Identifying Traces of Selection on Quantitative Traits in Plant and Animal Genomes

By

Timothy Mathes Beissinger

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Statistical and Quantitative Genetics)

at the

UNIVERSITY OF WISCONSIN—MADISON

2014

Date of final oral examination: 5/12/2014

The dissertation is approved by the following members of the Final Oral Committee:

Natalia de Leon, Associate Professor, Agronomy

Daniel Gianola, Professor, Animal Sciences, Biostatistics and Medical Informatics, Dairy Science

Shawn Kaeppler, Professor, Agronomy

Guilherme Rosa, Associate Professor, Animal Science and Biostatistics and Medical Informatics

Kent Weigel, Professor, Dairy Science

DEDICATION

This work is dedicated to my father, Dr. Richard Lynn Beissinger.

ACKNOWLEDGEMENTS

Science does not take place in a vacuum. Advancements in knowledge are only possible due to the hard work of earlier individuals. Therefore, I would like to first acknowledge the geneticists, statisticians, programmers, and other scientists whose work laid the foundation that this dissertation is based upon. Every publication found in the references of this document represents enormous amounts of thought and labor that have built positively on the collective understanding of the processes researched here.

More directly, there are specific individuals whose guidance and assistance has played a critical role in helping me complete this research. My co-advisor, mentor, teacher, and friend, Professor Natalia de Leon, inspired me to pursue this line of research. I began working for Natalia as an undergraduate assistant merely as means to earn money while majoring in seemingly separate fields as an undergraduate. Natalia's passion for genetics and her demonstration of the important contributions that I could make were the primer for my decision to pursue this research. Throughout my Ph.D., Natalia constantly pushed me to give and achieve more than I would have or could have without her guidance. The resources and projects that were available to me thanks to Natalia were pivotal to everything described in this dissertation. My co-advisor, Professor Daniel Gianola, provided an unquantifiable amount of guidance regarding interesting and desirable avenues of research for me to pursue. Dan's excellence and reputation as a scientist opened doors for me that would not have been open without his support. Dan provided suggestions and critiques of my research that improved it at all levels. A third person, Professor Shawn Kaeppler, was not an official co-advisor of mine but unofficially served as much more than that. Shawn was involved in nearly all of my research by providing

knowledge, guidance, and support. Without the help of these three people, none of the achievements described here would have been possible.

The other members of my committee, Professor Guilherme Rosa and Professor Kent Weigel, provided suggestions and critiques that steered my research in fruitful directions and whose assistance helped to make my Ph.D. a success. Professor Henner Simianer welcomed me into his group for a productive and enjoyable three months in Germany and gave suggestions, assistance, data, and collaboration during my stay there. In addition, I would like to thank collaborators including Professor C. Robin Buell and Dr. Candice Hirsch for their assistance toward making progress on several of the projects contained here. Candice Hirsch deserves special recognition for the example that she set as a student, postdoc, and now professor, always a few years ahead of me.

Science does not always take place at a computer or in a lab. As I learned during my Ph.D., it can take place in a corn field as well. Dustin Eilert, Julie Smith, Jimmy Flannery, Marina Runge, and Bob Vogelzang all demonstrated this by displaying remarkable dedication to make field projects and experiments a success. Several of the projects described here rely on data and materials that would not exist without the hard work of these individuals. The undergraduate and graduate students who sacrificed their evenings or summers to generate important data in the lab or corn field were critical to these data as well. In addition, I have learned that the friendships and camaraderie that can develop while working long hours, seven days a week, in a corn field in the heat of July are unparalleled.

I would also like to thank my fellow graduate students and the postdocs who have provided ideas, inspiration, humor, and friendship during my time as student. My Wisconsin

office mates, Jillian Foerster, German Muttoni, and more recently Joe Gage, as well as my Goettingen office mate Mahmoud Gholami, have worked with me on projects, discussed science, and generally kept things fun during school. They even steered my back toward science on the multiple occasions when I was convinced a more desirable career path would be driving long-distance trucks for the freedom of the road. Although he wasn't an office mate, James Johnson was equally important as a scientist and friend during school. Additional students and postdocs, including Nick Haase, Scott Stepflug and Rajan Sekhon were also meaningful colleagues and friends during my Ph.D. Also, I assistant-taught several courses during my degree, and the feedback and support that I received from graduate students taking these courses was tremendously useful for my improvement.

I have also been extremely fortunate to receive support from a variety of organizations during my degree that made graduate school a possibility. Monsanto provided me with three years of funding for my degree, and my relationship with Donn Cummings of Monsanto is one that I am extremely grateful for. I received one year of support from the University of Wisconsin Graduate School, which enabled my fifth and most productive year. In addition, Synbreed, a European plant and animal breeding project, provided funding for my research in Goettingen, Germany.

Finally, professional and academic advancement often depends on a solid and stable personal life. In my case, I am extremely lucky to have family and friends that I can rely on, depend on, and confide in. My wife, Renee Miller, and her unwavering support of my goals has been a constant source of encouragement. Renee has often believed that I can meet deadlines or complete projects that I didn't think were achievable, only for her to be proven right in the end. My mom, Dr. Janet Beissinger, has served as an example of how to live a fulfilling and

meaningful life, both personally and professionally. Both Renee and Janet have also been willing to not only listen to me describe my research, but have provided remarkably useful suggestions regarding science, time management, and more. My siblings and brother in law, Travis, Jenny, Danny, Abby, and Peter, have also provided support, encouragement, and welcomed distractions during my studies. In addition, I am lucky to have a wonderful set of in-laws, including Renee's parents Don and Beth and her siblings Rachel and Ben. Renee's family is surprisingly interested in my research, so much so that they keep copies of my publications on hand and show (or at least pretend to show) genuine interested in the science described.

To anybody that I failed to mention, I thank you doubly. Not only have you provided support, but you have done so without being properly acknowledged. The difference between success and failure can be paper thin, so no-matter how small or large your support or assistance seems, it may be what made this work possible. I will be forever indebted to all of you.

ABSTRACT

The process of selection produces distinct patterns of genetic variability in the genomes of organisms. These patterns may be leveraged for the purpose of identifying genetic variants that contribute to the variation of the selected trait or traits. Analysis of polymorphisms as potential selection-sites can provide interesting insight into evolutionary processes and the principles that govern heredity. Herein, methods to generate and analyze this type of information are investigated, developed, and applied. First, genotyping-by-sequencing, a technique for identifying and genotyping single nucleotide polymorphisms among individuals, is implemented and explored. Results show that this approach can be cost effective and efficient, but that unless sequencing is conducted to a high coverage, information for most markers will only be available for a subset of individuals. Next, a maize population that was selected for an increased ear number is analyzed. Twenty eight regions that were highly divergent between the base and selected population and putatively under selection are identified. Further analysis suggests that the majority of selection operated on standing genetic variation, and that fixation of favorable alleles was a rare occurrence. Thirdly, a window-based method for analysis of genetic differences among populations is developed. This method is based on the derivatives of cubic smoothing spline functions. It is demonstrated to increase the power of studies that seek to identify selection based on population divergence information by defining appropriate windows for analysis. Fourth, a method for identification of loci that have been jointly affected by selection is implemented in a diverse panel of chickens representing 72 distinct breeds. An assortment of selective sweeps impacting segments of DNA are identified, as well as up to three instances of cis-acting epistatic selection. Finally, these methods are placed into the broader context of scans for selection. Methods for identifying selected regions of genomes are described, and the prospects and limitations generally relating to these studies are discussed.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

PREFACE

This dissertation describes research pertaining to the identification of genetic variation that has been subjected to selection. For more than a century, selection has been understood as a process capable of dramatically shaping populations. Recent technological advancements have improved genotyping and computational technologies, and improvements in our ability to isolate and quantify selected variants have followed. At the same time, the stochastic processes at play during the evolution of populations have come to be better understood as well. We know that identifying selected polymorphisms is not simply a matter of finding regions of genomes that appear different today than they did before selection began, but it is necessary to thoroughly rule out the possibility that non-selective forces have allowed such changes to occur.

In Chapter One, a specific enabling technological advancement in genotyping is investigated. This is a protocol known as Genotyping-by-Sequencing (GBS). GBS involves sequencing targeted fragments of a genome and identifying single nucleotide polymorphisms (SNPs) that vary between sequenced individuals. We evaluated the GBS protocol to determine an ideal implementation of the process for goals including quantitative trait locus mapping and genome-wide association studies (which are useful for corroborating evidence found from studies of selection). We highlight reasonable marker goals for these types of studies, as well as useful targets for sequencing coverage that will make these marker goals achievable.

Chapter Two describes the study of a maize population that was subjected to artificial selection for 30 generations for an increase in ears per plant. Whole-genome pooled sequencing was employed to generate genotype information, which is similar to the GBS protocol because it

is based on next-generation sequencing technologies, but different because the whole genome was sequenced instead of only a targeted portion of it. We characterized the variability of allele frequencies at 1.2 million marker loci between pools of individuals sampled from the population before and after selection took place. We identified 28 regions that exceeded an outlier-based threshold for empirical significance, and we investigated these 28 regions in depth. Overall, the selected regions suggested that soft sweeps (on standing variation) were common, fixation of favorable alleles was rare, and that non-genic material may play a functional role in determining the number of ears per plant.

Some of the limitations we encountered during the analysis of the data described in Chapter Two were addressed in Chapter Three. Pooled sequencing provided highly error-prone allele frequency estimates, which had to be grouped over "windows" to isolate patterns of selection. This grouping, however, was performed in a relatively unsophisticated manner that involved the averaging of single-SNP estimates over sets of 25 adjacent SNPs. Chapter Three describes a novel method designed to remove some of the ambiguous aspects of this approach. In brief, a statistical technique that involves the inflection points of cubic smoothing splines fit to the data may be used to define the size and boundaries of windows over which single-SNP estimates should be averaged. We demonstrate via simulation that this approach identifies more true positives relative to false positives than other window averaging techniques.

In Chapter Four, we shift from focusing on the identification of selection that has operated on individual loci to the identification of joint selection on pairs of loci. Epistatic selection is a phenomenon that can impact pairs of loci, so finding cases of epistatic selection was one of our primary goals. To achieve this, patterns of linkage disequilibrium between populations were investigated in a panel of chickens that represented 72 distinct breeds. We

identified and computed a novel null distribution to test these patterns. We found that while this method did identify at least one putative case of epistatic selection, it also isolated several selective sweeps that most likely correspond to selection on individual loci but that generated patterns affecting linked locus pairs.

Finally, a fifth chapter is included. This chapter provides a general description of the existing methods that are often used to identify selection from genotypic information, and it places the research described in the preceding four chapters into the broader context of selection scans as a whole. The benefits of using selection studies to better understand populations are described, as are the limitations of these studies.

Chapter 1 MARKER DENSITY AND READ-DEPTH FOR GENOTYPING POPULATIONS USING GENOTYPING-BY-SEQUENCING

Authors: Timothy M Beissinger[*,§], Candice N Hirsch[†,**], Rajandeep S Sekhon[*,‡], Jillian M Foerster[*], James M Johnson[*], German Muttoni[*], Brieanne Vaillancourt[†,**], C. Robin Buell[†,**], Shawn M Kaeppler[*,‡], Natalia de Leon[*,‡]

Affiliations: [*] Department of Agronomy, University of Wisconsin Madison, WI 53706
[§] Department of Animal Sciences, University of Wisconsin, Madison, WI 53706
[†] Department of Plant Biology, Michigan State University, East Lansing, MI 48824
[**] DOE Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI 48824
[‡] DOE Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, WI 53706

Sequence data have been deposited in the NCBI Short Read Archive (BioProject accession PRJNA169551).

Publication information: This article was selected by the editorial board as a highlight of the April, 2013 issue of Genetics, Volume 193, no. 4, pp. 1073-1081.

1.1 Abstract

Genotyping-by-sequencing (GBS) approaches provide low-cost, high-density genotype information. However, GBS has unique technical considerations including a substantial amount of missing data and a non-uniform distribution of sequence reads. The goal of this study was to characterize technical variation using this method and to develop methods to optimize read-depth to obtain desired marker coverage. To empirically assess the distribution of fragments produced using GBS, approximately 8.69 Gb of GBS data was generated on the *Zea mays* reference inbred B73 utilizing *Ape*KI for genome reduction and single-end reads between 75 and 81 bp in length. We observed wide variation in sequence coverage across sites. Approximately 76% of potentially observable cut site-adjacent sequence fragments had no sequencing reads whereas a portion had substantially greater read depth than expected, up to 2,369 times the expected mean. The methods described in this manuscript facilitate determination of sequencing depth in the context of empirically defined read-depth to achieve desired marker density for genetic mapping studies.

1.2 Introduction

High-density genotypic information on large numbers of individuals is crucial for quantitative trait locus (QTL) mapping and association analysis. Cost-efficiency is an important component in generating genotypic data. Previous genotyping methods include markers such as microsatellites, amplified fragment length polymorphisms (AFLPs), and restriction fragment length polymorphisms, among others. The relatively high cost and limited marker density of these methods led to the use of single nucleotide polymorphisms (SNPs) as the current preferred genotyping system. As such, a large number of array-based SNP genotyping platforms are currently available (reviewed by Fan *et al.* 2006) as well as targeted or whole-genome sequencing-based technologies.

Sequencing-based approaches to SNP allele calling include whole-genome sequencing (Hillier *et al.* 2008), exome capture (Ng *et al.* 2009), RNA sequencing (Hansey *et al.* 2012), methylated DNA sequencing (Brunner *et al.* 2009), and restriction enzyme (RE) digestion (reviewed by Davey *et al.* 2011). RE-based approaches include restriction-site associated DNA sequencing (RAD-seq; Baird *et al.* 2008), complexity reduction of polymorphic sequences (CRoPS; van Orsouw *et al.* 2007), and genotyping-by-sequencing (GBS; Elshire *et al.* 2011). All of these methods represent efficient and cost-effective approaches to produce genetic information, but differ in their implementation. For instance, RAD-seq and GBS both involve sequencing DNA fragments adjacent to RE cut sites, yet while RAD-seq involves sequencing these fragments to high coverage, the focus of GBS is to sequence with low target coverage. Alternatively, CRoPS is based on sequencing DNA fragments that were originally generated as AFLP markers. In this study, we utilized the GBS protocol of Elshire *et al.* (2011) with minor modifications.

In brief, GBS utilizes RE digestion to preferentially target sites in low copy genomic regions, minimizing reads in repetitive sequences which are abundant in maize (Schnable *et al*. 2009) and which produce ambiguous SNP information. *Ape*KI is a RE frequently used for GBS in maize because it cuts retrotransposons infrequently and is partially methylation-sensitive, thereby preferentially generating fragments from low copy genic regions (Elshire *et al.* 2011), but additional enzymes can also be used for GBS. Cost-efficiency is achieved by multiplexing barcoded individuals (Baird *et al*. 2008; Elshire *et al.* 2011). Analysis pipelines rely on mapping the resulting fragments to a reference genome, if available. Otherwise, linkage relationships can be used to genetically map sequenced DNA in organisms without a sequenced reference genome. Recently, numerous researchers have successfully employed RE-based genotyping protocols to develop maps and/or map QTL in such species (Chutimanitsakun *et al.* 2011; Amores *et al.* 2011; Pfender *et al.* 2011; Baxter *et al.* 2011; Poland *et al.* 2012).

Although all of the sequencing-based genotyping strategies are capable of generating substantial amounts of data in a cost-efficient manner, however peculiarities of the genome structure can limit the utility of the data. For instance, features such as the presence of repetitive DNA make the unique alignment of sequence reads difficult or impossible. This is particularly problematic in plant species because of the high proportion of repeats frequently present (Treangen *et al.* 2011). Similarly, repetitive DNA that is not accounted for in reference genomes may allow repetitive sequences to be aligned uniquely and therefore cause false polymorphisms to be identified. Lastly, differences in guanine-cytosine (GC) content and other potential sources of sequencing biases can leave important genomic regions under or over represented (Minoche *et al.* 2011).

Addressing this, theoretical work has been done to determine the expected sequence coverage obtained from these technologies (Lander and Waterman, 1988; Wendl *et al.* 2006). Also, studies investigating desirable marker coverage for genotype-phenotype associations in the context of classical genotyping technologies have been performed (Piepho 2000). Herein, we describe theoretical and empirical considerations of using GBS (Elshire *et al.* 2011) for genetic analysis, with the goal of determining reasonable marker expectations and corresponding resource investments. GBS was conducted on the maize reference inbred, B73, using replicated DNA samples, barcodes, and independent sequencing lanes to gather empirical information. We then compared the theoretical coverage distribution to the actual distribution that was obtained through GBS. Next, we developed a theoretical tool to determine appropriate marker number for QTL mapping in bi-parental populations, as well as assessed marker number required for association mapping in diverse inbred populations. Lastly, we provided recommendations for the target number of raw sequence reads that should be generated to attain an effective density of markers. Although our results pertain to GBS in maize, simple adjustments make our techniques potentially applicable to a wide variety of protocols and species.

1.3 Materials and Methods

Library construction, sequencing, and read mapping: DNA was isolated from pooled leaf tissue from five to ten seedlings of the reference inbred line B73. Genomic DNA was extracted using a modified CTAB method (Saghai-Maroof *et al.* 1984). Multiple DNA extractions from B73 tissue were performed. Next, extracted DNA was barcoded and pooled following the procedure described by Elshire *et al*. (2011), with an additional gel-based size selection step to enrich for fragments of intermediate size. The size selection was incorporated because Illumina

reports that to optimize cluster formation the ideal fragment size range for single end libraries is 150-300 bp (http://www.illumina.com/support/faqs.ilmn). Additionally, the size selection step allows for further reduction of the effective genome size. Sequencing was conducted using Illumina TruSeq SBS 36 bp kits, versions 3-5, on eight lanes of the Illumina Genome analyzer II (GAII; Illumina, Inc; San Diego, CA, USA) at the University of Wisconsin Biotechnology Center (Madison, WI). For each library, 48 barcoded samples were pooled. The eight lanes of sequences were generated over multiple sequencing runs that were run to variable read lengths. For each lane of sequence, read quality was evaluated based on the Illumina purity filter, percent low quality reads, and distribution of Phred-like scores at each cycle. Lanes that had a lower quartile Phred-like score less than 20 prior to base 40 were not included in this analysis. Individual reads from sequencing lanes that passed this quality control (four lanes of 75bp reads, one lane of 76bp reads, and three lanes of 81bp reads) were then cleaned using the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html) with the fastx_clipper program requiring a minimum length of 20 bp after clipping. After running fastx_clipper with both adapter sequences across the eight lanes, 84% of the reads were retained with a minimum of 80% retention from one of the lanes. Sequences from each lane that passed this filtering were parsed to remove the barcode sequences using a custom Perl script requiring a perfect match to the barcode and the *Ape*KI cut site (i.e. GC[A/T]GC). Cleaned reads were mapped to the maize B73 version 2 pseudomolecules (AGPv2; http://ftp.maizesequence.org/; Schnable *et al*. 2009) using Bowtie version 0.12.7 (Langmead *et al.* 2009) allowing up to two mismatches and requiring a single best alignment to nuclear DNA.

Computational digestion and analysis: A custom Perl script was used to identify all *Ape*KI cut sites, irrespective of methylation state, in the maize B73 version 2 pseudomolecules (AGPv2; http://ftp.maizesequence.org/; Schnable *et al*. 2009) and the expected fragment sizes including the GC[A/T]G overhangs were determined. GC content for a 50bp window up- and down-stream of the cut sites excluding the GC[A/T]GC sequence was determined using a custom Perl script. All subsequent analyses were performed using standard functions implemented within R version 2.13 (R Development Core Team, 2011).

QTL Mapping: QTL mapping was performed for two data sets. In the first, previously published data from 283 individuals of the maize intermated B73 x Mo17 (IBM) population was analyzed (Eichten *et al.* 2011). The phenotype under study was plant height, and the total set of markers included 1,340 simple sequence repeats (SSRs). These markers were not generated using GBS, but the data set was chosen because the total number of SSRs and relative spacing of the markers approximates what could be expected per-individual from a low-coverage deployment GBS. First, QTL mapping was conducted with the full dataset using composite interval mapping with the software program R/qtl (Broman *et al.* 2003). The analysis included five covariates selected using forward selection, and the LOD-threshold was determined according to a Bonferroni-corrected 0.05 significance level. The total set of identified QTL was recorded. Next, randomly chosen marker subsets were used for QTL mapping. Subset sizes ranged from 100 markers to 1,300 markers, in increments of 100. For each marker subset size, QTL mapping was repeated 1,000 times with randomly selected marker subsets of the specified size. The proportion of QTL in the total set that were identified in each run was recorded.

Finally, the power of mapping with additional markers was evaluated based on the mean proportion of expected QTL that were identified at each marker subset size.

The second experiment involved a simulation study representing a non-intermated maize recombinant inbred (RI) population with 250 individuals. Ten QTL of equal effect were simulated, one on each chromosome, with an overall heritability of 0.5. A set of 11,917 markers were simulated, corresponding to what could be expected from a higher coverage deployment of GBS. QTL mapping was conducted in the same manner as for the plant height data described above, except in this case the true underlying QTL were known based on the simulation. Mapping was conducted at marker subset sizes ranging from 100 to 1,000, by 100, and then from 2,000 to 11,917, by 1000. Again, the power of additional markers was evaluated based on the mean proportion of known QTL that were identified at each marker subset size.

Determining appropriate target coverage for mapping purposes: A bootstrapping scheme was developed to determine the genotyping resources required to obtain reads from a specified number of distinct RE fragments for an individual DNA sample. First, the set of all B73 fragments that were successfully aligned to the B73 reference genome was considered representative of the comprehensive set of all sites with the potential of being both sequenced and aligned. Next, increasing numbers of fragments were sampled from this set, with replacement. The probability of sampling each fragment was made proportional to the number of times it was actually observed. Next, the number of additional unique fragment reads that were obtained at each round of sampling was counted to estimate how many total reads are required to obtain a desired number of distinct fragments per individual.

1.4 Results

Repeated sequencing of the maize inbred line B73: In total, we generated over 118 million GBS sequence reads from the reference inbred B73 to determine the distribution of reads throughout the genome. These reads corresponded to approximately 8.69 Gb of B73 sequence data before adapter and barcode clipping. There are approximately 3.9 million *Ape*KI sites in the B73 reference genome and our sequencing approach had the potential to capture up to 77 bp on either side of each cut site (up to 81 bp per read minus the 4-8bp barcode). It is expected, consequently, that this sequencing should provide approximately 14.3X coverage of the *Ape*KI target space assuming no size selection ($\frac{8,690,000,000}{3,936,260\times77\times2} = 14.3$). However, due to our additional step of size selection and technical bias for smaller fragments by the Illumina procedure, the expected coverage per observable position was substantially greater. Because gel-based size selection cannot perfectly isolate fragments of a particular size, we empirically estimated the experimentally-optimal fragment size by observing that 95% of the observed sequencing reads resulted from *Ape*KI fragments between 70 and 318 bp in length (here, we define *Ape*KI fragments as DNA segments between *Ape*KI cut sites). Since there are approximately 1.4 million optimally-sized *Ape*KI fragments predicted from the B73 reference genome, our sequencing provided an expected coverage of approximately 40.1X over the *Ape*KI reduced and size selected space ($\frac{8,690,000,000}{1,406,269\times77\times2} = 40.1$).

Using an informatics pipeline that allowed up to two mismatches in a read to map to the reference, we found that 43.9% of the B73 fragment reads had a single best-alignment to the reference, 46.7% could be aligned to multiple positions, and 9.3% could not be aligned to the

reference at all. The B73 fragments that did not align to the reference likely resulted from the requirement imposed of two or fewer mismatches for a read to be mapped, in the context of the relatively high error rate of the Illumina techonology used (Luo *et al.* 2012). Recent advances in Illumina sequencers, such as HiSeq, have reduced error rates relative to previous technologies which will improve the proportion of reads mapped. More permissive alignment algorithms may also reduce the proportion of fragments that cannot be aligned, but could increase spurious alignments in complex genomes such as maize. From the 3,936,260 potential *Ape*KI cut sites identified in the reference genome, only 35.7% (1,406,269) were expected to generate at least one fragment in the optimal size range of 70 bp to 318 bp. It was found that 27.4% (384,887) of these cut sites had at least one sequence read on at least one side of the cut site with a unique alignment, though some were sequenced many more times. Additionally, we obtained 174,954 uniquely aligned reads from *Ape*KI fragments that were larger or smaller than the 70 to 318 bp range (Figure 1.1). Lastly, we were able to uniquely align reads from 52,123 sites that were not predicted to be *Ape*KI-site-adjacent based on requiring a perfect cut site sequence in the reference genome. These unpredicted cut sites could be the result of errors in the reference, less than 100% restriction accuracy of *Ape*KI, or differences between our B73 source and that used to produce the reference sequence.

