), but when the number or SC alleles (*r*) increases from 1 to 2 in an individual, the ability to set seed increases drastically. .21

Figure 2.1. Average base quality before (top) and after (bottom) read processing of the forward reads from the PE library used to generate the *de novo* genome assembly. The graph was generated by FASTQC (S. Andrews, n.d.). 50

Figure 2.2. Linkage map for the KS backcross as constructed by Lepmap. Chromosome numbers here and in Figure 2.3 are identical 51

Figure 2.3. Linkage map for the FL backcross as constructed by Lepmap. Chromosome numbers here and in Figure 2.2 are identical. 52

Figure 2.4. The correspondence in marker position between the two backcrosses. There are the positions of all markers in common between LG1 in KS BC and LG1 in FL BC. 53

Figure 2.5. ALLMAPS output showing the inconsistency in marker order between the two backcrosses. The gray-striped tube in the middle represents the chromosome. The two linkage maps on either side of the chromosome represent linkage groups 18 from the KS BC and the FL BC unresolved maps..54

Figure 2.6. Genotype matrix for linkage group 1 in the KS backcross. Red represents a homozygous genotype, blue represents a heterozygous genotype, and white represents a missing value. Here, individuals are ordered along the y-axis in terms of the proportion of homozygous genotypes they have. 55

Figure 2.7. Genotype matrix for linkage group 1 in the KS segregant population. A subset of the individuals was used here for better resolution. Red represents a homozygous genotype, blue represents a heterozygous genotype, and white represents a missing value. 56

List of Tables

Table 1.1. Number of paired and singleton reads after trimming (Sickle) and correcting (Quake) raw reads. The percent of the original data preserved after each step is in parentheses. 22

Table 1.2. Assembly statistics for the original genome assembly as well as the assembly containing scaffolds longer than 1 kb. N50 is the length of the smallest scaffold in the set of longest scaffolds containing at least half of the total genome assembly, while L50 is the number of scaffolds in the set. 23

Table 1.3. Descriptive statistics for genetic map based on 92 F2 individuals. 24

Table 1.4. Two candidate genes that have been associated with SI in other species, also predicted in the region immediately around the peak of the QTL in *T. coronopifolia*. Along with the scaffold ID, scaffold-specific gene ID, I have included the AED score (strength of evidence for the predicted gene; the closer the value to 0, the higher the confidence in the prediction), the homologous *A. thaliana* locus (as found on TAIR), and the associated description and GO term. 25

Table 2.1. Numbers of paired-reads obtained from a paired-end (PE) and a mate-pair (MP) library from two sequencing efforts. 57

Table 2.2. Duplication level in in reads obtained from the two sequencing efforts as well as the combined dataset. 58

Table 2.3. Number of reads removed by each filter during quality control... 59

Table 2.4. Sample size (N), mean, and standard deviation (SD) for each phenotype for the two years the populations were phenotyped. The asterisk (*) denotes statistical difference between years. 60

Table 2.5. Sample size (N), mean, and standard deviation (SD) for each phenotype measured in the KS segregant population. 61

Table 2.6. Descriptive statistics of genome assembly. Half of the assembly is contained in scaffolds of length N50 or larger. 62

Table 2.7. Phenotype correlations in the KS backcross population. Asterisks indicate level of significance * $p<0.05$, ** $p<0.01$, *** $p<0.001$ 63

Table 2.8. Phenotype correlations in the FL backcross population. Asterisks indicate level of significance * $p<0.05$, ** $p<0.01$, *** $p<0.001$ 64

Table 2.9. Phenotype correlations in the segregant population. Asterisks indicate level of significance * $p<0.05$, ** $p<0.01$, *** $p<0.001$ 65

Table 2.10. Number of markers (N), length, average and maximum spacing for the KS backcross genetic map. Corresponding chromosome numbers here and in Table 2.11 refer to the same genomic sequence. 66

Table 2.11. Number of markers (N), length, average and maximum spacing for the FL backcross genetic map. Corresponding chromosome numbers here and in Table 2.10 refer to the same genomic sequence. 67

Table 2.12. QTL affecting the traits of interest in the 3 populations, with LOD score, the significance level at which they were detected, the % phenotypic variance they explain, and the effect of substituting a FL-derived allele for a KS-derived allele. 68

Table 3.1. For the Tolpis F2.25 and F2.27 populations: Number of markers discovered with at least one genotyped individual, number of markers with a sufficient number of genotyped individuals to be evaluated (the proportion of discovered markers that were evaluated), and number of "good" markers (the proportion of evaluated markers that were "good") for each pipeline ("A" = alignment pipeline, D" = de novo pipeline) with and without a read depth filter (RDF) for genotype calling. a. No filtering was done during genotyping b. a minimum of 6 reads were required for a genotype to be included in the analysis. 89

Table 3.2 For the Achroia KS BC and FL BC populations: Number of markers discovered with at least one genotyped individual, number of markers with a sufficient number of genotyped individuals to be evaluated (the proportion of discovered markers that were evaluated), number of markers containing >5% "impossible" genotypes (the proportion of evaluated markers with >5% "impossible" genotypes), and number of "good" markers (the proportion of evaluated markers that were "good") for each pipeline ("A" = alignment pipeline, D" = de novo pipeline) with and without a read depth filter (RDF) for genotype calling. a. No filtering was done during genotyping b. a minimum of 6 reads were required for a genotype to be included in the analysis. 90

Chapter 1

The Breakdown of Self-Incompatibility in an Insular Lineage of *Tolpis* (Asteraceae)

ABSTRACT

Angiosperms are the most diverse and abundant group of plants on planet Earth. One of the most frequently observed shifts in angiosperm diversification has been the one from cross-fertilization to self-fertilization. Such transitions are of particular interest in oceanic islands due to the disproportionate amount of phenotypic diversity they harbor. The genus *Tolpis* has a dozen or so described species, most of which are endemic to the Macaronesian islands. All species are self-incompatible (SI), except for *Tolpis coronopifolia*, which is self-compatible (SC), a direct result of the breakdown in SI. Like Brassicaceae, Asteraceae is thought to have a sporophytic SI (SSI), controlled by a single locus (*S*-locus) with multiple tightly-linked genes. In this study, I attempt to verify that a single locus is responsible for the breakdown of the SI response in *T. coronopifolia* by investigating the genetic architecture of self-seed, the capacity for a plant to set seed when given abundant self-pollen. I first constructed a genetic map for a mapping cross between *T. coronopifolia* (SC) and *T. santosii* (SI), and then performed a QTL analysis for association between genotype and self-set seed, a measure of self-compatibility. The analysis revealed a single large QTL affects percent self-seed set. Additionally, I assembled and annotated the genome of *T. coronopifolia*, and identified candidate genes in the QTL region. Using gene ontology terms associated with homologous proteins to the candidate genes, I identified two genes related to pollen-pistil interactions. One containing a serine/threonine kinase domain, similar to that identified as the female determinant of SI in Brassicaceae. A second, a peroxidase, has been hypothesized to facilitate pollen-stigma interactions. The draft genome and associated annotation will serve as genetic resources for the scientific community.

INTRODUCTION

Flowering plants are one of the most diverse and species-rich groups on Earth, with great ecological and economic impact. The angiosperm radiation is one of the greatest events of recent geological time. Their rapid rise and early diversification prompted Charles Darwin to call them “an abominable mystery” because they were a strong contradiction to his belief that “nature does not make a leap.” In this study, I attempt to shed light on the genetic basis of the transition from self-incompatibility to self-compatibility in an endemic species of flowering plants. A surprisingly large proportion (25%) of vascular plant species are endemic to oceanic islands, despite the fact that islands are only 5% of the total land surface on the planet. This extraordinary biodiversity is the result of three processes: dispersal, successful colonization, and diversification of the colonizers’ progeny (Carlquist 1974).

The breeding system of the colonizing individual is an important factor in the establishment and diversification of a new sexually reproducing population. Baker (1955) hypothesized that following a long-distance dispersal event, self-compatible (hereafter, “SC” will be used for “self-compatible” as well as “self-compatibility”) colonizers have a selective advantage in establishing new populations on remote oceanic islands compared to self-incompatible (hereafter, “SI” will be used for “self-incompatible” as well as “self-incompatibility”) colonizers, even though the levels of genetic variation in SC populations is reduced. Baker’s argument was based on the fact that a single propagule of a SC colonizer is all that is needed to establish a new population. With SI colonizers, at least two cross-compatible propagules would need to end up within a reasonable distance around the same time, and in circumstances that would favor gene exchange between the individuals. Although not impossible, that sequence of events is unlikely, and thus the establishment of a SI colonizer is a rare occurrence. Stebbins (1957) named Baker’s hypothesis “Baker’s Law,” noting that “it is of great significance for studies of the origin and migration of genera of flowering plants … that it deserves [this] recognition.” Seeing that SC individuals simply have better chances than SI individuals of establishing a new colony after a long-distance dispersal event, Baker (1967) suggested "Baker's law" be best called a "Baker's rule."

Breeding system is not a simple binary trait. Plants considered SI are sometimes able to self-fertilize (Levin 1996; Cheptou, Lepart, and Escarre 2002; Barrett 2003; Brennan et al. 2011), henceforth referred to as "pseudo-self-compatibility" (PSC; also used to designate pseudo-self-compatibility). Plants with this breeding system maximize the dual advantages of self-fertilization and outcrossing (Levin 1996): reproductive assurance (Baker 1955; Stebbins 1957) and transmission advantage (Fisher 1941), both associated with selfing, and the maintenance of genetic variation within the population which is a direct effect of cross-fertilization.

The SI response is controlled by a single highly polymorphic locus, the *S*-locus (De Nettancourt 2013; Boyes and Nasrallah 1993). Molecular mechanisms that inhibit self-fertilization have evolved independently multiple times during angiosperm diversification (Allen and Hiscock 2008). The two well-characterized SI systems are the gametophytic SI (GSI) and the sporophytic SI (SSI). In short, GSI is a non-self recognition system while SSI is a self-recognition system (Iwano and Takayama 2012). Given the scattered distribution of the two systems across angiosperm phylogeny, the high degree of within-family SI system conservation is surprising (Allen and Hiscock 2008). Gametophytic SI is has been the subject of numerous studies in the Solanaceae, Rosaceae, Scrophulariaceae, and sporophytic SI has been extensively studied in the Brassicaceae (Takayama and Isogai 2005).

The family Asteraceae has the highest number of endemic species on oceanic islands among flowering plant families (Crawford et al. 2011). The genus *Tolpis* has approximately 12 described species, ten of which are insular to Macaronesia, with the majority of those being endemic to the Canary Islands (Jarvis 1980; Park et al. 2001; Moore et al. 2002). A survey of the Canarian lineage of *Tolpis* showed that all but one species are either SI or PSC (Crawford et al. 2008). The only SC species is *Tolpis coronopifolia*

(Crawford et al. 2008). The SC trait is likely derived from a predominantly SI or PSC ancestor (Carlquist 1974; Anderson et al. 2001; Crawford et al. 2008; Igic, Lande, and Kohn 2008).

Asteraceae has a SSI system, which has been well described in *Senecio squalidus* (Asteraceae; Hiscock and Tabah 2003). In this system, the *S*-locus involves two linked genes that are inherited as a non-recombinant unit, an *S* allele/haplotype. The two-gene system is comprised of a pistil-expressed gene and a pollen-expressed gene. When the *S*-locus is functioning properly, if a pollen grain attaches to a stigma of the same *S* haplotype, SI response is triggered and pollen tube growth is arrested.

The SI response in *Tolpis* appears to be controlled by a single locus (Soto-Trejo et al. 2013), which is on par with the genetic basis of SI in *S. squalidus* (Hiscock 2000a). Furthermore, the SI allele in *Tolpis* appears to be partially dominant to the SC allele (Soto-Trejo et al. 2013). Self-compatibility appears to be caused by non-functional S-locus alleles, likely caused by loss-of-function mutations (Igic, Lande, and Kohn 2008). The primary goal of this study is to identify the genetic basis of the breakdown of SI in *T. coronopifolia*, and to investigate the genetic architecture of self-incompatibility in *Tolpis*. To my knowledge, this is the first study of this nature for an endemic species of an oceanic island. A secondary goal of the study is to provide genetic resources, a draft genome and draft annotation set, for *Tolpis* enthusiasts to aid in their future research endeavors.

MATERIALS AND METHODS

Overview

I assembled a draft genome for *T. coronopifolia*. The draft assembly was annotated based on transcripts and proteins from other angiosperms. Additionally, I constructed a genetic map based on a F2 mapping population and performed QTL analysis to identify the locus underlying breeding system (SI/SC). Based on the QTL analysis and the annotation, I found candidate genes in the region of the QTL.

Genomic Data

Genomic DNA was extracted from a single individual of the *Tolpis coronopifolia* (SC) to generate two genomic libraries for sequencing, one Illumina Paired-End (PE) with 300bp insert size (hereafter, "Run1") and one with 1-kb insert size (hereafter "Run2"). The DNA extraction was done by Patrick Monnahan, and the library construction was done by the Genome Sequencing Core (https://gsc.ku.edu/) at the University of Kansas, Lawrence. The reads in Run1 and Run2 were 75- and 100-bp long, respectively. In total, over 250 million pairs of reads (Table 1.1) were generated from the two runs. I used Sickle (https://github.com/najoshi/sickle) to trim the reads and discard any reads containing uncalled nucleotides (N's). Next, I used a maximum-likelihood based tool, Quake (Kelley, Schatz, and Salzberg 2010), to correct sequencing errors. I set the k-mer size parameter (*k*) in Quake to 19, and preserved singletons, reads whose pairing read was discarded, in a separate FASTQ file. Overall, quality control of the sequences removed a little over 2% of the raw data (Table 1.1).

Draft Genome Assembly

Most de Bruijn assemblers (Pevzner, Tang, and Waterman 2001) require that the user specify a *k*-mer size, where the optimal *k* depends on the repetitiveness of the genome, the heterozygosity, and technology-specific error rates (Chikhi and Medvedev 2014). Since these genome characteristics are unknown for *Tolpis*, I used KmerGenie (Chikhi and Medvedev 2014) to estimate an appropriate *k*-mer (73). To assemble the reads into scaffolds, I used ABySS (J. T. Simpson et al. 2009) because of its low error rate when assembling a human chromosome (Salzberg et al. 2012), its parallel processing ability and low memory requirements (J. T. Simpson et al. 2009). Using default parameters, I ran ABySS with *k*=73 on a single cluster node with 16 processors, and a total of 32GB of RAM. After the ABySS process was complete, I ran the Assemblathon 2 script "assemblathon_stats.pl" (Bradnam et al. 2013) to calculate descriptive statistics for the resulting assembly (Table 1.2). Due to the fragmented nature of the assembly, I used only scaffolds of length 1 kb or longer (hereafter "1-kb assembly"; assembly statistics in Table 1.2) in downstream analyses. That cutoff scaffold length had to be long enough to harbor a gene, and was calculated based on the average gene length in two species of Solanaceae (3.1kb), a closely related family of angiosperms.

Annotation

To annotate the 1-kb assembly, I used the MAKER2 pipeline (Holt and Yandell 2011). MAKER uses transcript and protein data to predict genic regions in the assembly. Due to lack of transcriptomic data from *Tolpis*, I did not train the gene finders using the normal pipeline; rather, MAKER2 generated gene and protein predictions solely based on alignments of ESTs and protein databases from related taxa. Namely, I provided the software with ESTs from *Helianthus annuus* (Asteraceae) and *Lactuca sativa* (Asteraceae), and protein sequences from 4 species: *Solanum lycopersicum* (Solanaceae)*, Solanum tuberosum* (Solanaceae), *Mimulus guttatus* (Phrymaceae), and *Arabidopsis thaliana* (Brassicaceae). All sequence files were downloaded from Phytozome (Goodstein et al. 2012) except for the Heliantus annuus ESTs, which were downloaded from UC Davis' The Compositae Genome Project Database (http://cgpdb.ucdavis.edu/asteraceae_assembly/).

Mapping Population and Genotyping

A mapping population was established and phenotypes were recorded as described in Crawford et al. (2008) and Crawford et al. (2013). In short, Dan Crawford used *T. coronopifolia* (SC) pollen to fertilize *T. santosii* (SI; Crawford, Mort, and Archibald 2013) individuals. Then, he used an individual from the F1 progeny with high self-set seed to give rise to an F2 population via self-pollination. He scored the members of the F2 for percent self-seed set (number of seeds divided by the total number of florets).

Based on the self-set seed, we additionally scored the progeny individuals as SC (the self-set seed ≥70%), or SI (the self-set seed < 70%). The three distributions are in Figure 1.1 (in Results section).

Genomic DNA was extracted from 101 individuals belonging to the F2 population as well as the parental *T. coronopifolia* species. MSG libraries were prepared according to the protocol described in Andolfatto et al. (2011). All molecular work was done by Nick McCool. Instead of *Mse*I restriction enzyme, *Ase*I was used because it cuts less frequently resulting in higher sequencing read coverage per locus. A step in the MSG library preparation is the ligation of a 6-bp barcode, unique to each individual, that allows for multiple individuals to be pooled and run on one sequencing lane. The resulting FASTQ file contains reads from all individuals. The FASTQ file was split by the unique inline barcode using custom Python scripts. I then aligned the sequenced reads from each individual, including *T. coronopifolia*, to the scaffolds of the 1-kb assembly using BWA (H. Li and Durbin 2009). I used GATK (McKenna et al. 2010) to call SNPs in the population.

I parsed the resulting VCF file (Danecek et al. 2011) using custom Python scripts. In addition to generating files for the genetic map construction software, the scripts polarized individual genotypes and filtered the SNPs. Since the draft genome is based on genomic data from the SC species, and data suggest that SI is dominant to SC (Crawford et al. 2008), I polarized individual genotypes in the following manner: at a given marker, individuals homozygous for the reference allele were classified as (*rr*), whereas individuals homozygous for the alternative allele as well as heterozygotes were classified as *RR* and *Rr*, respectively. I used capitalization to indicate dominance relationship at the SNPs. I only included SNP loci that meet all of the following criteria: 1) the locus has only two segregating alleles, 2) total read depth among samples is between 200 and 50000, 3) the genotype of the *T. coronopifolia* individual is called as homozygous for the reference allele, 4) at least 60% of the individuals have a genotype, and 5) the empirical heterozygosity for the locus is between 0.4 and 0.6. The SNPs that passed this filter (N=2,755) constitute the set of markers used to construct the linkage maps and to perform a

QTL analysis. For easier identification, I gave each marker an informative name containing the scaffold ID and nucleotide position of the polymorphism in the scaffold.

Linkage Map Construction

To build the linkage map, I used Lep-MAP2 (Rastas et al. 2013; Rastas et al. 2016), an open-source software package for constructing ultra-high-density linkage maps. In addition to being able to handle thousands of markers per chromosome, Lep-MAP2 is multi-threaded, which cut down on analysis runtime (Rastas et al. 2016). The software requires genotype data from one or more full-sib families (F_2 progeny along with their parents, F_1) in LINKAGE pedigree format (Lathrop et al. 1984). Since I didn't sequence an F1 individual, I inferred the hybrid genotypes based on the progeny individuals. For each SNP that made it into the final set of markers during VCF parsing, I noted the reference (SC) and the alternative (SI) alleles, and recorded the resulting genotype. The "artificial" F_1 parental genotypes were included in the LINKAGE file, once for each parent.

To split the markers into linkage groups (LGs), I first ran the *SeparateChromosomes* module, with a LOD (logarithm of odds) score limit of 12, and without a lower limit to the number of markers per linkage group. The subroutine performs a segregation distortion test by comparing expected Mendelian proportions and the distribution of genotypes in the progeny. The module identified 630 (out of 2,755) marker that deviate significantly (p<0.01) from the expected Mendelian ratios, and removed them. The remaining 2,125 markers were placed on 9 LGs (Table 1.3). Finally, I ordered the markers on each LG, separately, with the *OrderMarkers* module. Default values were used for all but two parameters, mapping function (useKosambi=0, so Haldane was used instead) and duplicate removal (removeDuplicates=1). Lep-MAP2 doesn't have a graphing module, so the output (marker order and recombination distances) was re-formatted for R/qtl (Arends et al. 2010). A graphical representation of the genetic map was generated in R/qtl (Figure 1.2).

QTL Mapping

To find evidence for linkage to a QTL, I performed a genome scan for a single QTL in R/qtl (Arends et al. 2010). First, I used standard interval mapping approach on the percent self-seed set, and ran 1000 permutations to establish a genome-wide likelihood of odds (LOD) threshold of significance (Churchill and Doerge 1994). Given the non-normal phenotype distribution (Figure 1.1a), I performed the same analysis on the log-transformed phenotype data, which identified the same QTL.

Since SI response in Tolpis is controlled by a single locus, with the SI allele being dominant to the SC allele (Soto-Trejo et al. 2013), I used a goodness-of-fit test at each marker in the final set (N=2,755) in F2 to identify scaffolds that are strongly associated with the SI/SC phenotype. For each marker, and using the binary phenotype scoring, I composed a 2x2 contingency table with genotype (*rr* and *RR*/*Rr*) and phenotype (SI and SC), and assigned each individual to a category based on their genotype/phenotype combination. I used Mendelian ratios for a single-gene inheritance for an intercross between two hybrids as the expected frequencies (0.25 *rr* and 0.75 *RR*/*Rr*). To asses if there is a significant difference between the observed and expected numbers for each of the 4 categories, I calculated a *G* statistic which was compared to the χ^2 distribution with 1 degree of freedom.

Identification of Putative Genes in the QTL region

I examined the predicted proteins on every scaffold (as indicated by marker name) located in the QTL region (within ~3.5cM of the peak) for homology to proteins in *A. thaliana* using the blast tool on The Arabidopsis Information Resource (TAIR). I blasted every protein sequence and inspected the best match for gene ontology (GO) terms related to pollen recognition or pollen-pistil interactions.

RESULTS

Phenotype Distributions

In the F2, self-seed set is a continuous trait with a skewed bimodal distribution (Figure 1.1a). When scored as SI or SC, the F2 individuals are 80.34% SI and 19.66% SC; that is approximately a 4:1 ratio (Figure 1.1b).

Genome Assembly and Annotation

The assembly statistics for the draft genome of *Tolpis coronopifolia* are presented in Table 1.2. The 1-kb assembly is a little under 61% of the total length of the original assembly but it has the benefit of consisting of a more manageable number of scaffolds (Table 1.2). Additionally, a preliminary set of gene annotations will be made publicly available as a resource for the *Tolpis* community. In total, MAKER2 found 29,320 genes, which is on par with the number of genes in *A. thaliana* (Swarbreck et al. 2008).

Linkage Map

Using 2,125 markers from 92 F2 individuals, I generated a marker-dense genetic map containing 9 LGs (Figure 1.2), which corresponds to the number of chromosomes for the genus (Jarvis 1980). Eight of the 9 LGs contain at least 200 markers. A detailed distribution of marker number across LGs, map lengths, and spacing are provided in Table 1.3.

QTL Mapping Analysis

The QTL analysis revealed a single QTL at 3.55cM on LG2, with a peak above the genome-wide threshold of ~4.75 (Figure 1.3a). The marker located at the peak of the QTL is called M8903214_17347 (hereafter "peak marker"). Results from standard interval mapping of the percent self-seed and from log-transformed percent self-seed, and results from the binary model for the SI/SC data are qualitatively the same but differ in the strength of evidence (Figure 1.3b). The locus explains between 52.09-80.20% of the variance in self-seed set, depending on the chosen analysis. In Figure 1.4, phenotype averages are shown for the 3 genotype groups (*RR*, *Rr*, *rr*). The trait appears to be nearly recessive, meaning that in order for a plant to be self-compatible, it needs to have two copies of the SC allele (*r*). Additionally, the SI phenotype is "leaky" meaning that, while SI plants are predominantly outcrossing, they are capable of setting self-seed at low frequency.

Candidate genes

I found a large number of genes in the scaffolds spanning the ~7cM around the peak of the QTL (Table 1.4). Of the 32 predicted protein sequences MAKER2 generated, I identified two genes of interest that could be a part of the *S*-locus (Table 1.4).

DISCUSSION

In this study, I investigated the genetic basis of the self-incompatibility breakdown in the insular lineage *Tolpis coronopifolia* (Asteraceae). One of the main results of my study is that there is a single large-effect QTL affecting breeding system, and the derived allele (SC) is nearly recessive. This finding agrees with previous research suggesting that there is a single locus of large effect determining SI response in the species (Soto-Trejo et al. 2013). Additionally, I identified 2 genes that are potentially involved with the breakdown of the SI response in *T. coronopifolia*. Finally, through this study, I generated considerable genetic resources, including a draft genome with an accompanying annotation and a gene-based genetic map, to advance future studies of *Tolpis*.

Sporophytic Self Incompatibility (SSI)

The QTL mapping results (Figure 1.3) indicate that the breakdown of the SI response is likelycontrolled by a single locus, a result that is in agreement with the described mechanisms of SSI not only in *Senecio* *squalidus* (Asteraceae) (Hiscock 2000b; Hiscock 2000a), but also in species of Brassicaceae (Hiscock and McInnis 2003). Studies have suggested the rejection of self-set seed is likely the derived state in most taxa of flowering plants (Hurka and Neuffer 1997; Beck, Al-Shehbaz, and Schaal 2006; Igic, Lande, and Kohn 2008). Despite the striking similarities between these SSI systems, molecular studies are providing evidence for divergent mechanisms of these systems.

In most well studied SSI systems, the *S* locus is a tightly linked multi-gene unit (Kusaba et al. 2001). The number of genes in the *S*-locus varies but there are two necessary components for recognition of self-pollen - a pollen-specific gene and a stigma-specific gene, the male and female determinant, respectively (Takayama and Isogai 2005). The two determinants have been identified in the Brassicaceae. The female determinant is the S-locus Receptor Kinase (SRK) gene (Stein et al. 1991; J. Nasrallah and Nasrallah 1993; Hatakeyama et al. 1998; Kusaba et al. 2001), while the male determinant is a S-locus Cysteine Rich (SCR) gene. A SRK protein consists of three domains: (i) an extracellular S-domain, (ii) a transmembrane domain, and (iii) an intracellular serine/threonine kinase domain. The annotation of the draft genome along with the QTL analysis identified two genes of interest. One of the genes is homologous to the intracellular serine/threonine kinase domain in *A. thaliana*, which points possibly to the identification of the female determinant gene of the *S*-locus in *T. coronopifolia*. I was unable to identify a SCR-homologous gene in the vicinity of the first candidate gene due to the disjointed nature of the assembly. Additionally, SCR proteins are small in size and highly divergent (Takayama et al. 2000; Watanabe et al. 2000) which makes their identification much more challenging. The second gene that I identified is a peroxidase specifically expressed in the stigma that has been associated with certain *S*-alleles (Hiscock et al. 2003; McInnis et al. 2005), but has yet to be explicitly linked to SI response. These homologies need to be explored by future functional genetics studies to confirm or refute their role in the SI response in *Tolpis*. Allen et al (2011) found no evidence of SRK being involved with the SI response in *S. squalidus*. Finally, a potentially causative relationship between the identified genes and SI breakdown is not evidence that either of the candidate genes are a part of the multigene until controlling pollen recognition.

The Genetics of the Transition to Selfing

In the present study, I have found that the breakdown of the SI response maps to a single QTL. In a similar interspecific cross between *Solanum lyconpersicum* (SC) and *Solanum habrochaites* (SI), Bernacchi and Tanksley (1997) found self-incompatibility to be strongly associated with a single QTL which they found to be in the S-locus. This result is also consistent with the findings of a recent study (Soto-Trejo et al. 2013), where SI and SC plants segregated in the progeny in a 3:1 ratio, as is expected from a single locus with a dominant allele. In *Capsella*, self-compatibility maps to a single locus, the described Brassicaceae *S*-locus (J. Nasrallah et al. 2007). Similar to *Tolpis*, *Capsella*'s SC character is derived (Hurka and Neuffer 1997). Unlike *Tolpis*, studies suggest that the SC allele is partially dominant to the SI allele. (Slotte et al. 2012).

In SI taxa, the emergence of self-compatibility is associated with the breakdown of the genetic mechanism that prevents self-fertilization (de Nettancourt 1977; Busch and Schoen 2008; Igic, Lande, and Kohn 2008; Pettengill and Moeller 2012). This shift in mating system is associated with specific changes in floral and reproductive traits named the selfing syndrome (Ornduff 1969). In addition to a high percent of selfed seed, selfing species have fewer, smaller, and less showy flowers, lower pollen-ovule ratios, and smaller anther-stigma separation relative to outcrossing relatives. *Tolpis coronopifolia* has several of those characteristics, including high self-seed set, smaller capitula, and fewer florets per capitulum (Crawford et al. 2008).

While not much more is known about the genetics of the appearance of non-functional *S*-alleles, there is extensive research on the subsequent changes in floral morphology. The association between selfing rate and floral characters has been observed in multiple genera, including *Arabidopsis* (Goodwillie et al.

2010), *Leptosiphon* (Goodwillie 1999), *Mimulus* (C. Ritland and Ritland 1989). Furthermore, studies have found that floral traits associated with the transition to selfing tend to be polygenic. In interspecific crosses of *Mimulus guttatus* and *Mimulus platycalyx*, Lin and Ritland (1997) examined 5 floral characters that are highly divergent between the selfing and outcrossing species; each character was associated with 1-3 QTL of small effect. In a similar cross between the *M. guttatus* and *M. nasutus*, Fishman *et al*. (2002) found at least 11 QTL associated with each floral trait they examined. Goodwillie *et al*. (2006) detected 3-7 QTL of moderate effect per floral character studied. In *Capsella*, morphological traits characteristic of the selfing syndrome were associated with 2-5 QTL per trait (Slotte et al. 2012). The traits measured in these studies are adaptive floral characters associated with the selfing syndrome, while the trait used in this study is the ability to set self-seed. It is possible that by acting on self-compatibility, selection resulted in correlated changes in these floral characters.

Pseudo Self-Compatibility and Baker's Law

An interesting result from this study is the variable dominance effects of the SC alleles; more specifically, a large portion of the plants that produce some self-set seed are heterozygous at the detected QTL (Figure 1.4). The ability of SI/SC heterozygotes to set self seed are likely to have propagated predominantly selfing lineages that are homozygous in the ancestral population of *T. coronopifolia*. Evidence suggest that SC has evolved from predominantly SI lineages multiple times in angiosperm diversification (Levin 1996; Cheptou, Lepart, and Escarre 2002; Barrett 2003; Brennan et al. 2011). While most plant species are predominantly self-incompatible (Goodwillie and Ness 2005), they are able to set self-seed at low or intermediate frequency, a phenomenon termed pseudo-self-compatibility (PSC; Levin 1996; Stephenson, Good, and Vogler 2000; Vogler and Stephenson 2001; Brennan, Harris, and Hiscock 2005; Mable et al. 2005; Stone, Sasuclark, and Blomberg 2006; Mena-Ali and Stephenson 2007). A population genetics study of *S. squalidus* has alluded to the fact that the SSI system is somewhat flexible in that such plants will set autogamous seed under stress conditions (Hiscock 2000b). The strength of SI is thought to be determined by genetic factors such as the dominance relationships between S-alleles, mutations that can make certain S-alleles nonfunctional, and the presence of modifier loci that influence the strength of S-alleles present in the population (Levin 1996; Stephenson, Good, and Vogler 2000; Good-Avila and Stephenson 2002). The results of a phylogenetic study further support that *Tolpis coronopifolia* evolved from PSC ancestors that colonized the Canary Islands (Archibald et al. 2006). This makes *Tolpis* an exception to Baker's law. Similar exceptions have been found in the Hawaiian silversword alliance (Asteraceae; L. R. Nielsen and Siegismund 2003) and in the Galápagos endemic genus *Scalesia* (Carr, Powell, and Kyhos 1986).

The ability to rely on self-fertilization when outcrossing is not an option (invasion of new geographical area, sudden loss of pollinator, etc) provides some of the benefits for colonizers discussed both by Baker (1955; 1967) and by Carlquist (1974). A single PSC colonizer capable of self-fertilization can propagate a population, which follows Baker's law. In addition to this reproductive assurance aspect, under favorable conditions, PSC populations will tend to utilize outcrossing as their main reproductive strategy (Levin 1996), which in turn would ensure that the levels of genetic diversity in the population are increased in comparison to the expected genetic diversity if the colonizer was strictly SC. Thus, Carlquist's argument that an SI colonizer will be more successful due to increased genetic diversity is also applicable to a PSC colonizer.

Despite the fact that the shift from cross-fertilization to self-fertilization is very frequent in angiosperm radiation, a disproportionately small percentage of plants utilize selfing as their main reproductive strategy (Goodwillie 2005). Stebbins (1957) hypothesized that it might be because this breeding system is an evolutionary "dead end" – transitions from outcrossing to selfing are not reversible, reduced genetic variation due to inbreeding depression diminishes the ability of selfing lineages to adapt to new environmental conditions, and, therefore, are at a greater risk of extinction. However, evidence supporting maladaptation of selfing species is somewhat lacking (Igic and Busch 2013). To elucidate the genetics of adaptation in selfing species, we need to understand both the genetic basis of floral characters associated with different breeding systems but also the genetic basis of selfing capacity of plants. In this study, I have attempted to make a small contribution towards the latter.

FIGURES

Figure 1.1. Distribution of F_2 phenotype: (a) percent self-seed set, (b) log of percent self-seed set, and (c) absolute frequency of SC and SI individuals.

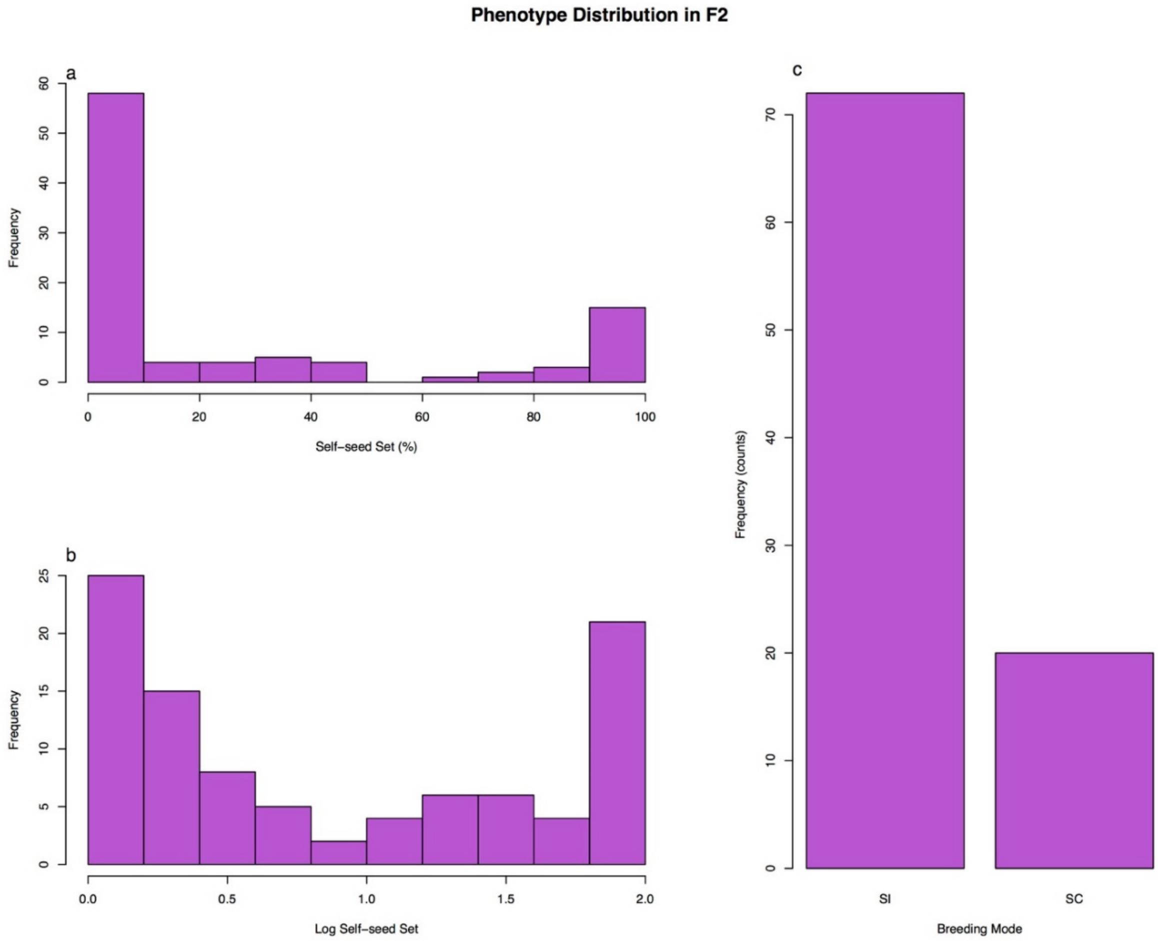


Figure 1.2. *Tolpis coronopifolia* (n=9) linkage map containing 9 linkage groups and a total of 2,125 markers from 92 individuals.

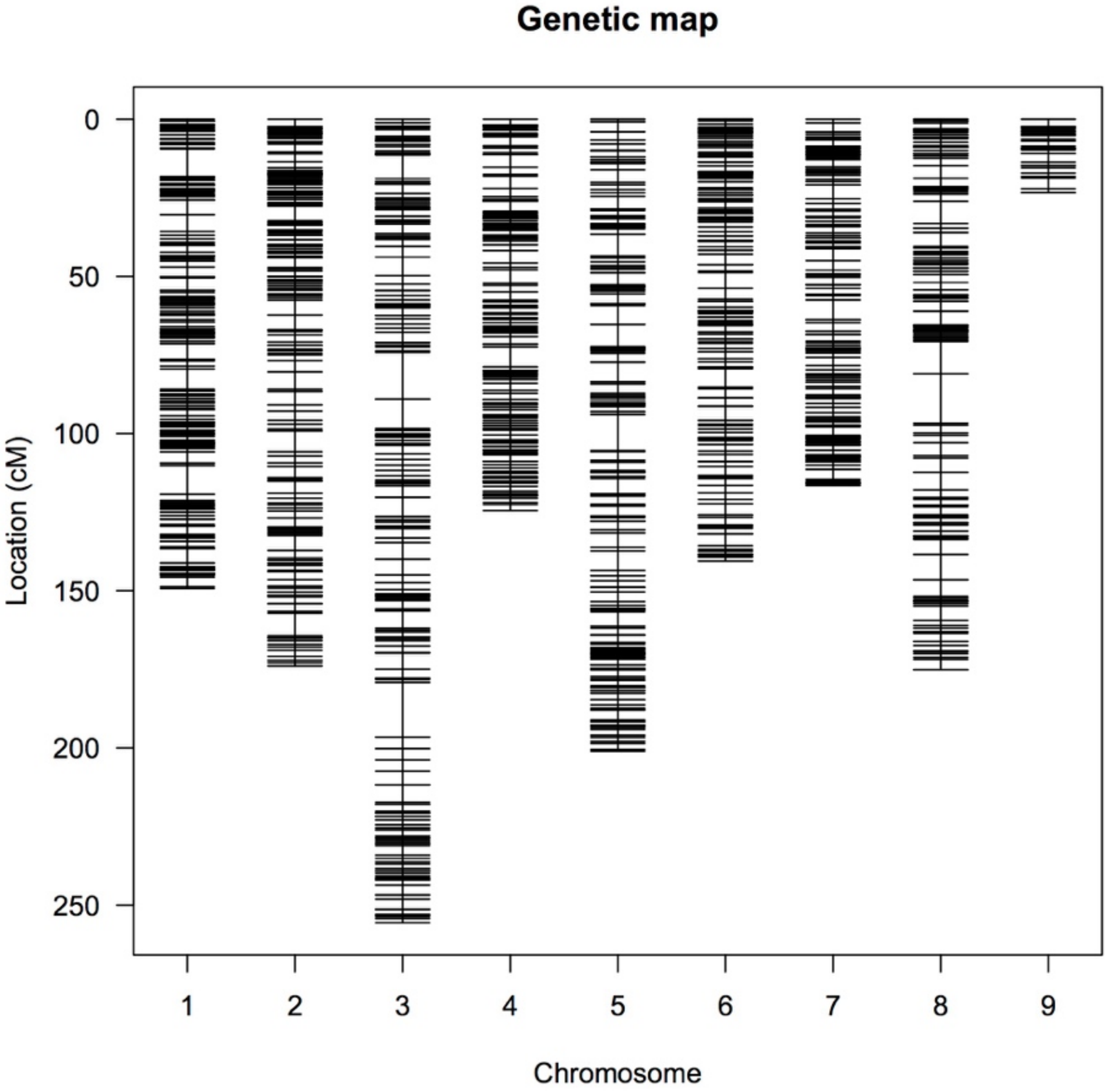


Figure 1.3. A QTL analysis revealed a single QTL around 3.5cM on LG2. (a) LOD scores for the 9 LGs using standard interval mapping of percent self-seed set. The dotted line is the genome-wide threshold established by a permutation test (N=1000) at α=0.01. (b) LOD scores from 3 analyses for LG2, which contains the QTL peak. The turquoise line is from standard interval mapping of percent self-seed, the violet is from standard interval mapping of log-transformed percent self-seed set, and the pink line is based on a QTL analysis using a binary model for the SI/SC scoring of the phenotype. The dotted line is the genome-wide threshold established by a permutation test (N=1000) at α=0.01, and is approximately the same for all analyses.

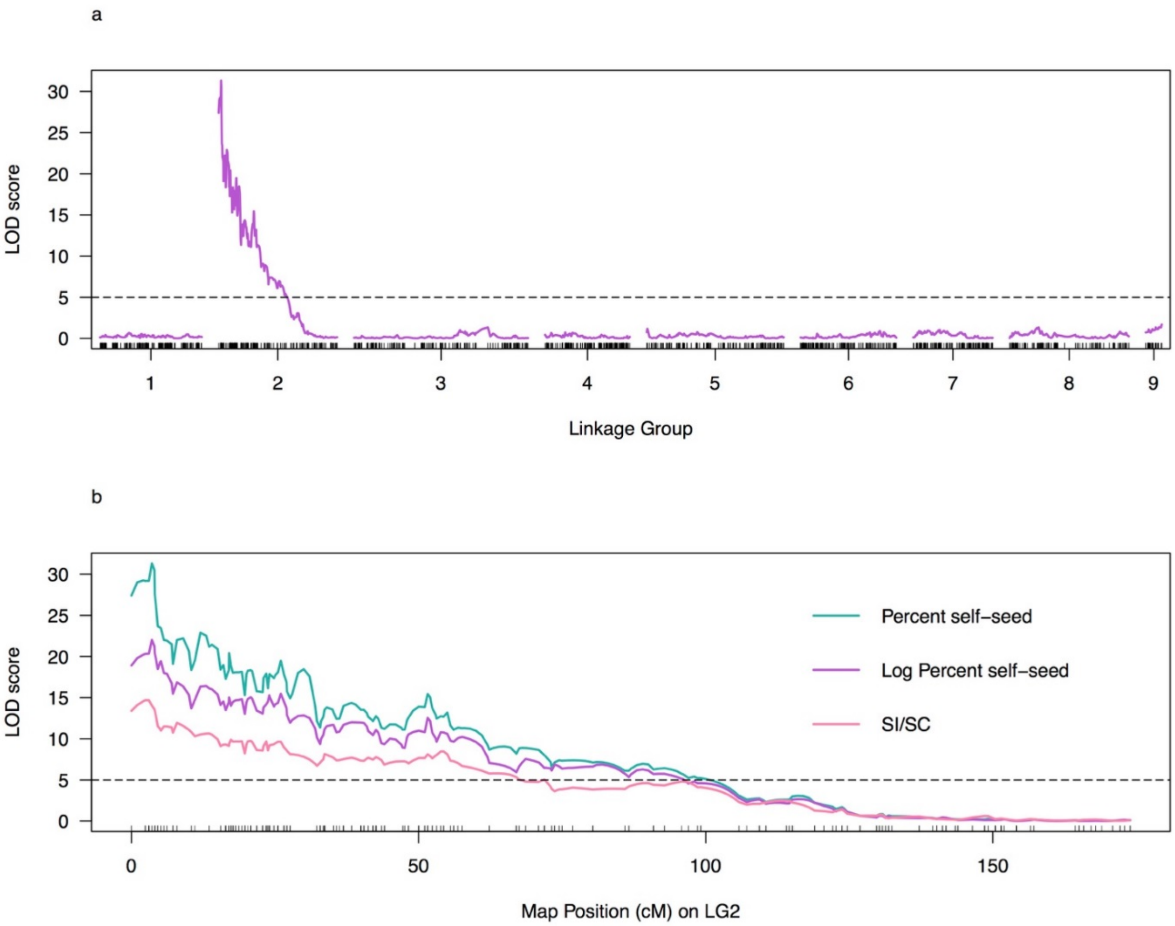
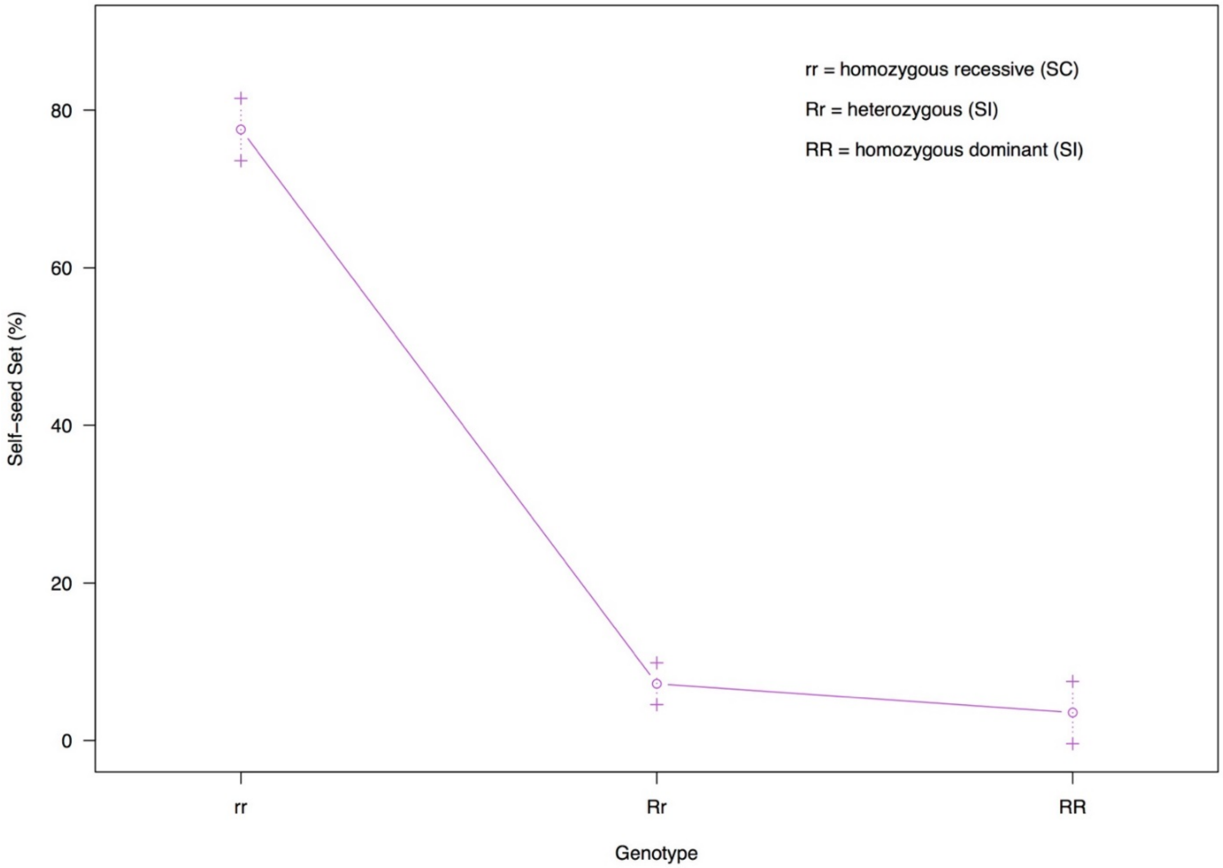


Figure 1.4. The effect plot for the peak marker at the QTL associated with SC in *T. coronopifolia*; phenotype average (+ symbols indicate the span of the standard error). A single copy of the SI allele (*R*) appears to be enough to inhibit self-fertilization (strong dominance effect), but when the number or SC alleles (*r*) increases from 1 to 2 in an individual, the ability to set seed increases drastically.



TABLES

Table 1.1. Number of paired and singleton reads after trimming (Sickle) and correcting (Quake) raw reads. The percent of the original data preserved after each step is in parentheses.

	Run1	*Run2*	*Singletons*	*Total*
Raw	116,878,047 (100.0)	136,733,917 (100.0)		253,611,965 (100.0)
Sickle	113,079,007 (96.7)	133,827,705 (97.9)	5,531,771	252,438,485 (99.5)
Quake	105,600,129 (90.4)	131,398,481 (96.1)	10,776,624	247,775,236 (97.7)

Table 1.2. Assembly statistics for the original genome assembly as well as the assembly containing scaffolds longer than 1 kb. N50 is the length of the smallest scaffold in the set of longest scaffolds containing at least half of the total genome assembly, while L50 is the number of scaffolds in the set.

	Original Assembly	1-kb Assembly
Total size of scaffolds (bp)	1,186,490,297	722,218,512
Number of scaffolds	3,977,270	111,059
Scaffolds > 1kb	110,995	110,995
Scaffolds > 10kb	22,042	22,042
Scaffolds > 100kb	4	4
Longest scaffold	108,991	108,991
Mean scaffold size	298	6,503
Median scaffold size	99	3675
N50 scaffold length	4,124	11,570
L50 scaffold count	51,125	18,025
%N	4.07	5.83

Table 1.3. Descriptive statistics for genetic map based on 92 F2 individuals.

Linkage Group	Number of Markers	Total Length (cM)	Average Spacing (cM)	Maximum Spacing (cM)
1	298	149.3	0.5	9.2
2	291	173.9	0.6	7.2
3	270	255.6	1	17.4
4	262	124.5	0.5	6.4
5	258	201	0.8	11.5
6	246	140.6	0.6	5.9
7	245	116.5	0.5	6.3
8	212	175.1	0.8	15.8
9	43	23.3	0.6	3.4
overall	2125	1359.9	0.6	17.4

Table 1.4. Two candidate genes that have been associated with SI in other species, also predicted in the region immediately around the peak of the QTL in *T. coronopifolia*. Along with the scaffold ID, scaffold-specific gene ID, I have included the AED score (strength of evidence for the predicted gene; the closer the value to 0, the higher the confidence in the prediction), the homologous *A. thaliana* locus (as found on TAIR), and the associated description and GO term.

Scaffold ID	Gene ID	AED score	A. thaliana locus	TAIR Description	GO term
8845966	0.2	0.1	AT1G71695	Peroxidase superfamily protein	hydrogen peroxide catabolic process, oxidation-reduction process, response to oxidative stress
8894954	0.6	0.3	AT4G21380	encodes a putative receptor-like serine/threonine protein kinases that is similar to Brassica self-incompatibility (S) locus	defense response, protein phosphorylation, recognition of pollen

Chapter 2

De novo Genome Assembly and QTL Mapping of Male Song Characters in *Achroia grisella*

ABSTRACT

The lesser waxmoth, *Achroia grisella* (Lepidoptera: Pyralidae), is a symbiont of the honeybee *Apis mellifera* with an interesting reproductive behavior - while in most moth species females use signals to attract males, in the lesser wax moth, the roles are reversed. Three male song characters, namely pulse rate (PR), peak amplitude (PA), and asynchrony interval (AI), appear to be important for female choice during mate selection. Multiple studies have been published aiming to determine the genetic architecture of these traits. And while loci have been detected as affecting the traits, genetic maps have been usually AFLP-based, which makes direct comparisons between similar studies difficult. Until now, no efforts have been made to construct a gene-based map that can be used for direct comparisons with other studies, and could be linked to a physical location in a genome sequence. Here, I have combined RAD-seq genotype data, from two large backcross populations, with QTL mapping, to investigate the genetic basis of three song characters important for mate pairing in *Achroia grisella* as well as two life history traits. Several QTL were detected at different significance levels. Additionally, I assembled a genome for the waxmoth, assembled transcripts from RNA-seq reads, and annotated the draft genome. Through homology to Z-linked proteins from two other species of Lepidoptera, I identified the Z chromosome in *Achroia*. Finally, I assigned the majority of the longest scaffolds in the assembly to chromosomes. An interesting outcome from this study is that *Achroia* has a surprisingly low frequency of crossover events, making it difficult to generate high-density linkage maps for individuals subject to a single round of recombination.