The number of reads from each fragment is expected to follow a Poisson distribution (Lander and Waterman, 1988), however the empirical data did not follow such expectation (Figure 1.2). Additionally, the median number of reads per sequence fragment was zero, which was substantially less than expected (approximately 40). The likely cause for this deviation is that many of the sequence fragments were observed thousands of times more than expected (up to 95,014 reads from a single end of an *Ape*KI fragment). At the same time, a disproportionately

large number (1,021,382 or 72.6%) of the predicted *Ape*KI fragments of size 70 to 318 bp had no observed sequencing reads from either end (Figure 1.2).

Analysis of the over-represented fragments relative to the B73 organelle reference genome revealed that a subset of these fragments were also present in DNA from organelles. Hence, they correspond to historical insertions of organellar DNA into the nuclear genome, an occurrence that has been documented previously (Lough *et al.* 2008). Other over-represented fragments are likely due to repeats in the maize genome that include *Ape*KI cut sites that were collapsed into a single, non-repetitive segment in the reference sequence. In these instances, the repeated fragments were mapped uniquely because the reference genome does not capture their repetitive nature. Highly over-represented sites (greater than 500 reads per site, corresponding to nearly a 0% probability based on the expected distribution) represented 0.5% of the sequenced *Ape*KI sites, but accounted for 41.7% of the total reads.

To further investigate the reason for the highly overrepresented sites, potential biases due to fragment GC content was explored. High or low content of GC within a fragment can affect read depth using Illumina sequencing technologies (Minoche *et al.* 2011). Technological advances have reduced this bias over time, but nevertheless, fragments extremely high or low in GC content are likely to be underrepresented. We assessed the GC content in windows of 50 bp adjacent to cut sites. The mean level of GC for these windows across B73 was 53.9% with a range of 0% to 100% (Figure 1.3A). It was observed that sites with GC content between 40% and 50% were more frequently sequenced using Illumina GAII. When GC content was outside of a seemingly ideal 10% - 70% range, the mean number of times that sites were sequenced decreased from 12.8 to 0.46 (Figure 1.3B). But, regions with such high or low GC accounted for less than 7% of all optimally-sized sequencing fragments in B73. Therefore, GC content bias

could explain some of the rarely-observed sites but such bias does not account for the highly over-represented sites.

The skewed distribution of sequencing reads we observed in B73 is not unique to this inbred line. Across three RI populations and an association panel (totaling approximately 1,500 diverse inbred lines), for which we performed a comparable GBS protocol, many of the same positions that were over represented through sequencing B73 also had disproportionately high coverage across the set of diverse lines (data not shown).

Marker number for QTL mapping: It is desirable to optimize marker density to maximize the efficient application of genotyping technologies. For the purpose of QTL mapping in bi-parental populations, marker density can be optimized by first recognizing that adding markers to stretches of the genome that correspond to the same parental genotype does not provide additional information. On the other hand, when markers flank a recombination event, additional markers in the region will increase resolution of the recombination position. Still, if the recombination event is not in the region where a true QTL lies, there is no practical use for increased marker density in this region. Thus, the probability of having both a QTL and a recombination event occur between two markers should be minimized (Figure 1.4).

Consider an individual with c chromosomes, n evenly spaced markers, r recombination events, and q QTL. A lower limit, $p_{0,}$ on the probability of both a recombination event and a QTL not occurring in a region with unknown parental genotype (i.e. between two markers) is given by:

$$p_0 = \frac{(n-r+c)^q}{(n+c)^q} \quad (1)$$

To derive (1), consider the available markers as dividing the target DNA into $n + c$ bins, which are the spaces between two markers or between a marker and a chromosome end. The numerator of (1) is the number of ways the QTL can be placed into these bins such that they are not in a bin where a recombination event occurred, while the denominator of (1) is the total number of ways that QTL can be placed into bins. An implicit assumption is that the probability of a QTL being located in any of the bins is equal, which is met when markers are evenly spaced. Although GBS markers are not perfectly evenly spaced, there is not large-scale clustering of *Ape*KI cut sites throughout the genome. More specifically, no one chromosome or chromosomal region has an abnormally high or low concentration of cut sites. This means that even spacing is a reasonable approximation on a genome-wide basis. Also, this estimation provides a lower limit because (1) assumes that every recombination event occurs in a unique bin—if there happen to be multiple recombination events in any single bin the numerator will be slightly increased. To obtain the number of markers (n) that will provide an expected proportion of p_0 individuals without any particular QTL flanked by markers of alternate parent genotypes, one may solve the above formula for n, with q=1, which is given by:

$$n = \frac{r-c(1-p_0)}{1-p_0} \quad (2)$$

Equation (2) was verified based on two QTL mapping experiments. The first experiment consisted of previously published data on plant height in the IBM population (Eichten *et al.* 2011), while the second was based on simulations for a non-intermated RI population of 250 individuals. In both cases, the optimal number of markers suggested by (2) was within 200 of the observed marker number that provided a maximally powerful test, and the difference between

the power of mapping at the marker level provided by (2) was only marginally lower than at the maximum (Figure 1.5). Furthermore, this confirms that the assumption of even spacing is robust against minor violations, as the IBM markers used were not evenly spaced. Interestingly, the simulated QTL mapping experiment data depicted a slight decrease in power with an unnecessarily dense set of markers. This is likely attributable to the Bonferroni correction for LOD threshold that was implemented.

Marker number for mapping with recombinant inbred and association populations: RI populations are commonly used for QTL mapping. Application of equation (2) requires knowledge of the genome-wide recombination rate (r) and the number of chromosomes (c). In the maize IBM RI population, for example, the average number of effective recombination events per individual is 57 (Fu *et al.* 2006). Since the IBM population was intermated four times before selfing (Lee *et al*. 2002), this value can be scaled to non-intermated standard RI populations by first multiplying by the reciprocal of the genetic map expansion factor incurred during the development of the IBM population and then by the expansion factor incurred during the development of a standard RI population. Respectively using the expansion factors $j/2 + (2^i-1)/2^i$ and $(2^{i+1}-1)/2^i$, for lines that have been inbred for i generations after being intermated for j generations (Teusher *et al.* 2006), on average, there are approximately 38 recombination events per individual in a non-intermated maize RI population.

Application of (2) to the IBM RI population given that $r = 57$ and $c = 10$, and allowing an expected $p_0 = 95\%$ of individuals without uncertain genotype at each QTL, we determined that the marker goal across the genome is $n = 1{,}130$. For a standard RI population, applying (2) with

the same parameters except that r = 38 results in the genome-wide marker goal being 750. We emphasize that these marker number requirements represent the minimum number of markers needed to produce an expected 95% of individuals with known genotypes at any particular QTL.

It is important to note that in RI populations, the genotyped markers need not be the same for each individual. In the case where several of the markers typed on each individual are distinct, as occurs in GBS, imputing markers that were not observed is the appropriate action. In the case of hundreds of individuals, each with mostly different markers typed, this will lead to a large proportion of imputed markers for the population. However, as long as the number of markers that were observed for each individual meets the values described above, mapping will have near the maximum power possible for the particular population under study. Moreover, although imputing between observed markers allows comparisons to be made between individuals with observed genotypes at different markers, it cannot increase mapping resolution (Figure 1.4). Therefore, a distinction must be made between the total number of GBS markers generated and a value that we deem the 'effective number of markers'. The effective number of markers for an individual is the number of markers with observed genotypes. Thus, a RI line typed at n markers but with unobserved genotypes at p markers before imputation, has an effective marker number with respect to resolution of n, not of $n + p$.

Also, the marker number suggested by (2) should be viewed as a minimum number of effective markers for mapping purposes. Additional markers will provide only minimal increased power to detect QTL, but they will reduce the proportion of individuals with uncertain parental genotypes due to recombination events near the QTL to fewer than the expected value of $1-p_0$. It is important to highlight that based on the estimation that maize contains approximately 39,500 genes (AGPv2; http://ftp.maizesequence.org/; Schnable *et al*. 2009), the marker numbers

suggested here will generate maps that have, on average, approximately 35 genes between markers in the IBM population, and approximately 53 genes between adjacent markers in non-intermated RI populations. However, improving QTL mapping resolution requires not just additional markers but also an increase in recombination events, which can only be achieved with an altered population structure or size.

Similarly to QTL mapping with a RI population, the goal of association mapping using a diverse set of inbred lines is to associate genotypes with phenotypes. However, these methods differ in that RI populations have simpler structure of relatedness generated by the expected recombination of regions originated from the two parents of the population, whereas more complex structure is present in a diverse population. Similarly, greater levels of linkage disequilibrium (LD) are expected to be present in RI populations compared to diverse populations. Therefore, it is well established that more markers are required for association mapping in order to capture markers within historical blocks of LD compared to structured bi-parental populations. In maize, for instance, it has been suggested that association mapping should be conducted with SNP markers spaced every 100-200 bp (Tenaillon *et al.* 2001). In other species, the required density of SNP markers for effective association mapping is dependent upon the level of historical LD.

Determination of read-depth required to achieve desirable marker density based on an empirical distribution of read coverage per marker position: The number of unique fragment reads that should be expected for a given total number of fragment reads was quantified to determine the optimal depth at which sequencing should be performed to achieve a desirable

marker coverage based on an empirically-determined distribution of reads. The quantification is based on a resampling of the data we generated from the maize reference inbred B73. The approach utilized is likely to provide a slight bias towards B73 fragments. For other lines or species, the proportion of reads that can be aligned to the reference genome is likely to vary. An adjustment for the proportion of fragments that can be aligned is needed for this method to be globally applicable to other maize lines, and a repetition of the process will apply for other species.

To quantify optimal target depth in a RI population, the number of unique fragments required per individual will vary with target marker number and average SNP density between the parents. But, given a target number of unique fragments per individual, a recommendation for the total number of reads that should be obtained for each sample is provided (Figure 1.6). Based on resampling observed B73 fragments, the expected number of unique fragments sequenced for a given number of total fragments sequenced incorporates the uneven coverage distribution. The focus is on unique fragments because these have the potential to contribute additional information. However, if the sequencing technology used is error prone, repetitive sequencing of sites may be required, and the approach utilized here can be modified to evaluate the expected number of fragments sequenced a specified number of times for a given total number of fragments sequenced.

As described above, diverse association panels require substantially more markers than do RI lines for effective mapping. With millions of *Ape*KI sites in maize, GBS based on that RE seemingly has the potential to generate marker densities near the target. However, based on the uneven coverage of sites that we observed, GBS would have to be performed with substantially greater depth than calculated simply by reads divided by target sites to obtain information at the

majority of the desired sites. For instance, from the approximately 8.69 Gb of B73 data generated in our study from 118 million reads, only 559,841 unique *Ape*KI fragments out of the 1.4 million expected to pass our size selection step were successfully sequenced and aligned. It appears that the additional approximately 840,000 fragments had an extremely low probability of being captured through sequencing. Given this, and the fact that LD decays over a span of only a few hundred bp in maize (Tenaillon *et al.* 2001), relying on downstream LD-based imputation for those sites that were missed is expected to be relatively ineffective. Instead, a reasonable approach is to minimize the amount of missing data by sequencing fewer sites at a higher target coverage, taking into account the variable sequencing depth that will be observed. Our resampling analysis suggests that using *Ape*KI-based GBS in maize, genotyping with a target of 23, 41, or 80 million reads is expected to result in missing data at approximately 30%, 20%, or 10% of sites, respectively, for a given individual (Figure 1.6). Determination of the appropriate target number of sequence reads in different species or by the use of alternative sequencing-based genotyping methods can be achieved by first sequencing a representative individual at high coverage and subsequently performing empirical resampling to identify the point of adequate coverage as suggested here.

1.5 Discussion

We have shown that the coverage of different sites throughout the maize genome as captured through the *Ape*KI-based GBS protocol is highly variable, although the reasons for the extreme variability are only partially understood. Therefore, sequencing approaches that succeed even when coverage is variable, or approaches that reduce the uneven coverage, are necessary.

Alternative sequencing approaches for genotyping individuals are abundant, including GBS with different or multiple enzymes (Poland *et al.* 2012), RAD-seq (Baird *et al.* 2008), and CRoPS (van Orsouw *et al.* 2007). In situations where highly variable levels of coverage are still observed, the strategy proposed here first operates on a single individual (B73 in this case) to be sequenced extensively. The variability of site-coverage in this individual will approximate the variability yet to be generated from later individuals. From the full set of sequenced fragments obtained from the first individual, including repeated fragments, random computer-based subsamples are drawn, with replacement. These are evaluated for the amount of additional sites observed as subsample size increases. The subsample size that provides enough site information for the desired marker number or level of missing data dictates the coverage that should be targeted.

Carrying out this strategy in maize suggested, for example, that acquiring 300,000 unique fragments per individual can be obtained by sequencing approximately 3.6 million total fragments per individual (Figure 1.6). But because of the variable coverage distribution, doubling the number of acquired unique fragments to 600,000 requires a more than nine-fold increase in total fragments sequenced, to approximately 27.9 million. Moreover, our results show that in order to minimize the level of missing data across individuals even more sequencing per individual is required (Figure 1.6). However, these target sequencing levels will vary and potentially be reduced in species with less repetitive genomes.

Although the sequencing coverage required to generate a dense marker set seems daunting, we have demonstrated that a substantial marker density is not required, or even useful, for the purpose of QTL mapping in RI populations. For this type of population structure, desirable marker coverage is given by Equation 2. Representative populations suggest that the

number of markers for efficient QTL mapping in bi-parental populations is on the order of hundreds to thousands. Substantially larger numbers of markers would be necessary, however, when association mapping is being conducted in a diverse population.

In summary, performing GBS on the maize inbred line B73 produced a highly skewed coverage of genomic positions, which is only partially accounted for by GC bias and duplicated positions. The result of the uneven coverage distribution is that no information is available at the majority of positions for which information was initially expected. Still, our findings suggest that even at relatively low coverage, GBS can produce enough information for powerful QTL mapping in bi-parental populations. However, obtaining dense genotyping resolution for downstream fine-mapping will require increased target coverage per individual. Using the method for association studies in maize, for example, will require genotyping at substantially greater target coverage. Therefore, researchers must be aware that in complex genomes, using simple approximations and standard distributions to determine target coverage vastly underestimates the sequence depth required to generate adequate data for complex analyses. Before large-scale sequencing commences, empirical enumerations of target coverage that account for potentially complicated genome compositions will lead to more complete and useful datasets relative to study goals.

1.6 Acknowledgements

This work was funded by the DOE Great Lakes Bioenergy Research Center (DOE BER Office of Science DE-FC02-07ER64494). T.B. and J.F. were supported by a gift to the University of Wisconsin-Madison Plant Breeding and Plant Genetics program from Monsanto.

G.M. was supported by a fellowship from DuPont-Pioneer Hi-Bred International, Inc. J.J. was supported by Hatch funds from the National Institute of Food and Agriculture, United States Department of Agriculture Project WIS01330.

1.7 Figures

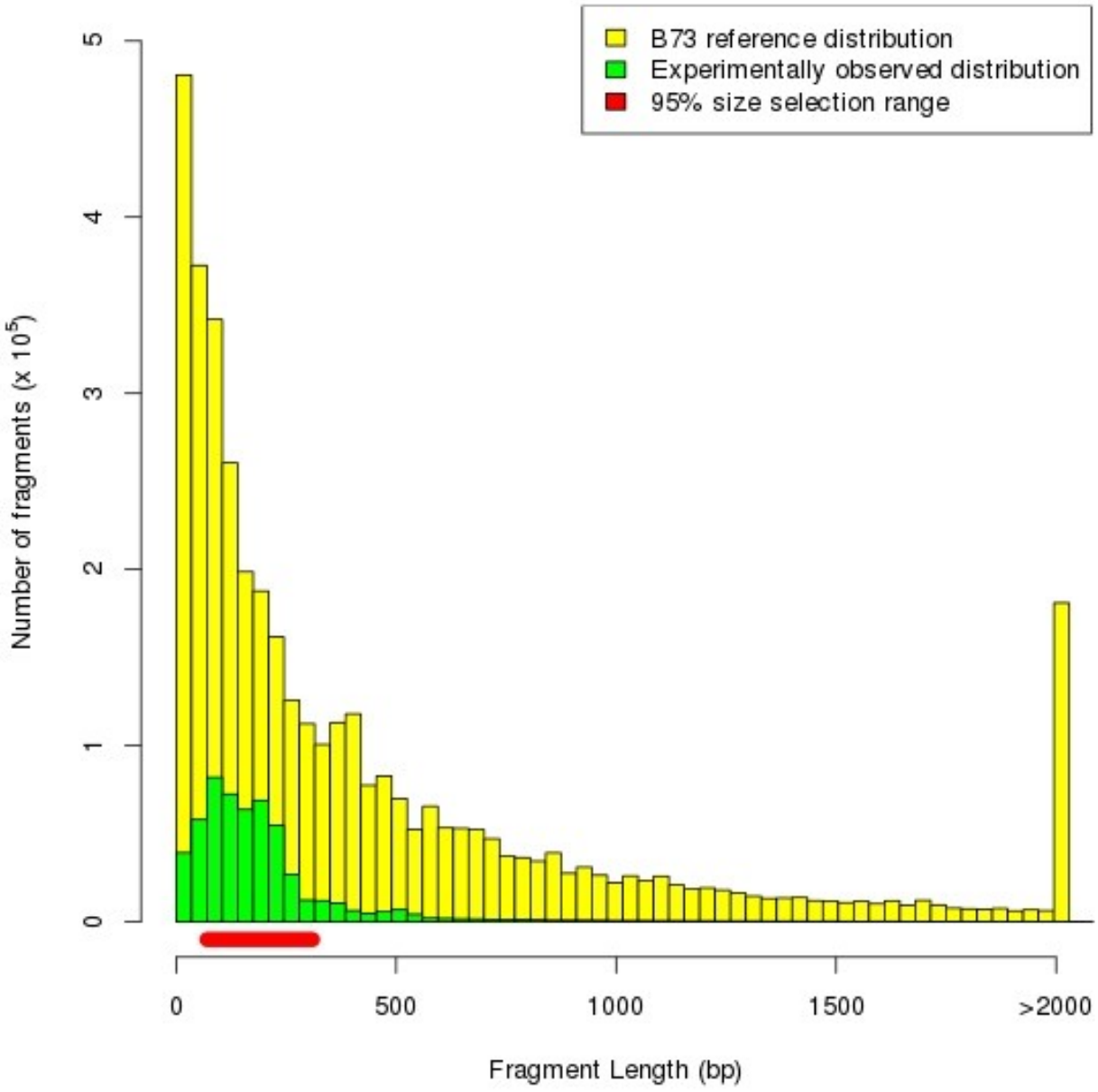


Figure 1.1: Distribution of the length of B73 ApeKI fragments

The length of ApeKI fragments expected based on an analysis of the reference genome and experimentally observed from approximately 8.69 Gb of B73 DNA sequence reads.

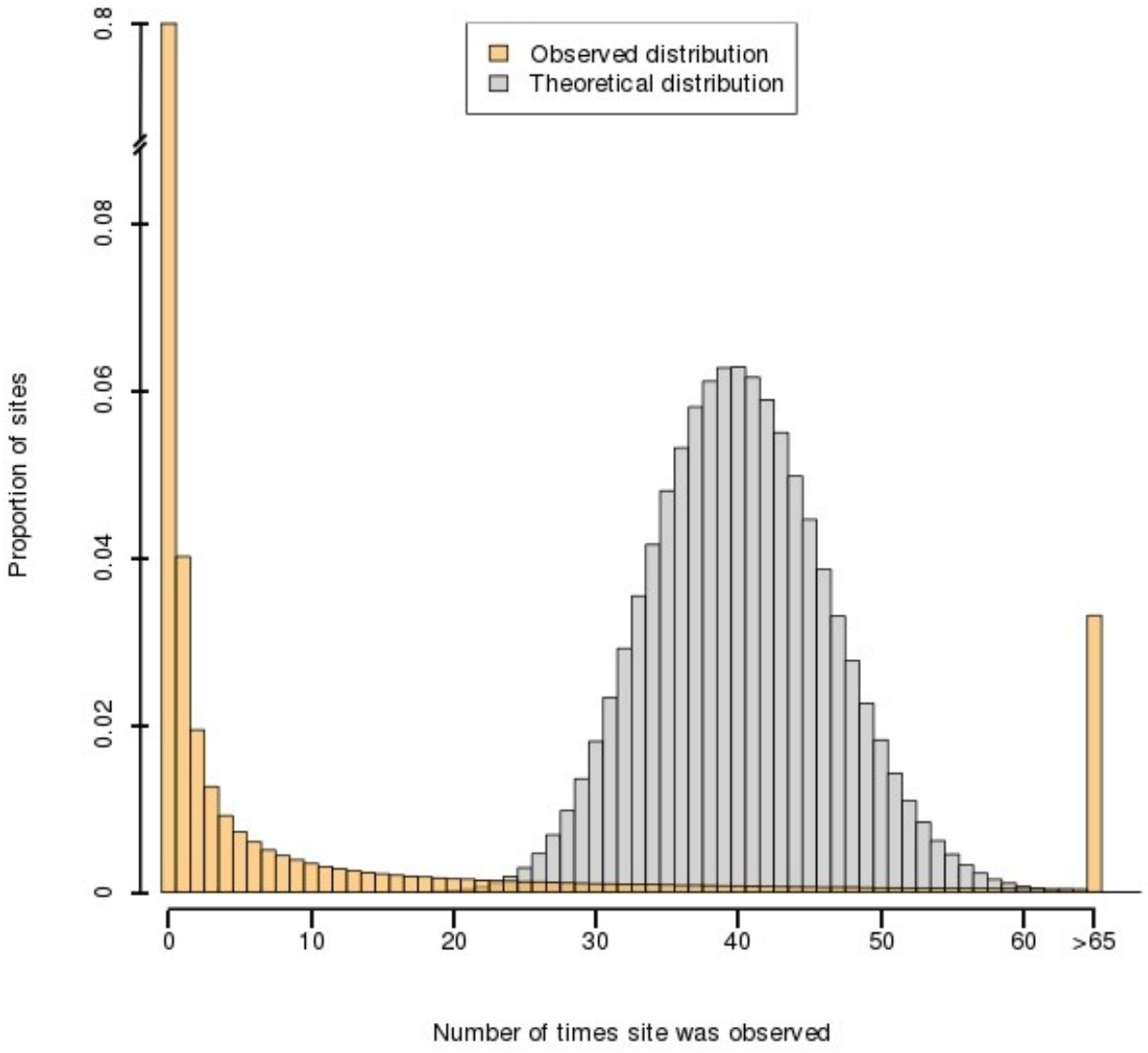


Figure 1.2: Observed and theoretical frequency distributions

The number of times that optimally-sized B73 ApeKI fragments were sequenced. Note the break in the vertical axis. "Sites" refers to DNA segments from either end of an ApeKI fragment. The number of reads per site is expected to follow a Poisson distribution with mean equal to the average coverage.