INTRODUCTION

Advancements in high-throughput sequencing technologies have opened new avenues of scientific inquiry in the field of evolutionary biology(Mardis 2008). Next generation sequencing (NGS) provides a vast amount of sequencing data at low and declining cost, and in a much shorter time than pre-existing sequencing technologies. As a result, the evolutionary genomics literature, which until recently featured studies predominantly on model organisms like flies, mammals, and *Arabidopsis*, is now becoming more inclusive to non-model species. The affordability of NGS has enabled groups that study non-model organisms not only to use these genomic approaches in their studies, but also to venture into new fields such as genomics and gene-based linkage maps(Ellegren 2014).

New methodologies, developed based on NGS technologies, have provided a framework for detecting and genotyping a large number of single nucleotide polymorphisms (SNPs). Restriction site Associated DNA sequencing (Baird et al. 2008) uses a restriction enzyme of choice to reduce genome complexity, modified Illumina adaptors to identify individuals within a sample, mechanical shearing of DNA fragments, and a size selection step during library preparation. Some aspect of RAD-seq have been modified by research groups to suit different research needs. For example, in both Genotyping by Sequencing (GBS; Elshire et al. 2011) and Multiplexed Shotgun Genotyping (MSG; Andolfatto et al. 2011), once barcoded adaptors have been ligated to the DNA fragments, no additional shearing takes place. While GBS lets PCR do the size selection, during MSG library preparation, size selection occurs before the PCR step. As a result, we are now able to construct gene-based linkage maps, which are of critical importance in the identification of quantitative trait loci (QTL) associated with traits of evolutionary significance. The advantage of gene-based linkage maps over classic AFLP-based maps is that markers can be linked to a physical location in a genome, which in turn allows for direct comparisons between studies of the same organism. Most importantly, the sequenced reads (RAD tags) do not require the existence of a reference genome to discover markers and call genotypes, which makes the methodology especially valuable for non-model organisms.

One such organism is the lesser waxmoth, *Achroia grisella* (Lepidoptera: Pyralidae), a symbiont of the honeybee *Apis mellifera* (Milum 1935) with an interesting reproductive behavior (Greenfield and Coffelt 1983). Generally, males of other moth species produce sounds only when in proximity with pheromone-emitting females (Greenfield 1981), *Achroia grisella* males attract females from a long-distance by producing ultrasonic sounds with the tymbals located on the sclerites that cover the bases of their forewings. The acoustic signals are generated during wing fanning as the downward motion of the wings pressures the tymbals. This downward stroke of the wings produces one pair of pulses in a period, while the upward stroke of the wing produces the second pair of pulses in a period as said motion removes the pressure from the tymbals. Features of these male signals have shown to be important for female choice during mate pairing(Jang and Greenfield 1996; Jang and Greenfield 1998b).

Studies have shown that three signal characteristics are important for female choice (Collins et al. 1999) – pulse rate (PR), peak amplitude (PA), and asynchrony interval (AI). Pulse rate is the rate at which signals are emitted. The peak amplitude is the greatest absolute value of sound pressure level recorded during a pulse. Finally, there is a small (100-400 μs) delay between pulses generated by strokes of the left and the right wings. That delay is termed the asynchrony interval. Females are attracted to signals that have greater PR, greater AI, and higher PA (Jang and Greenfield 1996; Jang and Greenfield 1998a; Limousin and Greenfield 2009). In other words, females prefer males that sing faster, more erratic, and louder.

While we know that there is substantial additive genetic variance (V_A) for these characters (Collins et al. 1999; Brandt and Greenfield 2004), most QTL studies have been able to explain only a small fraction of the phenotypic variance. Song characters are polygenic traits and a number of putative QTL have been associated with each trait (Limousin et al. 2012; Alem et al. 2013; Gleason et al. 2016). A locus on the Z chromosome was found to have an effect on PR (Gleason et al. 2016). These studies provide valuable insight into the genetics of the traits but some used AFLP-based maps (Limousin et al. 2012; Alem et al. 2013) or were based on a small number of markers (Gleason et al. 2016). Finally, since no efforts to sequence the genome have been made until now, it has been impossible to even speculate about putative causative genes.

In this study, I aimed to investigate the genetic architecture of three male song characters important to female choice as well as two life history traits (development time and weight). For this purpose, I mapped QTL using thousands of SNPs in two large backcross populations. Additionally, I assembled and annotated a draft genome, and the loci in the genetic map were linked to positions in the draft genome. Finally, during this study, I was able to generate genetic resources that will serve as a tool to help future research on this interesting non-model organism.

METHODS

Overview

I assembled a draft genome for *Achroia grisella* using short-read Illumina sequencing data, and used RNA-seq data from *Achroia* pupae and protein databases from related species to annotate it. Using protein homology, I identified the Z chromosome. I used a large set of SNPs from two backcross populations and a segregant population to form linkage groups, and to associate variants with male song characters. Finally, I used BLAST (Altschul et al. 1990) to link marker sequences to the genome scaffolds so that I can assign scaffolds to chromosomes.

Genomic Data

For the *de novo* assembly, two Illumina libraries were made with genomic DNA from an inbred line called "Kansas" (Gleason et al. 2016). One library was a short insert size paired end (SIPE, hereafter "PE") and the other was a long insert size mate pair (LIMP, hereafter "MP") library. For detailed descriptions of the two library types, see (Metzker 2009). The insert size ranges for the PE and MP libraries were 280-504bp and 3-5kb, respectively. Both libraries were constructed at Cofactor Genomics (Cofactor Genomics LLC; http://cofactorgenomics.com) with aliquots from a pooled DNA sample from multiple individuals. Multiple individuals were required to provide the necessary amount of DNA for library construction due to the limited amount of DNA that can be extracted from a single individual. All moths used for the pooled sample were males, the homogametic (ZZ) sex (Traut et al. 2007; Catchen, Braasch, and Postlethwait 2011), to ensure that autosomes and sex chromosomes were as uniformly covered as possible. Each library was sequenced three times on a lane of an Illumina HiSeq machine, once at Cofactor Genomics (Cofactor Genomics LLC; http://cofactorgenomics.com), and twice the Genome Sequencing Core (GSC; https://gsc.ku.edu/) at the University of Kansas. All sequenced reads were 101-bp long. In total, I obtained approximately 290 million and 405 million pairs of PE and MP reads, respectively (Table 2.1).

Since read quality has a major effect on the resulting assembly (Catchen, Braasch, and Postlethwait 2011; Salzberg et al. 2012), I preprocessed our FASTQ files using a number bioinformatic tools. Initially, I removed adaptor sequences from and trimmed the reads using SCYTHE (https://github.com/vsbuffalo/scythe) and SICKLE (https://github.com/najoshi/sickle), respectively. While running SICKLE, I removed all reads with length less than 80bp, and any reads with uncalled nucleotides (N's). Subsequently, I aligned the reads to the PhiX reference genome to remove any contaminant reads using Bowtie2 (Langmead and Salzberg 2012) with default settings and storing the unmapped reads. Finally, I corrected the set of uncontaminated, quality-trimmed sequences using QUAKE, a maximum-likelihood based tool for detecting and correcting sequencing errors (Kelley, Schatz, and Salzberg 2010). I used Quake with k=18, and preserved reads whose pairing read was discarded in a single (SE) FASTQ file. Preprocessing removed 10% of the original reads (). Even though the average per base quality didn't improve dramatically, the distribution of the metric become tighter around the mean (Figure 2.1).

To check if re-sequencing the libraries resulted in high duplication levels, I cleaned the PE reads generated at Cofactor Genomics (CG-PE) separately from the PE reads generated at KU (KU-PE), and additionally combined the two files into a single FASTQ (hereafter "combined-PE"). A custom python script was used to randomly sample 1 million reads from each of the three files (CG-PE, KU-PE, and combined-PE). I used a random sample of reads from the FASTQ files to run FastQC because I noticed that the reads in the beginning of the files were of lower quality than subsequent reads. I ran FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) on each sample and checked duplication levels. The expectation was that if the second sequencing run resulted in over-sequencing our library, the duplication level in the combined-PE sample would be much greater than the sum of the duplication levels in the CG-PE and KU-PE samples. Since FastQC only uses the first 200,000 reads to compute descriptive statistics, I wasn't surprised to see different numbers for the original file and the random sample. Duplication level in the combined-PE FastQC was not greater than in the CG-PE and KU-PE when summed (Table 2.2), so I used the combined-PE data as an input for the *de novo* assembly. Similarly, I followed these steps for the MP data, and combined CG-MP and KU-MP into a single FASTQ.

Genome Assembly and Evaluation

There are many *de novo* assemblers for large genomes, but no consensus as to which software is best. The results of recent genome assembly competitions (Earl et al. 2011; Salzberg et al. 2012; Bradnam et al. 2013) have suggested that the performance of an assembler depends not only on the graph algorithm employed by the software (see Miller, Koren, and Sutton 2010; Nagarajan and Pop 2013 for overview of graph algorithms for assembly) but also on the available data type and the characteristics of the genome. Most de Bruijn assemblers (Pevzner, Tang, and Waterman 2001)require that the user specify a *k*-mer size, where the optimal *k* depends on the repetitiveness of the genome, the heterozygosity, and technology-specific error rates (Chikhi and Medvedev 2014). I used KmerGenie (Chikhi and Medvedev 2014) to estimate an appropriate *k*-mer (93). To assemble the reads into scaffolds, I used ABySS (J. T. Simpson et al. 2009) because of its low error rate when assembling a human chromosome (Salzberg et al. 2012), and its parallel processing ability and low memory requirements (J. T. Simpson et al. 2009). Broadly, ABySS uses PE reads to assemble contigs, and then uses the MP reads to form scaffolds. I ran ABySS with default parameters and *k*=93 on a single cluster node with 16 processors, and a total of 32GB of RAM.

Despite the increased number of *de novo* sequence assemblers in the last decade, the best way to evaluate a draft assembly is not a trivial task. N50 is the shortest scaffold in the smallest set of ordered scaffolds comprising at least 50% of the assembly. The N50 scaffold length has been heavily used as a proxy to assembly quality but as studies have pointed out, it can be inadequate to fully characterize the quality of an assembly (Salzberg et al. 2012; Vezzi, Narzisi, and Mishra 2012). Therefore, an obvious shortcoming of the metric is its inability to account for misassembled sequences in the draft genome. As a result of the increased demand for statistically explicit evaluation of genome assemblies, likelihood-based measures like ALE (Clark et al. 2013), CGAL (Rahman and Pachter 2013), and LAP (Ghodsi et al. 2013) have been developed. However, there is no consensus yet as to which approach is the best, maybe due to the infancy of these tools. I used Cegma (Parra et al. 2009) with default parameters to assess the completeness of the draft assembly.

Transcriptomic Data

Given that the strain used for genome sequencing is no longer maintained, RNA was extracted and sequenced from two pupae from a different inbred strain, Louisiana (LA) line 112 (Zhou et al. 2008). The extraction was done by Stuart Macdonald. A standard, poly-A selected, unstranded TruSeq Illumina

RNAseq library was constructed for each individual. The two libraries were barcoded, pooled, qPCR quantified, and sequenced on a single lane of an Illumina High Output Paired End 100 flow cell. The run generated approximately 175 million reads, which were trimmed using SICKLE. Approximately 96% of the reads passed the filter criteria in SICKLE: (1) window-wise quality threshold parameter *q* was at least 40, and (2) minimum length post trimming was greater than 50.

All quality-trimmed reads were merged in a single FASTA file. To assemble transcripts, I used Trinity (Grabherr et al. 2011) default parameters except read normalization was turned on and the maximum read coverage for normalization was 50. The assembly was performed on a 16-core node with 256 GB of RAM, and it took a little less than 6 hours to complete. The assembler generated 96,420 transcripts with N50 of 2551 bp and mean scaffold length of 1178.45 bp.

Genome Annotation

To identify genic regions, I used Maker2 (Holt and Yandell 2011). I provided Maker2 with the transcripts assembled with Trinity, and protein databases from *Heliconius melpomene*, *Danaus plexippus*, and *Bombyx mori*. Additionally, I used the repeat database generated by RepeatModeler (Smit and Hubley 2010) containing short and long interspersed nuclear elements, long terminal repeat elements, small RNAs, and other unclassified repeats. I used two gene predictors - Augustus (Stanke and Morgenstern 2005) and SNAP (Korf 2004). For Augustus, I used a publicly available parameter set from the software developer for *H. melpomene* (http://bioinf.uni-greifswald.de/augustus/). Finally, for SNAP, I used an HMM file generated by bootstrap training of the gene predictor over three Maker runs.

Mapping Populations and Phenotypes

The mapping populations were derived from the Kansas (KS) and Florida (FL) inbred lines, as described in Gleason et al. (2016). In short, KS individuals were crossed to FL individuals to generate an F_1

population. A number of F_1 males were backcrossed to KS females, with each pair producing a number of offspring (family). Offspring from across all families composed the first mapping population (hereafter "Kansas Backcross" or "KS BC"). Similarly, a number of F_1 males were backcrossed to FL females giving rise to families with offspring, which are collectively referred to as the "Florida Backcross" or "FL BC". Since *Achroia* females show no germline meiotic crossing over (Suomalainen, Cook, and Turner 1973), only hybrid (F_1) males were used to produce recombinant offspring. In addition, F_1 females were crossed to *KS* males to produce segregant individuals (hereafter "Kansas Segregants" or "KS SG"). From the above-mentioned populations, I extracted DNA from 17 *KS* individuals, 14 FL individuals, 5 F_1 offspring, 447 KS recombinants, 465 FL recombinants, and 198 *KS* segregants.

All progeny individuals were phenotyped for 5 traits: pulse rate (PR), peak amplitude (PA), asynchrony interval, development time (DT), and weight (WT). Phenotypes for the KS segregant population were recorded in the course of one year, whereas the two backcrosses were phenotyped in two different years. The 5 traits were measured as described in Gleason et al. (2016).

Library Construction

To generate a genetic linkage map, a modified MSG protocol (Andolfatto et al. 2011) was used to generate sequence data for the individuals from the mapping populations, and genotype informative markers. The molecular work was done by Jenny Hackett. First, instead of *MseI*, genomic DNA was digested with *AseI* because the latter was expected to cut less frequently. Then, Illumina adaptors containing in-line barcodes were ligated to the DNA fragments, and groups of 48 unique barcodes (individuals) were pooled. During PCR amplification, a unique Illumina index is ligated to each fragment in the pooled sample. In practice, each individual was assigned to one of 24 unique Illumina indices, and within the index, to one of 48 unique barcodes. This protocol allows for a large number of DNA samples (up to 24x48) to be sequenced on the same lane simultaneously at a relatively low cost.

Marker Discovery

After demultiplexing the data with the *process_radtags* algorithm in Stacks (Catchen et al. 2011), all reads from the 17 KS individuals were pooled into a single sample (KS-pool); similarly for FL individuals (FL-pool). The KS-pool and FL-pool samples contained 16.3 million and 13.8 million reads, respectively. To *de novo* assemble loci for each parental pooled sample, I used *ustacks* (Catchen et al. 2011), another program of the Stacks pipeline. I parameterized *ustacks* to construct stacks with a minimum coverage (*m*) of 5, maximum number of differences between reads within a locus (*M*) of 2, no secondary alignments (*N*=0), maximum number of stacks per *de novo* locus (*max_locus_stacks*) equal to 2, and with Removal and Deleveraging algorithms enabled (*-r* and *–d*, respectively). The Removal algorithm excludes stacks that are highly repetitive while the deleveraging algorithm attempts to resolved over-merged stacks. In total, 148,448 and 143,201 stacks were assembled for *KS* and *FL*, respectively.

To establish homology of the KS loci with the FL loci, I used *cstacks* (Catchen et al. 2011), another Stacks component. In short, *cstacks* merges loci from the two parental pooled samples into a catalog. The most important *cstacks* parameter is the maximum distance allowed between loci (*n*) as it determines how loci are merged. I used *n* = 2 to allow for the alleles of potential heterozygous loci in parental individuals to merge. The resulting catalog contains parental alleles grouped into loci, and is the set of informative markers to be genotyped in the recombinants. Using a custom python script, I interrogated the catalog and identified 72,076 entries (*e.g.*, loci) in which each parent contributed a single monomorphic allele. I excluded all other loci because they would represent heterozygous loci or alleles sampled only in one of the two parental lines. Of the 72,076 loci, 26,905 loci (37.3%) were polymorphic and were used to genotype the recombinant and KS segregant progenies.

Genotyping Recombinants

To call genotypes in the recombinant populations, I first *de novo* assembled loci for each individual in the backcrosses using *ustacks* with similar parameterization to that used for the KS-pool and FL-pool samples (*m*=2, *N*=2, *-r*, *-d*, *-max_locus_stacks*=2). Each individual's loci were then matched against the catalog of informative loci generated by the parental lines. Finally, I used Stacks' *genotypes* program to make genotype calls for the two backcrosses and the KS segregants. On average, the 3 populations had a genotyping rate of 44.1%

Genetic Map Construction

To build the linkage map, I used Lep-MAP2 (Rastas et al. 2013; Rastas et al. 2016) an open-source software package for constructing ultra-high-density linkage maps. In addition to being able to handle thousands of markers per chromosome, Lep-MAP2 is multi-threaded, which minimizes analysis runtime (Rastas et al. 2016). The software requires genotype data from one or more full-sib families (progeny along with their parents) in LINKAGE pedigree format (Lathrop et al. 1984). The Stacks *genotypes* output file was converted to the appropriate format using custom Python scripts.

First, I used the *Filtering* module to remove loci with segregation distortion (*dataTolerance*=0.01), and to remove loci and individuals with scarce genotype data. For KS BC, I removed individual with more than 15,000 missing genotypes, and markers with more than 300 missing genotypes. These filters removed 18,773 makers (69.8%) and 108 individuals (24 %). For the FL BC, I filtered out individuals with more than 13000 missing genotypes and markers with more than 370 individuals. As a result, 20,935 markers (77.8%) were removed along with 152 FL BC individuals (32.7%). The reason for using different filtering criteria is to maintain similar levels of missing data in the genotype sets. To split the markers into linkage groups (LGs), I first ran the *SeparateChromosomes* module, with a LOD (logarithm of odds) score limit of 20, and a lower limit to the number of markers per linkage group of 20 markers. Next, each chromosome was ordered individually with *OrderMarkers* using the Kosambi function (useKosambi=1; Kosambi 1943), taking into account the achiasmatic meiosis in females (*initRecombination*=0.05 0, where the first number is the a priori probability of crossing over in males, and the second is the same probability in females), and removing identical markers within each chromosome (*removeDuplicates*=1).

QTL Mapping

To order and orient the assembled scaffolds based on the two genetic maps, I used ALLMAPS (Haibao Tang 2015). In the process, I compared the assignment of markers to LGs in the two maps. Based on overlapping markers, LGs in the FL BC were renamed to reflect the respective LGs in the KS BC. Eight (0.2%) of the 3,204 overlapping markers between the two backcrosses were removed because they were inconsistent in their placement - i.e., a marker was placed on different LGs in the two maps. I noticed that the order of the markers on a given LG was largely different between the two backcrosses due to lack of crossover events (See 'Genetic Map Construction' in Results for more details). Since the markers in these maps are in an order that isn't certain, I will be referring to the two maps as "unresolved" henceforth.

Since it was possible to map QTL to chromosomal level, I assigned each linkage group, in a given individual, the most frequent genotype for that linkage group (as long as it cleared a 90% threshold), and performed marker regression in R/qtl (Broman and Sen 2009). Effect sizes and percent phenotypic variance explained were also estimated.

Mapping Marker Sequences to the Genome Assembly

The sequences of the markers included in the unresolved genetic map were extracted from the *catalog* file (produced by *cstacks*), and formatted as a FASTA file using custom Python scripts. Every sequence header was formatted to reflect the LG and position of the marker in the KS map. Additionally, I made a nucleotide database from the genome assembly using *makeblastdb* on a local BLAST installation. The FASTA file containing marker sequences was then mapped to the assembled genome using BLASTN, with *max_target_seqs* set to 1, and *evalue* set to 1e-30. Effectively, the alignment places scaffolds onto LGs.

Identification of the Z chromosome

To identify the Z chromosome, I used custom Python scripts to extract protein sequences of Z-linked genes in two Lepidopteran species, *Bombyx mori* (N = 654) and *Melitaea cinxia* (N = 572). Each set of protein sequences was aligned to the assembled genome of *Achroia* using TBLASTN with *max_target_seqs* set to 1, and *evalue* set to 1e-50. The alignment identified scaffolds containing sequences with strong homology to Z-linked proteins. Of the 654 Z proteins in *B. mori*, 202 (30.9%) aligned to 123 *Achroia* scaffolds. Of the 572 Z proteins in *M. cinxia*, 136 (23.8%) aligned to 90 scaffolds in the *Achroia* assembly. The number of scaffolds identified by both alignments as homologous to Z was 80. Since I already had information linking scaffolds to LGs, I was able to identify the LG representing the Z chromosome in *Achroia*.

RESULTS

Phenotypes

Means and standard deviations for the 5 phenotypes in the backcrosses and the KS segregants are presented in Table 2.4 andTable 2.5, respectively. In the Kansas backcross, development time (DT), weight (WT), and pulse amplitude (PA) differed significantly between the two years (t-test, $p < 0.05$). In the Florida backcross, DT and PA differed significantly between year 1 and year 2 (t-test, $p < 0.05$). We found that WT and DT are significantly negatively correlated in all three populations (Table 2.7, Table 2.8, and Table 2.9). Additionally, PA is negatively correlated with DT in both backcrosses (Table 2.7, Table 2.8). PA is positively correlated WT in both backcrosses (Table 2.7, Table 2.8). This correlation could mean that larger males are able to emit songs with higher amplitude. Pulse rate (PR) correlated negatively with WT in the Florida backcross and the KS segregant population but the negative relationship in KS BC was not significant, possibly due to dominance. Finally, the only correlation between sexually selected traits was that between PR and PA in the FL BC. Although, the same relationship in the KS BC and the KS SG was also negative but not significant. Correlations between quantitative traits could be an indication of pleiotropic effect or of tightly linked genes controlling the traits.

Genetic Map Construction

Lepmap placed 8,091 markers on 30 linkage groups for the KS BC population, and 5,721 markers on 30 linkage groups for FL BC population (). Additionally, 12,801 markers in the KS SG population were also placed in 30 linkage groups. The number of linkage groups is consistent with the haploid number of chromosomes observed through karyotyping by Limousin et al. (2012). For the two backcross populations, the markers in every linkage group were ordered and the summary statistics for the KS BC and FL BC maps are in Table 2.10 and Table 2.11, respectively. Using homology and the expected (due to SG individuals being test males) “segregation distortion” of the Z in the KS SG population, I was also able to identify the Z chromosome.

In examining the marker order within between the two maps, I found that it was highly inconsistent in a particular way. Markers on the end of each linkage group were consistent between the two backcrosses, but markers in the middle of the linkage group were scrambled. Linkage group 1 is shown as an example in Figure 2.4. The same pattern was observed when I mapped the markers to the *Achroia* scaffolds, separately for every linkage group in each backcross, using ALLMAPS (Figure 2.5). When looking at the genotype calls for each linkage group, there seemed to be almost no recombination in the middle of the chromosome - LG 1 is shown in Figure 2.6 as an example. In Figure 2.6, the x-axis represents the markers along the length of the linkage group, while the y-axis represents individuals from KS BC. Red blocks are homozygous genotypes and blue blocks are heterozygous genotypes. A crossover event in an individual is represented by a set of neighboring blue blocks surrounded by red blocks, or vice versa. Here, it is clear that such events are observed predominantly at the ends of the linkage groups. The lack of crossing over explains why Lepmap can order markers at the ends of the chromosome better than in the middle. There isn’t enough information for the software to estimate recombination distances with high confidence for the markers that are in the middle of the linkage group. Interestingly, Gleason et al. (2016) observed a similar phenomenon in a much smaller set of markers.

I also considered genotyping error as the source of the problem. While genotyping error does not eliminate crossing over events, it is possible that Lepmap would interpret a small number of erroneous genotypes as evidence for crossing over, and force the markers into neighboring positions. So, if the error was systemic in the mapping software, I would observe the same pattern in the KS segregant population. That is not at all the case. The genotype matrix in KS SG were what is expected - i.e., for a given LG in an individual, all genotypes are exactly the same and there is no evidence for crossing over. Figure 2.7 shows a subset of the KS SG individuals for LG1. While there are single blocks of red surrounded by blue, and vice versa, they appear to be scattered along the linkage group which is indicative of random as opposed to a bias disguising errors into crossing over events. Additionally, I repeated the genotype calling step in all three populations with an increased read depth requirement of 8. The new genotype calls yielded a similar result. If the problem were solved with higher coverage, it could indicate systemic bias of genotyping error rate in areas of recombination. To my knowledge, no studies have even suggested that his might be possible. Thus, I concluded that genotyping error is unlikely to be the underlying cause of the marker order inconsistencies. It is worth noting that both Figure 2.6 and Figure 2.7 make it apparent that heterozygous genotypes are being under-called, which appears to be a RAD-seq technology problem (Gautier et al. 2013). For example, in Figure 2.7, the frequency of heterozygous genotypes in homozygous individuals (blue blocks in rows of red) is much smaller than the frequency of homozygous genotypes in heterozygotes (red blocks in rows of blue). Intuitively, it is much easier for a heterozygote to appear as a homozygote in RAD tags than it is for a homozygote to appear as a heterozygote.