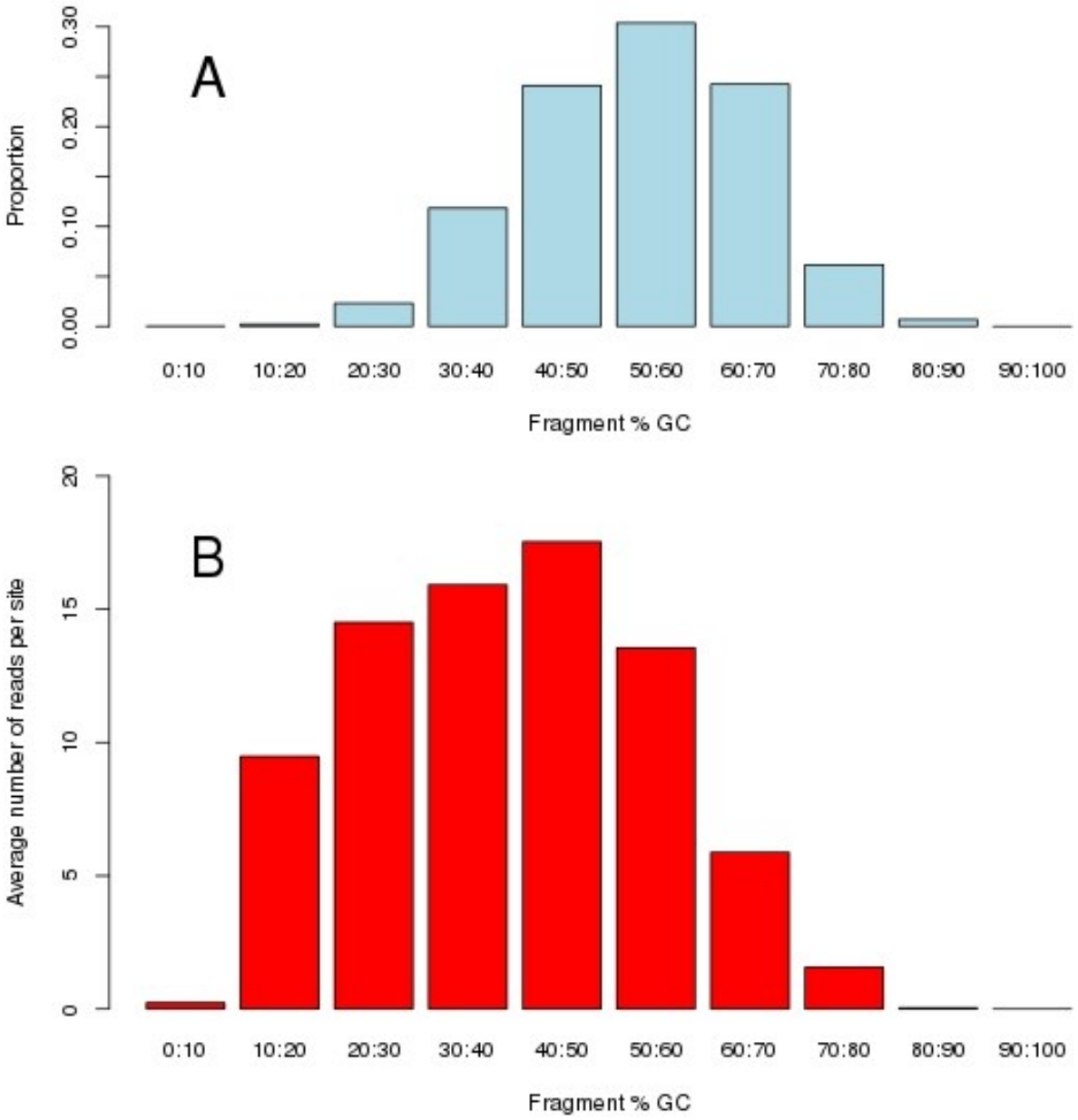


Figure 1.3: Distribution of GC content and coverage of optimally-sized (70-318 bp) sites.

A: The proportion of optimally-sized sequencing fragments with the specified GC content (computationally determined by analysis of the reference genome). B: Mean number of reads for optimally-sized B73 sequencing fragments with given GC content. Extremely high or low GC content negatively impacted read number per site, but the majority of fragments are in the intermediate GC range.

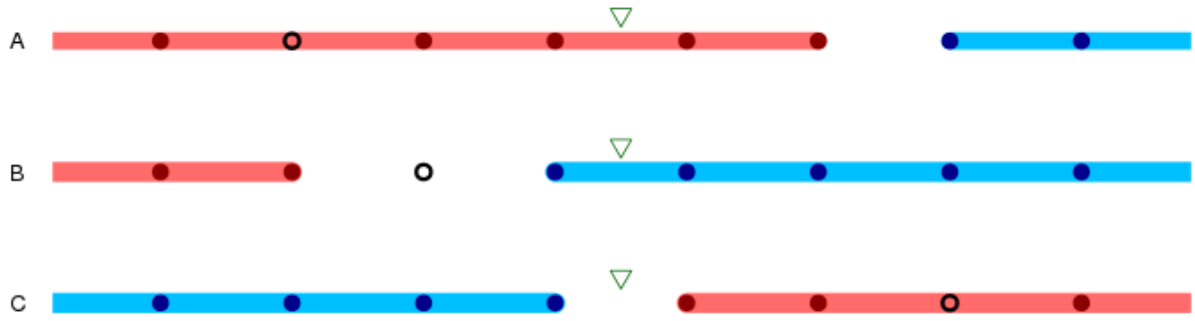


Figure 1.4: An example of genotypes for three hypothetical RI lines

Consider three RI lines, A, B, and C. Red circles correspond to observed marker genotypes from one of the parental lines, blue circles correspond to observed maker genotypes from the other parent, and open circles correspond to missing marker information. Red and blue shading illustrates that between two markers of the same parental genotype, genotypes can be inferred with great accuracy, even in the case of a missing marker genotype. However, genotypes between markers of alternate parental types remain unknown. The green arrows show the location of a 'true' quantitative trait locus (QTL). Notice that line C has unknown genotype at the QTL and therefore does not add power to a statistical test for QTL identification (although this individual would be particularly useful for downstream fine-mapping). Equations (1) and (2) provide the number of markers needed for the probability of occurrence of case C to be minimized.

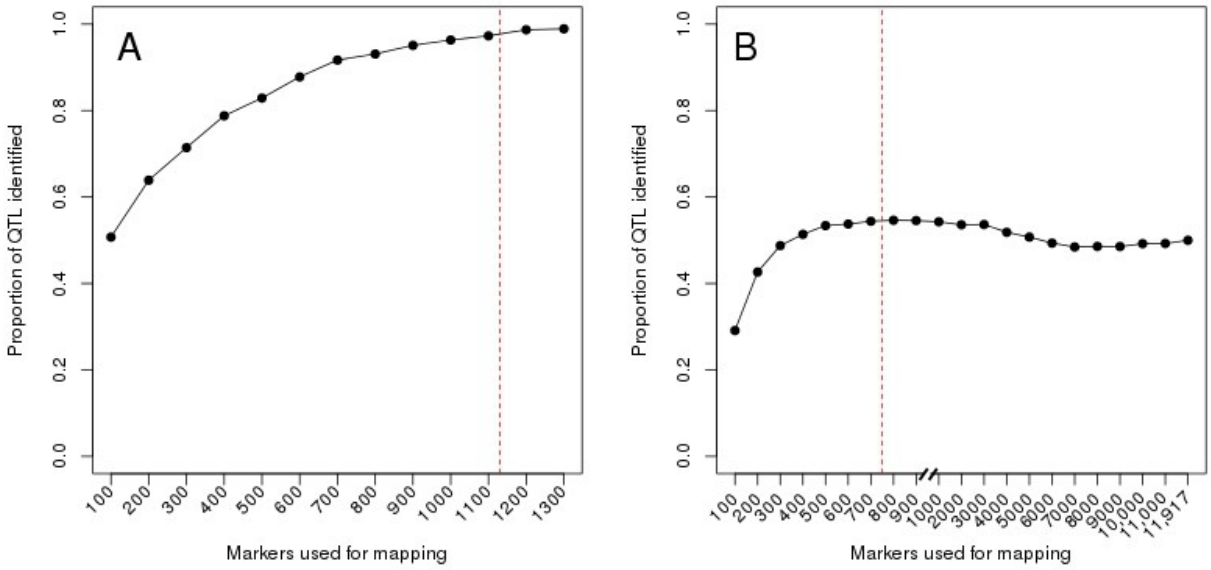


Figure 1.5: Validation of marker number estimate.

Two quantitative trait loci (QTL) mapping studies were performed to validate Equation (2), which estimates the number of markers required to maximize the power of a bi-parental QTL mapping study based on the number of chromosomes and level of recombination in a population. Depicted in both A and B is the mean proportion of QTL identified from 1,000 replicated mapping experiments at each marker subset level. A: For the Intermated B73 x Mo17 (IBM) RI population, the maximum number of QTL that could be identified was three, which was the number identified from mapping with the full data set. B: For the simulated RI population, which was not intermated before inbred development, the maximum number of QTL that could be identified were all 10 QTL simulated. In each plot, the red line depicts the number of markers suggested by Equation (2). For both experimental data from the IBM RI population, as well as data from a simulated non-intermated RI population, Equation (2) closely approximates the ideal marker number for maximal QTL identification.

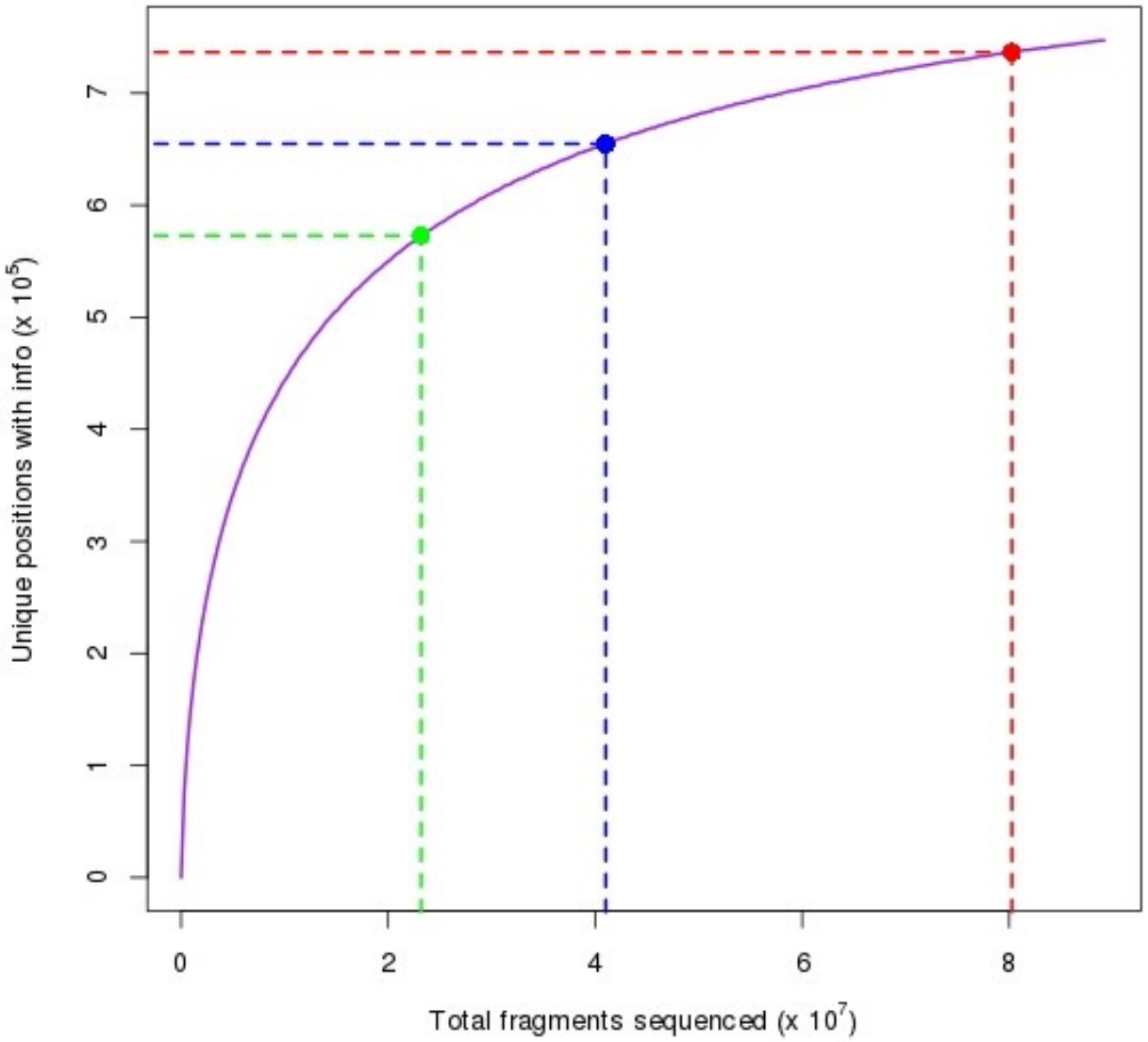


Figure 1.6: Resampling method to determine target sequencing depth.

A resampling analysis was conducted to determine the number of total fragment reads needed to achieve desirable levels of coverage. Plotted is how the number of uniquely identifiable sequenced DNA fragments resulting from sheared *Ape*KI fragments varies with the total number of sequenced DNA fragments. Results were generated based on empirically determined frequencies of fragment reads from approximately 8.69 Gb of B73 DNA sequence reads. The red, blue and green points highlight the number of total fragment reads necessary to observe 90%, 80% and 70% of the potential fragments, respectively.

1.8 References

Amores, A., J. Catchen, A. Ferrara, Q. Fontenot, J. H. Postlethwait, 2011 Genome evolution and meiotic maps by massively parallel DNA sequencing: spotted gar, an outgroup for the teleost genome duplication. Genetics 188: 799-808.

Baird, N. A., P. D. Etter, T. S. Atwood, M. C. Currey, A. L. Shiver *et al.,* 2008 Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS ONE 3(10): e3376.

Baxter, S. W., J. W. Davey, J. S. Johnston, A. M. Shelton, D. G. Heckel, 2011 Linkage mapping and comparative genomics using next-generation RAD sequencing of a non-model organism. PLoS ONE 6(4): e19315.

Broman, K.W., H. Wu, S. Sen, G. A. Churchill, 2003 R/qtl: QTL mapping in experimental crosses. Bioinformatics 19:889-890.

Brunner, A. L., D. S. Johnson, S. W. Kim, A. Valouev, T. E. Reddy *et al.,* 2009 Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. Genome Research 19: 1044-1056.

Chutimanitsakun, Y., R. W. Nipper, A. Cuesta-Marcos, L. Cistue, A. Corey *et al.,* 2011 Construction and application for QTL analysis of a restriction site associated DNA (RAD) linkage map in barley. BMC Genomics 12:4.

Davey, J. W., P. A. Hohenlohe, P. D. Etter, J. Q. Boone, J. M. Catchen, M. L. Blaxter, 2011 Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature Reviews Genetics 12: 499-510.

Eichten, S. R., J. M. Foerster, N. de Leon, Y. Kai, C. Yeh *et al.,* 2011 B73-Mo17 Near-isogenic lines demonstrate dispersed structural variation in maize. Plant Physiology 156(4): 1679-1690.

Elshire, R. J., J. C. Glaubitz, Q. Sun, J. A. Poland, K. Kawamoto *et al.,* 2011 A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS ONE 6(5): e19379.

Fan, J., M. S. Chee, and K. L. Gunderson, 2006 Highly parallel genomic assays. Nature Reviews Genetics 7: 632-644.

Fu, Y., T. Wen, Y. I. Ronin, H. D. Chen, L. Guo *et al.,* 2006 Genetic dissection of intermated recombinant inbred lines using a new genetic map of maize. Genetics 174: 1671-1683.

Hansey, C.N., B. Vaillancourt, R. S. Sekhon, N. de Leon, S. M. Kaeppler *et al.,* 2012 Maize (*Zea mays* L.) genome diversity as revealed by RNA-sequencing. PLoS ONE 7(3): e33071.

Hillier, L. W., G. T. Marth, A. R. Quinlan, D. Dooling, G. Gewell *et al.,* 2008 Whole-genome sequencing and variant discovery in *C. elegans*. Nature Methods 5(2): 183-188.

Lander, E. S., and M. S. Waterman, 1988 Genomic mapping by fingerprinting random clones: a mathematical analysis. Genomics 2: 231-239.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10(3): R25.

Lee, M., N. Sharapova, W. D. Beavis, D. Grant, M. Katt *et al*., 2002 Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant Mol. Biol. 48: 453-461.

Lough, A. N., L. M. Roark, K. Akio, T. S. Ream, J. C. Lamb *et al.,* 2008 Mitochondrial DNA transfer to the nucleus generates extensive insertion site variation in maize. Genetics 178: 47-55.

Luo, C., D. Tsementzi, N. Kyrpides, T. Read, and K. T. Konstantinidis, 2012 Direct comparisons of Illumina vs. Roche 454 sequencing technologies on the same microbial community DNA sample. PLoS ONE 7(2): e30087.

Minoche, A. E., J. C. Dohm, and H. Himmelbauer, 2011 Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. Genome Biology 12: R112.

Ng, S. B., Turner, E. H., Robertson, P. D., Flygare, S. D., Bigham, A. W., *et al.,* 2009 Targeted capture and massively parallel sequencing of 12 human exomes. Nature 461: 272-276.

Pfender, W. F., M. C. Saha, E. A. Johnson, M. B. Slabaugh, 2011 Mapping with RAD (restriction-site associated DNA) markers to rapidly identify QTL for stem rust resistance in *Lolium perenne*. Theor. Appl. Genet. 122: 1467-1480.

Piepho, H. P., 2000 Optimal marker density for interval mapping in a backcross population. Heredity 84: 437-440.

Poland, J. A., P. J. Brown, M. E. Sorrells, and J. Jannink, 2012 Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS ONE. 7(2): e32253.

R Development Core Team, 2011 R: A language and environment for statistical computing. http://www.r-project.org/. Vienna, Australia. ISBN: 3-900051-07-0.

Saghai-Maroof, M. A., K. M. Soliman, R. A. Jorgensen, and R. W. Allard, 1984 Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. 81: 8014–8018.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stein, F. Wei *et al.,* 2009 The B73 maize genome: complexity, diversity, and dynamics. Science 326(5956): 1112-1115.

Tenaillon, M. I., M. C. Sawkins, A. D. Long, R. L. Gaut, J. F. Doebley *et al.,* 2001 Patterns of DNA sequence polymorphism along chromosome 1 of maize (*Zea mays* ssp. *mays* L.). Proc. Natl. Acad. Sci. 98: 9161-9166.

Teuscher, F., V. Guiard, P. E. Rudolph, and G. A. Brockmann, 2005 The map expansion obtained with recombinant inbred strains and intermated recombinant inbred populations for finite generation designs. Genetics 170: 875-879.

Treangen, T. J., and S. L. Salzberg, 2011 Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13: 36-46.

Van Orsouw, N. J., R. C. J. Hogers, A. Janssen, F. Yalcin, S. Snoeijers *et al.,* 2007 Complexity reduction of polymorphic sequences (CRoPS): a novel approach for large-scale polymorphism discovery in complex genomes. PLoS ONE: 2(11): e1172.

Wendl, M. C., 2006 Occupancy modeling of coverage distribution for whole genome shotgun DNA sequencing. Bulletin of Mathematical Biology 68: 179-196.

Chapter 2 GENOME-WIDE SCAN FOR SELECTION FOLLOWING THIRTY GENERATIONS OF ARTIFICIAL SELECTION FOR INCREASED NUMBER OF EARS PER PLANT IN THE GOLDEN GLOW MAIZE POPULATION

Authors: Timothy M. Beissinger[*,§], Candice N. Hirsch[‡‡], Brieanne Vaillancourt[†,**], Shweta Deshpande[††], Kerrie Barry[††], C. Robin Buell[†,**], Shawn M. Kaeppler[*,§§], Daniel Gianola[§,‡], Natalia de Leon[*,§§]

Affiliations: [*] Department of Agronomy, University of Wisconsin, Madison, 53706

[§] Animal Sciences Department, University of Wisconsin, Madison, 53706

[†] Department of Plant Biology, Michigan State University, East Lansing, MI, 48824

[‡] Department of Dairy Science, University of Wisconsin, Madison, 53706

[**] Department of Energy Great Lakes Bioenergy Research Center, Michigan State University, East Lansing, MI, 48824

[§§] Department of Energy Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, 53706

[††] Department of Energy, Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA, 94598, USA

[‡‡] Department of Agronomy and Plant Genetics, University of Minnesota, Saint Paul, MN 55108

Sequence data have been deposited in the National Center for Biotechnology Information Sequence Read Archive (BioProject accession no. PRJNA94561)

Publication information: This article was published in the March, 2014 issue of Genetics, Volume 196, no. 3, pp. 829-840.

2.1 Abstract

A genome-wide scan to detect evidence of selection was conducted in the Golden Glow maize long-term selection population. The population had been subjected to selection for increased number of ears per plant for 30 generations, with an empirically estimated effective population size ranging from 384 to 667 individuals and an increase of more than threefold in the number of ears per plant. Allele frequencies at more than 1.2 million single nucleotide polymorphism (SNP) loci were estimated from pooled whole genome resequencing data, and F_{ST} values across sliding windows were employed to assess divergence between the population pre- and post-selection. Twenty-eight highly divergent regions were identified, with half of these regions providing gene-level resolution on potentially selected variants. Approximately 93% of the divergent regions do not demonstrate a significant decrease in heterozygosity, which suggests that they are not approaching fixation. Also, most regions display a pattern consistent with a soft-sweep model as opposed to a hard sweep model, suggesting that selection mostly operated on standing genetic variation. For at least 25% of the regions, results suggest that selection operated on variants located outside of currently annotated coding regions. These results provide insights into the underlying genetic effects of long-term artificial selection and identification of putative genetic elements underlying number of ears per plant in maize.

2.2 Introduction

Changes in allele frequency occur in populations undergoing selection (e.g. Wright 1931; Crow and Kimura 1970). Understanding the patterns of such changes can provide a wealth of information regarding the genetic factors that control traits under selection. By comparing the allelic composition of a population pre- and post-selection, the genetic control of a trait may be revealed through the discovery of altered allele frequencies, assuming that selection effects can be statistically separated from random genetic drift (e.g. Krimbas and Tsakas 1971; Parts *et al.* 2011). Additionally, an improved understanding of the processes that take place during selection will contribute to long-standing genetic questions. For instance, the relative levels of diversity around selected sites may demonstrate whether selection has operated primarily on long standing variation or relatively new mutations (Innan and Kim 2004; Przeworski *et al*. 2005; Hermisson and Pennings 2005). Work in this area has demonstrated that soft sweeps, or selection on standing variation, may often be found in cases of polygenic traits (Strasburg et al, 2012) and are expected to be common in human populations (reviewed by Pritchard et al, 2010). Another persistent question involves how rapidly selected sites approach fixation (Kimura, 1962), which can be addressed by an analysis of selected sites once they are identified. Kelly et al (2013) recently demonstrated a selection experiment in *Mimulus* for which numerous partial sweeps, or selective sweeps that have not reached fixation, were found. The relative importance of non-genic DNA is another long-standing question that may be addressed by the identification of selected regions (e.g. King and Wilson, 1975; Wray, 2007).

Previously, assessing allele frequencies in selected populations has only been feasible for an *a priori* set of candidates or for a limited set of random loci, due to the limited number of markers available and the cost of conducting the assays. However, the recent reduction in the

cost of DNA sequencing and single nucleotide polymorphism (SNP) detection (reviewed by Metzker 2010) now allows genome-wide characterization of allelic variants. Accordingly, multiple experiments utilizing high-density SNP or sequence data to identify selected sites in both naturally and artificially selected populations, and in both sexual and asexual species have been conducted (e.g. Akey *et al.* 2002; Parts *et al.* 2007; Bigham *et al.* 2010; Turner *et al.* 2011). The goals of these experiments ranged from localizing selected sites for unknown traits in natural populations (Voight *et al.* 2006) to identifying quantitative trait loci for specific traits in experimentally-derived populations or crosses (Parts *et al.* 2011).

The methods employed to identify selected sites in natural populations include assessment of variation between versus within populations (Lewontin and Krakauer, 1973; Akey *et al.* 2002), detection of abnormalities in the site frequency spectrum (SFS; Payseur *et al.* 2002), and assessment of local patterns of linkage disequilibrium (LD; Sabeti *et al.* 2002; Voight *et al.* 2006; Sabeti *et al.* 2007) to find recent selective sweeps (Maynard Smith and Haigh 1974). In natural populations, it is often difficult or impossible to evaluate the phenotypic effects of selected polymorphisms because many traits are simultaneously selected in such populations and the relative intensity of selection is unknown. In experimental populations, however, selection is often deliberately conducted under controlled conditions, allowing for better inference of the strength of selection and biological role of genes localized within potentially selected sites. Methods for identifying selection in these artificially selected populations may include any of the methods utilized for natural populations, but a benefit of these types of studies is that samples of the progenitor population are frequently available, which allows for direct measurement of allele frequency changes. Separation of selection versus genetic drift effects has been performed by comparing allele frequencies to simulations of drift (Wisser *et al*. 2008), and by developing

significance tests based on replicated or control populations (Parts *et al.* 2011; Turner *et al.* 2011).

Long-term breeding projects in agricultural species, both plants and animals, have generated excellent resources that can be leveraged for identifying loci that were impacted by artificial selection. In animals, for instance, Johansson *et al.* (2010) worked with a population of chickens divergently selected for body size and found that the majority of changes can be attributed to selection on standing genetic variation versus new mutations. Another study using chickens identified 82 putatively selected regions with reduced levels of heterozygosity (Qanbari *et al.* 2012). Similarly in cattle, Flori *et al.* (2009) found 13 regions that were under selection in recent history, a subset of which included genes previously known to impact milk production. Also, Pan *et al.* (2013) identified selected regions in cattle based on LD and then verified the functional roles of several genes based on a review of genome annotation, gene ontology enrichment analysis, and pathway enrichment analysis. Another interesting study in cattle was conducted by Qanbari *et al.* (2011), which employed a multi-faceted approach including both allele frequency- and LD-based methods to identify signatures of selection.

Several studies scanning for selection in agricultural crop species have also been conducted. For instance, Wright *et al.* (2005) looked for evidence of selection across a set of 774 maize genes and found that 2 to 4% had undergone selection. Recently, whole-genome studies have been conducted as well; both Jiao *et al.* (2012) and Hufford *et al.* (2012) looked for signatures of selection by investigating diverse sets of maize lines and highly dense marker sets. These studies have also been conducted with other important crops including soybeans (Lam *et al.* 2011), and rice (He *et al.* 2011). Often, plant species have the advantage that remnant seeds representing a population before selection began often remain available for years or decades

following the selection process itself (e.g. Odhiambo and Compton 1987). This characteristic was utilized by Wisser *et al.* (2008), who compared marker data gathered from samples before and after several generations of selection to identify loci affecting northern leaf blight resistance in closed populations of maize that had undergone selection.

Maize is an important crop species that has been subjected to artificial selection for approximately 9,000 years (Matsuoka *et al. 2002*). Modern research and breeding investments have provided numerous examples of existing maize populations that have been selected for a particular trait over time-spans ranging from only a few cycles to more than a hundred generations (e.g. Odhiambo and Compton 1987; Coors and Mardones 1989; Ross *et al*. 2006; Dudley 2007; Wisser *et al.* 2008). One such example involves the Golden Glow maize selection project (Coors and Mardones 1989) which has undergone selection for a specific yield component, prolificacy, defined as the number of ears per plant. Selection for an increase in number of ears per plant was accomplished using recurrent mass selection for 30 generations, maintaining a large effective population size (N_e) and strong selection intensity in the process. Selection succeeded in increasing the mean number of ears per plant from 1.6 at cycle 0 to 4.9 by cycle 24 (de Leon and Coors 2002). Number of ears per plant is a trait of particular interest to maize breeders because it is highly correlated with grain yield and density tolerance (Russell 1984; Carlone and Russell 1987; Subandi 1990; Duvick 1997; Ahmad *et al.* 2011). In fact, Coors and Mardones (1989) reported a correlation between ears per plant and grain yield per plant of 0.90 through cycle 12 of the Golden Glow population. Maita and Coors (1996) still found the correlation to be positive after 20 cycles of selection ($r = 0.71$), and reported that increased number of ears per plant may improve the population's ability to yield in stress conditions. Additionally, number of ears per plant is of interest as a model trait because it is correlated with

other important agronomic traits including lodging and moisture at harvest (Cross *et al.* 1987) and has been shown to be a secondary effect of maize domestication (Doebley *et al.* 1990). Overall, the combination of large N_e, strong selection intensity, substantial phenotypic response to selection, and practical and biological relevance of the trait make Golden Glow an ideal crop model population to evaluate allele frequency changes resulting from selection.

The objectives of this study were to: 1) Estimate SNP allele frequencies in the cycle 30 selected population relative to the initial population by pooled whole genome resequencing to scan for signatures of selection, and 2) evaluate the putatively selected regions to assess whether or not selected sites are approaching fixation, estimate the extent of selective sweeps and genetic hitchhiking, and explore the proportion of sites for which selection may have operated on intergenic as opposed to genic regions.

2.3 Materials and Methods

Germplasm: Selection for increased number of ears per plant in the Golden Glow maize population was initiated by J.H. Lonnquist at the University of Wisconsin-Madison in 1971. For the first 12 cycles of selection, selection intensity was maintained at approximately 2.5 to 5%. From the 13th cycle onward, the selection intensity was made stronger, to between approximately 0.5 and 1%. A complete description of the selection process was provided by de Leon and Coors (2002).

For the present experiment, 48 randomly chosen plants from each of cycles zero and 30 were utilized for analysis. To preserve population seed samples over the decades, remnant seed from the original cycles was occasionally increased through random mating of individual plants

utilizing large population sizes to minimize unwanted changes in allele frequency due to drift or unintentional selection. While genetic drift was minimized during this process, it could not be completely eliminated. The sample taken from cycle zero had incurred five generations of seed increase utilizing on average 110 individuals each generation, while that from cycle 30 had incurred two generations of increase utilizing on average 130 individuals each generation.

DNA extraction, SNP genotyping, and sequencing: DNA extraction for array-based SNP genotyping was performed for each individual sampled. Leaf tissue was harvested from 96 plants (48 from each population) followed by DNA extractions using the cetyl(trimethyl)ammonium bromide (CTAB) method (Saghai-Maroof *et al*. 1984). Genotyping was performed on the individual samples by Pioneer Hi-Bred International (Johnston, IA) using a 768 marker multiplex assay on the Illumina BeadArray platform (Jones *et al.* 2009). These array-based SNPs were only used for the determination of effective population size.

For the whole genome resequencing, an equal amount of tissue from 48 seedlings from each population cycle was harvested and pooled. From each pool, DNA was extracted using the CTAB method (Saghai-Maroof *et al*. 1984). Libraries with a target insert size of 270 bp were prepared according to the Illumina protocol (Illumina, Inc. San Diego, CA). Libraries were sequenced using the Illumina HiSeq (SanDiego, CA) at the Department of Energy Joint Genome Institute (Walnut Creek, CA) to generate 2x100 nucleotide pair-end (PE) sequence reads. Sequences are available in the Sequence Read Archive at the National Center for Biotechnology Information (BioProject number PRJNA94561). Sequence read quality was evaluated using the FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) and sequencing lanes with insufficient quality were not used in the analysis. In total, 555,078,520 read pairs from eight sequencing lanes of cycle zero and 652,901,808 read pairs from nine sequencing lanes of

cycle 30 were generated. Prior to mapping, reads from high quality lanes were cleaned using the fastx clipper program from the FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/index.html), which removed the Illumina adapter sequences, and required a minimum length of 20 bp after trimming.

Two mapping pipelines were used to establish a high confidence SNP set. In the "SE pipeline", all reads that passed through the cleaning step above were mapped as single-end (SE) reads using Bowtie version 0.12.7 (Langmead *et al.* 2009) to the B73 version 2 reference sequence (AGPv2; http://ftp.maizesequence.org; Schnable *et al*. 2009). An alignment was considered valid if there were two or fewer mismatches relative to the reference sequence (-v 2) and a read was required to only have one valid alignment (-m 1). All other parameters were set to the default values. In the "mixed pipeline" read pairs for which both reads passed through the cleaning step were mapped as PE and read pairs for which only one read passed the cleaning step were mapped as SE. For the PE mapping, in addition to the -m 1 and -v 2 options used in the SE pipeline mapping, the minimum insertion size was set to zero bp and the maximum insertion size was set to 1000 bp. All other parameters were set to the default values.

The same SNP detection pipeline was used for alignments from both the SE pipeline and the mixed pipeline. Within each population (cycle zero and cycle 30), all valid alignments were processed using SAMtools version 0.1.12a (Li *et al.* 2009) sort, merge, index, and pileup programs to generate unfiltered pileup files. For the pileup program, the -B option was used to disable BAQ computation. Nucleotide frequencies (A, T, C, and G) were determined at each position, requiring a quality score of at least 20 for a base within a read to be included. For a particular position, if at least two nucleotides were supported by at least two reads each across the two populations, that position was considered polymorphic. Only positions that were

identified as polymorphic in both the SE pipeline and the mixed pipeline were included for further analysis, and allele frequency estimates were based on the SE pipeline.

SNP filtering and estimating allele frequencies: In total, 8,128,042 SNPs were identified from sequencing, but the set of SNPs selected for analysis was filtered to include only high-confidence sites. Only SNPs with two alleles were included, due to the increased likelihood that multi-allelic SNPs included sequencing errors and because of complications related to assessing allele frequency changes at multi-allelic loci. It was required that every SNP location included in the analysis was observed at least 20 times in each population, and no more than 89 times based on the SE pipeline mapping. An observation of 89 corresponded to the mean SNP coverage plus two standard deviations. SNPs read at more than this level are more likely to be from organellar or repetitive DNA that is inaccurately represented by a single position in the reference genome. After filtering, 1,211,745 high-confidence SNPs were retained for analysis and the allele frequencies at each SNP were calculated in each population. Allele frequencies were computed according to their maximum likelihood estimate. Thus, the number of times a particular allele at a position was observed in the population was divided by the total number of times any allele was observed at that position.

Estimating effective population size: The effective population size over the course of the selection program was evaluated in two ways. First, because the number of breeding males and females used and selected in the experiment was known and held relatively constant across cycles, the simple relationship, $N_e = \frac{4N_m N_f}{N_m + N_f}$, where N_m and N_f are the number of mating males

and females, respectively, was used. Next, an estimate was made using the 768 array-based SNPs, according to the relationship $N_e = \frac{1}{2(1-\sqrt[t]{H_t/H_0})}$, where H_t and H_0 are the mean levels of heterozygosity in the t^{th} (t=30 in this case) and 0^{th} generation, respectively. Both equations for effective population size are provided by Crow and Kimura (1970).

Scan for selection: A genome-wide scan for selection similar to the Lewontin and Krakauer (1973) test was conducted. Note that because of the simple population structure involved in this experiment, with there being only two subpopulations and no migration between them, it is not expected that the shortcomings of the Lewontin-Krakauer test, with respect to migration (Nei and Maruyama, 1975) or correlations between subpopulations (Robertson, 1975), will be detrimental. Unfortunately, however, the implementation of pooled sequencing precludes the ability to accurately estimate LD in the population and therefore makes a simulation-based approach for establishing precise significance levels intractable. Instead, the scan described is used to classify genomic regions as empirically divergent or not divergent over the course of the experiment. The most divergent sites, based on a sliding-window estimate of F_{ST}, are highlighted as the most promising candidates for selection. This approach is justified based on the documented process of strong selection, coupled with the dramatic phenotypic changes that have accumulated during the experiment, which leads to the conclusion that selection is expected to have changed the allele frequency in regions of the Golden Glow population's genome associated with the selection target. As such, those sites demonstrating the greatest levels of divergence are those most likely to have been impacted by selection.

Because of the substantial sampling error that is inherent to pooled sequencing, a sliding window approach was implemented to evaluate divergence. First, SNP-specific estimates of F_{ST} were computed within R version 3.0.2 (R Core Team 2013) according to $F_{ST} = \frac{s^2}{\bar{p}(1-\bar{p})+s^2/r}$, where s^2 is the sample variance of allele frequency between populations, $\bar{p}$ is the mean allele frequency across populations, and r is the number of populations (Weir and Cockerham 1984). This formula assumes a large sample size, which is met by the previously described filtering step where loci observed fewer than 20 times were removed. The formula also corrects for bias based on the small number of populations sampled (in this case two). F_{ST} values were averaged over sliding windows of 25 SNPs. Thus, each SNP locus was assigned a new value based on the average of itself along with the 12 upstream and 12 downstream SNPs. A window size of 25 SNPs was chosen because such a size appeared to maximize signal from sites while minimizing noise from sampling and sequencing error. When smaller window sizes were employed, F_{ST} values behaved erratically, and employing larger window sizes led to F_{ST} values that were unrealistically homogenous. Such an approach for determining sliding window boundaries has been previously demonstrated (e.g. Myles *et al*. 2008; Akey 2009).

Outlying window-based F_{ST} values that exceeded the 99.9% or 99.99% level of the empirical window-based F_{ST} distribution were identified. It should be noted that these outlier levels were not chosen due to a connection with a specific level of significance, but instead because they provide a reasonable number of candidates for strong or extremely strong selection, respectively, which are used in downstream analyses. Once SNPs exceeding the outlier thresholds were identified, overall boundaries for divergent regions were defined by taking the set of all SNPs that exceeded the specified threshold and identifying groups of SNPs that likely correspond to the same divergent region. This was achieved by deeming outlying SNPs that were

within 5 megabases (MB) of one another as belonging to a single region. Five MB was used because at distances greater than this, the likelihood of LD between SNPs is minimal, yet when a distance of less than 5 MB was employed, it was clear that some of the resulting regions were likely correlated with the same selection event (Figure 2.4). Also, utilizing this relatively large window provides a conservative estimate of divergent region boundaries. It should be emphasized that with this approach the identified regions could be (and usually were) much smaller than 5 MB (Figure 2.5).

Each group of outlying SNPs was considered a divergent group, and the position of the 12^{th} SNP upstream of the group start and that of the 12^{th} SNP downstream of the group end was added to the group to define the boundaries of each divergent region. The up and downstream additions were incorporated because the sliding window method included information from up to the 12^{th} SNP distal from the region in either direction.

Testing for allele fixation: For each identified region that displayed evidence of divergence (exceeded the 99.9% F_{ST} level), a test was performed to determine if the pattern demonstrated was consistent with a drive toward fixation. To test for a loss of variability, the expected level of heterozygosity (based on the expectation from Hardy-Weinberg equilibrium) was computed for each SNP within a region at cycle zero and at cycle 30. Next, a t-test was performed to determine if the expected heterozygosity at cycle 30 was significantly different than the heterozygosity at cycle zero. A significant reduction in expected heterozygosity was interpreted as evidence for a tendency toward fixation at the divergent region, while no change or an increase in heterozygosity was interpreted as evidence that the region has not yet approached fixation.

Extent of hitchhiking: Genetic hitchhiking is the process by which the frequency of a neutral locus is altered due to being in LD with a locus under selection (Maynard Smith and Haigh 1974). Hitchhiking can occur in the instance of a "hard sweep", where selection operates on a newly arisen allele that is immediately beneficial, or from a "soft sweep", in which an allele previously segregating in the population becomes advantageous due to new selective pressures (Hermisson and Pennings 2005). It has been shown that the genomic footprint of a selective sweep is expected to extend substantially further from the selected locus under a hard sweep model than in a soft sweep (Innan and Kim 2004; Hermisson and Pennings 2005). To investigate whether the divergent, and potentially selected, regions identified by F_{ST} were consistent with soft or hard sweeps, the size of selected regions was used as an indicator of hitchhiking. A K-means clustering algorithm (Hartigan and Wong 1979) was performed to group divergent regions based on size. Two centers were employed, so that identified clusters were classified into one group of small size and one of large size regions, corresponding to those likely to depict hard sweeps and soft sweeps, respectively.

Also, data from the intermated B73 x Mo17 (IBM) population (Lee *et al.* 2002) was used to test if recombination rate is the main contributor to region size rather than the type of sweep experienced. Liu *et al.* (2009) estimated a map of cM/Mb in the IBM population. Although the IBM population will not necessarily display identical recombination patterns to the Golden Glow maize population at all positions in the genome, the overall patterns are expected to be similar. The physical position of each divergent region in the Golden Glow was anchored to the nearest physical position with a given cM/Mb in the Liu and co-workers (2009) map. In cases for which multiple IBM positions with reported cM/Mb were within a single Golden Glow selected region,

the level of cM/Mb across all of these positions was averaged. Thereby, every highly divergent region identified in the Golden Glow maize population was assigned a single value for cM/Mb. A significant product-moment correlation was tested for using both raw data as well as log-transformed data.

2.4 Results

Effective population size: A total of approximately 4,250 plants for cycles 1 to 12 and 14,250 plants for cycles 13 to 30 were evaluated in the selection plots, but approximately 1,000 males and 200 females were selected in each cycle leading to an N_e that is smaller than the total census size. Assuming plants were randomly mating over the course of the experiment, it was estimated based on population demographics that the effective population size was expected to be approximately 667. However, preferential pollen flow among neighboring plants and assortative mating among plants flowering on the same day may have prevented truly random mating, thus the N_e was also estimated from markers. Based on 768 array-based SNP markers, the effective population size estimate was 378. It is worth noting that due to the effects of selection, the marker-based estimate is expected to be biased downward. Therefore, true N_e for the Golden Glow population over the course of the selection experiment is likely somewhere between 378 and 667 individuals.

Twenty-eight genomic regions were identified as substantially divergent: To identify the specific genomic regions most likely to have been impacted by selection, an outlier-based approach that scanned for regions exceeding the 99.9% or 99.99% levels of the empirical

distribution, based on 25-SNP sliding windows, was employed. Using the 25-SNP sliding window statistic, specific genomic regions that were most likely to have been affected by selection were apparent (Appendix A). Twenty eight regions were identified as divergent at the 99.9% outlier level. Three of these regions also exceeded the 99.99% level (Figure 2.1; Table 2.1). Regions identified at the 99.9% level were found on all 10 of the maize chromosomes. The regions ranged in size from 4,251 bp to 9.2 MB and encompassed from zero to 73 predicted B73 5b annotated genes (http://ftp.maizesequence.org; Schnable *et al*. 2009). Assuming that there was limited unintentional selection for other traits during the course of the selection experiment, genes in these regions can be considered candidates for control of number of ears per plant in maize. Of the regions identified, 22 (79%) included five or fewer annotated genes, and 14 regions (50%) included one or zero annotated genes. As an example, a region on maize chromosome six that encompasses approximately 10 kb (AGPv2 position 119,682,711 – 119,692,810) falls entirely within a single predicted gene, GRMZM2G368678 (Figure 2.2). This gene is annotated as an "androgen induced inhibitor of proliferation" based on sequence similarity to the *Sorghum bicolor* gene Sb10g010710.1 and is expressed in shoot apical meristem and multiple other tissues (Sekhon *et al.* 2011).

Few regions show evidence of fixation: Each of the genomic regions that were identified as highly divergent was tested for a change in expected heterozygosity between cycle zero and cycle 30. A decrease in expected heterozygosity suggests that strong selection has taken place and that allele frequencies are being driven toward fixation. Conversely, an increase in expected heterozygosity at the selected site may be observed in the case where the initial favorable allele frequency was less than 0.5 and selection has taken place but was not strong enough or has not

been occurring for a long enough time to move the allele frequency close to fixation. Other explanations for an increase in heterozygosity can include over-dominance, complex linkage relationships between multiple selected sites, and variable selection environments. It is also possible that selection has occurred but there is no change in expected heterozygosity; this would be the case, for example, if allele frequency changed from 0.4 at cycle zero to 0.6 at cycle 30.

It was observed that only two of the 28 divergent regions demonstrated a statistically significant reduction in expected heterozygosity from cycle zero to cycle 30 (two-tailed Bonferroni-corrected p-value = 0.025/28 = 0.0009). However, 10 regions (35.7%) displayed a significant increase in expected heterozygosity. The change in the level of expected heterozygosity across the remaining 16 regions was not significant (Table 1). Examples of divergent regions that displayed an increase, decrease, and no change in expected heterozygosity are provided (Figure 2.3). These observations suggest that although selection was strong, it was not strong enough over the course of 30 generations to drive favorable alleles to fixation at the majority of sites that display evidence of strong divergence. Instead, most sites are still segregating in the population and, for a substantial subset of identified sites, the genetic variability in the population has increased as a result of selection. Kelly *et al.* (2013) demonstrated that increased heterozygosity across a region as a result of selection is expected in the situation where not only has selection led to more intermediate frequencies at the selected variant, but where the selected variant is also positively associated with rare variants in the region. Additionally, it may be predicted that the two regions which did display significant reductions in heterozygosity are those that experienced the strongest selection.

Selection mostly operated on standing variation: Studies have shown that selection on rare or new alleles is most likely to cause a hard sweep and lead to long-range hitchhiking, and that the hitchhiking pattern may be mostly or completely absent in the case of a soft sweep, when selection operates on an allele that was segregating in the population at the onset of selection (Innan and Kim 2004; Przeworski *et al*. 2005; Hermisson and Pennings 2005). More precisely, Hermisson and Pennings (2005) showed that a soft sweep from standing variation is expected to display a narrower footprint than that of a hard sweep. This is because for a new allele, LD between the favorable polymorphism and the genetic background in which it resides is likely to extend a longer range than if the allele were at an intermediate frequency and therefore was present in a variety of different haplotypes. To investigate the prevalence of hard versus soft sweeps during this selection program, the size of divergent regions was used to indicate the extent of hitchhiking and thus the type of sweep that may have occurred; larger regions suggest longer-range hitchhiking and therefore indicate a hard sweep (or the possibility of favorable alleles at multiple loci in proximity to each other), while smaller regions suggest less hitchhiking and likely a soft sweep.

Substantial variability in the size of divergent regions was observed (Table 1), with a range of 4,251 bp to 9.2 Mb. The median region size was 69.3 kb. A K-means clustering algorithm with two centers was employed to separate regions into two groups based on size (Hartigan and Wong 1979). The results were that 26 of the 28 regions were placed into a small-size cluster and only two regions were placed into a large-size cluster (Figure 2.5). The median region size for the small size cluster was 61.2 kb, while that of the large size cluster was 6.8 Mb. Because of the method by which regions were identified, the possibility that the large regions are

resultant of multiple independently selected sites in close proximity cannot be ruled out, although this appears to be less likely due to the relatively small size of the gap assumed (Figure 2.4).

There is also the possibility that the size of regions is heavily influenced by the variability of recombination rates across the genome. This was tested by utilizing a recombination map developed from the IBM population (Lee *et al.* 2002). No evidence for a correlation between recombination rate and region size was found in the raw data (ρ = -0.126, p-value = 0.5215) or by utilizing log-transformed region sizes (ρ = 0.172, p-value = 0.3807; Figure 2.6). Therefore, it is likely that the two large regions correspond to a hard sweep model with a large amount of hitchhiking due to selection on rare, relatively new variants, and the remaining 26 regions demonstrate selection on standing variation. Consequently, the vast majority of selection is not consistent with selection on new variants but instead on existing genomic variants that were already segregating well before cycle zero. This observation is also consistent with the large phenotypic response seen in a relatively small number of generations of selection; rare variants (e.g. less than p=0.01), even of substantial magnitude, would take multiple generations of selection before they began to contribute meaningful variation to the selection response.

Selection on genes or intergenic regions: Functional alleles can involve changes in the coding sequence, transcriptional control of genes by nearby promoter and controlling elements, and non-translated controlling sequences. To determine the potential importance of genic versus non-genic variants underlying phenotypic variation for number of ears per plant, the position of the 28 divergent regions was classified as non-genic or genic (containing one or more annotated gene models). Of these, seven (25%) regions neither contain currently annotated genes nor are

located within 5 kb of a 5b reference gene (Schnable *et al.* 2009). This suggests either that the population harbors selected genes that are not present or annotated in the reference sequence or that a sizeable subset of the selection in the Golden Glow population has operated on non-genic regions, or a combination thereof.

2.5 Discussion

This analysis provides insight into the genetic processes that take place during long-term experimental selection as well as the genetic control of number of ears per plant in maize, based on a long-term selection program. A total of 28 highly divergent regions were identified, and therefore likely to have been under selection, with representation on all ten of the maize chromosomes. Among these, 22 contain five or fewer annotated gene models and 14 contain one or zero annotated genes. Moreover, evidence from past studies helps to corroborate the potential role of some of the identified regions. For instance, GRMZM2G368678 is annotated as an androgen induced inhibitor of proliferation based on sequence similarity to a *S. bicolor* gene. Separately, a quantitative trait locus (QTL) study was performed by de Leon *et al.* (2005) based on a mapping population derived from the Golden Glow population at cycle 23. The study identified a QTL on chromosome six for ear number that closely corresponds to one of the divergent regions that was identified. Additionally, the maize gene *zcn15* was found within a divergent region on a different area of chromosome 6 that contains a total of five annotated genes. Danilevskaya *et al.* (2008) report that this is among the most favorable candidates for function as a promoter of the floral transition. Additional research is needed to determine if the

genes and regions putatively subjected to selection are directly involved in meristem function resulting in increased number of ears per plant in this population.

A detailed analysis of F_{ST} values and expected heterozygosity across the selected regions demonstrated that in the majority of cases, selection did not drive variants toward fixation. Such a result appears to coincide with findings from other studies involving numerous different species. In *Drosophila*, for example, studies based on reverse evolution (Teotónio *et al.*, 2009) as well as long-term evolution (Burke *et al.* 2010) have shown little or no evidence of fixation or substantial changes in diversity. Also, Parts *et al.* (2011) found minimal evidence of fixation in yeast populations that had been selected for heat tolerance for 288 generations. Interestingly, for ten of the identified Golden Glow regions, heterozygosity significantly increased as a result of selection, compared to only two where it decreased. Although this observation is consistent with certain models of selection (Crow and Kimura 1970), it is often forgotten as a potential consequence of selection. Specifically, increased regional heterozygosity after selection is expected if the selected variant has been driven to an intermediate frequency and is positively associated with rare alleles at neighboring loci, which have also, therefore, been driven to more intermediate frequencies (Kelly *et al.* 2013). One possibility for a positive association between rare alleles is that occasional historical outcrossing has introduced haplotypes consisting of an abundance of rare alleles into the population.