The lack of useful information and the stochastic nature of the ordering process itself would likely result in a new marker order with every new iteration of the process. Even in large mapping populations like the ones I have used in this study, calculating recombination distance depends on observing at least some new combinations of alleles. Given the lack of confidence in estimated marker order within LGs, and the relative lack of clear crossover events within chromosomes, I concluded that it would be futile to attempt to map QTL to sub-chromosomal level. Therefore, I ignored all crossing over and assigned each chromosome in every individual a single genotype based on 3 filters: for a given chromosome in an individual, (1) at least 80% of the markers have genotype calls, (2) the minimum number of called markers for that chromosome is 30, and (3) at least 90% of the markers on that chromosome must have the same genotype, which also becomes the chromosome-wide genotype. While the total number of markers is significantly reduced, the simplified genotype matrix (henceforth "haplotype matrix") is much more representative of the data than the original matrix containing thousands of genotypes per individual. Additionally, assuming that all QTL act in the same direction (i.e., KS alleles all have positive effect on the phenotype, or all have negative effect on the phenotype), testing for associations between phenotype and chromosome increases my power to detect loci since multiple small-effect QTL on a chromosome are now considered collectively. The drawback of this manipulation is that I can only map QTL to the chromosome level. However, considering the number of chromosomes (N=30) and the number of genes (est. N=13,191), mapping a QTL to a chromosome using this method allows me a higher gene mapping resolution than in an equivalent backcross or F2 QTL study in *Drosophila* (Mackay 2001).

QTL mapping

To detect phenotype-genotype associations, I used marker regression in R/qtl, with 1000 permutations to set genome-wide LOD thresholds for significance. Three thresholds we established at three different significance levels: (1) α=0.05, (2) α=0.10, and (3) α=0.20. The first is the universally accepted significance level for detection of QTL while the other two allowed me to detect QTL that were borderline significant and to explore any associations, even weak, with the Z chromosome.

The QTL mapping identified a number of chromosomes affecting the three main traits of interest: pulse rate (PR), peak amplitude (PA), and asynchrony interval (AI) (Table 2.12). For PR, two linkage groups were strongly associated (α=0.05) with the phenotype in the RK backcross. I also found one linkage group in the RF backcross that affects PR. For PA, 3 putative QTL were detected (α=0.20)- two loci on two separate linkage groups in FL BC, and one in KS SG. For AI, I found two weaker QTL on two distinct linkage groups in the KS BC population. For the life history traits, DT and WT, I detected QTLs in all three populations. In the KS BC population, I found 3 QTL of varying strength for WT and 1 strong QTL for DT. Notably, the same linkage group (LG7) is associated with both WT and DT (Table 2.12), which is not surprising considering that I also found a strong correlation between the two traits. In the FL BC population, I detected one putative QTL for each phenotype. The two were on separate chromosomes, and different from the chromosomes identified in the KS BC. Finally, in the KS SG population, two weaker QTL were detected for DT, and one strong QTL was detected for WT. Overall, the fraction of genetic variance explained by the QTL is fairly small (0.25-4.92%) except for the DT QTL which explains 6.15% of the phenotypic variance in the KS BC.

Aside from co-localization of DT and WT on LG7, not one QTL for any trait is co-localized within population, and no QTL overlaps across populations. There are a few explanations for the lack of co-localization. First of all, the power to detect QTL of small effects is quite low (Table 2.12), and therefore the chances of detecting the same QTLs in all three populations are fairly low. Additionally, the real number of QTL underlying each trait are likely to be higher than the findings of this study. Beavis (1994) showed that any finite sample size will underestimate QTL number and overestimate QTL effect.
The detection of QTL in one population but not the other could also be due to dominance but given the experimental design, there is no way to assess variance attributed to dominance. Finally, an important artifact of the way I mapped QTL is that if there is some number of small-effect QTL on a chromosome, and the direction of the individual effects sums to a number close to 0, none of these loci would be detected as contributing significantly to the phenotypic variance.

Genome Assembly and Annotation

The assembled genome of *A. grisella* is quantitatively described in Table 2.6. The total length of the assembly is within the expected 400-600-Mb genome size (J. Gleason, pers. comm). Half of the 418 Mb are in scaffolds of length 87.3 Kb or longer. The GC content (32.4%) is on par with other sequenced genomes in the order of Lepidoptera (Zhan et al. 2011). Based on a Cegma run, the assembled scaffolds contain 196 (79.03%) of the 248 core eukaryotic genes. The annotation pipeline identified 13,191 genes with max AED of 0.5 (Eilbeck et al. 2009). The number of identified genes is close to the number of predicted protein coding genes *in H. melpomene* 12,669 {12,669; HeliconiusGenomeConsortium:2012dt} and *B. mori* 12,901 {12,901; Xia:2004uo}. Using BLAST, I mapped the set of markers used for the construction of the unresolved genetic map to the genome assembly. On average, 75% of markers per linkage group in both backcrosses mapped to the genome. While the target scaffolds are a very small proportion of the assembled scaffolds (~4%), the total length of those scaffolds is 60% of the total assembly length. Thus, our genetic markers placed the majority of long scaffolds onto chromosomes.

DISCUSSION

The goal of this study was to determine the genetic basis of 3 male song characters, important to female choice during mate pairing. In addition to mapping several QTL, I generated genomic resources for this non-model organism including a draft genome and annotation. Furthermore, I assigned assembly scaffolds to chromosomes and identified the Z chromosome through protein homology.

Developmental Traits

In line with another QTL study, DT and WT are negatively correlated (Gleason et al. 2016) meaning that the longer it takes a moth to emerge from the pupal stage, the smaller it is. QTL associated with developmental time (DT) and weight (WT) were found in all 3 populations in this study. Across all populations, I found a single significant DT QTL in the KS BC explaining ~6.15% of the phenotypic variation, while the strongest QTL for WT (in KS SG) explains only ~5% of the variance in weight. I also found two significant WT QTL, one in the KS BC population and the other in the KS SG population. While neither of these significant QTL for WT was co-localized with a significant DT QTL, a putative WT QTL was found on the same linkage group as the significant DT QTL in the KS BC. This is of interest because the two traits were strongly negatively correlated and the QTL effect are in the opposite direction in all three populations, which could suggest that a pleiotropic gene is present, or that multiple genes affecting the traits are in very close proximity. Two other QTL studies had similar findings - DT and WT were strongly negatively correlated, and QTLs were co-localized (Alem et al. 2013; Gleason et al. 2016). And while we can't make direct comparisons to Alem et al. because they used AFLP markers, we can make direct and meaningful comparisons with the Gleason et al. study because the backcross individuals in that experiment are a subset of out KS BC population.

Male Song Characters

I detected numerous QTL for male song characters. Overall, the traits appear to be polygenic because the percentage of phenotypic variance explained by each QTL is fairly low. For PR, we found two significant QTL, both in the KS BC, similar to the QTL number found in previous studies (Limousin et al. 2012; Alem et al. 2013; Gleason et al. 2016). However, compared to those studies, the QTLs here are of much lower effect size. Specifically, Limousin et al. (2012) reported that their QTLs explain over 50% of the phenotypic variance observed in their population, and concluded that PR is likely controlled by a number of loci of moderate effect and not by a few loci of large effect. My findings suggest that PR is affected by multiple loci of small effect, collectively explaining ~ 3.6% of the phenotypic variation, an estimate much closer to what Gleason et al. (2016) reported (~4.4%).

Using homology to *Bombyx mori*, Gleason et al. (2016) suggested that one of the PR QTL they detected is on the Z chromosome. I have found no evidence of a PR QTL on the Z. However, that should not be interpreted necessarily as a contradiction. There are a number of reasonable explanations as to why my analysis has not found this connection. It could very well be the case there are small-effect PR QTL with opposite effects on the Z, and that my chromosomal-level QTL detection approach is unable to detect those signals because they cancel each other out. It is also conceivable that my analysis, having increased power due to larger sample size, is correctly inferring the absence of significant QTL, while the reported link in the Gleason et al. study is simply an overestimation of the effect of a small QTL (Beavis 1998). Since this study provides novel genomic resources for *Achroia grisella*, it would be useful to determine if the QTL found on the Z in Gleason et al. (2016) also maps to the scaffolds assigned to the Z in my study.

For peak amplitude, I found 3 putative QTL which is on par with other QTL studies in *Achroia* (Limousin et al. 2012; Alem et al. 2013; Gleason et al. 2016). However, the QTL detected in this study explain only ~ 3% of the variance in PA, a smaller number than that reported in the AFLP-based QTL studies (Limousin et al. 2012; Alem et al. 2013) but similar to that reported by Gleason et al. (2016). Interestingly, none of the PA QTL I found were in the KS BC, which is shared between the two studies. Here, again, a few explanations are plausible (see earlier discussion on lack of PR QTL on Z) but a good starting point would be to map the ESTs from the other study to the *Achroia* genomic scaffolds that have been assigned to chromosomes. For the PR and PA of male song, the only evidence of co-localization comes from a PR QTL in the KS BC and a PA QTL in the KS SG population, both of which are found on linkage group 11. Limousin et al. (2012) found two strong QTL for these male characters to be present on the same linkage group. Unfortunately, since the study utilized AFLPs, direct comparisons regarding location of the QTL between my study and theirs are not possible.

Here, I found putative QTL for asynchrony interval in the KS BC only, which accounted for a small fraction of the phenotypic variance, qualitatively similar to results reported by Limousin et al.(2012). Gleason et al. (2016), on the other hand, were unable to detect any QTL affecting AI, which could be explained by the fact that the QTL mapping in that study was based on a much smaller number of markers

(N=75). It has been suggested that male size is directly responsible for AI (Brandt and Greenfield 2004) but the two traits do not correlate or share QTL, as would be expected. These results are not surprising considering the low heritability of the trait (Collins et al. 1999).

The disparity in variance explained as well as the replicability of detected QTL across populations and the co-localization of QTL could be attributed to environmental variance, genetic factors only detectable in a specific genetic background, or power to detect QTL in our populations. Furthermore, findings of a QTL analysis are dependent on the type of experimental cross and the number of individuals in the population (Beavis 1998; Xu 2003). Finally, my approach to mapping QTL on the chromosome level may have contributed to these disagreements. Ignoring crossover events all together, even if they are rare, can result in signals from small-effect QTL with opposite directions canceling out, and can lead to false negatives.

Choices Have Consequences

The method I used to detect QTL was a way to deal with the fact that without a valid genetic map, one with consistent estimates of marker positions, QTL mapping would not even be possible. Since genotyping error is not expected to eliminate crossing over events, the most likely explanation for the lack of information, in the middle of our linkage groups, leading to semi-stochastic marker ordering is that *Achroia* has a very low crossing over rate. My approach relies on the fact that any given marker on a linkage group in a progeny individual, is representative of a large portion of the LG it is placed on, and by ignoring the order of the markers and assigning each LG a genotype based on the that LG's genotype distribution, I have lost the ability to map QTL to sub-chromosomal level but greatly increased my power because the number of tests performed have greatly decreased.

The low recombination frequency I observed has been reported previously by AFLP- and EST-based studies (Limousin et al. 2012; Alem et al. 2013; Gleason et al. 2016), and some of the estimates of LG length (12-66cM; Alem et al. 2013), suggest that there are chromosomes that don't have a crossover event in every meiosis. Since I expected that a large number of recombinant progeny, like the KS BC, would be enough to compensate for infrequent crossing over events, I chose to use a restriction enzyme that ended up cutting more frequently than would be ideal for the QTL mapping part of this study.

When it comes to applications of RAD-seq, the finite amount of data produced from sequencing technologies forces a choice between read depth and number of loci. One of the goals of this study was to generate genomic resources that will aid the community in future research endeavors, so a natural instinct was to improve the contiguity of the assembled genome using an ultra-dense genetic map – i.e., the more markers in the map, the more scaffolds can be ordered and oriented into chromosomes. So, the expectation that with more markers, I can significantly improve the assembled genome also made me consider the common-cutter restriction enzyme. Of course, this wouldn't be a consideration if I had used PacBio (Rhoads and Au 2015) or Oxford Nanopore (Deamer and Akeson 2000) data to close the gaps.

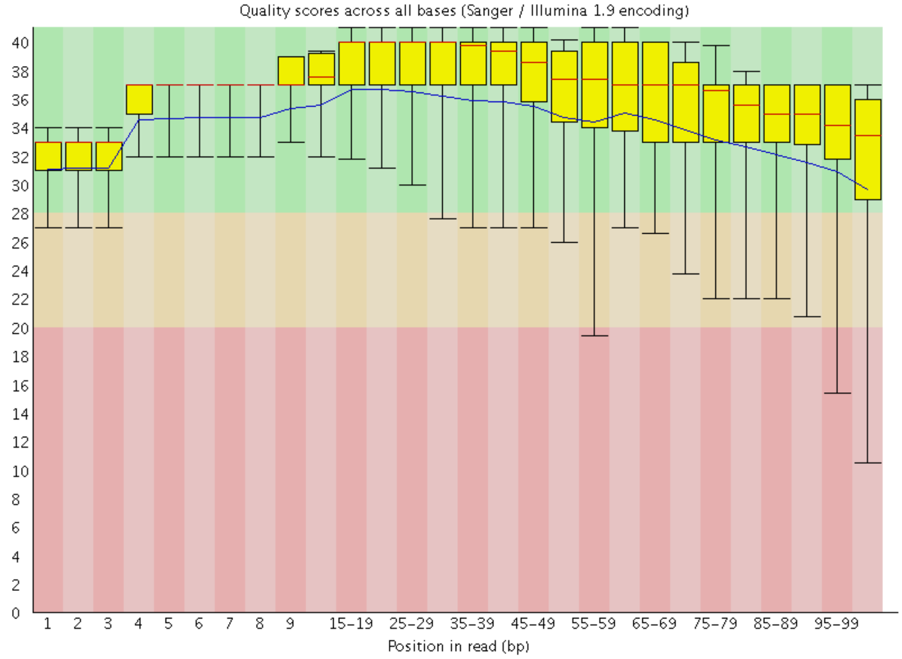
On the other hand, the frequency with which *AseI* cuts DNA allowed for tens of thousands of markers for more than 1,000 individuals to be genotyped and used in forming linkage groups. Of course, the number of markers is inversely proportional to the coverage at each marker, so the data were somewhat scarce. The extent of missing data and the bias in genotyping error in heterozygous individuals, could be attributed to the nature of the RAD-seq method I used, which resembles more low-depth genome resequencing than RAD sequencing. So, both missing data and error rate could have been different had I used a different sequencing method.

Conclusion

I was unable to map QTL to sub-chromosomal level but I did map QTL to chromosomes. I have also generated a reference genome sequence, annotated it, assigned the majority of the longest scaffolds to chromosomes based on the unresolved linkage map, and have identified the Z chromosome. The above mentioned genomic resources will be made available to the scientific community in hopes that they will open new avenues of research for non-model organisms.

FIGURES

Figure 2.1. Average base quality before (top) and after (bottom) read processing of the forward reads from the PE library used to generate the *de novo* genome assembly. The graph was generated by FASTQC (S. Andrews, n.d.).



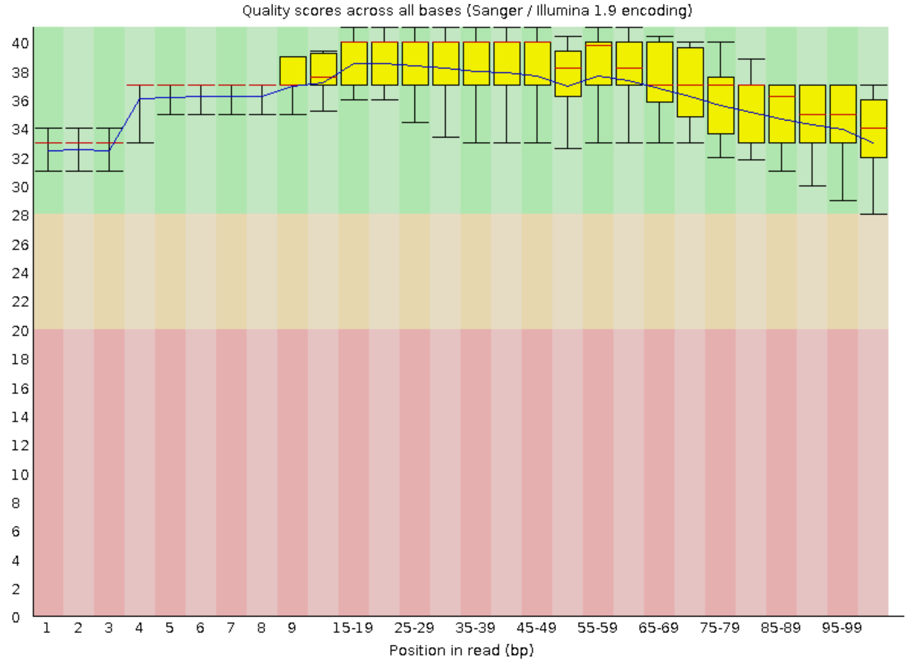


Figure 2.2. Linkage map for the KS backcross as constructed by Lepmap. Chromosome numbers here and in Figure 2.3 are identical

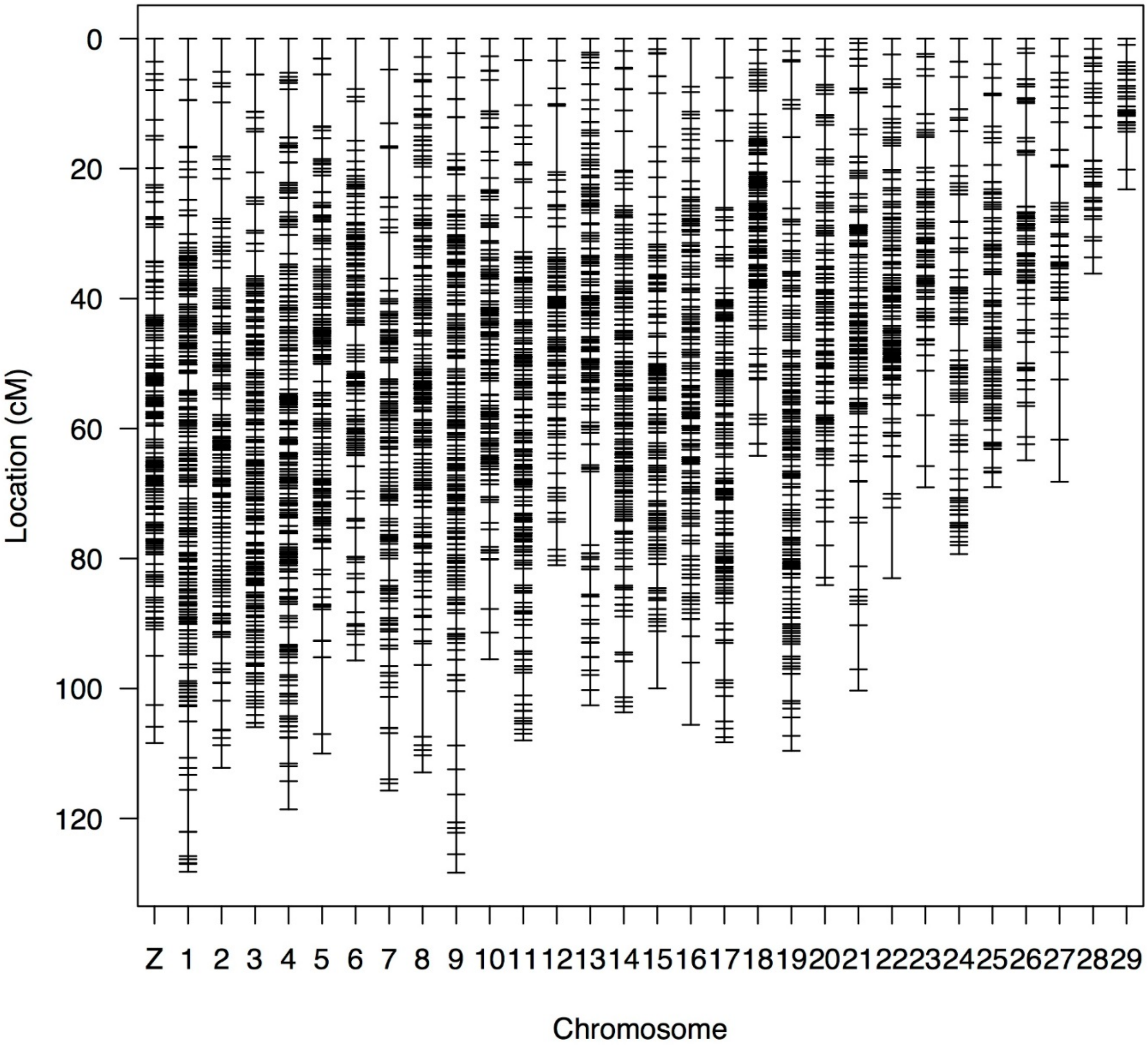


Figure 2.3. Linkage map for the FL backcross as constructed by Lepmap. Chromosome numbers here and in Figure 2.2 are identical.

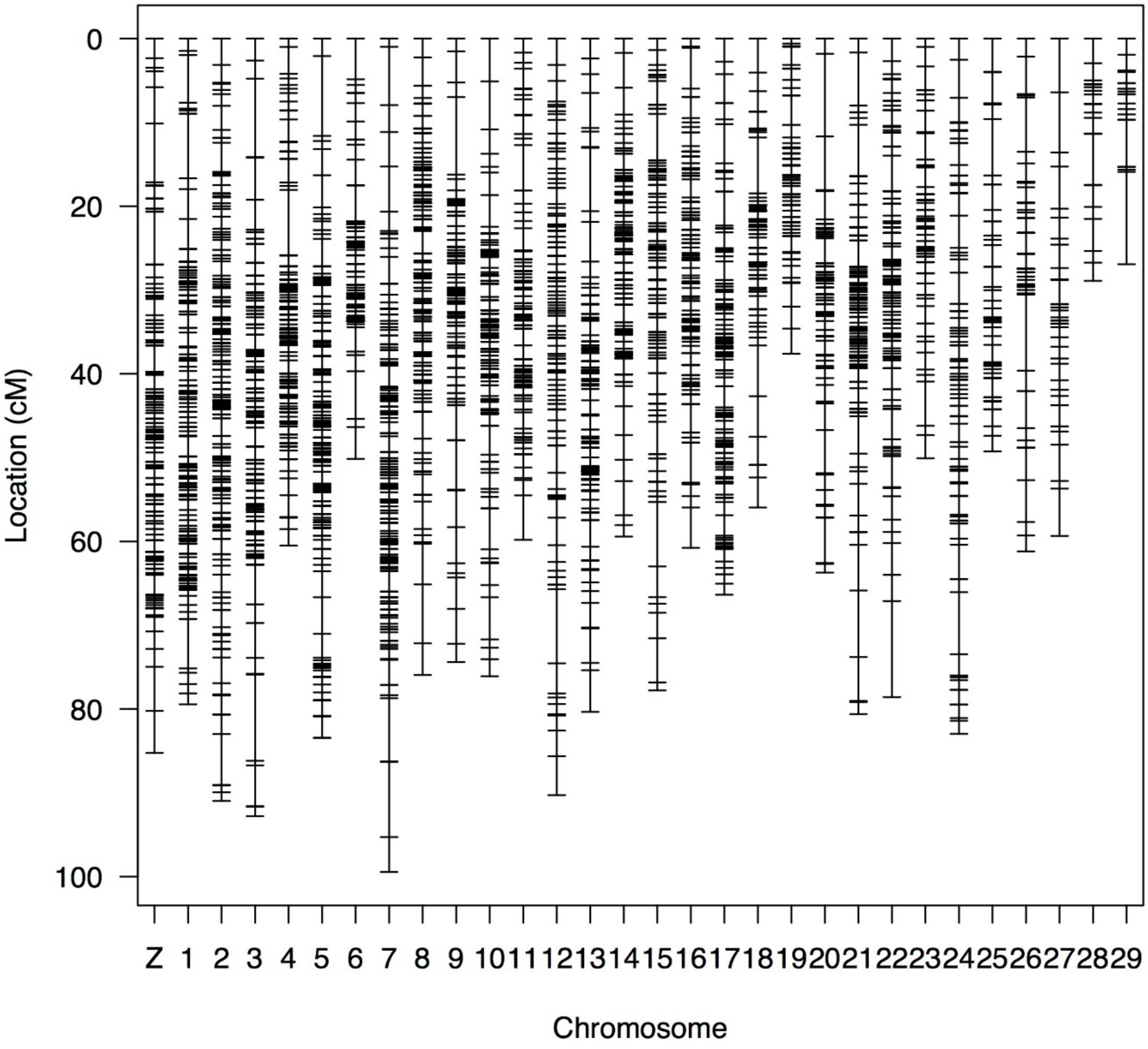


Figure 2.4. The correspondence in marker position between the two backcrosses. There are the positions of all markers in common between LG1 in KS BC and LG1 in FL BC.