The finding of increased heterozygosity resulting from selection may be important when choosing appropriate methods to use to scan for selection. For instance, methods that scan for selective sweeps by looking for a loss of variability (e.g. Kim and Stephan 2002) would have no power to detect selection from such a signature. Therefore, it is important to take into consideration that selection is an ongoing process and that, even in a simple fully-additive model,

selected loci for which the favorable allele has initial frequency of less than 0.5 will show an increase in heterozygosity before it begins to decrease as the allele moves closer to fixation, and that the same may be observed at neighboring sites depending on the initial haplotype structure of the population. Another possibility is that over-dominant gene-action is present for several selected sites, driving alleles to equilibrium at an intermediate frequency instead of to fixation. If such is the case, it would provide evidence in favor of the over-dominant theory to explain heterosis (reviewed by Schnable and Springer 2013). More likely, however, is that this result is simply a function of 30 generations being too short a time for a substantial loss in heterozygosity for all but the most strongly selected sites.

It is notable that regions with a substantial amount of long-range hitchhiking, demonstrating hard sweeps, were rare in this experiment. Selection on relatively new mutations or rare alleles, which are in high LD with the genetic background in which they reside, is the situation that leads to long-range hitchhiking. Conversely, selection on relatively common alleles, the model of a soft-sweep, is not expected to display a substantial pattern of hitchhiking (Hermisson and Pennings 2005; Przeworski *et al*. 2005). Two of the 28 divergent regions identified in this experiment were clustered separately from the remaining 26 regions, suggesting that they may be cases of long-range hitchhiking and thus hard sweeps. This implies that the majority of selection operated on standing variation for which beneficial alleles were segregating in the Golden Glow population before selection began, suggesting that although most of the polymorphisms capable of generating a high number of ears per plant were present at cycle zero, it was not until selection incrementally increased the frequency of these variants within individuals that highly prolific phenotypes emerged.

This observation is consistent with one made by Coop *et al.* (2009), regarding human populations. The authors found that while positive selection in the human genome may be common, such selection driving new mutations to fixation is exceedingly rare. Similarly, Innan and Kim (2004) investigated selective sweeps on standing variation, as may have occurred during a domestication event. The authors focused particularly on maize. Their finding was that selection on standing variation may not be identifiable, because genetic variation at linked loci surrounding the selected site will not necessarily be reduced. This finding agrees with one reported by Teshima *et al.* (2006), who found that for an initially neutral mutation that had drifted to frequency 0.05 when it became beneficial, allele frequencies in the selected population for loci surrounding the selected site are likely to be intermediate. The onset of selection at cycle zero of the Golden Glow experiment parallels what occurs at the onset of domestication, where the fitness of individuals suddenly and dramatically changes due to new selective pressures, thus the patterns of variation may be similar. However, because in this study the main approach for identifying selection was allele frequency divergence between the selected and non-selected populations rather than finding regions with reduced variation, this approach is not limited by the potential lack of reduced variation. Yet upon further exploring each of the selected regions to identify those consistent with long-range hitchhiking, the findings here match expectations; the overwhelming tendency was that selection modified allele frequencies at isolated sites rather than across wide spans, suggesting that most of the observed selection operated on standing genetic variation.

Lastly, several of the divergent sites (25%) contain no currently annotated genes nor are in close proximity to any annotated gene models. While this could be due to Golden Glow genes that are not present in the reference genome, it is also possible that these are instances of

selection on non-genic DNA. The possibility of expression-controlling regions leading to major phenotypic differences between organisms was discussed decades ago by authors such as King and Wilson (1975). Since then, a multitude of studies have identified such regions across a wide array of species (reviewed by Wray, 2007). In maize, the expression of the *tb1* gene has been shown to be affected by intergenic sequence tens of kilobases away from the gene itself (Clark *et al.,* 2006). Similarly, the finding appears to be pervasive in domesticated animal species, where studies involving horses (Gu *et al*, 2009), cattle (Qanbari *et al.* 2011), and chickens (Qanbari *et al.* 2012) have all identified selection in gene-poor regions. Likewise, in human populations it has been observed that at a minimum, 14% of selected regions identified across multiple studies result from selection on non-coding material (Akey 2009). In *Drosophila*, various regulatory changes that modify phenotype have been found (Sucena and Stern 2000; Prud'homme *et al.* 2006). The incomplete nature of the maize reference genome (Schnable *et al.* 2009), coupled with this study's inability to precisely isolate the causative sites that were selected down at nucleotide level, precludes firm conclusions regarding the proportion of selection that operated on non-genic material. For instance, even within selected regions that do include genes it is possible that the causative variant was not one of those genes but instead a regulatory variant. The findings here imply that at least for a subset of sites, non-coding polymorphisms are selectively relevant.

In summary, important insight into the putative control of number of ears per plant was gained by scanning for signatures of selection based on differences in allele frequency between selected and unselected cycles in a maize population subjected to artificial selection for a number of generations. Furthermore, the findings show that, at least for the Golden Glow population, soft-sweeps appear to be more common than hard-sweeps, the rate of allele fixation is relatively

slow for regions under selection, and changes in allele frequencies in non-coding polymorphisms that have effects on the phenotype can be generated by selection.

2.6 Acknowledgements

This work was funded by the DOE Great Lakes Bioenergy Research center (DOE BER Office of Science DE-FC02-07ER64494). The work conducted by the U.S. Department of Energy Joint Genome Institute is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Simulations were performed using resources and the computing assistance of the UW-Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison and the Wisconsin Alumni Research Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science. DuPont-Pioneer provided SNP genotyping with the Illumina Golden Gate assay. T.B. was supported by the University of Wisconsin Graduate School and by a gift to the University of Wisconsin-Madison Plant Breeding and Plant Genetics program from Monsanto.

2.7 Figures

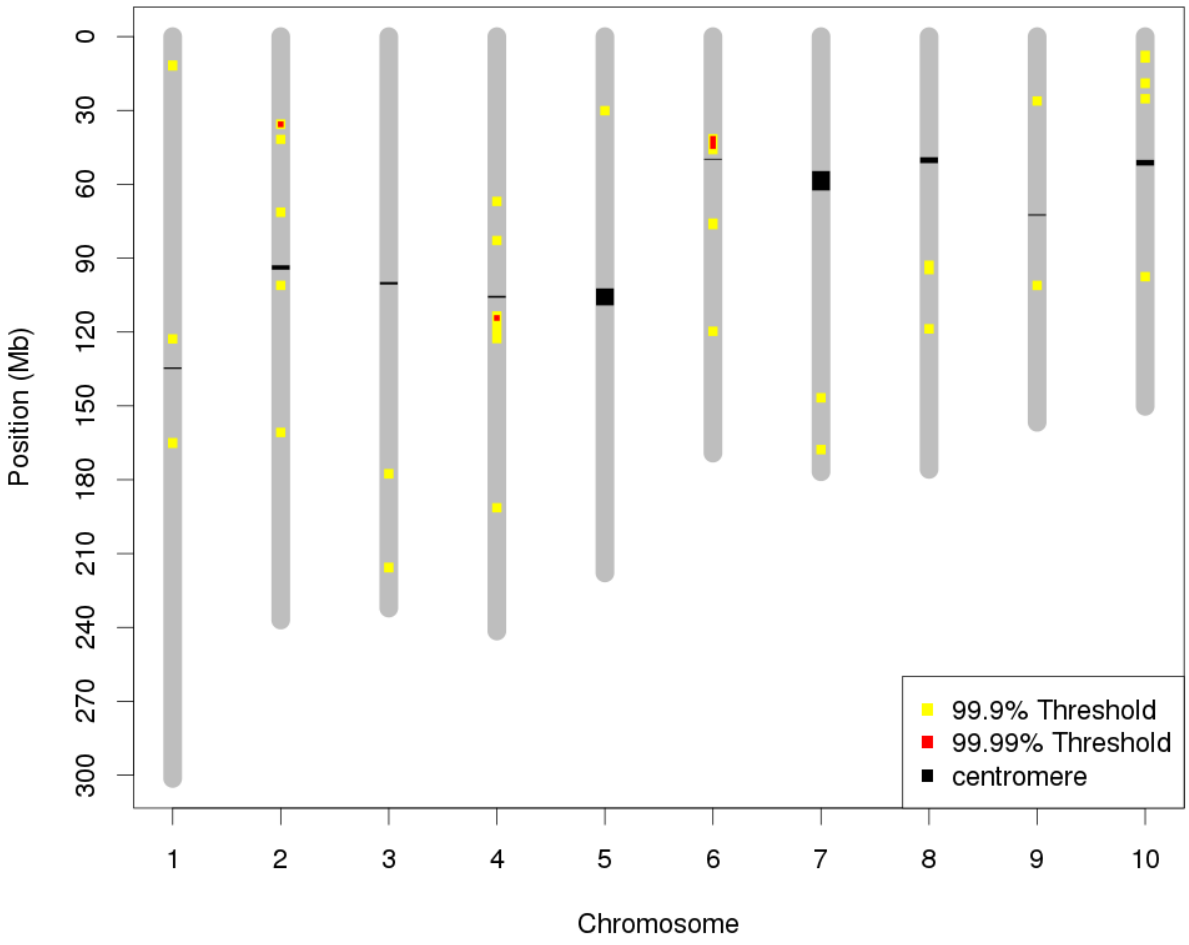


Figure 2.1: Location of identified regions

Physical location of 28 regions identified as divergent and potentially under selection for number of ears per plant based on changes in allele frequency between cycle 0 and cycle 30 of the Golden Glow maize population using 25 bp sliding windows estimates of F_{ST}.

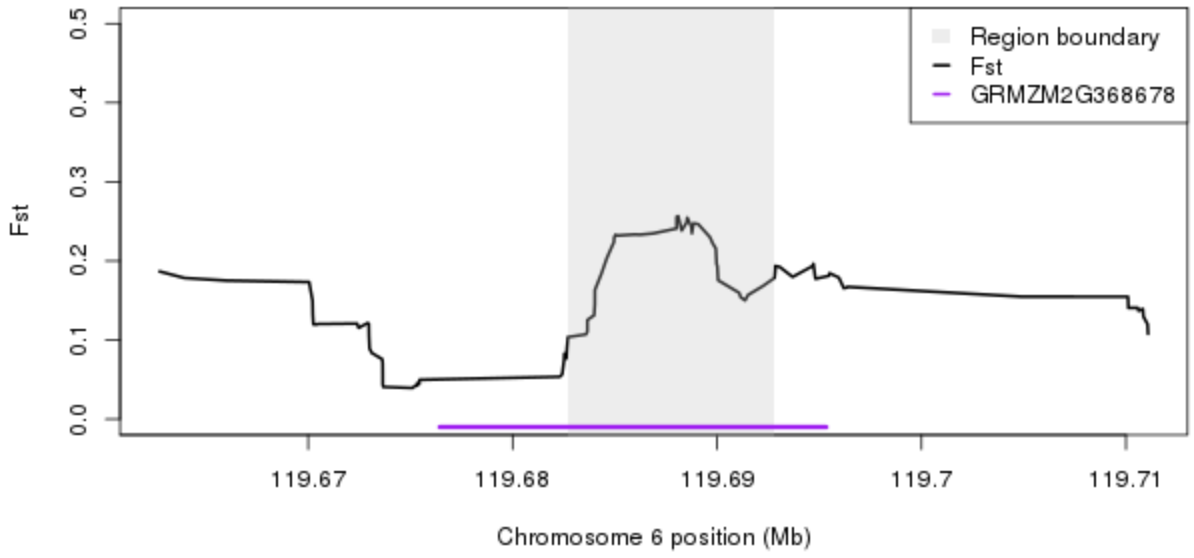


Figure 2.2: F_{ST} for an example region.

F_{ST} across a single significantly selected region of chromosome 6. The selected region is included entirely within a predicted gene, GRMZM2G368678, which has been annotated as an androgen induced inhibitor of proliferation in *Sorghum bicolor*.

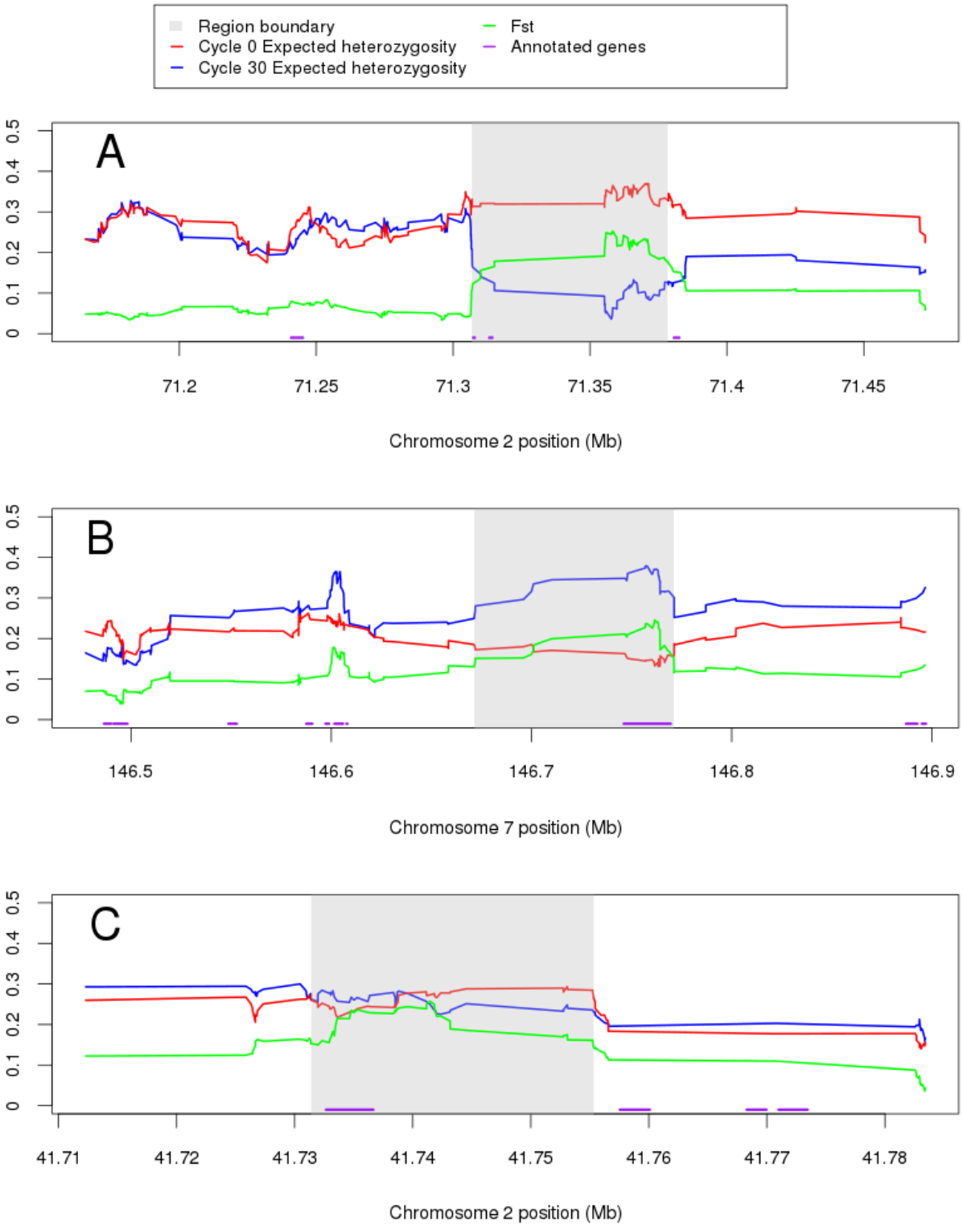


Figure 2.3: Expected heterogyzosity for three regions.

Expected heterozygosity and F_{ST} for three example selected regions of the Golden Glow maize population. Expected heterozygosity was calculated using individual-SNP values of 2p(1-p),

where p is the allele frequency of the minor allele. For the purpose of plotting, values were averaged over 25-SNP sliding windows. A) Expected heterozygosity and F_{ST} over a region that demonstrates a loss in variability between cycle zero and cycle 30. B) Expected heterozygosity and F_{ST} over a region that demonstrates a gain in variability between cycle zero and cycle 30. C) Expected heterozygosity and F_{ST} over a region that demonstrates an insignificant change in variability between cycle zero and cycle 30, even while F_{ST} increased, demonstrating that allele frequencies were changing.

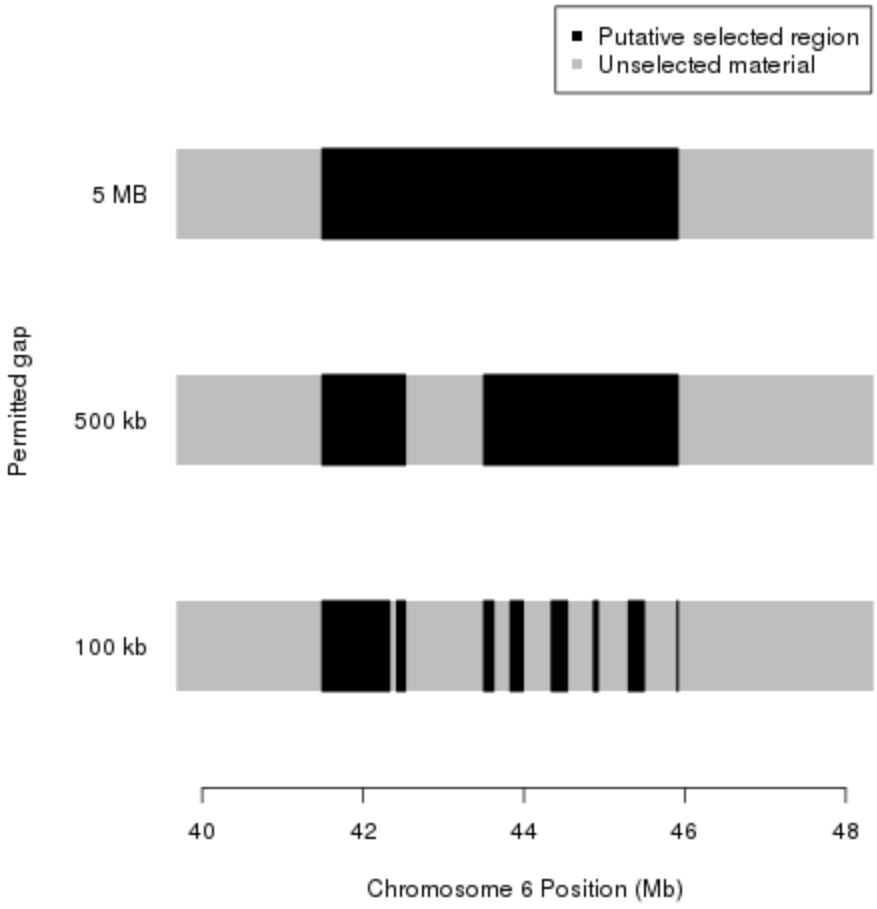


Figure 2.4: A justification for the utilized gap size.

An illustration of the rationale behind using a gap of up to 5 MB between divergent SNPs within a single declared region. Depicted is a region on chromosome six that was identified as potentially under selection. The Grey bar represents the genome, while the black bars show the declared selected region. The top example shows that when a 5 MB gap was permitted, a single region was identified. As the permitted gap between SNPs belonging to the region was decreased, multiple regions were identified. It is more likely that there was a single selected variant and that LD across the region resulted in high F_{ST} values for multiple SNPs than that multiple independent selected variants were in such close proximity to one another. Limitations in our ability to accurately estimate LD from pooled sequencing data prevent us from estimating precise gap width per region.

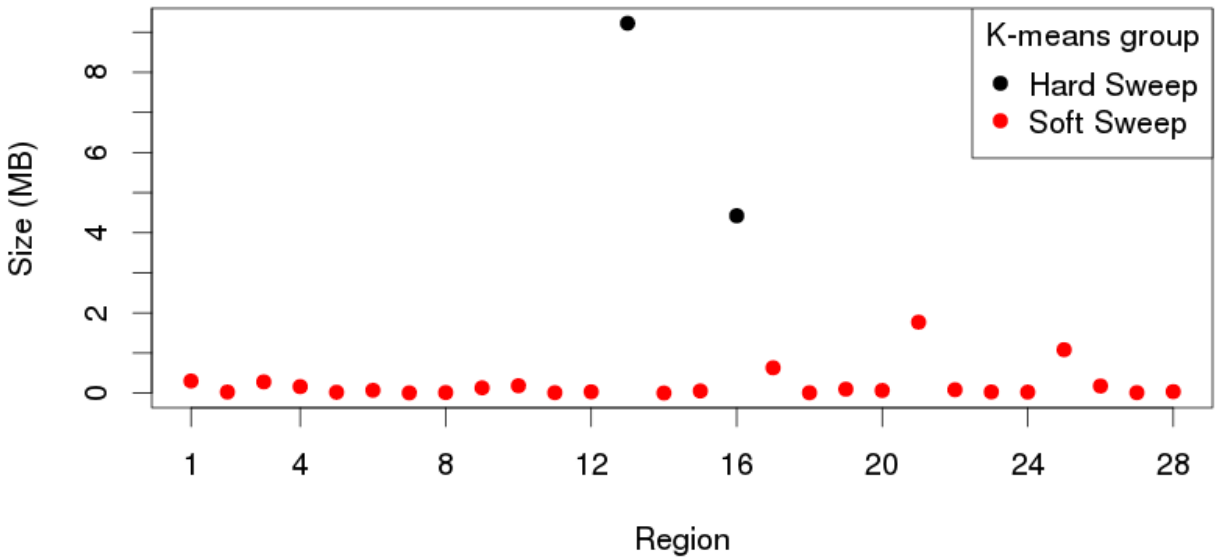


Figure 2.5: Hard and soft sweeps.

The physical size of each region identified as potentially under selection based on F_{ST} is plotted, along with the size-group that the region was placed in according to k-means clustering. Regions are numbered in the same order as Table 1.

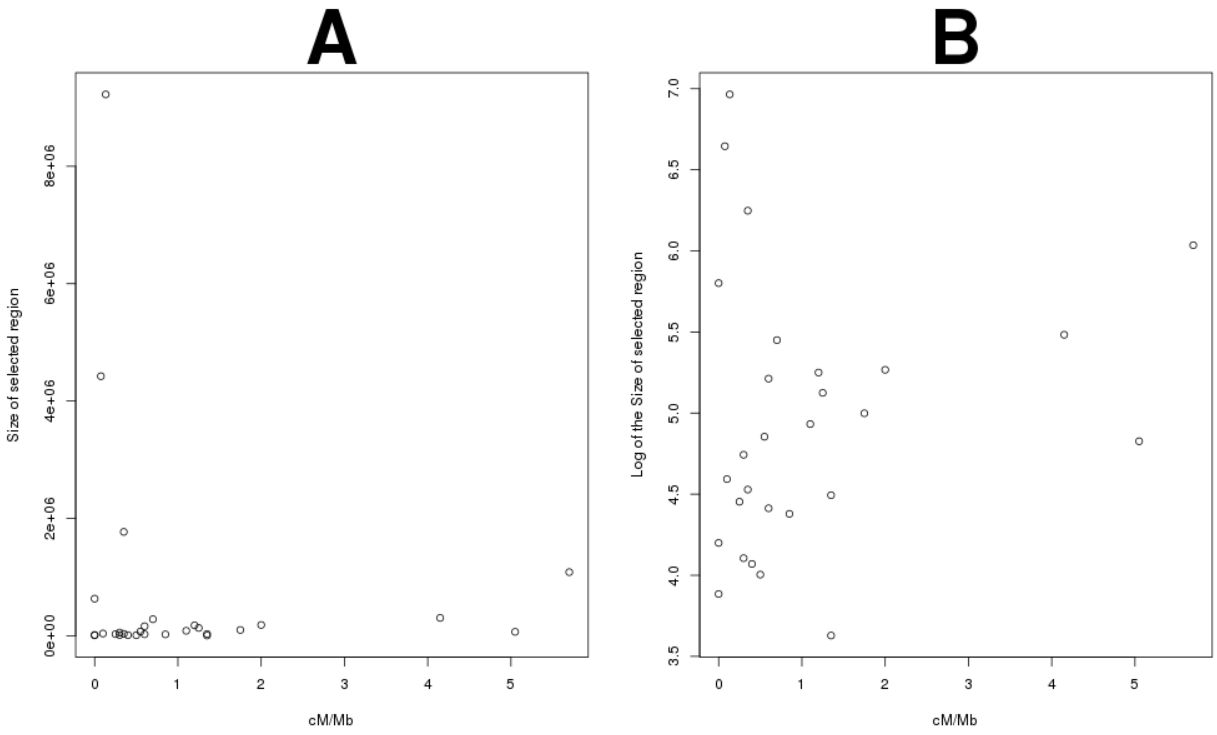


Figure 2.6: Recombination rate vs. physical size.

Recombination rate in the IBM population, a set of recombinant inbred lines derived from the maize inbreds B73 and Mo17, is plotted against the physical size of potentially selected regions that were found in the Golden Glow maize population. Each point depicts a putatively selected region. A: The physical size of each region was compared to the approximate recombination rate for that area of the genome as approximated by the IBM population (correlation = -0.126, p-value = 0.5215). B: The log of the physical size of each region was compared to the approximate recombination rate for that area of the genome as approximated by the IBM population (correlation = 0.172, p-value = 0.3807).

2.8 Tables

Table 2.1: Information for putatively selected regions.

Position, size, and expected heterozygosity information for each of the 28 highly divergent regions identified. The test significance cell for regions that displayed significant changes in heterozygosity (p-value=0.0009) is highlighted in grey.

Region Number	Chromosome	Start Position	End Position	99.99% Significance	Genes Contained	Heterozygosity test significance	Mean change in heterozygosity
1	1	11,588,371	11,892,655	N	8	*0.000*	0.132
2	1	122,802,601	122,831,005	N	0	0.473	0.024
3	1	164,947,151	165,229,053	N	12	0.191	0.052
4	2	35,519,192	35,682,346	Y	3	*0.000*	0.173
5	2	41,731,365	41,755,299	N	2	0.595	-0.019
6	2	71,306,928	71,378,431	N	3	*0.000*	-0.253
7	2	101,062,088	101,069,759	N	0	0.071	0.076
8	2	160,786,800	160,802,631	N	2	0.608	0.026
9	3	177,548,249	177,681,538	N	2	0.026	-0.047
10	3	215,594,013	215,778,968	N	4	0.014	-0.111
11	4	66,924,240	66,935,990	N	0	*0.000*	0.196
12	4	82,825,221	82,858,997	N	0	0.006	-0.131
13	4	113,455,144	122,680,452	Y	73	*0.000*	0.080
14	4	191,396,139	191,400,390	N	1	0.298	0.051
15	5	30,083,952	30,139,317	N	1	0.868	0.005
16	6	41,490,195	45,914,266	Y	42	*0.000*	0.122
17	6	75,749,792	76,382,768	N	5	0.003	0.099
18	6	119,682,711	119,692,810	N	1	*0.000*	0.229
19	7	146,671,419	146,771,150	N	1	*0.000*	0.211
20	7	167,742,364	167,809,449	N	1	0.484	-0.034
21	8	92,876,772	94,647,137	N	26	*0.000*	0.025
22	8	118,681,864	118,767,444	N	3	0.106	0.069
23	9	26,149,935	26,181,104	N	0	0.809	-0.010
24	9	101,071,793	101,097,690	N	1	*0.000*	-0.243
25	10	7,635,223	8,719,903	N	13	*0.000*	0.056
26	10	18,846,988	19,024,881	N	1	0.931	-0.004
27	10	25,251,913	25,264,660	N	0	0.032	0.089
28	10	97,503,134	97,542,318	N	0	*0.000*	0.171

2.9 References

Ahmad, M., R. Ahmad, M. Ishaque, and A. U. Malik, 2011 Why do maize hybrids respond differently to variations in plant density? Crop Environ. 2: 52–60.

Akey, J. M., 2009 Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Research 19: 711–722.

Akey, J. M., G. Zhang, K. Zhang, L. Jin, and M. D. Shriver, 2002 Interrogating a high-density SNP map for signatures of natural selection. Genome Research 12: 1805–1814.

Bigham, A., M. Bauchet, D. Pinto, Z. Mao, J. M. Akey, *et al.*, 2010 Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genetics 6: e1001116.

Burke, M. K., J. P. Dunham, P. Shahrenstani, K. R. Thornton, M. R. Rose, *et al.,* 2010 Genome-wide analysis of a long-term evolution experiment with *Drosophila.* Nature 467: 587-590.

Carlone, M. R. and W. A. Russell, 1987 Response to plant densities and nitrogen levels for four maize cultivars from different eras of breeding. Crop Science 28: 465–470.

Clark, R. M., T. N. Wagler, P. Quijada, and J. Doebley, 2006 A distant upstream enhancer at the maize domestication gene *tb1* has pleiotropic effects on plant and inflorescent architecture. Nature Genetics 38(5): 594-597.

Coop, G., J. K. Pickrell, J. Novembre, S. Kudaravalli, J. Li, *et al.*, 2009 The role of geopgraphy in human adaptation. PLoS Genetics 5: e1000500.

Coors, J. G. and M. C. Mardones, 1989 Twelve cycles of mass selection for prolificacy in maize I. Direct and correlated responses. Crop Science 29: 262–266.

Cross, H. Z., J. T. Kamen, and L. Brun, 1987 Plant density, maturity and prolificacy effects on early maize. Can. J. Plant Sci. 67: 35–42.

Crow, J. F. and M. Kimura, 1970 *An introduction to population genetic theory.* Harper and Row Publishers, New York.

Danilevskaya, O. N., X. Meng, Z. Hou, E. V. Ananiev, and C. R. Simmons, 2008 A genomic and expression compendium of the expanded PEBP gene family from maize. Plant Physiology 146: 250–264.

de Leon, N. and J. G. Coors, 2002 Twenty-four cycles of mass selection for prolificacy in the Golden Glow maize population. Crop Sci. 42: 325–333.

de Leon, N. J. G. Coors, and S. M. Kaeppler, 2005 Genetic control of Prolificacy and Related Traits in the Golden Glow Maize Population II: Genotypic analysis. Crop Sci. 45: 1370-1378.

Doebley, J., A. Stec, J. Wendel, and M. Edwards, 1990 Genetic and morphological analysis of a maize-teosinte F2 population: Implications for the origin of maize. Proc. Natl. Acad. Sci. 87: 9888–9892.

Dudley, J. W., 2007 From Means to QTL: The Illinois long-term selection experiment as a case study in quantitation genetics. Crop Science 47: S20 – S31.

Duvick, D. N., 1997 What is yield? pp. 332–335 in *Developing Drought and Low N-Tolerante Maize. Proceeding of a Symposium*, edited by G.O. Edmeades, M. Bänziger, H. R. Mickelson, and C. B. Peña-Valdivia CIMMYT, El Batan, Mexico.

Flori, L., S. Fritz, F. Jaffrezic, M. Boussaha, I. Gut, *et al.*, 2009 The genome response to artificial selection: a case study in dairy cattle. PLoS One 4: e6595.

Gu, J., N. Orr, S. D. Park, L. M. Katz, G. Sulimova, *et al.*, 2009 A genome scan for positive selection in thoroughbred horses. PLoS One 4: e5767.

Hartigan, J. A. and M. A. Wong, 1979 A K-means clustering algorithm. Applied Statistics 28: 100–108.

Hermisson, J. and P. S. Pennings, 2005 Soft sweeps: Molecular population genetics of adaptation from standing genetic variaton. Genetics 169: 2335–2352.

Hufford, M. B., X. Xu, J. Heerwaarden, T. Pyhajarvi, J. Chia, *et al.*, 2012 Comparative population genomics of maize domestication and improvement. Nature Genetics 44(7): 808-811.

Innan, H. and Y. Kim, 2004 Pattern of polymorphism after strong artificial selection in a domestication event. Proc. Natl. Acad. Sci. 101: 10667–10672.

Jiao, Y., H. Zhao, L. Ren, W. Song, B. Zeng, *et al.*, 2012, Genome-wide genetic changes during modern breeding of maize. Nature Genetics 44(7): 812-815.

Johansson, A. M., M. E. Pettersson, P. B. Siegel, and O. Carlborg, 2010 Genome-wide effects of long-term divergent selection. PLoS Genetics 6: e1001188.

Jones, E., W. Chu, M. Ayele, J. Ho, E. Bruggeman, *et al.*, 2009 Development of single nucleotide polymorphism (SNP) markers for use in commercial maize (Zea mays L.) germplasm. Molecular Breeding 24: 165–176.

Kelly, J. K., B. Koseva, and J. P. Mojica, 2013 The genomic signal of partial sweeps in *Mimulus guttatus*. Genome Biol Evol 5(8): 1457-1469.

Kim, Y. and W. Stephan, 2002 Detecting a local signature of genetic hitchhiking along a recombining chromosome. Genetics 160: 765–777.

Kimura, M., 1962 On the probability of fixation of mutant genes in a population. Genetics 47: 713-719.

King, M. and A. C. Wilson, 1975 Evolution at two levels in humans and chimpanzees. Science 188: 107–116.

Krimbas, C. B. and S. Tsakas, 1971 The genetics of *Dacus oleae*. V. Changes of esterase polymorphism in a natural population following insecticide control – selection or drift? Evolution 25: 454–460.

Langmead, B., C. Trapnell, M. Pop, and S. L. Salzberg, 2009 Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10: R25.

Lam, H., X. Xu, X. Liu, W. Chen, G. Yang, *et al.,* 2010, Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nature Genetics 42(12): 1053-1059.

Lewontin, R. C. and J. Krakauer, 1973 Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. Genetics 74: 175-195.

Lee, M., N. Sharapova, W. D. Beavis, D. Grant, M. Katt, *et al.*, 2002 Expanding the genetic map of maize with the intermated B73 x Mo17 (IBM) population. Plant Mol. Biol. 48: 453-461.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, *et al.*, 2009 The sequence alignment/map format and SAMtools. Bioinformatics 25: 2078–2079.

Liu, S., C. Yeh, T. Ji, K. Ying, H. Wu, *et al.*, 2009 *Mu* transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. PLoS Genetics 5(11): e1000733.

Maita, R. and J. G. Coors, 1996 Twenty cycles of biparental mass selection for prolificacy in the open-pollinated maize population Golden Glow. Crop Science 36: 1527–1532.

Matsuoka, Y., Y. Vigouroux, M. M. Goodman, J. Sanchez G., E. Buckler*et al.*, 2002 A single domestication for maize shown by multilocus microsatellite genotyping. Proc. Natl. Acad. Sci. 99: 6080–6084.

Maynard Smith, J. and J. Haigh, 1974 The hitch-hiking effect of a favourable gene. Genet. Res. 23: 23–35.

Metzker, M. L., 2010 Sequencing technologies – the next generation. Nature Reviews Genetics 11: 31–46.

Myles, S., K. Tang, M. Somel, R. E. Green, J. Kelso, *et al.*, 2008 Identification and analysis of genomic regions with large between-population differentiation in humans. Annals of human genetics 72: 99–110.

Nei, M., T Maruyama, 1975 Letters to the editors: Lewontin-Krakauer test for neutral genes. Genetics 80: 395.

Odhiambo, M. O. and W. A. Compton, 1987 Twenty cycles of divergent mass selection for seed size in corn. Crop Science 27: 1113–1116.

Pan, D., S. Zhang, J. Jiang, Q. Zhang, and J. Liu, 2013 Genome-wide deterction of selective signature in Chinese Holstein Cattle. PLoS One 8: e60440.

Parts, L., F. A. Cubillos, J. Warringer, K. Jain, F. Salinas, *et al.*, 2011 Revealing the genetic structure of a trait by sequencing a population under selection. Genome Research 21: 1131–1138.

Payseur, B. A., A. D. Cutter, and M. W. Nachman, 2002 Searching for evidence of positive selection in the human genome using patterns of microsatellite variability. Mol. Biol. Evol. 19: 1143–1153.

Pritchard, J. K., J. K. Pickrell, and G. Coop, 2010 The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current Biology 20: R206-R215.

Prud'homme, B., N. Gompel, A. Rokas, V. A. Kassner, T. M. Williams, *et al.*, 2006 Repeated morphological evolution through cis-regulatory changes in a pleiotropic gene. Nature 440: 1050–1053.

Przeworski, M., G. Coop, and J. D. Wall, 2005 The signature of positive selection on standing genetic variation. Evolution 59: 2312–2323.

Qanbari, S., D. Gianola, B. Hayes, F. Schenkel, S. Miller, *et al.*, 2011 Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. BMC Genomics 12: 318.

Qanbari, S., T. M. Strom, G. Haberer, S. Weigend, A. A. Gheyas, *et al.*, 2012 A high resolution genome-wide scan for significant selective sweeps: An application to pooled sequence data in laying chickens. PLoS One 7: e49525.

R Core Team, 2013 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austraia. URL http://www.R-project.org/

Robertson, A. 1975 Gene frequency distributions as a test of selective neutrality. Genetics 81: 775-785.

Ross, A. J., A. R. Hallauer, and M. Lee, 2006 Genetic analysis of traits correlated with maize ear length. Maydica 51: 301–313.

Russell, W. A., 1984 Agronomic performance of maize cultivars representing defferent eras of breeding. Maydica 24: 375–390.

Sabeti, P. C., D. E. Reich, J. M. Higgins, H. Z. P. Levine, D. J. Richter, *et al.,* 2002, Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

Sabeti, P. C., P. Varilly, B. Fry, J. Lohmueller, E. Hostetter, C. Cotsapas, *et al.*, 2007 Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

Saghai-Maroof, M. A., K. M. Soliman, R. W. Jorgensen, and R. W. Allard, 1984 Ribosomal DNA spacer-length polymorphisms in barley: Mendelian inheritance, chromosomal location, and population dynamics. Proc. Natl. Acad. Sci. 81: 8014–8018.

Schnable, P. S. and N. M. Springer, 2013 Progress toward understanding heterosis in crop plants. Annual Reviews of Plan Biology 64: 10.1146/annurev–arplant–042110–103827.

Schnable, P. S., D. Ware, R. S. Fulton, J. C. Stean, F. Wei, *et al.*, 2009 The B73 maize genome: complexity, diversity and dynamics. Science 326: 1112–1115.

Sekhon, R. S., H. Lin, K. L. Childs, C. N. Hansey, C. R. Buell, *et al.*, 2011 Genome-wide atlas of transcription during maize development. The Plant Journal 66: 553–563.

Strasburg, J. L., N. A. Sherman, K. M. Wright, L. C. Moyle, J. H. Willis, *et al*., 2012 What can patterns of differentiation across plant genomes tell us about adaptation and speciation? Phil. Trans. R. Soc. B 367: 364-373.

Subandi, 1990 Ten cycles of selection for prolificacy in a composite variety of maize. Indonesian Journal of Crop Science 5: 1–11.

Sucena, E. and D. L. Stern, 2000 Divergence of larval morphology between Drosophila sechellia and its sibling species caused by cis-regulatory evolution of ovo/shaven-baby. Proc. Natl. Acad. Sci. 97: 4530–4534.

Teotónio, H., I. M. Chelo, M. Bradic, M. R. Rose, and A. D. Long, 2009 Experimental evolution reveals natural selection on standing genetic variation. Nature Genetics 41(2): 251-257.

Teshima, K. M., G. Coop, and M. Przeworski, 2006 How reliable are empirical genome scans for selective sweeps? Genome Research 16: 702–712.

Turner, T. L., A. D. Stewart, A. T. Fields, W. R. Rice, and A. M. Tarone, 2011 Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in Drosophila melanogaster. PLoS Genetics 7: e1001336.

Voight, B. F., S. Kudaravalli, X. Wen, and J. K. Pritchard, 2006 A Map of Recent Positive Selection in the Human Genome. PLoS Biology 4: e72.

Weir, B. S. and C. C. Cockerham, 1984 Estimating F-statistics for the analysis of population structure. Evolution 38(6): 1358-1370.

Wisser, R. J., S. C. Murray, J. M. Kolkman, H. Ceballos, and R. J. Nelson, 2008 Selection mapping of loci for quantitative disease resistance in a diverse maize population. Genetics 180: 583–599.

Wray, G. A., 2007 The evolutionary significance of *cis*-regulatory mutations. Nature Reviews Genetics 8: 206-216.

Wright, S., 1931 Evolution in Mendelian populations. Genetics 16: 97–159.

Sright, S. I., I. V. Bi, S. G. Schroeder, M. Yamasaki, J. F. Doebley, *et al*., 2005 The effects of artificial selection on the maize genome. Science 308: 1310-1314.

Ziwen He, W. Zhai, H. Wen, T. Tang, Y. Wang, et al., 2011 Two evolutionary histories in the genome of rice: the roles of domestication genes. PLoS Genetics 7(6): e1002100.

Chapter 3 Defining Window-Boundaries for Genomic Analyses Using Smoothing Spline Techniques, With Applications to Population Diversity Data

Authors: Timothy M. Beissinger[*,§], Guilherme J. M. Rosa[§,**], Shawn M. Kaeppler[*,§§], Daniel Gianola[§,‡], Natalia de Leon[*,§§]

Affiliations: [*] Department of Agronomy, University of Wisconsin, Madison, 53706

[§] Animal Sciences Department, University of Wisconsin, Madison, 53706

[‡] Department of Dairy Science, University of Wisconsin, Madison, 53706

[**] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, 53792

[§§] Department of Energy Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, 53706

Note: The "GenWin" R function referred to throughout this chapter will be made available to the public as an R package on CRAN.

3.1 Abstract

High-density genomic sequence data is often analyzed by combining information over windows of adjacent markers. Interpretation of data grouped in windows versus by individual data point may increase statistical power, simplify computation, reduce technical and sampling noise, and reduce the total number of tests performed. However, use of adjacent marker information can result in over- or under-smoothing, undesirable window boundary specifications, or highly correlated test statistics. We introduce a method for defining windows based on statistically guided breakpoints in the data as a basis for analysis of multiple adjacent data points. This method involves first fitting a cubic smoothing spline to the data and then identifying the inflection points of the fitted spline, to serve as boundaries of adjacent windows. This technique does not require prior knowledge of linkage disequilibrium, and therefore is applicable to data collected from individual or pooled sequencing experiments. Moreover, in contrast to existing methods, an arbitrary choice of window size is not necessary, as these are determined empirically and allowed to vary along the genome. Simulations applying the method to pooled sequencing F_{ST} data demonstrate that the method often generates better than twice the discovery to false positive ratio than that of currently available approaches. Additionally, a comparison of the approach to a previous study involving pooled sequencing F_{ST} data from maize suggests that outlying windows are better isolated from their neighbors than when using a standard sliding window approach. This method can be implemented using an R function called "GenWin".

3.2 Background

A persistent question that arises during the analysis of high-density genotyping or sequencing information is how to best analyze noisy data. For instance, it is particularly pertinent when analyzing sequence data from pooled samples of populations where, depending on the number of individuals pooled and the level of coverage per site, estimates of individual base pair (bp) allele frequencies can be very imprecise (Zhu *et al.*, 2012). To account for this variability, methods based on estimating parameters over "windows" have been used successfully to diminish error while retaining true signal for objectives such as identifying evidence of selection in populations (Turner *et al*., 2011; Qanbari *et al*., 2012; Kelly *et al*., 2013; Beissinger *et al*., 2014). A brief general description of these window-based techniques is that observations from individual genetic markers, often single nucleotide polymorphisms (SNPs), are treated as samples that are representative of a phenomenon that affects a small region of the genome, as opposed to one that affects SNPs independently. In the context of identifying selection signatures, genetic hitchhiking (Maynard Smith and Haigh, 1974) makes this approximation quite reasonable, but it may be sensible in other settings as well since, with increasingly dense markers, linkage disequilibrium (LD) between SNPs within any particular region is likely to be substantial. Therefore, a summary statistic may be computed for an entire window instead of for an individual SNP. This summary statistic can be as simple as taking the mean of single-SNP estimates (Qanbari *et al*, 2012), or it can take a more complex form such as an aggregated measurement of divergence according to the Fisher's angular transformation (Fisher and Ford, 1947; Kelley et al, 2013). By utilizing a sample of observations that are considered each to be an estimate of the same phenomenon, as opposed to treating observations individually, sampling error may be markedly reduced while retaining true signal. An inherent assumption of these

methods is that the individual marker estimates within a window are independently and identically distributed.

Two types of approaches for delineating window boundaries are commonly utilized. These may be referred to as "distinct windows", where markers in different windows do not overlap, and "sliding windows", where they do. When using distinct windows, the genome is divided into disjoint segments of equal length, with length defined according to either number of SNPs (Johansson *et al*, 2010; Kelly *et al*, 2013), or number of bp (Hider *et al*, 2013). A summary statistic, such as the mean, is then calculated over all SNPs within a defined window. Distinct windows often succeed at reducing the sampling error of estimates while reducing the number of statistical tests being performed, but the placement of windows is random or sequential, so power may be lost if window placement inadvertently splits one meaningful region into adjacent windows. In a sliding window approach, a window length (again in number of bp or SNPs) is defined and windows are incrementally advanced along the genome, a single or a few SNPs at a time, ensuring that every possible window is considered (Burke *et* al, 2010; Rubin *et al*, 2010). Using such an approach, however, the number of tests is not dramatically reduced since a new window is defined for every SNP or every few SNPs. Also, highly-correlated statistics are generated since each window overlaps with those neighboring it.

Beyond the limitations mentioned above, in the case of either distinct or sliding windows, determining the proper window size is typically subjective, with researchers often only loosely justifying their choice of size (Akey *et al*, 2009; Burke *et al*, 2010), or acknowledging that their choice is "arbitrary" (Turner *et al*, 2011; Myles *et al*, 2008). This is dissatisfying for two reasons. Firstly, there should be an optimum window size that balances noise reduction with signal identification to maximize power, and identifying this optimum point would be ideal.

Secondly, a subjective definition of window size typically leads to the use of a uniform window size across the genome, which is not appropriate since various genetic parameters, especially recombination rate and LD, vary along a chromosome.

To address these problems, we have developed an empirically-driven framework for determining appropriate window sizes while simultaneously defining their ideal boundaries. Our method retains the benefits of distinct windows and therefore reduces the total number of tests and generates values that are not inherently correlated, while also borrowing from sliding-windows by reducing the risk of erroneously dividing signal between adjacent windows. Additionally, the ideal window size is allowed to vary along the genome, fitting with the biological framework. The method is based on first fitting a cubic smoothing spline (Wahba, 1990) to single-SNP estimates of a value such as F_{ST} (Weir and Cockerham, 1984). Various forms of smoothing splines have been used to analyze genomic information previously (Zhang *et al*, 2003; Pintus *et al*, 2013), but their use for the purpose of defining windows has not been previously reported. The smoothness of the spline is chosen via leave-one-out cross-validation, to ensure it optimally predicts single-SNP values. The second derivative of the spline is then computed and inflection points are identified. Finally, inflection points are treated as window boundaries and a distinct-window analysis proceeds. By utilizing inflection points to define window boundaries, it is assumed that any peak in the spline (a predicted peak of the underlying phenomenon), is placed in a single window instead of split across windows. Cross-validation of the spline's smoothness leads to window-size determination that is likely to be appropriate. Additionally, although a uniform smoothness is chosen for the fitted spline, this does not explicitly restrict the location of its inflection points, thereby allowing non-uniformity of window sizes.

In this manuscript, we describe a smoothing spline-based approach for defining windows using genomic data. In addition, we apply the method to both simulated and real data to identify signatures of selection, and demonstrate the method's advantages over previously used techniques. Although we present this method in the context of F_{ST}-based studies, the method should be applicable to other several other methods that require the pooling of genotypic data over windows. This method has been implemented in a freely available R function, GenWin (Appendix B).

3.3 Methods

Spline technique

A series of individual markers along a chromosome provide several error-prone estimates of specific statistics such as F_{ST}. Therefore, observations from individual markers may be treated as estimates of an underlying continuous function, *f*, that specifies the true value of the statistic of interest at every position. Within this framework, various smoothing spline methodologies (Wahba, 1990) may be used to estimate *f* and therefore its value at any position, $f(t_i)$, where t_i is the chromosomal position in bp of marker i. If *f* is assumed to be continuous and twice differentiable, it may be approximated via a cubic smoothing spline (Silverman, 1985). The cubic smoothing spline estimate, $\hat{f}$, of a function *f,* is defined as the solution that minimizes *S(f)*, where

$$S(f) = \sum\{Y_i - f(t_i)\}^2 + \lambda \int f''(x)^2 dx \, .$$

Here, Y_i is the observed realization of the function and $\hat{f}$ is restricted to be a member of the class of twice-differentiable functions. This formulation seeks to minimize the sum of squared errors of estimates obtained using $\hat{f}$, while ensuring that $\hat{f}$ is fairly smooth. This is achieved by penalizing the sum of squared errors by the integral of the squared derivative of *f*, at a rate determined by a smoothing parameter, λ. This parameter may be chosen via cross validation, so that the minimizer of *S(f)* is the function that provides the best predictive ability of the observed data. It has been shown (Reinsch, 1967) that $\hat{f}$ is a piecewise-cubic polynomial, where pieces are joined at marker positions and that even at these positions the first and second derivatives of $\hat{f}$ are continuous.