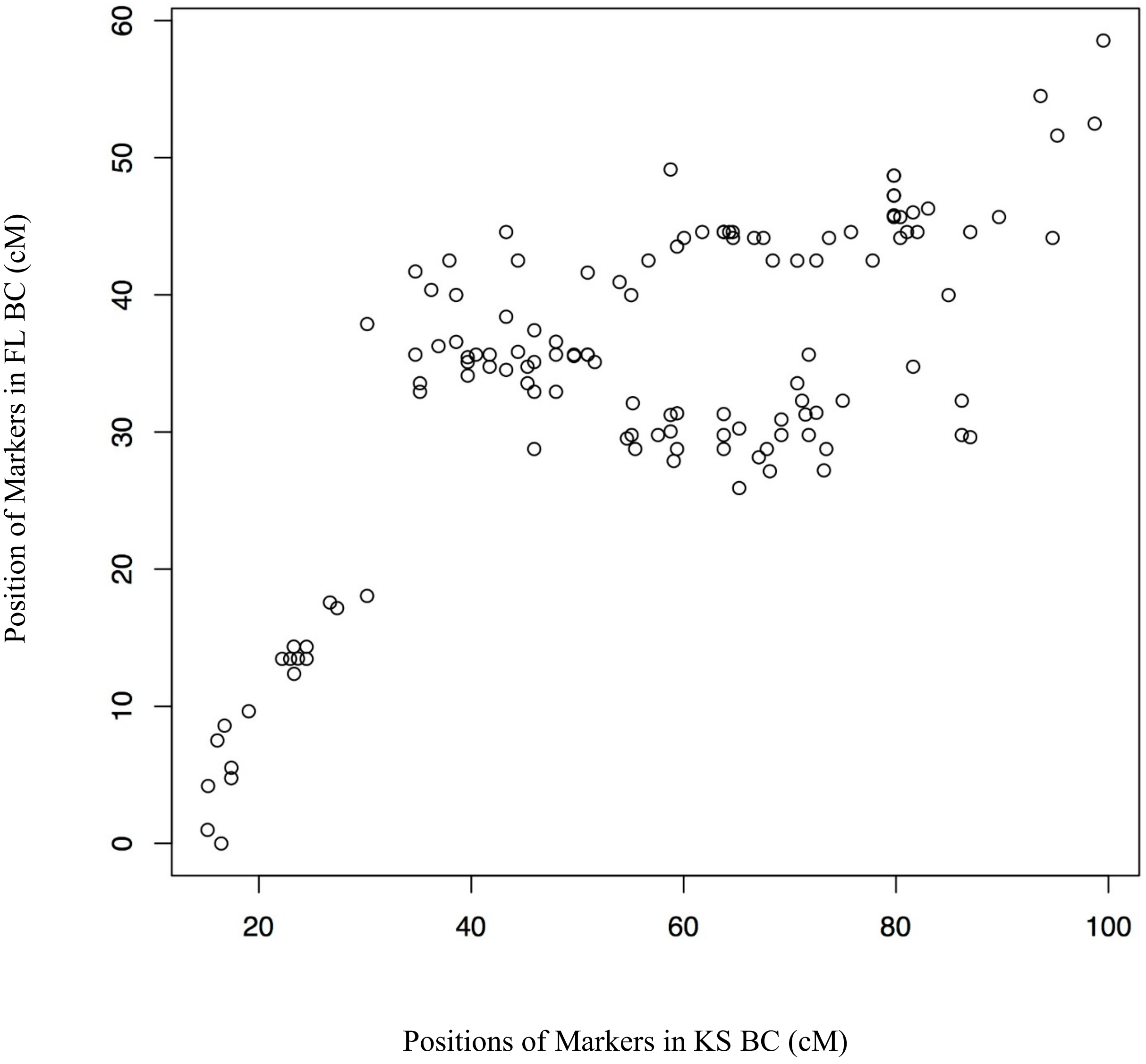


Figure 2.5. ALLMAPS output showing the inconsistency in marker order between the two backcrosses. The gray-striped tube in the middle represents the chromosome. The two linkage maps on either side of the chromosome represent linkage groups 18 from the KS BC and the FL BC unresolved maps.

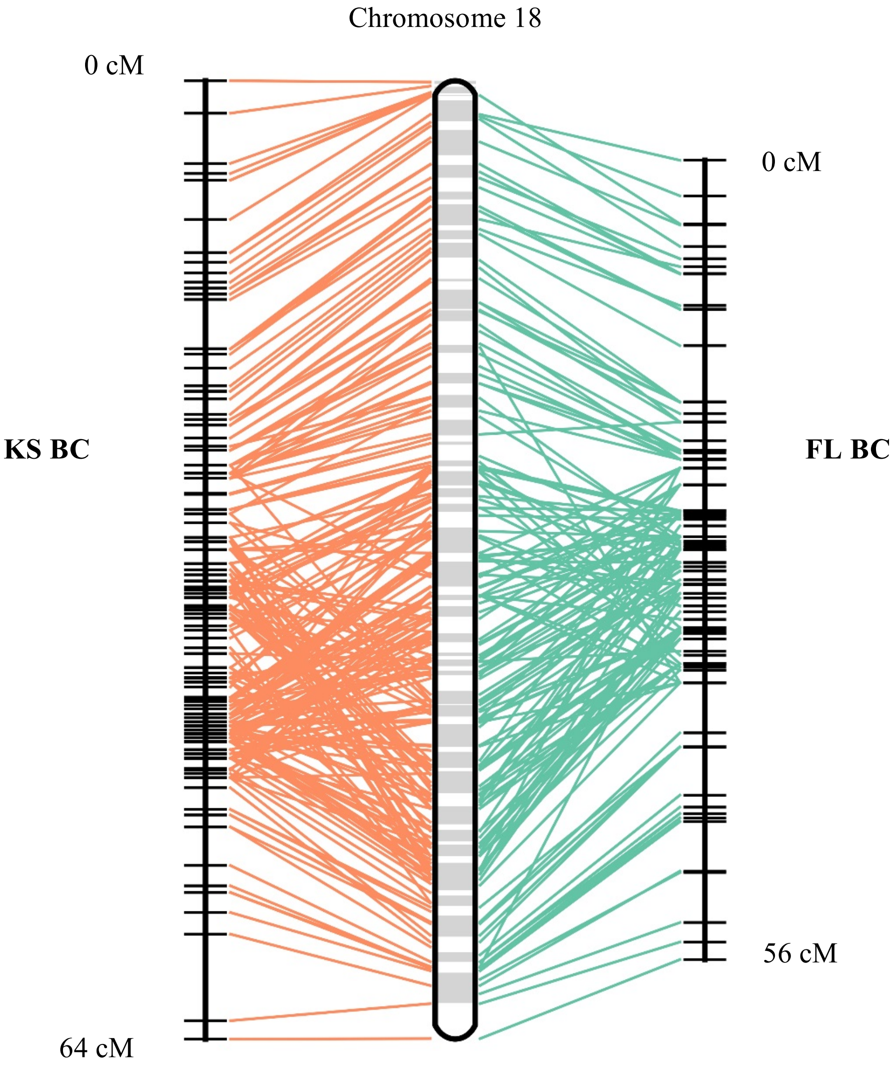


Figure 2.6. Genotype matrix for linkage group 1 in the KS backcross. Red represents a homozygous genotype, blue represents a heterozygous genotype, and white represents a missing value. Here, individuals are ordered along the y-axis in terms of the proportion of homozygous genotypes they have.

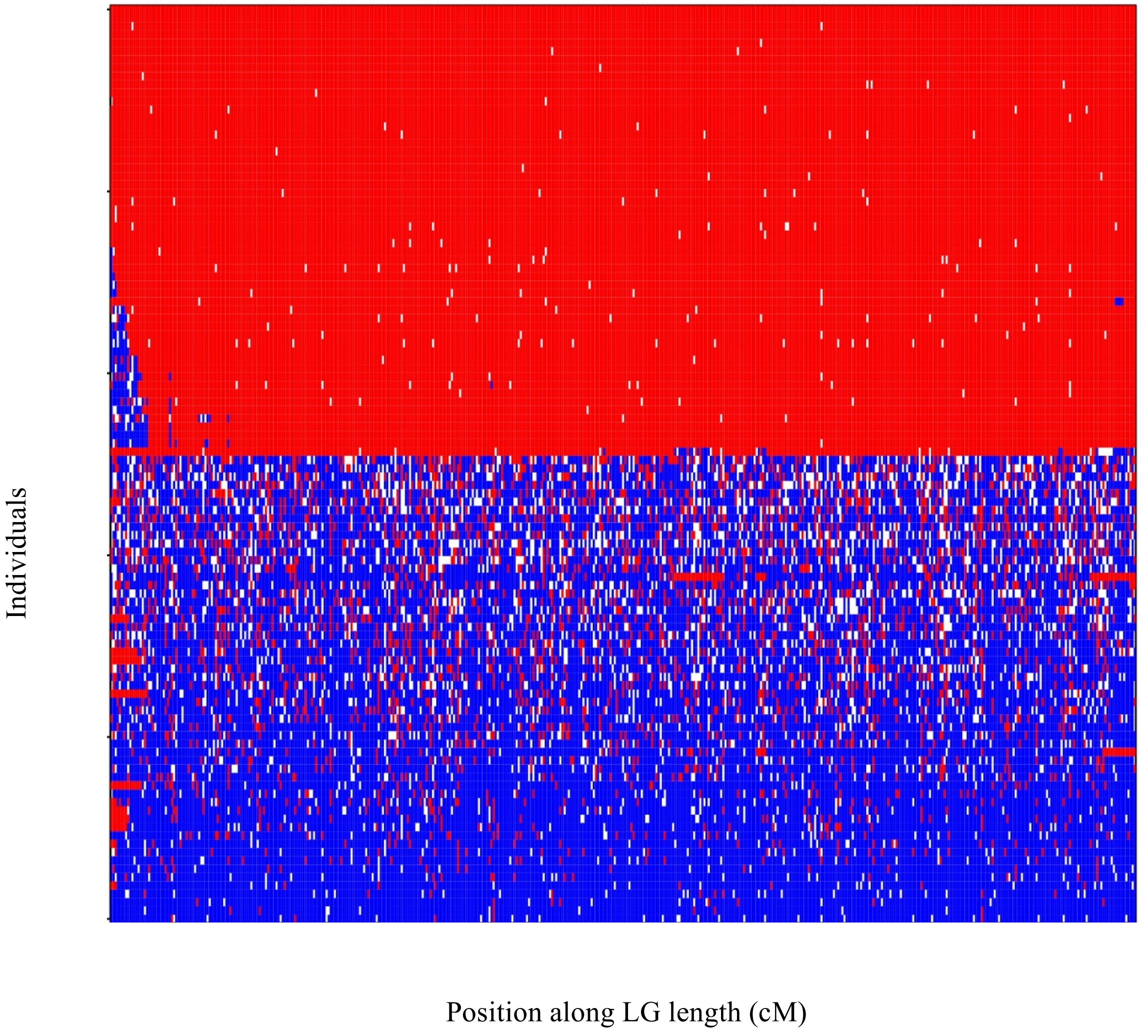
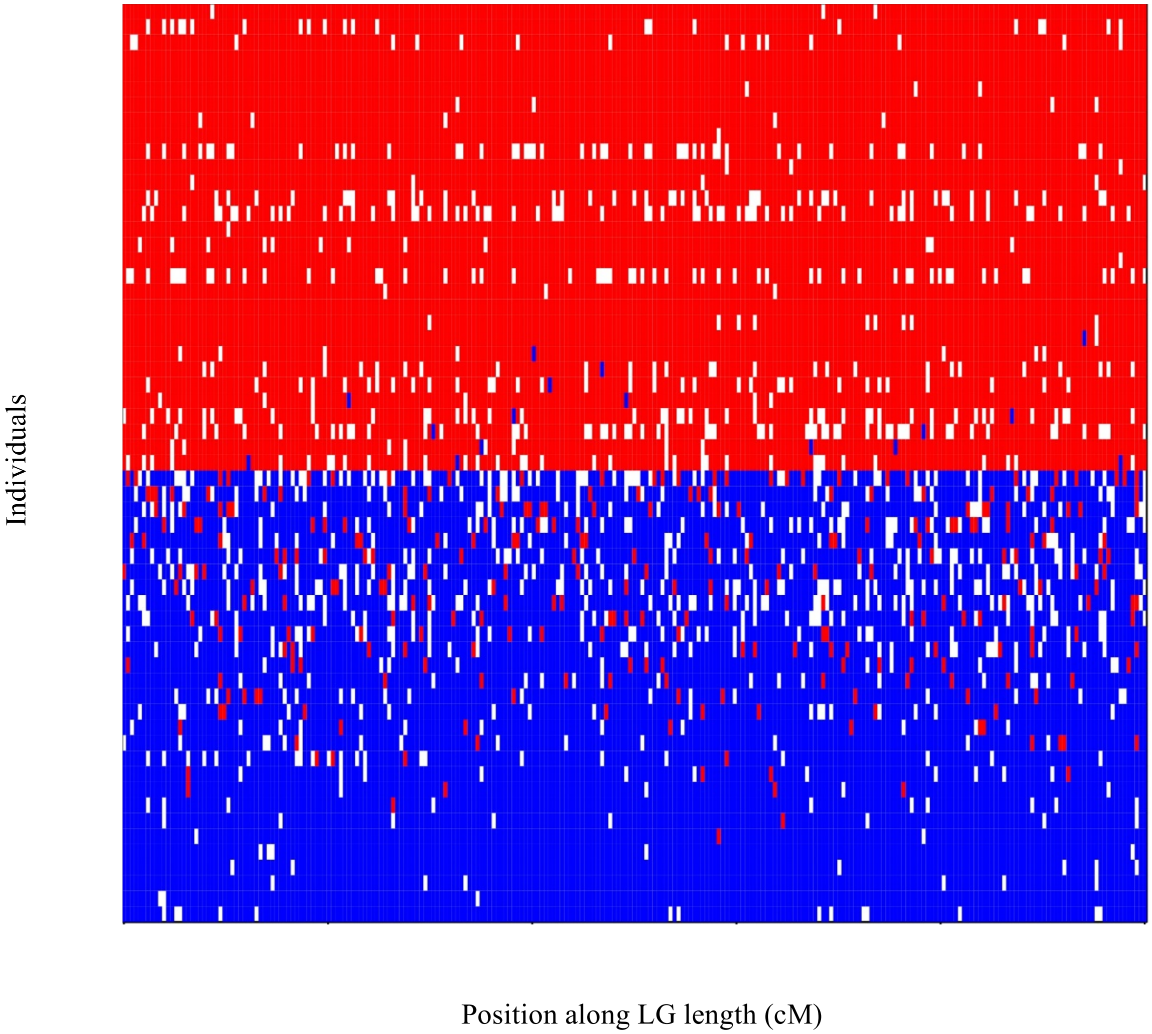


Figure 2.7. Genotype matrix for linkage group 1 in the KS segregant population. A subset of the individuals was used here for better resolution. Red represents a homozygous genotype, blue represents a heterozygous genotype, and white represents a missing value.



Tables

Table 2.1. Numbers of paired-reads obtained from a paired-end (PE) and a mate-pair (MP) library from two sequencing efforts.

	PE	*MP*
GSC (KU)	129,297,129	233,774,466
Cofactor	161,387,784	171,403,749
Total	290,684,913	405,178,215

Table 2.2. Duplication level in in reads obtained from the two sequencing efforts as well as the combined dataset.

	PE Forward	*PE Reverse*	*MP Reverse*	*MP Forward*
KU	2.15	1.95	4.90	2.73
Cofactor	8.94	8.75	7.12	5.14
combined	2.43	2.27	5.16	2.93

Table 2.3. Number of reads removed by each filter during quality control.

	PE	*MP*	*SE*	*Total Reads*
Raw	290,684,913 (100.0)	405,178,215 (100)	-	695,863,128 (100)
Sickle	251,070,361 (86.4)	326,649,635 (80.6)	69,701,500	647,421,496 (93.0)
PhiX	241,504,981 (83.1)	318,932,115 (78.7)	68,367,821	628,804,917 (90.4)
Quake	239,540,116 (82.4)	299,172,619 (73.8)	87,719,032	626,431,767 (90.0)

Table 2.4. Sample size (N), mean, and standard deviation (SD) for each phenotype for the two years the populations were phenotyped. The asterisk (*) denotes statistical difference between years.

		YEAR 1		**YEAR 2**	
PHENOTYPE	Population	N	Mean (SD)	N	Mean (SD)
DT	KS BC	238	42.5 (4.12) *	208	43.79 (4.91) *
WT	KS BC	237	14.37 (1.94) *	208	11.82 (1.63) *
PR	KS BC	238	75.15 (6.32)	208	75.04 (5.47)
PA	KS BC	238	76.1 (16.19) *	208	65.33 (12.9) *
AI	KS BC	238	76.1 (16.19)	208	65.33 (12.9)
DT	FL BC	236	41.64 (3.78) *	220	38.47 (2.94) *
WT	FL BC	236	15.14 (2.34)	228	15.54 (2.18)
PR	FL BC	236	73.66 (6.39)	228	74.27 (6.45)
PA	FL BC	236	70.05 (14.48) *	228	68.18 (13.15) *
AI	FL BC	236	658.53 (349.85)	228	725.35 (401.1)

Table 2.5. Sample size (N), mean, and standard deviation (SD) for each phenotype measured in the KS segregant population.

Phenotype	*N*	*Mean (SD)*
DT	198	41.82 (2.61)
WT	198	14.93 (2.34)
PR	198	75.66 (5.55)
PA	198	77.26 (16.48)
AI	198	684.4 (342.35)

Table 2.6. Descriptive statistics of genome assembly. Half of the assembly is contained in scaffolds of length N50 or larger.

Number of scaffolds	***74,159***
Total size of scaffolds	418,422,425
Longest scaffold	731,388
Number of scaffolds > 1K nt	12,067 (16.3%)
Number of scaffolds > 10K nt	6,202 (8.4%)
Number of scaffolds > 100K nt	1,117 (1.5%)
Mean scaffold size	5,642
Median scaffold size	185
N50 scaffold length	87,338
GC% content	32.4
scaffold %N	2.22

Table 2.7. Phenotype correlations in the KS backcross population. Asterisks indicate level of significance * p<0.05, ** p<0.01, *** p<0.001

	DT	*WT*	*PR*	*PA*
WT	-0.216 ***			
PR	-0.014	-0.081		
PA	-0.166 ***	0.522 ***	-0.051	
AI	0.057	0.048	-0.084	0.06

Table 2.8. Phenotype correlations in the FL backcross population. Asterisks indicate level of significance * p<0.05, ** p<0.01, *** p<0.001

	DT	*WT*	*PR*	*PA*
WT	-0.240 ***			
PR	0.017	-0.094 *		
PA	-0.112 *	0.287 ***	-0.113 *	
AI	-0.076	0.068	-0.023	0.025

Table 2.9. Phenotype correlations in the segregant population. Asterisks indicate level of significance * p<0.05, ** p<0.01, *** p<0.001

	DT	*WT*	*PR*	*PA*
WT	-0.386***			
PR	0.06	-0.168*		
PA	-0.11	0.118	-0.093	
AI	-0.049	0.044	-0.005	-0.039

Table 2.10. Number of markers (N), length, average and maximum spacing for the KS backcross genetic map. Corresponding chromosome numbers here and in Table 2.11 refer to the same genomic sequence.

Chromosome	*N*	*Length (cM)*	*Average Spacing*	*Maximum Spacing*
Z	301	108.4	0.4	7.6
1	437	128.2	0.3	7.1
2	307	112.2	0.4	8.3
3	491	106	0.2	6.3
4	452	118.6	0.3	7.4
5	343	110	0.3	11.8
6	307	95.7	0.3	7.8
7	296	115.7	0.4	8.3
8	339	112.9	0.3	11.1
9	377	128.4	0.3	8.4
10	277	95.5	0.3	7.7
11	365	108	0.3	6.9
12	199	81	0.4	10.2
13	339	102.6	0.3	11.3
14	293	103.7	0.4	6
15	272	100	0.4	8.8
16	298	105.6	0.4	9.6
17	294	108.3	0.4	10.3
18	324	64.2	0.2	5.4
19	319	109.6	0.3	6.9
20	222	84.1	0.4	5
21	227	100.3	0.4	6.8
22	293	83	0.3	10.9
23	135	69	0.5	7.8
24	147	79.3	0.5	5.3
25	145	69	0.5	4.8
26	130	64.9	0.5	8
27	67	68.2	1	9.3
28	50	36.2	0.7	5.1
29	45	23.2	0.5	5.8
overall	8091	2791.6	0.3	11.8

Table 2.11. Number of markers (N), length, average and maximum spacing for the FL backcross genetic map. Corresponding chromosome numbers here and in Table 2.10 refer to the same genomic sequence.

Chromosome	*N*	*Length (cM)*	*Average Spacing*	*Maximum Spacing*
Z	352	85.2	0.2	7
1	332	79.4	0.2	7.7
2	361	91	0.3	6.1
3	277	92.8	0.3	10.3
4	206	60.5	0.3	7.8
5	287	83.5	0.3	9.5
6	172	50.2	0.3	5.7
7	313	99.5	0.3	9
8	325	75.9	0.2	7
9	186	74.4	0.4	9.3
10	224	76.1	0.3	5.7
11	201	59.8	0.3	5.4
12	206	90.3	0.4	8.9
13	177	80.3	0.5	7.6
14	245	59.4	0.2	4.2
15	211	77.8	0.4	7.7
16	179	60.8	0.3	4.9
17	226	66.4	0.3	4.7
18	138	56	0.4	6.7
19	130	37.6	0.3	3.5
20	124	63.7	0.5	9.9
21	196	80.6	0.4	7.9
22	202	78.6	0.4	11.5
23	108	50.1	0.5	5.3
24	115	83	0.7	7.4
25	67	49.3	0.7	6.7
26	60	61.2	1	9.1
27	53	59.4	1.1	7.2
28	24	28.9	1.3	6.1
29	24	26.9	1.2	11
overall	5721	2038.5	0.4	11.5

Table 2.12. QTL affecting the traits of interest in the 3 populations, with LOD score, the significance level at which they were detected, the % phenotypic variance they explain, and the effect of substituting a FL-derived allele for a KS-derived allele.

Population	*Phenotype*	*LG*	*LOD*	*Significance level (α)*	*% Variance Explained*	*Effect*	*Power*
KS BC	DT	7	8.71	0.05	6.15	-2.3072	0.93
KS BC	WT	7	1.96	0.2	0.5	0.3696	0.01
KS BC	WT	12	2.33	0.1	0.25	0.2094	0.00
KS BC	WT	27	1.56	0.05	2.61	0.7911	0.43
KS BC	PR	11	2.59	0.05	2.11	-1.9514	0.31
KS BC	PR	13	2.27	0.05	1.5	-1.6839	0.16
KS BC	AI	20	2.09	0.2	1.39	-97.89	0.11
KS BC	AI	22	1.69	0.1	1.69	105.95	0.16
FL BC	DT	14	1.84	0.1	0.81	-0.7813	0.04
FL BC	WT	8	2.14	0.1	1.27	-0.5771	0.10
FL BC	PR	15	1.86	0.1	0.95	1.4367	0.06
FL BC	PA	4	1.74	0.2	1.38	3.656	0.12
FL BC	PA	5	1.67	0.2	1.21	3.3135	0.08
KS SG	DT	12	1.97	0.2	2.89	-0.9879	0.09
KS SG	DT	19	1.68	0.1	3.36	1.0684	0.14
KS SG	WT	28	2.27	0.05	4.92	1.113	0.28
KS SG	PA	11	1.77	0.2	3.07	6.558	0.12

Chapter 3

Draft Genome Assemblies in Non-Model Organisms:

Do they facilitate or hinder the construction of genomic loci from RAD-seq data?

ABSTRACT

The low cost and high output of Next Generation Sequencing (NGS) technologies have led to a great number of published draft genomes of non-model organisms. Additionally, thanks to refinements in laboratory techniques, the field of evolutionary genetics is becoming more accessible to non-model study systems. RAD-seq, a genomic library preparation technique that uses a restriction enzyme to reduce genome complexity, has become one of the most commonly used sequencing methods to study natural populations and non-model organisms. By digesting genomic DNA with a restriction enzyme of choice, and ligating a sample-specific identifiable barcode, RAD-seq has enabled scientists to interrogate portions of the genomes in hundreds of individuals. The sequenced data ("tags") can be analyzed by either using a reference genome to which reads are aligned, or by relying on software to assemble loci *de novo*. Studies have shown that, when available, a quality reference genome is a source of information for the proper assembly of loci. For species that do not have a reference genome, many approaches have been suggested for experimental design, as well as for quality control of the *de novo*-assembled loci. In this chapter, I ask whether a fragmented genome sequence aids in the assembly of loci for the study of variation in natural populations. I used two mapping populations from two non-model species with scaffolded genomes. The RAD tags from each population were analyzed using two pipelines that only differed in whether loci were assembled through alignment of tags to a reference or were assembled *de novo*. I leveraged the fact that these were mapping populations, and was able to determine what proportion of the markers were consistent with expectations from Mendelian genetics. I found that even a fragmented genome assembly can aid in re-construction of loci, and the proportion of markers that were inconsistent with the expected genotypic frequencies was disconcertingly high. Finally, I discuss how the two main kinds of assembly errors affect population genetic summary statistics generated from RAD tags.

INTRODUCTION

High-throughput sequencing has revolutionized the field of evolutionary genetics. In the past, the literature in the field was predominantly focused on model organisms like *Drosophila* (K. A. Hughes 1995), *Escherichia coli* (Lenski, Rose, and Simpson 1991), maize (Wang et al. 1999), mouse (Alibert et al. 1994), *Xenopus* (M. K. Hughes and Hughes 1993), and *Arabidopsis* (Valvekens, Van Montagu, and Van Lijsebettens 1988) because of the availability of genetic resources for these organisms. The development of next-generation-sequencing (NGS) technologies has made it possible for scientists to explore questions in evolutionary genetics of non-model organism and also in related fields like population genetics and phylogenetics.

With the speed of sequencing increasing and the cost decreasing, it has also become easier for groups working on non-model species to sequence the genome of their study organisms. While the quality of published genomes varies from "standard draft" to "finished" genomes (Chain et al. 2009), the number of genome assemblies being published is increasing rapidly. Currently, the National Center for Biotechnology Information (NCBI - https://www.ncbi.nlm.nih.gov) lists assembly data for over 4,300 organisms. Knowing the genome sequence of an organism is essential to determining base composition, number and type of genes and transposable elements, and other characteristics important to questions in evolutionary genetics, population genetics, and phylogenetics.

Population genomics, the comparison of the genetic make-up of individuals in a population using large numbers of loci[1] scattered across the genome, makes inferences about evolutionary processes that are acting on the population. Unlike earlier technologies, NGS allows for the rapid sequencing of thousands,

[1] Throughout Chapter 3, I use the term "locus" to mean a 100-bp stretch of DNA in the genome, regardless of whether the exact location of the fragment in the genome is known.

if not hundreds of thousands, of loci in an individual. Advances in lab techniques have further enabled scientists to include hundreds of individuals in their studies. One such technique is Restriction site Associated DNA Sequencing (RAD-seq; Baird et al. 2008; Hohenlohe et al. 2010; Emerson et al. 2010). RAD-seq is a 'reduced representation' procedure because the library contains only a subset of the genome, specifically sequences that flank restriction enzyme cut sites.

The first step in RAD-seq is to digest the genomic DNA using a restriction enzyme. The resulting fragments are ligated to Illumina adaptors that are modified to allow binding and amplification only of fragments containing the restriction site. These adaptors also contain barcodes, short identifying sequences unique to each individual, that are used to bioinformatically separate the sequenced reads. After tagged fragments from multiple individuals are pooled, the DNA is randomly sheared and size selected. The resulting library is sequenced on the Illumina platform, and the sequenced reads contain the sample identifier, the restriction site, and a short region downstream of the restriction site.

This protocol has been modified by research groups to suit their particular needs while reducing cost and effort. As a term, RAD-seq was used to describe the method Baird et al. (2008) developed. Currently, the term is used to describe a family of methods that digest DNA with restriction enzymes and sequence the resulting reads. These methods include modifications to the number and type of enzymes used, the size selection step, barcoding, and/or adaptor ligation (K. R. Andrews et al. 2016). For example, genotyping by sequencing (GBS; Elshire et al. 2011) uses a common-cutter enzyme, and short fragments are size selected during PCR amplification. Multiplexed shotgun genotyping (MSG; Andolfatto et al. 2011) uses a common-cutter to digest the DNA, and after barcoded fragments are pooled, they are size selected. As a result, the original RAD-seq method produces fragments containing a single restriction enzyme cut site whereas GBS and MSG produce fragments flanked by two cut sites. This difference is important only when opting to do paired end sequencing. In MSG and GBS, the reverse reads start at the second cut site and extend the exact same number of nucleotides into the sequence, covering the same genomic location.

In contrast, in RAD-seq, the second reads start at the randomly cut site which is a variable physical distance from the restriction enzyme cute site. As such, the reverse reads don't align at the same genomic location.

Regardless of the specific protocol, the bioinformatics pipelines used to analyze the data have the same general steps. First, reads are de-multiplexed based on the in-line barcodes, and the barcodes are removed. Quality filtering can be done to exclude fragments lacking the restriction enzyme cut site(s), and trim the 3' end if the base quality is low. If a reference genome exists, the reads can be aligned to the reference genome and loci will be determined based on the alignment. If a reference genome is not available, loci can be constructed *de novo* by assembling reads with a small number of pairwise differences into groups. These groups represent genomic loci, and the differences represent segregating alleles. Statistical models are used to distinguish between true polymorphism and sequencing error. Once loci are discovered, single nucleotide polymorphisms are detected, and genotypes are called using maximum likelihood (Hohenlohe et al. 2010) or Bayesian methods (R. Nielsen et al. 2012). Finally, depending on the aims of the study and the downstream analytical methods, filtering the genotype data set might be necessary to remove markers and/or individuals with a low number of genotypes.

RAD-seq methods have been applied in both experimental crosses and in natural populations. In experimental crosses, one can leverage the large number of SNPs to generate ultra-dense genetic maps in order to answer quantitative genetics questions surrounding interesting traits, or to help scaffold draft genome assemblies. In natural populations, RAD-seq markers have been used to estimate genomic diversity, inbreeding, population structure (Barley et al. 2015), effective population size (Pujolar et al. 2013), and to detect signatures of selection (Hohenlohe et al. 2010), and introgression (Eaton and Ree 2013).

The sources of error and bias for RAD-seq specifically, and NGS generally, were recently reviewed by Mastretta-Yanes et al. (see Table 1; 2014). A few are of particular interest here. Allele dropout is the failure of a restriction enzyme to cut the DNA due to a mutation in the recognition site. When an allele contains a cut site with a polymorphism, and is consequently not sequenced, that allele is called a null allele in the context of RAD tags (Gautier et al. 2013; Arnold et al. 2013). Another important source of error and bias is the PCR step during library preparation. PCR duplicates, a disproportionate number of copies of a DNA fragment in the PCR product of the sample, can bias allele frequencies at a locus, result in genotyping error, or introduce erroneous alleles due to PCR errors (Pompanon et al. 2005). Furthermore, PCR has the tendency to preferentially amplify fragments based on their GC content (Benjamini and Speed 2012). This bias leads to high variance in read depth across loci.

From a bioinformatics standpoint, there are a few additional major sources of error and bias. Regions containing sequences of high similarity, like paralogous or repetitive regions, are often grouped into a single locus. This phenomenon is called over-merging, and can lead to allele frequency estimates that are completely spurious. Many analytical pipelines, including Stacks (Catchen et al. 2011; Catchen et al. 2013) and PyRAD (Eaton 2014), cannot properly deal with indels, and fail to construct loci containing this kind of polymorphism. As a result, allele frequencies of indel-containing loci are not estimated properly. Finally, the biggest source of bias is mapping reads to a reference genome. The extent of this problem depends on how divergent the study population is from the reference genome. If a subset of individuals in a sample is more closely related to the individual(s) sequenced for the genome assembly, the alleles from those individuals will align to the reference genome better than the remaining sample. This could lead to biased estimates of allele frequency.

There are challenges associated specifically with using RAD-seq in natural populations. When we use experimental crosses, we can leverage the *a priori* knowledge of the genetic composition of a population to sanity check our markers. When the genotype frequencies of the founding populations are known, principles of Mendelian genetics help us form expectations about the genotype and allele frequencies in offspring generations. Therefore, markers from experimental crosses should be consistent with these expectations. In natural populations, on the other hand, no such expectations can be formed because we don't know the genetic make-up of the founding populations. Thus, in natural populations, we largely rely on analytical tools to correctly construct genetic markers.

Parameters used in *de novo* assembly of loci has proven to influence downstream analyses and inferences (Ilut, Nydam, and Hare 2014). The optimization of these parameters depends on the amount of polymorphism in the genome under investigation, the read depth, and the frequency of sequencing error (Catchen et al. 2013). Catchen et al. (2013) optimized the parameters for the threesprine stickleback and were able to detect erroneous loci using the high quality reference genome (Jones et al. 2012). Because such genomes are not available for many non-model taxa that cannot be brought into the lab for controlled crosses, strategies to optimize the *de novo* assembly of loci have been proposed (Mastretta-Yanes et al. 2014). However, many such species have draft genome assemblies that contain the majority of the genetic information on scaffolds that are not grouped into chromosomes. To my knowledge, there are no studies focusing on whether a poorly assembled or highly-fragmented genome is valuable in discovery of genomic loci. Given the increased emergence of draft assemblies and the importance of correctly identifying segregating alleles at a locus, it is important to determine if a lower-quality genome aids or hinders the re-construction of loci from RAD-seq data in natural populations.

In this chapter, I aim to assess the quality of markers constructed using alignment of RAD tags to the draft genomes of two non-model organisms, and compare it to the quality of markers constructed in the absence of reference sequences. To accomplish the goal, I used data from two mapping populations from each species, and discovered markers using both de novo and reference-based assembly of loci. Additionally, I explore the consequences of different assembly-related errors on population genomic statistics commonly estimated in applications of RAD-seq in studies of natural populations.

METHODS

Study Organisms

The genus *Tolpis* has a dozen or so described species, most of which are endemic to the Macaronesian islands. As an insular species on an oceanic island, *Tolpis* is the subject of studies aiming to understand the patterns and processes of evolution (Crawford et al. 2006; Crawford et al. 2008; Crawford et al. 2010; Soto-Trejo et al. 2013; Crawford et al. 2015; Mort et al. 2015). More recently, the breeding system of the genus has been of particular interest because of a recent transition from self-incompatibility (most *Tolpis* species) to self-compatibility in *Tolpis coronopifolia* (Crawford et al. 2008). As part of a study on this transition (see Chapter 1- cite), I assembled a draft genome for *T. coronopifolia* using two Illumina short-read pair-end libraries, one with a 300-bp and one with 1000-bp insert size. The assembly contains a total of 1.2Gb across 3,977,270 scaffolds, with a N50 of 4,124 bp. See Table 1.2 for more details.

The other organism is the lesser wax moth. *Achroia grisella* (Lepidoptera: Pyralidae) is an obligate symbiont of the western honey bee, *Apis melifera* (Milum 1935). The species exhibits an interesting behavior during mate pairing (Greenfield and Coffelt 1983). In most species of moth, females attract males but in *A. grisella,* males attract females by emitting ultrasound signals through asynchronous forewing movement (Spangler and Takessian 1986). Studies have shown that certain signal characters make these calls attractive to females (Jang and Greenfield 1996; Jang and Greenfield 1998b), which leads to interesting questions surrounding sexual selection in the species. Recently, I assembled a draft genome for the species as a part of a study on the genetic architecture of male song characters. I used an Illumina short-read paired-end library and a mate-pair library with insert sizes of of 280-500 bp and 3-5 kb, respectively. The assembled genome is 418Mb across 74,159 scaffolds, with a N50 of 87,338 bp. See Table 1.2 for more details.

Mapping populations and Genomic Library Preparation

Two *Tolpis* mapping populations were established as described in Crawford et al. (2008) and Crawford et al. (Crawford, Mort, and Archibald 2013). In short, *T. coronopifolia* (SC) pollen was used to fertilize *T. santosii* (SI; Crawford, Mort, and Archibald 2013) individuals, and an individual from the resulting F1 was used to give rise to two F2 populations (namely, F2.25 and F2.27) via self-pollination. Genomic DNA was extracted from individuals from both populations, N=101 and N=84 from F2.25 and F2.27, respectively. MSG libraries were prepared according to the protocol described in Andolfatto et al. (2011) with one modification: instead of *Mse*I restriction enzyme, *Ase*I was used because it cuts less frequently resulting in higher coverage per genomic location. From the sequencing effort, I obtained approximately 220.5 and 140 million reads for the F2.25 and F2.27 populations, respectively. Unlike the original RAD-seq method (Baird et al. 2008), MSG results in reads of equal length that have the same start and end location in the genome.

Two *Achroia* mapping populations were established through backcrossing to the two inbred lines, Kansas and Florida, that are described in Gleason et al. (2016) Due to achiasmatic meiosis in females, F1 males were crossed to females from the Kansas and Florida inbred lines to generate the Kansas Backcross (hereafter “KS BC”) and the Florida Backcross (hereafter “FL BC”), respectively. Genomic DNA was extracted from individuals from both backcrosses, 447 and 465 from KS BC and FL BC, respectively. Modified MSG libraries were constructed to allow for sequencing of all individuals simultaneously. The library was run on a single lane of Illumina, and resulted in approximately 341 and 255 million reads for the KS BC and FL BC populations, respectively.

Marker Discovery

To discover markers, I used Stacks (Catchen et al. 2011; Catchen et al. 2013), an open source software for marker discovery from GBS data. Briefly, Stacks is comprised of 4 programs: (1) pstacks/ustacks, which puts together stacks of identical reads and then assembles loci from the stacks for every individual in a population using either raw GBS reads, or an alignment file (SAM/BAM), (2) cstacks, which makes a catalog of loci that are present in the population, (3) sstacks, which searches for matches between the loci generated for each individual and the catalog, and (4) populations, which calls genotypes in the population and calculates population genetics statistics. While there are other software packages that do comparable bioinformatic work like PyRAD (Eaton 2014) and TASSEL (Bradbury et al. 2007), I chose Stacks because it is well-maintained, widely used for population genomic studies, and user friendly.

To determine if a draft genome assembly affects the quality of markers generated, I used the MSG reads from the 2 Tolpis and the 2 *Achroia* populations to discover markers through two pipelines that differed only in whether or not the genome assembly was used. In the first pipeline, the MSG reads were aligned to the assembled genome, and the alignment file was used as the input by *pstacks* for the construction of loci (hereafter, the "alignment pipeline"), and in the second pipeline, the reads were used by *ustacks* to assemble loci de novo (hereafter, the "*de novo* pipeline"). Once loci were generated by the *pstacks* and *ustacks*, the remaining Stacks components were used in an identical manner between the two pipelines, and the complete pipelines were replicated in both organisms.

For the alignment pipeline in the F2.25 population, I masked repeats in the *Tolpis* genome assembly using RepeatModeler (Smit and Hubley 2010). Then, I aligned the reads from each individual to the masked genome assembly using BWA aligner (H. Li and Durbin 2009) with default settings, and generated a sorted BAM file using SAMtools (H. Li et al. 2009; H. Li 2011). The BAM file served as the input to *pstacks*, and the only parameter I specified was the minimum depth of coverage for a stack to be considered a locus ($m = 3$). Then, I used all *Tolpis* individuals to create a catalog of loci in the population in *cstacks* with defaults parameters. Loci from all individuals were then matched against the catalog with *sstacks*, and, finally, *populations* was used to call genotypes for all individuals at all possible loci in the F2.25 population and export the information to a VCF file (Danecek et al. 2011). This process was repeated for the F2.27 population. Similarly, the *Achroia* populations were used in the same manner but with the alignment being to the *Achroia* draft genome.

For the *de novo* pipeline in the F2.25, I used individuals' reads in *ustacks* to generate loci in absence of the *Tolpis* draft assembly. In ustacks, I set the minimum stack depth (*m*) to 3 to be comparable to the alignment pipeline. Additionally, within *ustacks*, I used the Removal algorithm *(-r*) to ensure that highly repetitive stacks are excluded from the analysis, the Deleveraging algorithm *(-d*) to resolve loci that contain too many nucleotide differences, and excluded secondary reads from haplotype calling (-*H*). Finally, I also set the maximum number of stacks per locus to 2 (*max_locus_stacks*=2) because I expected that, in the mapping populations, there would be only two alleles segregating per locus. The remaining Stacks components (*cstacks*, *sstacks*, *populations*) were used the same way as in the alignment pipeline. I repeated the marker discovery process in the F2.27 population and the two *Achroia* populations.

The two pipelines employed here are representative of the protocol one would follow when studying natural populations, when a reference genome is available and in the absence of one. The only parameter value that cannot be set when employing this pipeline in natural population is the maximum number of segregating alleles per locus, a value that varies from locus to locus, and population to population.

Evaluation

The VCF files generated by populations included genotype calls for every individual at all loci in the catalog. To replicate a pipeline one is likely to use for natural populations, I excluded markers with fewer than 16 genotyped individuals. At that stage, I used expectations from Mendelian genetics to determine if a marker is "good" or "bad." For the *Tolpis* populations, I tested the genotype frequencies at each remaining marker for deviations from the 1:2:1 ratio expected from an intercross of F_1 heterozygous individuals. I calculated a *G* statistic and compared it to the χ^2 distribution with 2 degrees of freedom. I classified a marker as "good" if the genotype frequencies were consistent with the 1:2:1. Conversely, a marker, where the observed genotype frequencies deviated significantly from the expected, was considered "bad."

In the *Achroia* populations, markers with fewer than 16 genotyped individuals were excluded. Since only two genotypes are possible in these populations (heterozygous and homozygous for the recurrent parent), I also excluded markers containing more than 5% "impossible" homozygous genotypes. In a backcross population, Mendelian genetics predicts two genotypes: a heterozygote, and a homozygous for the recurrent parent, and the absence of the homozygous genotype for the second parent. For example, for any given marker in the KS BC, a homozygous genotype for the FL allele was considered "impossible." For the remaining markers, I calculated a *G* statistic to test for deviations from the 1:1 ratio expected in a backcross; the statistic was compared to the χ^2 distribution with 1 degree of freedom. A marker, where observed genotype frequencies were significantly different from the expected were called "bad." Markers with genotype frequencies consistent with the expected were classified as "good."

Finally, because coverage has been found to have a profound effect on the genotyping error rate (Hohenlohe, Catchen, and Cresko 2012; Davey, Cezard, and Utrilla 2013; Catchen et al. 2013), I also implemented a simple read depth filter for genotype calling (hereafter, "RDF"). The filter requires at least 6 reads for a genotype to be called. If a genotype is called based on less than 6 reads, it gets converted to a missing genotype (NA). The evaluation of markers as "good" or "bad" was performed twice per population. The first time, RDF was off (0) and no genotypes were discarded due to low coverage. The second time the filter was on (1), and genotypes were discarded before the marker evaluation process. The evaluation script, including the RDF, were coded in Python 2.7.

RESULTS

Overall, using the alignment of RAD-seq reads to a reference sequence as the starting point for construction of loci increased the proportion of good loci in the final data sets for all populations (Tables Table 3.1a,b, Table 3.2a,b). This result holds true regardless of whether or not RDF was utilized. However, the proportion of "good" loci is surprisingly low. In the F2.25, the percent of reads that are "good" is 5.7% in the de novo pipeline and 17.1% in the alignment pipeline (Tables Table 3.1a). These number increase to 9.3% and 24.1%, respectively, when RDF was on Tables Table 3.1b. That is, at most, only about a 1/4th of the markers with least 16 genotyped individuals are consistent with the expected 1:2:1 ratio. The observed pattern in F2.27 is similar (Table 3.1). In Achroia, the proportion of "good" loci in the data is close to 0 when the RDF is off (Table 3.2a). When RDF is used, the proportion jumps up to 23.0% and 28.1% for the de novo and alignment pipelines, respectively, in the KS BC (Table 3.2b). A similar pattern is seen in the FL BC (Table 3.2)

The two pipelines generated different numbers of loci per population (see Tables 3.1-2). In the KS BC population, the alignment pipeline constructed almost twice as many markers as the *de novo* pipeline, regardless of RDF. Similar for the FL BC. The opposite pattern is observed in the *Tolpis* populations where the *de novo* pipeline discovered 3-5 times the number of loci compared to the alignment pipeline. Here too, RDF didn't seem to change that pattern. It is worth mentioning that with RDF on, the number of markers decreased in the 4 populations compared to when RDF was off.

In terms of the proportion of "bad" markers, the alignment pipeline resulted in a smaller fraction of bad markers in both F2.25 and F2.27. In both populations of *Achroia*, the proportion of "bad" markers in the alignment pipeline was smaller compared to the *de novo* pipeline but only when the RDF was off. When RDF was on, the alignment pipeline produced higher proportion of bad markers compared to the *de novo* pipeline. Finally, the proportion of markers with too many "impossible" genotypes in the KS BC was

35.5% and 55.2% for the alignment and *de novo* pipelines, respectively. Those numbers shrink to 7.8% and 10.4%, respectively, when the read depth filter is utilized.

Discussion

Effect of mapping RAD tags to a reference genome

Does using a draft genome assembly aid in the re-construction of genetic loci when studying natural populations of non-model organisms? To answer that question, I used MSG data from 4 mapping populations. This is essentially an experiment to determine if marker identification pipelines can identify loci that survive the test of Mendelian segregation. Segregation is directly evident in mapping populations but difficult to evaluate in natural samples (unless one samples both parents and offspring). I treat the mapping populations as if they were samples from natural populations, and then examined the quality of genetic markers produced by two bioinformatic pipelines that are commonly used in population genomics studies. In one of the pipelines, I discovered markers based on the alignment of reads to a draft assembly, and in the other pipeline, the markers were created *de novo*. Overall, the use of the genome assembly consistently produced a larger proportion of “good” markers. However, that proportion was unexpectedly low, sometimes less than 1%.

In all populations, all else being equal, the loci generated by the alignment contained a higher percentage of “good” reads. This makes sense because aligning sequence reads to a reference sequence allows alleles from different loci but with high sequence similarity to be assigned correctly. In that kind of situation, the degree of fragmentation of the assembly should not make a difference unless the loci are completely or partially missing from the assembly. In this study, both genome assemblies are fairly fragmented. However, the total assembly size for both is approximately the expected size of the genome (see Results in Chapter 1 and Chapter 2).

The highest proportion of "good" reads was in the alignment pipeline of the KS BC, and it was still only about 28%. That means that, at best, only 1 in 3 markers are behaving as one would expect given that the individuals are from a backcross. This is particularly troubling for two reasons. First, when sampling natural populations, one has a limited number of sanity checks to use in determining if a marker is good or bad. So, using these pipelines with default parameters is almost certainly leading to a data set with a high percentage of markers that are not consistent with Mendelian expectations. The consequences of using erroneous markers are discussed later. Second, here, I used a parameter in the *de novo* pipeline to restrict the number of alleles per locus (*max_locus_stacks*) to 2. While that is true for the populations that I was dealing with, natural populations don't have that kind of prior knowledge. So, even with a favorable parameter, the pipeline was unable to properly construct loci.

Previous studies have suggested that read depth is a factor in genotyping error rate (Hohenlohe, Catchen, and Cresko 2012; Davey, Cezard, and Utrilla 2013; Catchen et al. 2013). Here, the *Achroia* populations provide an excellent illustration that low read depth can lead to spurious genotype calls. When calling genotypes without requiring a minimum number of reads, the proportion of "good" loci in the KS BC is miniscule (0.5-1%, Table 3.2). When a read depth filter (RDF) of 6 is applied to the same data, the proportion increases to 17.8-21.9%. Removing genotypes with low read coverage decreases the number of genotyped individuals per marker. However, those genotypes have an inherently higher error rate. For a locus with two segregating alleles, the probability of sampling only one of them in 5 trials is a 0.03125, a heterozygote will appear as a homozygote about 3% of the time when the coverage is 5 reads have been sequenced. That probability increases as the number of reads decreases. With only two reads, the probability of sampling only one of the two alleles is 0.25. Therefore, a read depth filter for genotype calling seems a necessary step in bioinformatics pipelines that aim to accurately estimate allele or genotype frequencies. Depending on the read coverage per locus, the minimum read depth required for a genotype call can be less than or greater than 6 depending on how many loci one can afford to discard based on the goal of the study.

Another interesting observation is the total number of loci assembled through the alignment and *de novo* pipelines. In an ideal world, where one has access to a high-quality reference genome, it would be possible to optimize the input parameters for the construction of loci in both pipelines. Then, the number of loci produced in the two pipelines should be approximately the same. Here, we see large differences between the two pipelines, and the pattern in the two species is exactly the opposite. In *Achroia*, the alignment pipeline produced twice as many loci as the *de novo* pipeline. In *Tolpis*, the de novo pipeline produced 3-4 times the number of loci generated by the alignment. If the genome assemblies of the two species are largely error-free, these numbers indicate that the *de novo* pipeline in Achroia is over-merging alleles whereas the *de novo* pipeline in *Tolpis* is under-merging alleles. Without additional data, it is impossible to decide where the misassembly occurred, in constructing the loci *de novo*, or in mapping reads to a poorly assembled genome.

Effects of assembly errors on population genetic estimates

A large portion of loci in my data sets are "bad," many of them likely due to misassembly of alleles into loci. The way in which a locus fails to assemble properly is important because the resulting erroneous loci are used to estimate basic population statistics, which are in turn used to calculate various population genetic parameters. Therefore, it is essential to understand the indirect effects of misassemblies of markers on statistics commonly used when studying natural population.

The estimates I will be discussing are the the building blocks of populations genetics: nucleotide polymorphism (S), nucleotide diversity (π), observed heterozygosity (H_o), and the inbreeding coefficient (F). These estimates are necessary when testing for deviations from selective neutrality, quantifying population differentiation and migration, or determining effective population size and linkage disequilibrium, and other processes of evolutionary interest. Nucleotide polymorphism (or nucleotide heterozygosity) is simply the number of segregating nucleotide positions in a set of aligned sequences. The nucleotide diversity, on the other hand, is the average number of nucleotide mismatches per possible pairwise comparison in a set of aligned sequences (Nei and Li 1979). The observed heterozygosity is the proportion of heterozygous individuals in the population. The inbreeding coefficient is defined as the deviation of observed heterozygosity (H_O) from the expected (H_E) under Hardy-Weinberg equilibrium, $F = \frac{H_E - H_O}{H_E}$. All of these parameters are estimated in the context of a locus in a population. In reality, when sampling natural populations, questions usually revolve around multiple populations or subpopulations and any evolutionary forces acting on them. At the larger scale, an important estimate is that of genetic differentiation among subpopulations, or F_{ST}. I chose to focus on F_{ST} because it is a widely used population genetic parameter in studies of natural populations. Here, F_{ST} is defined as the variance of allele frequency between the subpopulations as a fraction of the expected heterozygosity under Hardy-Weinberg when subpopulation allele frequency is equal to the average population allele frequency in the total population. Namely, $F_{ST} = \frac{\sigma^2}{\bar{p}(1-\bar{p})}$, where σ^2 is the variance in allele frequency among subpopulations, and $\bar{p}$ is the average frequency of the allele in the total population. For sake of simplicity, in the following examples, subpopulations are in Hardy-Weinberg equilibrium.

The two main ways in which a locus is not assembled correctly are over- and under-merging of alleles. Under-merged alleles are segregating alleles at a polymorphic locus that do not get grouped together as a marker; instead they are split into multiple loci. Consider as an example a population that is polymorphic for locus *A*[*] with 2 segregating alleles (*a.1* and *a.2*) at some intermediate frequency. Here, the average number of nucleotide mismatches is greater than 0, and the number of segregating nucleotide sites is ≥ 1.

[*] I use capital letters to represent true loci, capital letters with a prime symbol (′) to represent misassembled loci, and small letters to represent alleles.

Thus, nucleotide polymorphism (S) and nucleotide diversity (π) are both > 0. In this instance, under-merging would result in the two alleles being split into two monomorphic loci (*A.1′* and *A.2′*) for all heterozygous individuals. The allele frequency for the *a.1* allele at the *A.1′* locus is now 1, and, similarly, the frequency of *a.2* at the *A.2′* locus is also 1. Therefore, the number of segregating nucleotide sites and the number of nucleotide mismatches is 0 for the two loci, which in turn leads to $\pi = 0$ and S=0 for both *A.1′* and *A.2′*. It follows that splitting a polymorphic locus into two monomorphic loci results in underestimating genetic variation at the nucleotide level. Finally, since all heterozygotes for locus *A* now appear as homozygotes, the heterozygosity estimate in the population will be 0, and the estimated inbreeding coefficient is F = 1.

Over-merging occurs when artificial polymorphic loci are assembled by grouping alleles from different loci. The artificially combined loci may be either monomorphic or polymorphic. The simplest example is to consider a single population with two truly monomorphic loci (*A* and *B* with alleles *a* and *b*, respectively). Since the loci are monomorphic, S and π are 0. With over-merging, non-homologous alleles with low sequence divergence are grouped into an erroneous polymorphic marker (*AB′*) with two segregating alleles (*a* and *b*). The merge of the two non-homologous alleles results in creation of at least one spurious segregating site and at least one pairwise nucleotide difference. Therefore, S and π are going to be estimated as ≥ 0. Importantly, when the genotypes are called in the population, individuals with enough reads for both loci will appear as heterozygotes, and individuals with reads only for one of the two loci will appear as homozygotes for the respective allele. So, over-merging of monomorphic loci introduces artificial polymorphism that leads to inflated values of S, π, H_o, and an underestimated F.