An overview of the smoothing spline method for defining sliding windows is provided in Figure 3.1. In the first step, a smoothing spline, $\hat{f}$, is fitted to the raw data measured on individual SNPs. Next, this fitted spline is used to identify the positions at which to split the data for window-based analysis. Specifically, the inflection points of the fitted spline (positions where $\hat{f}'' = 0$) are taken as window boundaries. Because $\hat{f}$ will necessarily be concave-down at every local maxima, this ensures that every potential peak in the spline and therefore every predicted peak in the underlying data, falls into a single window rather than being split between windows. Additionally, large windows are created in regions where $\hat{f}$ is mostly flat, which implies a low amount of signal relative to noise, and small windows are created in regions where $\hat{f}$ is rougher, indicating higher signal relative to noise. Once windows have been defined, analyses may proceed as with any methodology that involves genomic windows. However, the non-uniformity of window sizes must be appropriately accounted for in such an analysis. Certain statistics naturally account for this variability. For instance, a simple t-test to assess changes in expected heterozygosity between two populations will appropriately handle differences in the number of

observations per window. However, certain situations require a more cautious treatment of the variability in window sizes. F_{ST}-based scans for selection, for example, often utilize outlier-based thresholds to identify potentially interesting regions with high F_{ST} values (Akey, 2009) . In this setting, seemingly high values from small windows may be less informative than seemingly intermediate values from large windows, due to the greater sampling error associated with fewer markers being included smaller windows. A reasonable approximation is to consider individual markers as independent and identically distributed observations with some underlying mean value across the window. Then, a T-test like statistic, *W,* may be generated such that $W = (\bar{X} - \mu) \Big/ \sqrt{s^2/n}$, where $\bar{X}$ is the mean value over the window, μ is the mean value over the entire dataset, s^2 is the sample variance of F_{ST} across the entire dataset, and *n* is the number of observations in the window. Thus, each window will have a specific value of *W*, which may be used to compare among windows of varying size and to identify outliers. Similarly to a T-test statistic, the *W* statistic penalizes the mean across each window according to its difference from the grand mean, the overall variability of the data, and most importantly the number of individual observations used to compute that mean. Moreover, although *W* follows the form of a t-statistic, it is not expected to follow a T-distribution, e.g., for selection scans where multiple generations of genetic drift adds variability between populations. Still, *W* scales means computed from windows of unequal size so that comparisons are possible and outliers may be identified.

R function

The function "GenWin" is an R (R Core Team, 2013) function that may be used to implement the smoothing spline method described above (Appendix B). A cubic smoothing

spline is fitted to single-SNP estimates of some parameter of interest, e.g. F_{ST}, accounting for the position of each estimate in bp. GenWin depends on the "pspline" package (Ramsey and Ripley, 2013) for rapidly fitting a cubic smoothing spline to data. The smoothing parameter may be chosen via cross validation (CV) or generalized cross validation (GCV) (Craven and Wahba, 1978). The inflection points of the spline are identified by isolation of the points where the second derivative switches sign, and the resolution over which inflection points are computed may be specified by the user. Computing second derivatives at every bp slows down computation and may lead to erratic window boundaries, while doing this every few thousand bps may allow properties of the fitted spline to be missed. After inflection points are identified, these are used to define window boundaries, and a variety of statistics for each window, including the *W* statistic described above, can be returned. Plotting of the spline fitted to the raw data is optional. Only one chromosome should be analyzed at a time, as generating a function that is continuous between the end of one chromosome and the beginning of another is not biologically sensible. The GenWin function takes between a few seconds and a few minutes on a typical workstation to analyze 100,000 markers that may lead to several thousand windows, depending on the smoothness of the fitted spline.

Simulations

The software QMSIM (Sargolzaei and Schenkel, 2009) was used to simulate an artificially selected population suitable for testing the spline-window method. A diploid, ten chromosome species was simulated, with each chromosome being 200 centiMorgans (cM) and 100 MB in length. First, 5,000 historical generations with 5,000 random mating individuals per

generation were simulated to establish a base population to undergo selection. No selection took place during the 5,000 historical generations. Next, 100 replications of selection were conducted, where selection was carried out on a trait affected by 30 QTL and with a heritability of 0.5. Selection based on "high" phenotypes was carried out for 30 discrete generations, with 500 males and 500 females selected to contribute gametes each generation, and a litter size of 50 individuals per female (i.e. census population size = 25,000). Three QTL and 100,000 markers were simulated on each chromosome. Marker positions were assigned randomly, to a resolution of 10^{-5} cM, while for each chromosome QTL were placed at precisely 50 cM, 100 cM, and 150 cM. QTL effects were randomly sampled from a normal distribution. Markers and QTL were both di-allelic, and recurrent mutation was permitted during the historical generations at a rate of 2.5×10^{-5}.

The output from QMSIM included allele frequencies for each of the 1,000,000, markers pre- and post-selection, in each of the 100 replicated populations. Binomial sampling was conducted within R (R Core Team, 2013) to further simulate a set of pooled genotyping data for analysis. For every marker within each simulated replicate, in each of the pre-selection and post-selection populations, 100 individuals (200 gametes) were sampled to create a simulated set of individuals for genotyping, and then 50 samples were drawn from those samples to approximate pooled sequencing to 50X coverage. It has been shown that pooled sequencing is well approximated by binomial sampling (Zhu *et al*, 2012). Thus, this process generated a set of estimated allele frequencies corresponding to a population that had undergone thirty generations of selection, followed by sampling 100 individuals to be sequenced, with a sequencing depth equivalent to a 50X coverage.

Often, a study comparable to this simulated one would be analyzed using outlier-thresholds to identify potentially selected sites (e.g., Amaral *et al,* 2011). However, since this population was fully derived via simulation, we were able to use the simulated population without selection (e.g. the "null" model) to define significance thresholds. For this model, the only changes to the previously described protocol were that individuals were selected independently of their trait performance, and 20 replicated populations (again including 1,000,000 di-allelic markers per population) were simulated.

To evaluate the spline-window method's performance compared to either sliding or distinct windows of various sizes, F_{ST} values between the pre- and post-selection populations were computed for each marker according to $F_{ST} = \frac{s^2}{\bar{p}(1-\bar{p})+s^2/r}$, where s^2 is the sample variance of allele frequency between populations, $\bar{p}$ is the mean allele frequency across populations, and *r=2* is the number of populations (Weir and Cockerham, 1984). Sliding window and distinct window values were computed for windows of five, 10, 25, 50, 100, 250, and 500 SNPs. Additionally, the spline-window method was used with *W* statistics for comparison between windows of unequal size, utilizing GCV and a resolution of 100 bp. Significance thresholds for each method were determined using simulations of no selection by identifying the maximum observed value in each unselected replication and taking the 95% quantile of these values. Finally, the simulations were analyzed for true positive (i.e. detection) and false positive rates. Windows that exceeded the simulated significance thresholds were deemed true positives if they fell within 5 cM of a simulated QTL, and were deemed false positives otherwise.

Empirical data analysis

Maize data previously published by Beissinger et al (2014) was re-analyzed. The data involved a maize population subjected to artificial selection for thirty generations for an increase in the number of ears per plant. Pooled sequencing was conducted pre- and post-selection, and estimated F_{ST} values for approximately 1.2 million SNP markers were available. In the previous analysis, sliding windows of 25 SNPs were used, and a 99.9% outlier threshold was employed to identify the most divergent regions of the genome, likely to have been under selection pressure. For this re-analysis, the spline-window method was applied and GCV was used to choose the smoothing parameter. Again, a 99.9% outlier threshold was employed to identify outlying *W* statistics signifying regions with a likely selection signature. For consistency with the previous analysis, outlying windows within 5 MB of one another were grouped together as likely corresponding to the same selection event. Results from the spline-window analysis were compared to those of the previously published sliding-window analysis to determine the degree of overlap between the methods.

3.4 Results

Simulations

Simulations showed that both sliding windows and distinct windows of five or 10 SNPs identified markedly fewer QTL than the other methods (Table 3.1). With such small windows, the data were extremely variable and therefore significance thresholds set by the no-selection simulations were so high that it was extremely difficult for them to be exceeded. The positive aspect of this, however, is that these four methods identified fewer false positives than their respective class of method (i.e. sliding or distinct) that included more SNPs per window. In fact,

the distinct window methods of size five or 10 SNPs identified the fewest false positives of all methods investigated. For these methods, however, low false positive rates came at the expense of reasonable detection rates. For instance, either sliding or distinct windows of only five SNPs identified, on average, fewer than 25% of the simulated QTL, which was substantially fewer than the other methods. Conversely, all sliding and distinct window implementations of 25 or more SNPs, as well as the spline-window method, managed to identify a comparable number of QTL on average, all with mean detection rates greater than 50% of the total number of QTL but less than 66.67%. It should be noted that depending on QTL allele frequencies at the onset of selection, it is not expected for every QTL to be detectable, so maximum detection rates below 66.67% are not necessarily surprising. These methods had large variability in the number of false positives that were identified, with the sliding window methods showing especially large numbers of false positives. For this reason, the ratio of detected QTLs to false positives was used to evaluate the performance of each method. Excluding the five or 10 SNP window methods due to their low detection rates, as discussed previously, the ratio of detection rate to false positive rate of the spline-window method (4.7) was more than double that of the second best performing technique, distinct windows of 25 SNPs (2.18).

Two important remarks should be made. First, distinct windows of five or 10 SNPs had the most favorable ratio of detection rate to false-positive rate of all the methods (61.83 and 12.31, respectively), but the cost of this benefit was identifying notably fewer QTL in total. Therefore, researchers interested in only the very most promising sites may be best served by adopting a relatively small window size and taking only the very most outlying windows as worthy of further study. This approach is likely to find the most extreme QTL and by limiting the search to only the very most outlying sites, the expectation is that false positives are rare. The

second point is that, for any method, the ideal window size results from an interplay between the true amount of signal in the data and the amount of error that results from sampling and genotyping. Therefore, there is no single 'ideal' window size that will hold across experiments, but instead the best window size will vary depending on the genetic structure underlying the trait under study and the genotyping methods applied. The spline-window method provides a useful alternative by letting the variability in the data itself determine the appropriate window size.

Real data analysis

The spline-window analysis of the previously published maize data identified 23 unique regions exceeding a 99.9% empirical outlier level which are expected to be associated with selection (Table 3.2). Within these 23 regions, 17 overlapped with those identified in the previously published analysis, while six were novel discoveries that had not been identified using sliding windows (in one case two spline-based regions corresponded to a single previously reported region). In addition, 12 of the regions that had been identified previously were no-longer outliers according to the spline-window analysis. As expected a substantial amount of variability in the size of the windows was observed using the spline method. While the previous analysis restricted all windows to precisely 25 SNPs, the spline approach suggested that approximately 64.3% of windows should be smaller than 25 and 34.0% should be bigger, with only 1.6% of windows being estimated at exactly 25 SNPs. Moreover, 10% of windows exceeded 51 SNPs and the maximum fitted window size was 349 SNPs, which implies that a large amount of variability in the noisiness of the data will be inappropriately accounted for if a single window size is used.

Additionally, the spline-window approach appeared to be superior at separating outlying regions from the background variability of the data compared to the previous analysis that employed 25-SNP sliding windows. While sliding windows necessarily lead to correlations between adjacent windows, causing an outlying window to be surrounded by other windows that are also outlying or nearly outlying, spline-based windows do not share this property. Specifically, defining windows with splines allows each significant or outlying window region to have a clearly-defined start and stop at the underlying inflection points of the spline. Therefore, except in cases of selective-sweeps that spanned several megabases, outliers identified based on the inflection points of the fitted spline generally were well distinguished from their neighbors (Figure 3.2).

3.5 Discussion

This study demonstrates that arbitrarily defining a window size for the analysis of high-throughput genotypic data and then proceeding to analyze an experiment across sliding or distinct windows of the specified size could present limitations. Small window sizes will tend to diminish the potential discovery rate of the study, while large window sizes will tend to create an abundance of false positives. The spline-window method avoids both extremes by using patterns in the data to define windows. These windows are placed at interesting features and their size is determined by the variability present in the data itself.

The results of the simulation analysis established that the spline-window method achieved a balance between discovery rate and false positive rate that was considerably better than any of the other methods assessed that identified a comparable number of QTLs. Our re-

analysis of previously published data demonstrated that this technique performs well in experimental situations. The simulated scenario utilized here is expected to resemble the genetic architecture of the empirical data set previously analyzed. Based on that assumption, results suggest that the detection rates of the spline-window and 25 SNP sliding window methods should be similar, while the false positive rate of the sliding window method is substantially higher than that of the spline window approach. This is consistent with the spline-window method identifying 23 selected regions and the previous analysis identifying 28, with 17 regions overlapping.

The spline-window method has potential to be used across multiple types of studies, even though we have presented it in the context of F_{ST}-based scans for selection. This method may apply in various situations where noisy genomic data are divided into windows for analysis. For example, the d_i statistic of Akey et al (2009), evaluations of heterozygosity (e.g.Turner *et al*, 2011), and similar metrics that can be computed for individual loci or across windows fit into this framework very well, since the spline can be fitted to single-locus estimates of any statistic before window-based smoothing is performed. Extending this approach to statistics such as Tajima's D (Tajima, 1989) or EHH (Sabeti *et al*, 2002) and variants thereof, may be possible as well, but this is not straight forward since these statistics must be computed across windows in the first place, and therefore the values that the spline should be fitted to are not as clear.

There are some areas of study that may prove fruitful for improving and extending this method. The first involves the smoothing parameter, λ. When this parameter is chosen via cross-validation, a single value is employed for each chromosome. Since recombination rates, and therefore levels of genomic variability, can vary substantially along a chromosome, this approach may be benefit from an extension to include multiple smoothing parameters fitted simultaneously

to different regions within a chromosome. For example, an ideal spline may be smoother (resulting from a larger smoothing parameter) in centromeric and peri-centromeric regions than elsewhere. Secondly, it is difficult to adequately estimate LD from pooled sequencing data. There has been progress toward this goal (e.g., Feder *et al*, 2012), but the short length of sequencing reads that is currently feasible, relative to typical distances of LD decay, represents a substantial limitation for applying such methods. Spline-windows may be heavily dependent on underlying levels of LD and recombination, and therefore there may be the possibility to extend this general approach as a means to assess LD in pooled sequencing situations.

3.6 Conclusions

An important component of the analysis of data from studies that involve high-density genomic sequence information is how to best group regions of the genome for analysis. This is particularly relevant for identification of selection signatures based on pooled sequencing data, where estimates of features, such as F_{ST}, contain substantial sampling error. We proposed a spline-based method that simultaneously defines ideal boundaries and variable sizes for windows of observations that may be analyzed together. Simulations coupled with empirical data analysis demonstrated that this method is comparably powerful to existing methods but less susceptible to false positives. We have made this method freely and publicly available in the R function "GenWin".

3.7 Figures

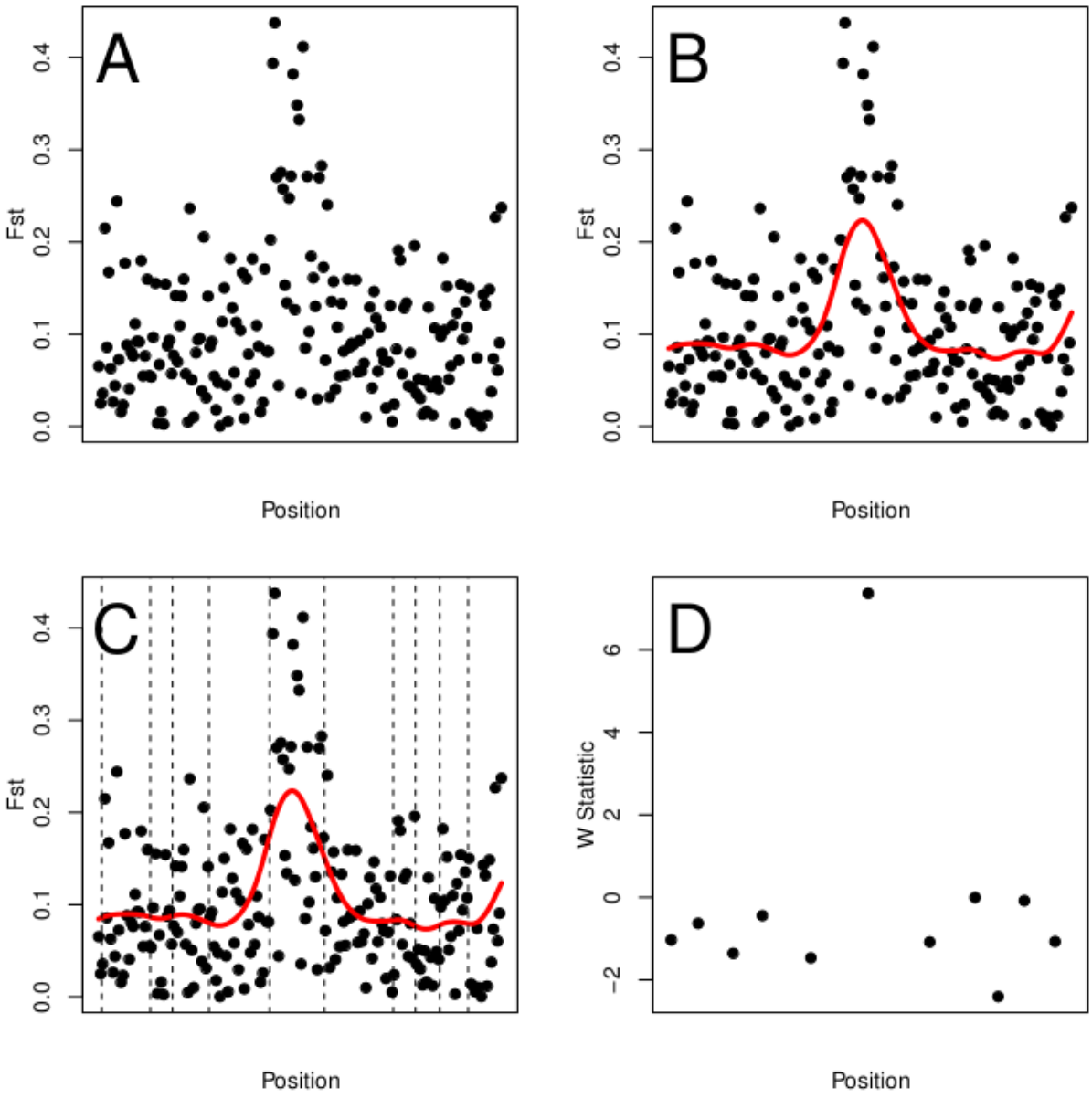


Figure 3.1: Step-by-step depiction of the spline-window method.

The steps of the spline-window method are depicted, using a simulated set of 200 markers across a portion of a chromosome. A) Raw data (F_{ST}) computed on individual markers, is shown. B) A cubic smoothing spline, depicted by the red line, is fitted to the data. C) Inflection points of the spline are depicted by the dashed vertical lines. D) Inflection points of the spline are used to define window boundaries, and a statistic such as W is computed.

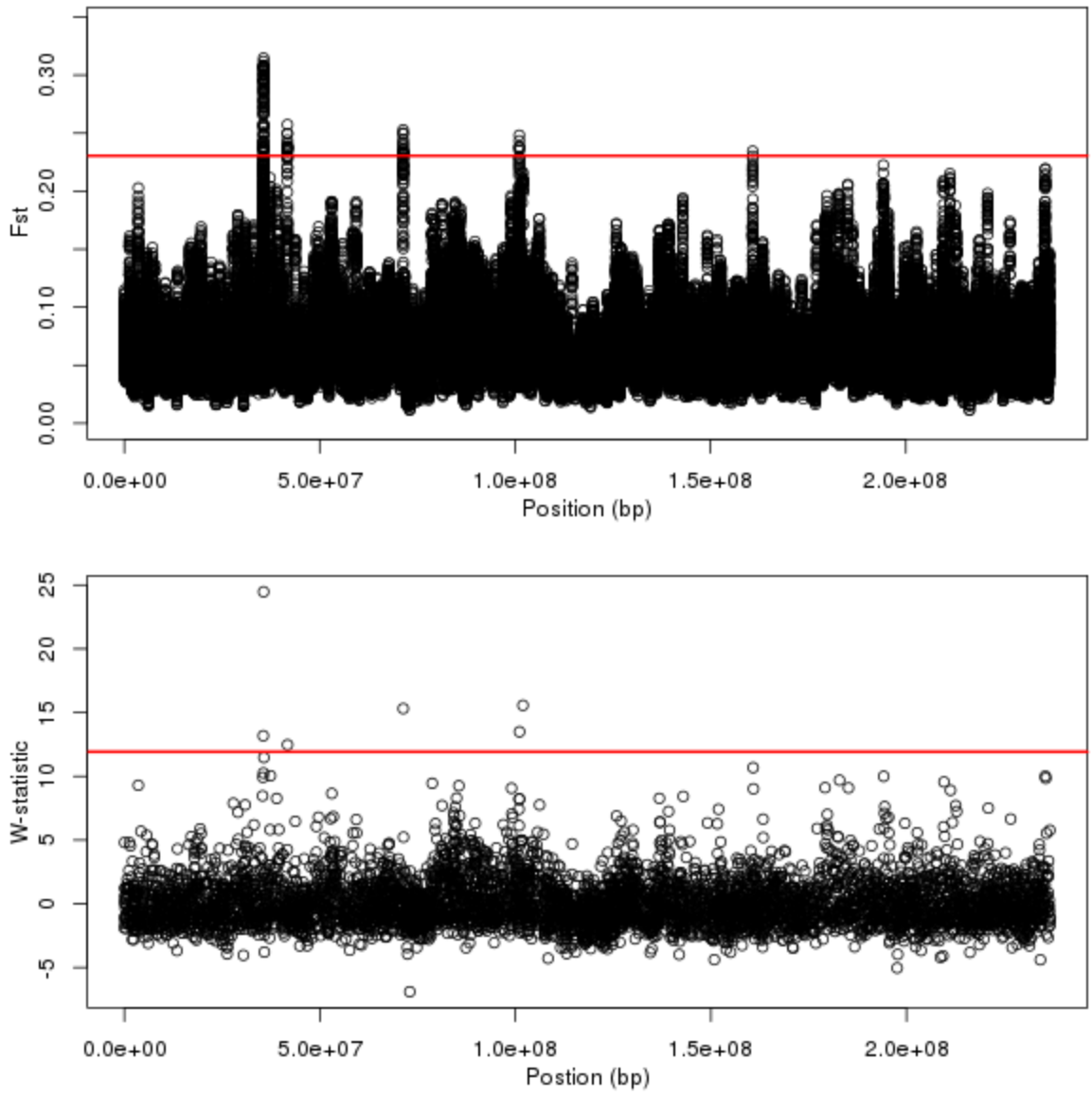


Figure 3.2: A comparison of the spline-window method and a sliding window approach.
A comparison of regions exceeding a 99.9% threshold using 25-SNP sliding windows and spline windows, based on empirical data. The data analyzed are from Beissinger et al. (2014), which was a study of a 30-generation artificial selection experiment for maize ear number. Top (A): Adapted from Beissinger et al. (2014), F_{ST} values and their outlier threshold (red line) found on maize chromosome 2 using 25-SNP sliding windows. Bottom (B): W statistics and their outlier threshold (red line) found using the spline-window method for the same data set.

3.8 Tables

Table 3.1: Power of various window methods.

Results from applying an assortment of window-methods to a simulated selection experiment involving 30 QTL, 30 generations of selection, and pooled sequencing at 1,000,000 markers to estimate allele frequencies. Mean number of QTL detected (out of 30), mean false positive rate, and ratio of detection rate to false positive rate across simulations is provided for each of the methods evaluated.

Method	Mean Detected	Mean False Positives	Detection Ratio
Sliding-5	6.3	6.4	1.0
Sliding-10	13.3	4.2	3.2
Sliding-25	17.6	75.1	0.2
Sliding-50	18.4	232.6	0.1
Sliding-100	18.2	488.2	0.0
Sliding-250	18.4	2,082.5	0.0
Sliding-500	19.5	9,065.1	0.0
Distinct-5	7.4	0.1	61.8
Distinct-10	12.2	1.0	12.3
Distinct-25	16.8	7.7	2.2
Distinct-50	18.0	11.3	1.6
Distinct-100	17.7	9.5	1.9
Distinct-250	19.5	31.5	0.6
Distinct-500	18.3	28.5	0.6
Spline Windows	16.0	3.4	4.7

Table 3.2: Regions detected using spline windows compared to sliding windows.