Here is a simple example of two subpopulations (*P1* and *P2*) of equal size with a biallelic locus (*A*), and with one of the segregating alleles being present in both subpopulations. So, *p.1* and *p.2* are segregating in *P1* at some intermediate frequency, and *p.2* and *p.3* are segregating in *P2*. Consider the case of under-merging, where the *de novo* assembly of loci fails to recognize *p.1*, *p.2*, and *p.3* as homologous alleles, and instead clusters *p.1* and *p.3* into one polymorphic locus (*A.1′*) and *p.2* into its own monomorphic locus (*A.2′*). Heterozygotes from *P1* will be called homozygotes for both *A.1′* and *A.2′*. Similarly, heterozygotes from the second subpopulation, *P2*, will appear homozygous at the two artificial loci, *A.1′* and *A.2′*. It is immediately obvious that the heterozygosity in the assembled loci is 0. Additionally, *p.1* and *p.2* are fixed in *P1*, and *p.2* and *p.3* are fixed in *P2*. Since *A.2′* appears monomorphic for *p.2* in both populations, $F_{ST} = 0$ for said locus. Since *A.1′* is fixed for different alleles in the two populations, the estimate of F_{ST} for this locus is 1. Without specific allele frequencies for the initial subpopulations, I cannot estimate the genetic differentiation. However, given that the alleles in the subpopulations are of intermediate frequency, I expect that the true value of F_{ST} would be a value greater than 0 but also less than 1. Therefore, the incorrect assembly of loci has an effect on the F_{ST} estimate, and the direction and magnitude of that effect depends on the true allele frequencies in the two subpopulations.

Conclusion

Even when pipelines for marker discovery are implemented with default parameters and without any organism-specific optimization, I found that a reference genome sequence helps improve the final set of markers by increasing the proportion of “good” markers relative when compared to the final set produced in the absence of a reference sequence. Furthermore, I found that using a read depth threshold for genotype calling is necessary to reduce noise stemming from low-coverage genotypes. Nevertheless, because of the relatively low proportion of “good” markers in the final data sets, I would urge caution when using default parameters for bioinformatics pipelines reconstructing loci from reduced genome complexity libraries because misassemblies introduce error in the estimation of important population statistics. A natural next step is to understand how to optimize input parameters to minimize misassemblies, and to explore ways to utilize even a highly-fragmented genome assembly to aid in the proper re-construction of loci. While it is true that RAD-seq has inherent problems (Arnold et al. 2013;

Mastretta-Yanes et al. 2014), the range of applications in non-model organisms is wide (Ekblom and Galindo 2011; Ellegren 2014; da Fonseca et al. 2016) and it will remain an important technique for the foreseeable future.

TABLES

Table 3.1. For the Tolpis F2.25 and F2.27 populations: Number of markers discovered with at least one genotyped individual, number of markers with a sufficient number of genotyped individuals to be evaluated (the proportion of discovered markers that were evaluated), and number of "good" markers (the proportion of evaluated markers that were "good") for each pipeline ("A" = alignment pipeline, D" = de novo pipeline) with and without a read depth filter (RDF) for genotype calling. a. No filtering was done during genotyping b. a minimum of 6 reads were required for a genotype to be included in the analysis.

a.

Population	**Pipeline**	**Number of Markers with Called Genotypes**	**Number of Markers Tested**	**Good markers**
F2.25	A	99,717	49,784 (50%)	8,498 (17.1%)
F2.25	D	297,142	189,350 (64%)	10,856 (5.7%)
F2.27	A	59,034	18,711 (32%)	6,072 (32.5%)
F2.27	D	212,843	92108 (43%)	10,718 (11.6%)

b.

Population	**Pipeline**	**Number of Markers with Called Genotypes**	**Number of Markers Tested**	**Good markers**
F2.25	A	77,082	40,897 (53%)	9,873 (24.1%)
F2.25	D	282,655	156,160 (55%)	14,570 (9.3%)
F2.27	A	46,215	12,857 (28%)	4,354 (33.9%)
F2.27	D	204,822	65,317 (32%)	10,790 (16.5%)

Table 3.2 For the Achroia KS BC and FL BC populations: Number of markers discovered with at least one genotyped individual, number of markers with a sufficient number of genotyped individuals to be evaluated (the proportion of discovered markers that were evaluated), number of markers containing >5% "impossible" genotypes (the proportion of evaluated markers with >5% "impossible" genotypes), and number of "good" markers (the proportion of evaluated markers that were "good") for each pipeline ("A" = alignment pipeline, D" = de novo pipeline) with and without a read depth filter (RDF) for genotype calling. a. No filtering was done during genotyping b. a minimum of 6 reads were required for a genotype to be included in the analysis.

a.

Population	**Pipeline**	**Number of Markers Discovered**	**Markers Tested**	**Number of markers with "impossible" genotypes**	**Good markers**
KS BC	A	80,734	63,109 (78%)	22428 (35.5%)	795 (1.3%)
KS BC	D	41,001	32,301 (79%)	17817 (55.2%)	191 (0.6%)
FL BC	A	66,085	50,014 (76%)	29155 (58.3%)	919 (1.8%)
FL BC	D	36,587	29,010 (79%)	19418 (66.9%)	62 (0.2%)

b.

Population	**Pipeline**	**Number of Markers Discovered**	**Markers Tested**	**Number of markers with "impossible" genotypes**	**Good markers**
KS BC	A	66666	52042 (78%)	4064 (7.8%)	14628 (28.1%)
KS BC	D	35376	27470 (78%)	2848 (10.4%)	6310 (23.0%)
FL BC	A	54006	39923 (74%)	6602 (16.5%)	8685 (21.8%)
FL BC	D	31510	24162 (77%)	3914 (16.2%)	4381 (18.1%)

Literature Cited

Alem, S, R Streiff, B Courtois, S Zenboudji, D Limousin, and M D Greenfield. 2013. “Genetic Architecture of Sensory Exploitation: QTL Mapping of Female and Male Receiver Traits in an Acoustic Moth..” *Journal of Evolutionary Biology* 26 (12): 2581–96. doi:10.1111/jeb.12252.

Alibert, Paul, Sabrina Renaud, Barbara Dod, Francois Bonhomme, and Jean-Christophe Auffray. 1994. “Fluctuating Asymmetry in the Mus Musculus Hybrid Zone: a Heterotic Effect in Disrupted Co-Adapted Genomes.” *Proc. R. Soc. B* 258 (1351). The Royal Society: 53–59. doi:10.1098/rspb.1994.0141.

Allen, A M, and S J Hiscock. 2008. “Evolution and Phylogeny of Self-Incompatibility Systems in Angiosperms.” In *Self-Incompatibility in Flowering Plants*, 73–101. Springer Berlin Heidelberg. doi:10.1007/978-3-540-68486-2_4.

Allen, Alexandra M, Christopher J Thorogood, Matthew J Hegarty, Christian Lexer, and Simon J Hiscock. 2011. “Pollen–Pistil Interactions and Self-Incompatibility in the Asteraceae: New Insights From Studies of Senecio Squalidus (Oxford Ragwort).” *Annals of Botany* 108 (4). Oxford University Press: 687–98. doi:10.1093/aob/mcr147.

Altschul, Stephen F, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. 1990. “Basic Local Alignment Search Tool.” *Journal of Molecular Biology* 215 (3): 403–10. doi:10.1016/S0022-2836(05)80360-2.

Anderson, Gregory J, Gabriel Bernardello, Tod F Stuessy, and Daniel J Crawford. 2001. “Breeding System and Pollination of Selected Plants Endemic to Juan Fernández Islands.” *American Journal of Botany* 88 (2). Botanical Society of America: 220–33.

Andolfatto, Peter, Dan Davison, Deniz Erezyilmaz, Tina T Hu, Joshua Mast, Tomoko Sunayama-Morita, and David L Stern. 2011. “Multiplexed Shotgun Genotyping for Rapid and Efficient Genetic Mapping..” *Genome Research* 21 (4): 610–17. doi:10.1101/gr.115402.110.

Andrews, Kimberly R, Jeffrey M Good, Michael R Miller, Gordon Luikart, and Paul A Hohenlohe. 2016. “Harnessing the Power of RADseq for Ecological and Evolutionary Genomics..” *Nature Reviews Genetics* 17 (2): 81–92. doi:10.1038/nrg.2015.28.

Andrews, S. n.d. “FastQC: a Quality Control Tool for High Throughput Sequence Data.” http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Archibald, Jenny K, Daniel J Crawford, Arnoldo Santos-Guerra, and Mark E Mort. 2006. “The Utility of Automated Analysis of Inter-Simple Sequence Repeat (ISSR) Loci for Resolving Relationships in the Canary Island Species of Tolpis (Asteraceae)..” *American Journal of Botany* 93 (8): 1154–62. doi:10.3732/ajb.93.8.1154.

Arends, Danny, Pjotr Prins, Ritsert C Jansen, and Karl W Broman. 2010. “R/Qtl: High-Throughput Multiple QTL Mapping.” *Bioinformatics (Oxford, England)*.

Arnold, B, R B Corbett-Detig, D Hartl, and K Bomblies. 2013. “RADseq Underestimates Diversity and Introduces Genealogical Biases Due to Nonrandom Haplotype Sampling..” *Molecular Ecology* 22 (11): 3179–90. doi:10.1111/mec.12276.

Baird, Nathan A, Paul D Etter, Tressa S Atwood, Mark C Currey, Anthony L Shiver, Zachary A Lewis, Eric U Selker, William A Cresko, and Eric A Johnson. 2008. “Rapid SNP Discovery and Genetic Mapping Using Sequenced RAD Markers..” *PloS One* 3 (10). Public Library of Science: e3376. doi:10.1371/journal.pone.0003376.

Baker, H G. 1955. “Self-Compatibility and Establishment After ‘Long-Distance’ Dispersal.” *Evolution* 9 (3): 347. doi:10.2307/2405656.

Baker, Herbert G. 1967. “Support for Baker's Law-as a Rule.” *Evolution* 21 (4): 853. doi:10.2307/2406780.

Barley, Anthony J, Patrick J Monnahan, Robert C Thomson, L Lee Grismer, and Rafe M Brown. 2015. “Sun Skink Landscape Genomics: Assessing the Roles of Micro-Evolutionary Processes in Shaping

Genetic and Phenotypic Diversity Across a Heterogeneous and Fragmented Landscape.” *Molecular Ecology* 24 (8): 1696–1712. doi:10.1111/mec.13151.

Barrett, Spencer C H. 2003. “Mating Strategies in Flowering Plants: the Outcrossing-Selfing Paradigm and Beyond..” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358 (1434). The Royal Society: 991–1004. doi:10.1098/rstb.2003.1301.

Beavis, William. 1998. “QTL Analyses: Power, Precision, and Accuracy .” In *Molecular Dissection of Complex Traits*, edited by Andrew H Peterson, 145–62. Molecular dissection of complex traits.

Beck, J B, I A Al-Shehbaz, and B A Schaal. 2006. “Leavenworthia (Brassicaceae) Revisited: Testing Classic Systematic and Mating System Hypotheses.” *Systematic Botany* 31 (1): 151–59. doi:10.1600/036364406775971732.

Benjamini, Yuval, and Terence P Speed. 2012. “Summarizing and Correcting the GC Content Bias in High-Throughput Sequencing..” *Nucleic Acids Research* 40 (10): e72. doi:10.1093/nar/gks001.

Bernacchi, Dario, and Steven D Tanksley. 1997. “An Interspecific Backcross of Lycopersicon Esculentum × L. Hirsutum: Linkage Analysis and a QTL Study of Sexual Compatibility Factors and Floral Traits.” *Genetics* 147 (2). Genetics: 861–77.

Boyes, D C, and J B Nasrallah. 1993. “Physical Linkage of the SLG and SRK Genes at the Self-Incompatibility Locus of Brassica Oleracea..” *Molecular & General Genetics : MGG* 236 (2-3): 369–73.

Bradbury, Peter J, Zhiwu Zhang, Dallas E Kroon, Terry M Casstevens, Yogesh Ramdoss, and Edward S Buckler. 2007. “TASSEL: Software for Association Mapping of Complex Traits in Diverse Samples..” *Bioinformatics (Oxford, England)* 23 (19): 2633–35. doi:10.1093/bioinformatics/btm308.

Bradnam, Keith R, Joseph N Fass, Anton Alexandrov, Paul Baranay, Michael Bechner, İnanç Birol, Sébastien Boisvert, et al. 2013. “Assemblathon 2: Evaluating De Novo Methods of Genome Assembly in Three Vertebrate Species..” *GigaScience* 2 (1): 10. doi:10.1186/2047-217X-2-10.

Brandt, L S E, and M D Greenfield. 2004. “Condition-Dependent Traits and the Capture of Genetic Variance in Male Advertisement Song..” *Journal of Evolutionary Biology* 17 (4): 821–28. doi:10.1111/j.1420-9101.2004.00716.x.

Brennan, A C, D A Tabah, S A Harris, and S J Hiscock. 2011. “Sporophytic Self-Incompatibility in Senecio Squalidus (Asteraceae): S Allele Dominance Interactions and Modifiers of Cross-Compatibility and Selfing Rates.” *Heredity* 106 (1). Nature Publishing Group: 113–23. doi:10.1038/hdy.2010.29.

Brennan, A C, S A Harris, and S J Hiscock. 2005. “Modes and Rates of Selfing and Associated Inbreeding Depression in the Self-Incompatible Plant Senecio Squalidus (Asteraceae): a Successful Colonizing Species in the British Isles.” *The New Phytologist* 168 (2). Blackwell Science Ltd: 475–86. doi:10.1111/j.1469-8137.2005.01517.x.

Broman, Karl W, and Śaunak Sen. 2009. *A Guide to QTL Mapping with R/Qtl*. New York, NY: Springer New York. doi:10.1007/978-0-387-92125-9.

Busch, Jeremiah W, and Daniel J Schoen. 2008. “The Evolution of Self-Incompatibility When Mates Are Limiting.” *Trends in Plant Science* 13 (3): 128–36. doi:10.1016/i.tplants.2008.01.002.

Carlquist, Sherwin. 1974. *Island Biology*. Columbia University Press.

Carr, Gerald D, Elizabeth A Powell, and Donald W Kyhos. 1986. “Self-Incompatibility in the Hawaiian Madiinae (Compositae): an Exception to Baker's Rule.” *Evolution* 40 (2): 430. doi:10.2307/2408823.

Catchen, Julian M, Angel Amores, Paul Hohenlohe, William Cresko, and John H Postlethwait. 2011. “Stacks: Building and Genotyping Loci De Novo From Short-Read Sequences.” *G3: Genes*.

Catchen, Julian M, Ingo Braasch, and John H Postlethwait. 2011. “Conserved Synteny and the Zebrafish Genome..” In, 104:259–85. Methods in cell biology. doi:10.1016/B978-0-12-374814-0.00015-X.

Catchen, Julian, Paul A Hohenlohe, Susan Bassham, Angel Amores, and William A Cresko. 2013. “Stacks: an Analysis Tool Set for Population Genomics..” *Molecular Ecology* 22 (11): 3124–40. doi:10.1111/mec.12354.

Chain, P S G, D V Grafham, R S Fulton, M G FitzGerald, J Hostetler, D Muzny, J Ali, et al. 2009. “Genome Project Standards in a New Era of Sequencing.” *Science* 326 (5950). NIH Public Access:

236–37. doi:10.1126/science.1180614.

Cheptou, P O, J Lepart, and J Escarre. 2002. “Mating System Variation Along a Successional Gradient in the Allogamous and Colonizing Plant Crepis Sancta (Asteraceae).” *Journal of Evolutionary Biology* 15 (5). Blackwell Science Ltd: 753–62. doi:10.1046/j.1420-9101.2002.00443.x.

Chikhi, Rayan, and Paul Medvedev. 2014. “Informed and Automated K-Mer Size Selection for Genome Assembly..” *Bioinformatics (Oxford, England)* 30 (1): 31–37. doi:10.1093/bioinformatics/btt310.

Churchill, G A, and R W Doerge. 1994. “Empirical Threshold Values for Quantitative Trait Mapping..” *Genetics*.

Clark, S C, R Egan, P I Frazier, and Z Wang. 2013. “ALE: a Generic Assembly Likelihood Evaluation Framework for Assessing the Accuracy of Genome and Metagenome Assemblies.” *Bioinformatics (Oxford, England)* 29 (4): 435–43. doi:10.1093/bioinformatics/bts723.

Collins, R D, Y Jang, K Reinhold, and M D Greenfield. 1999. “Quantitative Genetics of Ultrasonic Advertisement Signalling in the Lesser Waxmoth Achroia Grisella (Lepidoptera: Pyralidae)..” *Heredity* 83 (Pt 6) (December): 644–51.

Crawford, D J, J K Archibald, A Santos-Guerra, and M E Mort. 2006. “Allozyme Diversity Within and Divergence Among Species ofTolpis(Asteraceae-Lactuceae) in the Canary Islands: Systematic, Evolutionary, and Biogeographical Implications.” *American Journal of Botany* 93 (4): 656–64. doi:10.3732/ajb.93.4.656.

Crawford, Daniel J, Gregory J Anderson, Lurdes Borges Silva, Miguel Menezes de Sequeira, Mónica Moura, Arnoldo Santos-Guerra, John K Kelly, and Mark E Mort. 2015. “Breeding Systems in Tolpis (Asteraceae) in the Macaronesian Islands: the Azores, Madeira and the Canaries.” *Plant Systematics and Evolution* 301 (8). Springer Vienna: 1981–93. doi:10.1007/s00606-015-1210-5.

Crawford, Daniel J, Jenny K Archibald, Danielle Stoermer, Mark E Mort, John K Kelly, and Arnoldo Santos-Guerra. 2008. “A Test of Baker’s Law: Breeding Systems and the Radiation of Tolpis(Asteraceae) in the Canary Islands.” *International Journal of Plant Sciences* 169 (6): 782–91. doi:10.1086/533604.

Crawford, Daniel J, Jenny K Archibald, John K Kelly, Mark E Mort, and Arnoldo Santos-Guerra. 2010. “Mixed Mating in the ‘Obligately Outcrossing’Tolpis (Asteraceae) of the Canary Islands.” *Plant Species Biology* 25 (2): 114–19. doi:10.1111/j.1442-1984.2010.00275.x.

Crawford, Daniel J, Mark E Mort, and Jenny Archibald. 2013. “Tolpis Santosii (Asteraceae: Cichorieae), a New Species From La Palma, Canary Islands.” *Vieraea*, no. 41. Organismo Autónomo Complejo Insular de Museos y Centros. Cabildo de Tenerife: 169–76.

da Fonseca, Rute R, Anders Albrechtsen, Gonçalo Espregueira Themudo, Jazmín Ramos-Madrigal, Jonas Andreas Sibbesen, Lasse Maretty, M Lisandra Zepeda-Mendoza, Paula F Campos, Rasmus Heller, and Ricardo J Pereira. 2016. “Next-Generation Biology: Sequencing and Data Analysis Approaches for Non-Model Organisms..” *Marine Genomics* 30 (December): 3–13. doi:10.1016/j.margen.2016.04.012.

Danecek, P, A Auton, G Abecasis, C A Albers, E Banks, M A DePristo, R E Handsaker, et al. 2011. “The Variant Call Format and VCFtools.” *Bioinformatics (Oxford, England)* 27 (15): 2156–58. doi:10.1093/bioinformatics/btr330.

Davey, J W, T Cezard, and P Fuentes Utrilla. 2013. “Special Features of RAD Sequencing Data: Implications for Genotyping - Davey - 2012 - Molecular Ecology - Wiley Online Library.” *Molecular ….*

De Nettancourt, D. 2013. *Incompatibility in Angiosperms*. doi:10.1007/978-3-662-12051-4.

de Nettancourt, Dr Dreux. 1977. “The Genetic Basis of Self-Incompatibility.” In *Incompatibility in Angiosperms*, 3:28–57. Monographs on Theoretical and Applied Genetics. Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-662-12051-4_2.

Deamer, D W, and M Akeson. 2000. “Nanopores and Nucleic Acids: Prospects for Ultrarapid Sequencing..” *Trends in Biotechnology* 18 (4): 147–51.

Earl, Dent, Keith Bradnam, John St John, Aaron Darling, Dawei Lin, Joseph Fass, Hung On Ken Yu, et al. 2011. “Assemblathon 1: a Competitive Assessment of De Novo Short Read Assembly Methods..”

Genome Research 21 (12): 2224–41. doi:10.1101/gr.126599.111.

Eaton, Deren A R. 2014. "PyRAD: Assembly of De Novo RADseq Loci for Phylogenetic Analyses.." *Bioinformatics (Oxford, England)* 30 (13): 1844–49. doi:10.1093/bioinformatics/btu121.

Eaton, Deren A R, and Richard H Ree. 2013. "Inferring Phylogeny and Introgression Using RADseq Data: an Example From Flowering Plants (Pedicularis: Orobanchaceae).." *Systematic Biology* 62 (5): 689–706. doi:10.1093/sysbio/syt032.

Eilbeck, Karen, Barry Moore, Carson Holt, and Mark Yandell. 2009. "Quantitative Measures for the Management and Comparison of Annotated Genomes." *BMC Bioinformatics* 10 (1). BioMed Central Ltd: 67. doi:10.1186/1471-2105-10-67.

Ekblom, R, and J Galindo. 2011. "Applications of Next Generation Sequencing in Molecular Ecology of Non-Model Organisms.." *Heredity* 107 (1): 1–15. doi:10.1038/hdy.2010.152.

Ellegren, Hans. 2014. "Genome Sequencing and Population Genomics in Non-Model Organisms." *Trends in Ecology & Evolution* 29 (1): 51–63. doi:10.1016/j.tree.2013.09.008.

Elshire, Robert J, Jeffrey C Glaubitz, Qi Sun, Jesse A Poland, Ken Kawamoto, Edward S Buckler, and Sharon E Mitchell. 2011. "A Robust, Simple Genotyping-by-Sequencing (GBS) Approach for High Diversity Species.." Edited by Laszlo Orban. *PloS One* 6 (5): e19379. doi:10.1371/journal.pone.0019379.

Emerson, Kevin J, Clayton R Merz, Julian M Catchen, Paul A Hohenlohe, William A Cresko, William E Bradshaw, and Christina M Holzapfel. 2010. "Resolving Postglacial Phylogeography Using High-Throughput Sequencing.." *Proceedings of the National Academy of Sciences of the United States of America* 107 (37). National Acad Sciences: 16196–200. doi:10.1073/pnas.1006538107.

Fisher, R A. 1941. "Average Excess and Average Effect of a Gene Substitution." *Annals of Eugenics*. doi:10.1111/j.1469-1809.1941.tb02272.x/pdf.

Fishman, Lila, Alan J Kelly, and John H Willis. 2002. "Minor Quantitative Trait Loci Underlie Floral Traits Associated with Mating System Divergence in Mimulus.." *Evolution* 56 (11): 2138–55. doi:10.1554/0014-3820(2002)056[2138:MQTLUF]2.0.CO;2.

Gautier, Mathieu, Karim Gharbi, Timothee Cezard, Julien Foucaud, Carole Kerdelhué, Pierre Pudlo, Jean-Marie Cornuet, and Arnaud Estoup. 2013. "The Effect of RAD Allele Dropout on the Estimation of Genetic Variation Within and Between Populations.." *Molecular Ecology* 22 (11): 3165–78. doi:10.1111/mec.12089.

Ghodsi, M, C M Hill, I Astrovskaya, and H Lin. 2013. "De Novo Likelihood-Based Measures for Comparing Genome Assemblies." *BMC Research ….*

Gleason, Jennifer M, Yihong Zhou, Jennifer L Hackett, Bethany R Harris, and Michael D Greenfield. 2016. "Development of a Genomic Resource and Quantitative Trait Loci Mapping of Male Calling Traits in the Lesser Wax Moth, Achroia Grisella.." *PloS One* 11 (1). doi:10.1371/journal.pone.0147014.

Good-Avila, Sara V, and Andrew G Stephenson. 2002. "The Inheritance of Modifiers Conferring Self-Fertility in the Partially Self-Incompatible Perennial, Campanula Rapunculoides L. (Campanulaceae) ." *Evolution* 56 (2). Blackwell Publishing Ltd: 263–72. doi:10.1111/j.0014-3820.2002.tb01336.x.

Goodstein, David M, Shengqiang Shu, Russell Howson, Rochak Neupane, Richard D Hayes, Joni Fazo, Therese Mitros, et al. 2012. "Phytozome: a Comparative Platform for Green Plant Genomics.." *Nucleic Acids Research* 40 (Database issue): D1178–86. doi:10.1093/nar/gkr944.

Goodwillie, C. 1999. "Multiple Origins of Self-Compatibility in Linanthus Section Leptosiphon (Polemoniaceae): Phylogenetic Evidence From Internal-Transcribed-Spacer Sequence Data." *Evolution* 53 (5): 1387–95. doi:10.2307/2640885.

Goodwillie, C, C Ritland, and K Ritland. 2006. "The Genetic Basis of Floral Traits Associated with Mating System Evolution in Leptosiphon (Polemoniaceae): an Analysis of Quantitative Trait Loci." *Evolution* 60 (3): 491–504.

Goodwillie, Carol, and Jennifer M Ness. 2005. "Correlated Evolution in Floral Morphology and the Timing of Self-Compatibility in Leptosiphon Jepsonii(Polemoniaceae)." *International Journal of Plant Sciences* 166 (5): 741–51. doi:10.1086/431801.

Goodwillie, Carol, Risa D Sargent, Christopher G Eckert, Elizabeth Elle, Monica A Geber, Mark O Johnston, Susan Kalisz, et al. 2010. “Correlated Evolution of Mating System and Floral Display Traits in Flowering Plants and Its Implications for the Distribution of Mating System Variation.” *The New Phytologist* 185 (1). Blackwell Publishing Ltd: 311–21. doi:10.1111/j.1469-8137.2009.03043.x.

Grabherr, Manfred G, Brian J Haas, Moran Yassour, Joshua Z Levin, Dawn A Thompson, Ido Amit, Xian Adiconis, et al. 2011. “Trinity: Reconstructing a Full-Length Transcriptome Without a Reference Genome From RNA-Seq Data.” *Nature Biotechnology* 29 (7): 644–52. doi:10.1038/nbt.1883.

Greenfield, Michael D. 1981. “Moth Sex Pheromones: an Evolutionary Perspective.” *The Florida Entomologist* 64 (1): 4. doi:10.2307/3494597.

Greenfield, Michael D, and James A Coffelt. 1983. “Reproductive Behaviour of the Lesser Waxmoth, Achroia Grisella (Pyralidae: Galleriinae): Signalling, Pair Formation, Male Interactions, and Mate Guarding.” *Behaviour* 84 (3/4). BRILL: 287–315. doi:10.2307/4534248?ref=search-gateway:115b4a28732346ba7a587ccaf03d047a.

Haibao Tang, Xingtan Zhang Chenyong Miao Jisen Zhang Ray Ming James C Schnable Patrick S Schnable Eric Lyons Jianguo Lu. 2015. “ALLMAPS: Robust Scaffold Ordering Based on Multiple Maps.” *Genome Biology* 16 (1). BioMed Central. doi:10.1186/s13059-014-0573-1.

Hatakeyama, K, T Takasaki, M Watanabe, and K Hinata. 1998. “Molecular Characterization of S Locus Genes, SLG and SRK, in a Pollen-Recessive Self-Incompatibility Haplotype of Brassica Rapa L..” *Genetics* 149 (3): 1587–97.

Hiscock, S J. 2000a. “Genetic Control of Self-Incompatibility in Senecio Squalidus L. (Asteraceae): a Successful Colonizing Species..” *Heredity* 85 (Pt 1) (July): 10–19.

Hiscock, Simon J. 2000b. “Self-Incompatibility in Senecio Squalidus L. (Asteraceae).” *Annals of Botany* 85 (suppl 1). Oxford University Press: 181–90.

Hiscock, Simon J, and David A Tabah. 2003. “The Different Mechanisms of Sporophytic Self–Incompatibility.” *Philosophical Transactions of the Royal Society of London B: Biological Sciences* 358 (1434). The Royal Society: 1037–45. doi:10.1098/rstb.2003.1297.

Hiscock, Simon J, and Stephanie M McInnis. 2003. “Pollen Recognition and Rejection During the Sporophytic Self-Incompatibility Response: Brassica and Beyond.” *Trends in Plant Science* 8 (12): 606–13. doi:10.1016/j.tplants.2003.10.007.

Hiscock, Simon J, Stephanie M McInnis, David A Tabah, Catherine A Henderson, and Adrian C Brennan. 2003. “Sporophytic Self-Incompatibility in Senecio Squalidus L (Asteraceae)--the Search for S..” *Journal of Experimental Botany* 54 (380): 169–74.

Hohenlohe, Paul A, Julian Catchen, and William A Cresko. 2012. “Population Genomic Analysis of Model and Nonmodel Organisms Using Sequenced RAD Tags..” *Methods in Molecular Biology (Clifton, N.J.)* 888: 235–60. doi:10.1007/978-1-61779-870-2_14.

Hohenlohe, Paul A, Susan Bassham, Paul D Etter, Nicholas Stiffler, Eric A Johnson, and William A Cresko. 2010. “Population Genomics of Parallel Adaptation in Threespine Stickleback Using Sequenced RAD Tags..” *PLOS Genetics* 6 (2). Public Library of Science: e1000862. doi:10.1371/journal.pgen.1000862.

Holt, Carson, and Mark Yandell. 2011. “MAKER2: an Annotation Pipeline and Genome-Database Management Tool for Second-Generation Genome Projects.” *BMC Bioinformatics* 12 (1). BioMed Central Ltd: 491. doi:10.1101/gr.403602.

Hughes, Kimberly A. 1995. “The Evolutionary Genetics of Male Life-History Characters in Drosophila Melanogaster.” *Evolution* 49 (3): 521. doi:10.2307/2410276.

Hughes, M K, and A L Hughes. 1993. “Evolution of Duplicate Genes in a Tetraploid Animal, Xenopus Laevis..” *Molecular Biology and Evolution* 10 (6). Oxford University Press: 1360–69. doi:10.1093/oxfordjournals.molbev.a040080.

Hurka, Herbert, and Barbara Neuffer. 1997. “Evolutionary Processes in the genusCapsella (Brassicaceae).” *Plant Systematics and Evolution* 206 (1-4). Springer-Verlag: 295–316. doi:10.1007/BF00987954.

Igic, Boris, and Jeremiah W Busch. 2013. “Is Self-Fertilization an Evolutionary Dead End?.” *New*

Phytologist 198 (2): 386–97. doi:10.1111/nph.12182.

Igic, Boris, Russell Lande, and Joshua R Kohn. 2008. “Loss of Self-Incompatibility and Its Evolutionary Consequences.” *International Journal of Plant Sciences* 169 (1): 93–104. doi:10.1086/523362.

Ilut, Daniel C, Marie L Nydam, and Matthew P Hare. 2014. “Defining Loci in Restriction-Based Reduced Representation Genomic Data From Nonmodel Species: Sources of Bias and Diagnostics for Optimal Clustering.” *BioMed Research International* 2014 (June). Hindawi Publishing Corporation. doi:10.1155/2014/675158.

Iwano, Megumi, and Seiji Takayama. 2012. “Self/Non-Self Discrimination in Angiosperm Self-Incompatibility.” *Current Opinion in Plant Biology* 15 (1): 78–83. doi:10.1016/j.pbi.2011.09.003.

Jang, Y, and M D Greenfield. 1998a. “Absolute Versus Relative Measurements of Sexual Selection: Assessing the Contributions of Ultrasonic Signal Characters to Mate Attraction in Lesser Wax Moths, Achroia ....” *Evolution*.

Jang, Yikweon, and Michael D Greenfield. 1996. “Ultrasonic Communication and Sexual Selection in Wax Moths: Female Choice Based on Energy and Asynchrony of Male Signals.” *Animal Behaviour* 51 (5): 1095–1106. doi:10.1006/anbe.1996.0111.

Jang, Yikweon, and Michael D Greenfield. 1998b. “Absolute Versus Relative Measurements of Sexual Selection: Assessing the Contributions of Ultrasonic Signal Characters to Mate Attraction in Lesser Wax Moths, Achroia Grisella (Lepidoptera: Pyralidae).” *Evolution* 52 (5): 1383. doi:10.2307/2411308.

Jarvis, C E. 1980. “Systematic Studies in the Genus *Tolpis* Adanson.” *Ph.D. Dissertation, University of Reading*. Reading, UK.

Jones, Felicity C, Manfred G Grabherr, Yingguang Frank Chan, Pamela Russell, Evan Mauceli, Jeremy Johnson, Ross Swofford, et al. 2012. “The Genomic Basis of Adaptive Evolution in Threespine Sticklebacks.” *Nature* 484 (7392). Nature Research: 55–61. doi:10.1038/nature10944.

Kelley, D R, M C Schatz, and S L Salzberg. 2010. “Quake: Quality-Aware Detection and Correction of Sequencing Errors.” *Genome Biology*.

Korf, Ian. 2004. “Gene Finding in Novel Genomes.” *BMC Bioinformatics* 5 (1). BioMed Central Ltd: 59. doi:10.1186/1471-2105-5-59.

Kosambi, D D. 1943. “The Estimation of Map Distances From Recombination Values.” *Annals of Eugenics* 12 (1). Blackwell Publishing Ltd: 172–75. doi:10.1111/j.1469-1809.1943.tb02321.x.

Kusaba, M, K Dwyer, J Hendershot, J Vrebalov, J B Nasrallah, and M E Nasrallah. 2001. “Self-Incompatibility in the Genus Arabidopsis: Characterization of the S Locus in the Outcrossing a. Lyrata and Its Autogamous Relative a. Thaliana..” *The Plant Cell* 13 (3): 627–43.

Langmead, Ben, and Steven L Salzberg. 2012. “Fast Gapped-Read Alignment with Bowtie 2..” *Nature Methods* 9 (4): 357–59. doi:10.1038/nmeth.1923.

Lathrop, G M, J M Lalouel, C Julier, and J Ott. 1984. “Strategies for Multilocus Linkage Analysis in Humans..” *Proceedings of the National Academy of Sciences of the United States of America* 81 (11). National Acad Sciences: 3443–46.

Lenski, R E, M R Rose, and S C Simpson. 1991. “Long-Term Experimental Evolution in Escherichia Coli. I. Adaptation and Divergence During 2,000 Generations.” *The American Naturalist* 138 (6): 1315–41. doi:10.1086/285289.

Levin, D A. 1996. “The Evolutionary Significance of Pseudo-Self-Fertility.” *American Naturalist*. doi:10.2307/2463457.

Li, H. 2011. “A Statistical Framework for SNP Calling, Mutation Discovery, Association Mapping and Population Genetical Parameter Estimation From Sequencing Data.” *Bioinformatics (Oxford, England)* 27 (21): 2987–93. doi:10.1093/bioinformatics/btr509.

Li, H, and R Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics (Oxford, England)* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

Li, H, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, 1000 Genome Project Data Processing Subgroup. 2009. “The Sequence Alignment/Map Format and SAMtools.” *Bioinformatics (Oxford, England)* 25 (16): 2078–79. doi:10.1093/bioinformatics/btp352.

Limousin, D, and M D Greenfield. 2009. “Evaluation of Amplitude in Male Song: Female Waxmoths Respond to Fortissimo Notes..” *The Journal of Experimental Biology* 212 (Pt 24): 4091–4100. doi:10.1242/jeb.035345.

Limousin, Denis, Réjane Streiff, Brigitte Courtois, Virginie Dupuy, Sylvain Alem, and Michael D Greenfield. 2012. “Genetic Architecture of Sexual Selection: QTL Mapping of Male Song and Female Receiver Traits in an Acoustic Moth..” *PloS One* 7 (9): e44554. doi:10.1371/journal.pone.0044554.

Lin, J Z, and K Ritland. 1997. “Quantitative Trait Loci Differentiating the Outbreeding Mimulus Guttatus From the Inbreeding M-Platycalyx.” *Genetics* 146 (3). Genetics: 1115–21.

Mable, B K, AVRS Dart, C D Berardo, and L Witham. 2005. “Breakdown of Self-Incompatibility in the Perennial Arabidopsis Lyrata (Brassicaceae) and Its Genetic Consequences.” *Evolution* 59 (7): 1437–48. doi:10.1111/j.0014-3820.2005.tb01794.x.

Mackay, T F. 2001. “The Genetic Architecture of Quantitative Traits..” *Annual Review of Genetics* 35 (1): 303–39. doi:10.1146/annurev.genet.35.102401.090633.

Mardis, Elaine R. 2008. “The Impact of Next-Generation Sequencing Technology on Genetics.” *Trends in Genetics* 24 (3): 133–41. doi:10.1016/j.tig.2007.12.007.

Mastretta-Yanes, A, N Arrigo, N Alvarez, T H Jorgensen, D Piñero, and B C Emerson. 2014. “Restriction Site-Associated DNA Sequencing, Genotyping Error Estimation and De Novoassembly Optimization for Population Genetic Inference.” *Molecular Ecology Resources* 15 (1): 28–41. doi:10.1111/1755-0998.12291.

McInnis, Stephanie M, Liliana M Costa, José F Gutiérrez-Marcos, Catherine A Henderson, and Simon J Hiscock. 2005. “Isolation and Characterization of a Polymorphic Stigma-Specific Class III Peroxidase Gene From Senecio Squalidus L. (Asteraceae).” *Plant Molecular Biology* 57 (5). Kluwer Academic Publishers: 659–77. doi:10.1007/s11103-005-1426-9.

McKenna, Aaron, Matthew Hanna, Eric Banks, Andrey Sivachenko, Kristian Cibulskis, Andrew Kernytsky, Kiran Garimella, et al. 2010. “The Genome Analysis Toolkit: a MapReduce Offramework for Analyzing Next-Generation DNA Sequencing Data.” *Genome Research* 20 (9): 1297–1303. doi:10.1101/gr.107524.110.

Mena-Ali, Jorge I, and Andrew G Stephenson. 2007. “Segregation Analyses of Partial Self-Incompatibility in Self and Cross Progeny of Solanum Carolinense Reveal a Leaky S-Allele.” *Genetics* 177 (1). Genetics: 501–10. doi:10.1534/genetics.107.073775.

Metzker, Michael L. 2009. “Sequencing Technologies — the Next Generation.” *Nature Reviews Genetics* 11 (1): 31–46. doi:10.1038/nrg2626.

Miller, Jason R, Sergey Koren, and Granger Sutton. 2010. “Assembly Algorithms for Next-Generation Sequencing Data..” *Genomics* 95 (6): 315–27. doi:10.1016/j.ygeno.2010.03.001.

Milum, V G. 1935. “Lesser vs Greater Wax-Moth.” *Gleanings in Bee Culture* 64 (January): 662–66.

Moore, Michael J, Javier Francisco-Ortega, Arnoldo Santos-Guerra, and Robert K Jansen. 2002. “Chloroplast DNA Evidence for the Roles of Island Colonization and Extinction in Tolpis (Asteraceae: Lactuceae)..” *American Journal of Botany* 89 (3). Botanical Society of America: 518–26. doi:10.3732/ajb.89.3.518.

Mort, Mark E, Daniel J Crawford, John K Kelly, Arnoldo Santos-Guerra, Miguel Menezes de Sequeira, Mónica Moura, and Juli Caujapé-Castells. 2015. “Multiplexed-Shotgun-Genotyping Data Resolve Phylogeny Within a Very Recently Derived Insular Lineage..” *American Journal of Botany* 102 (4): 634–41. doi:10.3732/ajb.1400551.

Nagarajan, Niranjan, and Mihai Pop. 2013. “Sequence Assembly Demystified.” *Nature Reviews Genetics* 14 (3). Nature Publishing Group: 157–67. doi:10.1038/nrg3367.

Nasrallah, June, and Mikhail Nasrallah. 1993. “Pollen-Stigma Signaling in the Sporophytic Self-Incompatibility Response.” *The Plant Cell* 5 (10): 1325. doi:10.2307/3869785.

Nasrallah, June, Pei Liu, Susan Sherman-Broyles, Renate Schmidt, and Mikhail Nasrallah. 2007. “Epigenetic Mechanisms for Breakdown of Self-Incompatibility in Interspecific Hybrids.” *Genetics* 175 (4). Genetics: 1965–73. doi:10.1534/genetics.106.069393.

Nei, M, and W H Li. 1979. “Mathematical Model for Studying Genetic Variation in Terms of Restriction

Endonucleases..” *Proceedings of the National Academy of Sciences of the United States of America* 76 (10). National Academy of Sciences: 5269–73.

Nielsen, L R, and H R Siegismund. 2003. “Partial Self-Incompatibility in the Polyploid Endemic Species Scalesia Affinis (Asteraceae) From the Galápagos: Remnants of a Self-Incompatibility System?.” *Botanical Journal of the ….* doi:10.1046/j.1095-8339.2003.00168.x.

Nielsen, Rasmus, Thorfinn Korneliussen, Anders Albrechtsen, Yingrui Li, and Jun Wang. 2012. “SNP Calling, Genotype Calling, and Sample Allele Frequency Estimation From New-Generation Sequencing Data..” *PloS One* 7 (7). Public Library of Science: e37558. doi:10.1371/journal.pone.0037558.

Ornduff, Robert. 1969. “Reproductive Biology in Relation to Systematics.” *Taxon* 18 (2). International Association for Plant Taxonomy (IAPT): 121–33. doi:10.2307/1218671?ref=search-gateway:93d6a1675815723f22f1014929c1ed4b.

Park, S J, E J Korompai, J Francisco-Ortega, A Santos-Guerra, and R K Jansen. 2001. “Phylogenetic Relationships of Tolpis (Asteraceae: Lactuceae) Based on ndhF Sequence Data.” *Plant Systematics and Evolution* 226 (1-2). Springer-Verlag: 23–33. doi:10.1007/s006060170071.

Parra, G, K Bradnam, Z Ning, T Keane, and I Korf. 2009. “Assessing the Gene Space in Draft Genomes.” *Nucleic Acids Research* 37 (1). Oxford University Press: 289–97. doi:10.1093/nar/gkn916.

Pettengill, James B, and David A Moeller. 2012. “Tempo and Mode of Mating System Evolution Between Incipient Clarkia Species..” *Evolution* 66 (4): 1210–25. doi:10.1111/j.1558-5646.2011.01521.x.

Pevzner, Pavel A, Haixu Tang, and Michael S Waterman. 2001. “An Eulerian Path Approach to DNA Fragment Assembly.” *Proceedings of the ….*

Pompanon, François, Aurélie Bonin, Eva Bellemain, and Pierre Taberlet. 2005. “Genotyping Errors: Causes, Consequences and Solutions..” *Nature Reviews Genetics* 6 (11): 847–59. doi:10.1038/nrg1707.

Pujolar, J M, M W Jacobsen, J Frydenberg, T D Als, P F Larsen, G E Maes, L Zane, J B Jian, L Cheng, and M M Hansen. 2013. “A Resource of Genome-Wide Single-Nucleotide Polymorphisms Generated by RAD Tag Sequencing in the Critically Endangered European Eel.” *Molecular Ecology Resources* 13 (4): 706–14. doi:10.1111/1755-0998.12117.

Rahman, Atif, and Lior Pachter. 2013. “CGAL: Computing Genome Assembly Likelihoods..” *Genome Biology* 14 (1): R8. doi:10.1186/gb-2013-14-1-r8.

Rastas, Pasi, Federico C F Calboli, Baocheng Guo, Takahito Shikano, and Juha Merilä. 2016. “Construction of Ultradense Linkage Maps with Lep-MAP2: Stickleback F2 Recombinant Crosses as an Example..” *Genome Biology and Evolution* 8 (1): 78–93. doi:10.1093/gbe/evv250.

Rastas, Pasi, Lars Paulin, Ilkka Hanski, Rainer Lehtonen, and Petri Auvinen. 2013. “Lep-MAP: Fast and Accurate Linkage Map Construction for Large SNP Datasets..” *Bioinformatics (Oxford, England)* 29 (24): 3128–34. doi:10.1093/bioinformatics/btt563.

Rhoads, Anthony, and Kin Fai Au. 2015. “PacBio Sequencing and Its Applications.” *Genomics, Proteomics & Bioinformatics* 13 (5). Beijing Institute of Genomics, Chinese Academy of Sciences and Genetics Society of China: 278–89. doi:10.1016/j.gpb.2015.08.002.

Ritland, Carol, and Kermit Ritland. 1989. “Variation of Sex Allocation Among Eight Taxa of the Mimulus Guttatus Species Complex (Scrophulariaceae).” *American Journal of Botany* 76 (12). Botanical Society of America: 1731–39. doi:10.2307/2444472?ref=search-gateway:b0f8a5054a62c4016a726c60951db733.

Salzberg, Steven L, Adam M Phillippy, Aleksey Zimin, Daniela Puiu, Tanja Magoc, Sergey Koren, Todd J Treangen, et al. 2012. “GAGE: a Critical Evaluation of Genome Assemblies and Assembly Algorithms..” *Genome Research* 22 (3): 557–67. doi:10.1101/gr.131383.111.

Simpson, Jared T, Kim Wong, Shaun D Jackman, Jacqueline E Schein, Steven J M Jones, and İnanç Birol. 2009. “ABySS: a Parallel Assembler for Short Read Sequence Data.” *Genome*.

Slotte, T, K M Hazzouri, D Stern, and P Andolfatto. 2012. “Genetic Architecture and Adaptive Significance of the Selfing Syndrome in Capsella.” …. doi:10.2307/41503448.

Smit, AFA, and R Hubley. 2010. “RepeatModeler Open-1.0.” Repeat Masker Website. http://www.repeatmasker.org.

Soto-Trejo, Fabiola, John K Kelly, Jenny K Archibald, Mark E Mort, Arnoldo Santos-Guerra, and Daniel J Crawford. 2013. “The Genetics of Self-Compatibility and Associated Floral Characters in Tolpis(Asteraceae) in the Canary Islands.” *International Journal of Plant Sciences* 174 (2). University of Chicago PressChicago, IL: 171–78. doi:10.1086/668788.

Spangler, Hayward G, and Alex Takessian. 1986. “Further Observations on Sound Production by the Lesser Wax Moth, Achroia Grisella (F.) (Lepidoptera: Pyralidae).” *Journal of the Kansas Entomological Society* 59 (3). Allen Press: 555–57. doi:10.2307/25084819?ref=search-gateway:115b4a28732346ba7a587ccaf03d047a.

Stanke, M, and B Morgenstern. 2005. “AUGUSTUS: a Web Server for Gene Prediction in Eukaryotes That Allows User-Defined Constraints.” *Nucleic Acids Research* 33 (Web Server): W465–67. doi:10.1093/nar/gki458.

Stebbins, G Ledyard. 1957. “Self Fertilization and Population Variability in the Higher Plants.” *The American Naturalist* 91 (861): 337–54. doi:10.1086/281999.

Stein, J C, B Howlett, D C Boyes, M E Nasrallah, and J B Nasrallah. 1991. “Molecular Cloning of a Putative Receptor Protein Kinase Gene Encoded at the Self-Incompatibility Locus of Brassica Oleracea..” *Proceedings of the National Academy of Sciences of the United States of America* 88 (19). National Academy of Sciences: 8816–20.

Stephenson, Andrew G, Sara V Good, and Donna W Vogler. 2000. “Interrelationships Among Inbreeding Depression, Plasticity in the Self-Incompatibility System, and the Breeding System of Campanula Rapunculoides L. (Campanulaceae).” *Annals of Botany* 85 (suppl 1). Oxford University Press: 211–19.

Stone, J L, M A Sasuclark, and C P Blomberg. 2006. “Variation in the Self-Incompatibility Response Within and Among Populations of the Tropical Shrub Witheringia Solanacea (Solanaceae).” *American Journal of Botany* 93 (4). Botanical Society of America: 592–98. doi:10.3732/ajb.93.4.592.

Suomalainen, Esko, Laurence M Cook, and John R G Turner. 1973. “Achiasmatic Oogenesis in the Heliconiine Butterflies.” *Hereditas* 74 (2): 302–4. doi:10.1111/j.1601-5223.1973.tb01134.x.

Swarbreck, David, Christopher Wilks, Philippe Lamesch, Tanya Z Berardini, Margarita Garcia-Hernandez, Hartmut Foerster, Donghui Li, et al. 2008. “The Arabidopsis Information Resource (TAIR): Gene Structure and Function Annotation..” *Nucleic Acids Research* 36 (Database issue): D1009–14. doi:10.1093/nar/gkm965.

Takayama, S, H Shiba, M Iwano, H Shimosato, F S Che, N Kai, M Watanabe, G Suzuki, K Hinata, and A Isogai. 2000. “The Pollen Determinant of Self-Incompatibility in Brassica Campestris..” *Proceedings of the National Academy of Sciences of the United States of America* 97 (4). National Acad Sciences: 1920–25. doi:10.1073/pnas.040556397.

Takayama, Seiji, and Akira Isogai. 2005. “Self-Incompatibility in Plants..” *Annual Review of Plant Biology* 56 (1): 467–89. doi:10.1146/annurev.arplant.56.032604.144249.

Traut, Walther, Ken Sahara, Atsuo Yoshido, František Marec, Iva Fuková, Hong-Bin Zhang, Cheng-Cang Wu, Marian R Goldsmith, and Yuji Yasukochi. 2007. “Conserved Synteny of Genes Between Chromosome 15 of Bombyx Moriand a Chromosome of Manduca Sexta Shown by Five-Color BAC-FISH.” *Genome* 50 (11): 1061–65. doi:10.1139/G07-082.

Valvekens, Dirk, Marc Van Montagu, and Mieke Van Lijsebettens. 1988. “Agrobacterium Tumefaciens-Mediated Transformation of Arabidopsis Thaliana Root Explants by Using Kanamycin Selection.” *Proceedings of the National Academy of Sciences of the United States of America* 85 (15). National Acad Sciences: 5536–40. doi:10.1073/pnas.85.15.5536.

Vezzi, F, G Narzisi, and B Mishra. 2012. “Feature-by-Feature – Evaluating De Novo Sequence Assembly.” *PloS One*.

Vogler, D W, and A G Stephenson. 2001. “The Potential for Mixed Mating in a Self-Incompatible Plant.” *International Journal of Plant Sciences* 162 (4): 801–5. doi:10.1086/320787.

Wang, Rong-Lin, Adrian Stec, Jody Hey, Lewis Lukens, and John Doebley. 1999. “The Limits of

Selection During Maize Domestication." *Nature* 398 (6724). Nature Publishing Group: 236–39. doi:10.1038/18435.

Watanabe, Masao, Akiko Ito, Yoshinobu Takada, Chie Ninomiya, Tomohiro Kakizaki, Yoshihito Takahata, Katsunori Hatakeyama, et al. 2000. "Highly Divergent Sequences of the Pollen Self-Incompatibility (S) Gene in Class-I S Haplotypes of Brassica Campestris (Syn. Rapa) L." *FEBS Letters* 473 (2): 139–44. doi:10.1016/S0014-5793(00)01514-3.

Xu, Shizhong. 2003. "Theoretical Basis of the Beavis Effect.." *Genetics* 165 (4): 2259–68.

Zhan, Shuai, Christine Merlin, Jeffrey L Boore, and Steven M Reppert. 2011. "The Monarch Butterfly Genome Yields Insights Into Long-Distance Migration." *Cell* 147 (5): 1171–85. doi:10.1016/j.cell.2011.09.052.

Zhou, Yihong, Heidi K Kuster, Jeffrey S Pettis, Robert G Danka, Jennifer M Gleason, and Michael D Greenfield. 2008. "Reaction Norm Variants for Male Calling Song in Populations of Achroia Grisella (Lepidoptera: Pyralidae): Toward a Resolution of the Lek Paradox.." *Evolution* 62 (6). Blackwell Publishing Inc: 1317–34. doi:10.1111/j.1558-5646.2008.00371.x.