A comparison of regions exceeding a 99.9% threshold using 25-SNP sliding windows and spline windows, based on empirical data. The data analyzed are from Beissinger et al. (2014), which was a study of a 30-generation artificial selection experiment for maize ear number. Previously published outlying regions identified as putatively controlling number of ears by plant based on 25-SNP sliding windows are compared with those identified applying the spline-window method to the same data set.

	25-SNP sliding window outlier regions		**Spline-window outlier regions**	
Chromosome	Start Position	End Position	Start Position	End Position
1	11,588,371	11,892,655	11686850	11872650
1			54485850	54564950
1	122,802,601	122,831,005	122,790,650	124,093,750
1	164,947,151	165,229,053		
2	35,519,192	35,682,346	35,520,750	35,648,950
2	41,731,365	41,755,299	41,728,850	41,770,550
2	71,306,928	71,378,431	71,314,050	71,377,150
2	101,062,088	101,069,759	101,037,150	102,026,750
2	160,786,800	160,802,631		
3	177,548,249	177,681,538	177,671,050	177,749,050
3			207,464,650	211,847,850
3	215,594,013	215,778,968		
4	66,924,240	66,935,990		
4	82,825,221	82,858,997	82,818,050	82,860,750
4	113,455,144	122,680,452	113,401,750 120,298,350	114,347,650 122,682,750
4			140,791,850	140,834,650
4	191,396,139	191,400,390		
5			24,460,850	24,539,450
5	30,083,952	30,139,317	30,034,650	30,120,950
6	41,490,195	45,914,266	41,517,550	45,921,450
6	75,749,792	76,382,768	76,072,450	76,176,350
6			86,671,650	86,727,750
6	119,682,711	119,692,810	119,683,750	119,707,650
7	146,671,419	146,771,150		
7	167,742,364	167,809,449		
8	92,876,772	94,647,137	94,633,950	94,680,950
8	118,681,864	118,767,444		
9	26,149,935	26,181,104	25,947,850	26,183,950
9	101,071,793	101,097,690		
10	7,635,223	8,719,903	8,703,450	8,718,950
10	18,846,988	19,024,881		
10	25,251,913	25,264,660		
10	97,503,134	97,542,318		
10			136,171,150	136,259,150

3.9 References

Akey, JM: **Construction genomic maps of positive selection in humans: Where do we go from here?** *Genome Res.* 2009 **19**(5): 711-722.

Akey JM, Ruhe AL, Akey DT, Wong AK, Connelly CF, et al: **Tracking footprints of artificial selection in the dog genome.** *Proc. Natl. Acad. Sci.* 2010 **107**(3): 1160-1165.

Amaral AJ, Ferretti LF, Megens J, Crooijmans RPMA, Nie H, et al: **Genome-wide footprints of pig domestication and selection revealed through massive parallel sequencing of pooled DNA.** *PLoS ONE* 2011 **6**(4): e14782.

Beissinger TM, Hirsch CN, Vaillancourt B, Deshpande S, Barry K, et al: **A genome-wide scan for evidence of selection in a maize population under long-term artificial selection for ear number.** *Genetics* 2014 **196**: 829-840.

Burke MK, Dunham JP, Shahrestani P, Thornton KR, Dose MR *et al*: **Genome-wide analysis of a long-term evolution experiment with *Drosophila.*** *Nature* 2010 **467**: 587-590.

Craven P and Wahba G: **Smoothing noisy data with spline functions.** *Numerische Mathematik* 1978 **31**(4): 377-403.

Fisher RA, Ford EB: **The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula*.** *Heredity* 1947 **1**: 143-174.

Hider JL, Gittelman RM, Shah T, Edwards M, Rosenbloom A, et al: **Exploring signatures of positive selection in pigmentation candidate genes in populations of East Asian Ancestry.** *BMC Evol. Biol.* 2013 **13**: 150.

Johansson AM, Pettersson ME, Siegel PB, Carlborg Ö: **Genome-wide effects of long-term divergent selection.** *PLoS Genet.* 2010 **6**(11): e1001188.

Kelly JK, Koseva B, Mojica J.P.: **The genomic signal of partial sweeps in *Mimulus guttatus*.** *Gen. Biol. And Evol.* 2013 **5**(8): 1457-1469.

Maynard Smith J, Haigh J: **The hitch-hiking effect of a favourable gene.** *Genet. Res.* 1974 **23:** 23-35.

Myles S, Tang K, Somel M, Green RE, Kelso J, et al: **Identification and analysis of genomic regions with large between-population differentiation in humans.** *Ann. Hum. Gen.* 2008 **72**(1): 99-110.

Pintus E, Sorbolini S, Albera A, Gaspa G, Dimauro C: **Use of locally weighted scatterplot smoothing (LOWESS) regression to study selection signatures in Piedmontese and Italian Brown cattle breeds.** *Animal Genetics* 2013 **45:** 1-11.

Qanbari S, Strom TM, Haberer G, Weigend S, Gheyas AA, et al: **A high resolution genome-wide scan for significant selective sweeps: An application to pooled sequencing data in laying chickens.** *PLoS ONE* 2012 **7**(11): e49525.

R Core Team: **A language and environment for statistical computing.** Vienna: R Foundation for Statistical Computing; 2013.

Ramsey J and Ripley B: **pspline: Penalized Smoothing Splines. R backage version 1.0-16.** http://CRAN.R-project.org/package=pspline

Reinsch, C: **Smoothing by spline functions**. *Numer. Math.* 1967 **10**: 177-183.

Rubin C, Zody M, Eriksson J, Meadows JRS, Sherwood E, *et al*: **Whole-genome resequencing reveals loci under selection during chicken domestication.** *Nature* 2010 **464**: 587-591.

Sabeti PC, Reich DE, Higgins JM, Levine HZP, Richter DJ: **Detecting recent positive selection in the human genome from haplotype structure.** *Nature* 2002 **419:** 832-837.

Sargolzaei M and Schenkel FS: **A large-scale genome simulator for livestock.** *Bioinformatics* 2009 **25:** 680-681.

Silverman BW: **Some aspects of the spline smoothing approach to non-parametric regression curve fitting**. *Jounal of the Royal Statistical Society*. *Series B (Methodological)* 1985 **47**(1): 1-52.

Tajima F: **Statistical method for testing the neutral mutation hypothesis by DNA polymorphism.** *Genetics* 1989 **123:**585-595.

Turner TL, Stewart AD, Fields AT, Rice WR, Tarone AM: **Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*.** *PLoS GEN.* 2011 **7**(3): e1001336.

Weir BS, Cockerham CC: **Estimating F-statistics for the analysis of population structure.** *Evolution* 1984 **38**(6): 1358-1370.

Whaba G: *Spline Models For Observational Data.* Philadelphia: Society for Industrial and Applied Mathematics; 1990.

Zhang Z, Roeder K, Wallstrom G, Devlin B: **Integration of association statistics over genomic regions using Bayesian adaptive regression splines.** *Hum. Genomics* 2003 **1**(1): 20-29.

Zhu Y, Bergland AO, González J, Petrov DA: **Empirical validation of pooled whole genome population re-sequencing in *Drosophila melanogaster*.** *PLoS ONE* 2012 **7**(7): e41901.

Chapter 4 Using the Variability of Linkage Disequilibrium Between Subpopulations to Scan for Selection in a Diverse Panel of Chickens

Authors: Timothy M. Beissinger[*,§], Mahmoud Gholami[†], Malena Erbe[†], Steffen Weigend[‡], Annett Weigend[‡], Natalia de Leon[*,§§], Daniel Gianola[§,‡‡,**], Henner Simianer[†]

Affiliations:

[*] Department of Agronomy, University of Wisconsin, Madison, 53706

[§] Animal Sciences Department, University of Wisconsin, Madison, 53706

[†] Animal Breeding and Genetics Group, Georg-August-University Goettingen, Germany

[‡] Institute of Farm Animal Genetics, Friedrich Loeffler Institut, Neustadt-Mariensee, Germany

[‡‡] Department of Dairy Science, University of Wisconsin, Madison, 53706

[**] Department of Biostatistics and Medical Informatics, University of Wisconsin, Madison, 53792

[§§] Department of Energy Great Lakes Bioenergy Research Center, University of Wisconsin, Madison, 53706

4.1 Abstract

A whole-genome scan for identifying selection acting on pairs of linked loci is suggested and implemented. The scan is based on D'^2_{IS}, one of Ohta's 1982 measures of between-population linkage disequilibrium (LD). An approximate empirical null distribution for the statistic is derived. Although the partition of LD into between population components was originally used to investigate epistatic selection, we demonstrate that the value of D'^2_{IS} may also be influenced by a single-locus selective sweep with linkage but no epistasis. The proposed scan is implemented in a diverse panel of chickens including 72 distinct breeds. 1,553 locus-pairs are identified as having putatively been under single-locus or epistatic selection. These pairs of loci generally cluster to form overlapping or neighboring signals of selection. Known variants that were expected to have been under selection in the panel are identified, as well as an assortment of novel regions that have been putatively under selection.

4.2 Introduction

A variety of patterns are generated in the genomes of organisms that are undergoing selection. Such patterns depend on a multitude of factors including the demographic history of the population or populations in question, the type of selection that is taking place, and the relative importance of the variant or variants that are being selected. For many of these factors, appropriate tests for selection have been developed and are in wide use. For instance, in cases of directed evolution with experimental populations, especially for those with biological replication, changes in allele frequency may be directly measured to identify single-locus selection (Wisser *et al.* 2008; Turner *et al.* 2011; Parts *et al.* 2011). In a related test, which is particularly useful if samples from pre-selection populations are not available, patterns of nucleotide variability between versus within populations may be leveraged to similarly identify selection at a single locus (Lewontin and Krakauer 1973; Akey *et al.* 2002; Beissinger *et al.* 2013). This type of test may theoretically be able to distinguish between directional selection and balancing selection, since the former is expected to drive allele frequencies toward fixation while the latter will maintain an elevated level of variability (Akey *et al.* 2002). Additionally, selection on an individual locus is known to reduce genetic variability at linked sites (Maynard Smith and Haigh 1974), which has led to an assortment of tests for selection based on data that are observed either within a single population (Sabeti *et al.* 2002; Voight *et al.* 2006) or between populations (Tang *et al.* 2007; Sabeti *et al.* 2007). This class of tests, however, is most powerful for detecting recent strong selection since the length of any haplotypes that show reduced variability, which are the basis of these tests, will decay over time.

An alternative type of selection that is also expected to generate interesting genomic patterns is epistatic selection. In an epistatic selection scenario, the favored phenotype is

dependent on an interacting set of alleles at more than one locus. Therefore, a non-random association, or linkage disequilibrium (LD), between alleles at such interacting loci is expected to increase. If these loci are genetically linked, the LD between them may grow over generations (Kimura 1965). Ohta (1982a; b) created a set of statistics to partition LD into between and within sub-population components, in a manner analogous to Wright's F_{ST} statistics (Wright 1949), which are frequently used to identify single-locus selection based on sub-population variability. According to Ohta, comparisons between these statistics, which are denoted D^2_{IT}, D^2_{IS}, D^2_{ST}, D'^2_{IS}, and D'^2_{ST}, may putatively suggest whether epistatic selection or random genetic drift is the driving force behind observed levels of LD between pairs of markers. Although this type of test was originally implemented in a purely theoretical setting, software has been developed and implemented to apply these statistics to experimental data (Black IV and Krafsur 1985; Garnier-Gere and Dillmann 1992). Interesting examples applying these statistics to evaluate loci for evidence of epistatic selection depict an array of findings. For instance, Song et al. (2009) found that overall drift is more important in shaping LD patterns than is epistatic selection in *Boechera stricta*, as did Vitalis et al. (2002) in *Marsilea strigosa*. Alternatively, however, Miyashita et al. (1993), investigated the patterns of LD across a specific region of *Drosophila melanogaster* and found that epistatic selection is likely to be involved. More recent extensions of Ohta's theory can be found as well. Storz and Kelly (2008), for example, used a similar approach to measure between-population components of the Z_{nS} statistic (Kelly 1997). A related statistic was employed by Ma et al. (2010). Both of these methods were based on expected haplotype frequencies and did not incorporate information regarding gametic phase. Therefore, they represent summaries of the covariance in allele frequencies between loci.

Interestingly, although they have been extensively used since their introduction, Ohta's D-statistics have not been extended to or implemented in a framework that involves dense genome-wide marker or sequence data on the scale which is commonly employed today. The greatest limitation of the statistics is that they were designed for equilibrium model situations under which testing statistical significance was not relevant. Therefore, although an estimate of the relative contributions of random processes compared to epistatic selection for a particular pair of loci may be made, previous research has not addressed the uncertainty of these contributions. Secondly, although D statistics are analogous to Wright's *F* statistics, Ohta's original estimators assume that the effects of epistatic selection should be similar in all subpopulations, which is contrary to the form in which F_{ST} tests are typically (though not exclusively) employed. Modifications have been made that address this discrepancy and that allow high subpopulation differentiation to serve as a signature of selection as well (Black IV and Krafsur 1985), but the difficulty of assigning significance remains. A third and final impediment to applying these statistics in whole-genome scans for epistatic selection involves complications related to distinguishing a signature generated by epistasis from that of a large sweeps resulting from hitchhiking (Maynard Smith and Haigh 1974; Sabeti *et al.* 2002).

Herein we extend the subdivision of LD into its between and within subpopulation components to a framework that may be applied in a whole-genome scan for selection. In doing so, we follow the computational approach of Black IV and Krafsur (1985) in employing Δ_{ij}, the Burrows composite measure of LD (Cockerham and Weir 1977), to compute LD when gametic phase is ambiguous. We identify one of Ohta's statistics, D'^{2}_{IS}, as the most informative for testing pairs of linked loci that are jointly impacted by selection at either the single-locus or epistatic level. We then identify an empirically-based null distribution for these statistics that

measures most of the variability expected due to chance alone, while excluding that which may be generated as a result of linked selection. Using this distribution to define significance thresholds, our proposed scan is employed to identify genetically linked pairs of loci that demonstrate significant divergence in their gametic frequencies between subpopulations relative to the total population. We highlight that such pairs of loci can be generated by either epistatic selection or through an ongoing selection event at a single locus that is in LD with its neighbors, an important limitation of D-statistics that, to our knowledge, has not been noted before. However, by evaluating the pairwise patterns of D'^2_{IS} over regions, inferences about the type of selection that is taking place may be possible.

This genome-scan was employed using a highly-diverse panel of chickens, consisting of 72 breeds with at least 18 individuals per breed genotyped at nearly 600,000 SNP markers. Multiple previously identified regions known to impact relevant traits were shown to contain significant locus pairs, as well as several novel, putatively selected genomic regions. Regions that suggested patterns consistent with single locus sweeps vs. epistatic selection were explored.

4.3 Background and Theory

Variance components of LD in subdivided populations

When a population of individuals is divided into subpopulations with limited migration, the variance of LD for that population may be divided into subcomponents as well. Ohta (1982a; b) denoted these variance components D^2_{IT}, D^2_{IS}, D^2_{ST}, D'^2_{IS}, and D'^2_{ST}, defined as follows. Consider two loci, A and B. We may use k as a subpopulation index. Therefore, let $x_{i,k}$ and $y_{j,k}$ be the frequency of the i^{th} and j^{th} allele at loci A and B, respectively, in the k^{th} subpopulation. Next,

define $g_{ij,k}$ is the gametic frequency of genotypes A_iB_j in the k^{th} subpopulation. Standard "bar" notation may be used to represent averages, so that $\overline{g_{ıJ}}$, $\overline{x_ı}$, and $\overline{y_J}$ denote the averages over subpopulations. Observe, therefore, that $\overline{g_{ıJ}}$, $\overline{x_ı}$, and $\overline{y_J}$ correspond to the gametic and allele frequencies in the total population (if the total population is balanced). According to Ohta (1982a; b), the variance components of LD may be defined according to:

$$\boldsymbol{D_{IT}^2} = E\left\{\sum_{i,j}\left(g_{ij,k} - \overline{x_ı}\,\overline{y_J}\right)^2\right\}$$

$$\boldsymbol{D_{IS}^2} = E\left\{\sum_{i,j}\left(g_{ij,k} - x_{i,k}y_{j,k}\right)^2\right\}$$

$$\boldsymbol{D_{ST}^2} = E\left\{\sum_{i,j}\left(x_{i,k}y_{j,k} - \overline{x_ı}\,\overline{y_J}\right)^2\right\}$$

$$\boldsymbol{D'^2_{IS}} = E\left\{\sum_{i,j}\left(g_{ij,k} - \overline{g_{ıJ}}\right)^2\right\}$$

$$\boldsymbol{D'^2_{ST}} = E\left\{\sum_{i,j}\left(\overline{g_{ıJ}} - \overline{x_ı}\,\overline{y_J}\right)^2\right\} \quad ,$$

where the expectation is taken with respect to the distribution over alleles and subpopulations. In other words, these variance components are based on treating $x_{i,k}$, $y_{j,k,}$ and $g_{ij,k}$ as random independent and identically distributed random variables each corresponding to a distribution. However, the complicated nature of drift makes specifying that distribution difficult or impossible for all but the simplest scenarios. D_{IT}^2 is a measure of the correlation of A_i and B_j on the same gametes of a subpopulation relative to the expectation according to allele frequencies in the total population, D_{IS}^2 measures the expected variance of LD for subpopulations, D_{ST}^2 is the

expected correlation of A_i and B_j in a subpopulation relative to their expected correlation in the total population, D'^2_{IS} is the correlation of A_i and B_j on the same gamete in a subpopulation relative to that of the total population, and D'^2_{ST} is the variance, computed over alleles only, of the LD of the total population. Observe that although the subscripts are the same (IT, IS, and IT), Ohta's use of subscripts differs substantially from Wright's (Wright, 1949).

If interest lies in identifying pairs of loci that display highly variable LD between subpopulations, D^2_{IT}, D^2_{ST}, and D'^2_{IS} may be appropriate since each of these components measures a population-specific measure compared to a total-population measure. However, it follows from Ohta (1982a; b) that $D^2_{IT} = D'^2_{IS} + D'^2_{ST}$, so the information contained in D^2_{IT} that is relevant to this goal fully resides in D'^2_{IS}. Notice also that D^2_{ST} depends only on expected haplotype frequencies and does not incorporate actual gametic information (there is no $g_{ij,k}$ or $\overline{g_{ij}}$ term), precluding its relevance as a test of non-random association between populations. Moreover, D'^2_{IS} can be manipulated in the following manner:

$$\begin{aligned}
\boldsymbol{D'^2_{IS}} &= E\left\{\sum_{i,j}\left(g_{ij,k} - \overline{g_{ij}}\right)^2\right\} \\
&= E\left\{\sum_{i,j}\left[\left(g_{ij,k} - \bar{x}_i\bar{y}_j\right) - \left(\overline{g_{ij}} - \bar{x}_i\bar{y}_j\right)\right]^2\right\} \\
&= E\left\{\sum_{i,j}\left[d_{ij,k} - D_{ij}\right]^2\right\} \ ,
\end{aligned}$$

where $d_{ij,k}$ measures the covariance of alleles at loci A and B in subpopulations relative to that in the total population and and D_{ij} measures the covariance of alleles at loci A and B in the total population. This demonstrates that by measuring the variability of gametic frequencies in

subpopulations relative to the total population, D'^2_{IS} can be considered a measure of the variability of LD measured at two scales. These observations collectively demonstrate that D'^2_{IS} is the preferred statistic to use for identifying pairs of loci jointly undergoing selection.

When the variance components of LD in subdivided populations were introduced, Ohta (1982a; b) suggested that, for a pair of loci, if $D'^2_{IS} > D'^2_{ST}$ and $D^2_{ST} > D^2_{IS}$, limited migration, and therefore drift, is expected to be more important than epistatic selection in generating LD. Conversely, $D'^2_{IS} < D'^2_{ST}$ and $D^2_{ST} < D^2_{IS}$ was said to imply that the same allelic combinations are favorable across subpopulations and therefore epistatic selection is responsible for observed levels of LD. However, these two conditions are based on the assumption of functional combinations being favored in all subpopulations, leaving out the possibility that such combinations may vary between subpopulations. Later, Black IV and Krafsur (1985) proposed that under dispersive epistatic selection between populations it will hold that $D'^2_{IS} > D'^2_{ST}$ and $D^2_{ST} < D^2_{IS}$. But, neither Ohta's nor Black and Krafsur's conditions involve theoretical or empirical distributions for calculating strict or approximate significance of epistatic selection, nor do they incorporate the possibility that selection on an individual locus, with some amount of linkage across a region, may be responsible for the observed values.

Genome scan and null distribution

To build upon Ohta's (1982a; b) and Black and Krafsur (1985) test, an approach designed for a whole-genome scan was employed. It is capable of identifying either epistatic selection on pairs of markers, or sweeps impacting pairs of loci due to linkage. D'^2_{IS} is used as the basis for this scan. An approximate null distribution, which depicts the expected variability of D'^2_{IS} in a

scenario without selection, was identified. To derive this null distribution, notice that when two loci, A and B, are unlinked the expected value of D'^2_{IS} will not include the effects of joint selection on the pair. This is because, firstly, single-locus selection on either locus will not systematically impact the correlation of alleles at loci A and B. Moreover, still considering two loci that are unlinked, even epistasis such that a particular allelic combination is beneficial will not indefinitely increase LD between the loci unless the selection coefficient is extremely large (which would rapidly lead to fixation), a phenomenon deemed quasi-linkage equilibrium by Kimura (1965), and later studied in more depth by others (Nagylaki 1993). Hence, even when there is epistasis between loci, D'^2_{IS} for unlinked loci will not portray excessively high values as a result of selection. However, when recombination between loci is not random, i.e. A and B are linked to at least a certain extent, dispersive single-locus selection with linkage will elevate levels of D'^2_{IS}, as will dispersive epistatic selection for a favorable allele combination. In the case of single locus selection, this signature is due to hitchhiking (Maynard Smith and Haigh 1974) and will be temporary because, over generations, recombination will break apart the expected correlation of A_i and B_j. In the case of epistatic selection the signature will last for generations because the correlation of favorable allele pairs will grow over time (Kimura 1965).

Therefore, patterns of D'^2_{IS} observed for unlinked loci do not reflect selective sweeps or epistatic selection, although they do retain much of their dependency on demographic factors such as population size, mutation rate, and migration rate. There is only one potential contributor to D'^2_{IS} under a model of drift that may be missed for unlinked locus pairs compared to linked pairs; for a period of time after a variant appears but before recombination breaks down its association with the background in which it arose, the pattern of drift shown by that mutation will be correlated to those of its neighbors. Although this precludes the expected distribution of

D'^2_{IS} for drifting unlinked loci from being identical to that of linked loci, such a distribution is still quite useful to set a boundary that excludes much of the variability expected by chance alone. The result is that an empirical, population-specific null distribution that accounts for random processes including mutation, migration, sampling error, genotyping error, and the majority of drift, but excludes selection may be developed for D'^2_{IS} by computing the statistic over pairs of unlinked loci. Using this null distribution, critical thresholds depicting the most outlying values expected for all but the most extreme case of drift are identified. To identify pairs of loci that have been subjected to selection, D'^2_{IS} values are computed for pairs of linked loci and compared to the identified critical thresholds. Notably, in a practical setting D'^2_{IS} does depend somewhat on the number of subpopulations included in its calculation, because fewer populations allow more sampling error, and this must be accounted for when determining the critical thresholds of drift (see methods).

Distinguishing patterns generated due to a large selective sweep from epistatic selection

Although D'^2_{IS} is capable of identifying dispersive cases of both linked epistatic selection and single-locus selection with linkage, the value of this statistic for an isolated pair of loci cannot alone be used to distinguish the type of selection in place. However, in certain cases the overall pattern of D'^2_{IS} across a region, which is directly dependent upon gametic frequencies, may indicate whether epistasis is likely to be at play. This results from the fact that, over generations, the extent of increased LD for pairs of loci surrounding an individual locus that is undergoing selection will diminish, since it is not being systematically maintained. When linked epistatic selection is taking place, however, increased LD between the two linked and epistatic

loci will be preserved for as long as the advantage of the variant persists (Kimura 1965). Even though the recombination rate between the loci must be smaller than the epistatic selection coefficient for this pattern to appear at all (Ohta 1982a), over generations the probability of recombination, and even double recombination events, between the loci increases. Therefore, when epistatic selection is strong it is plausible that a specific pair of loci may be held in tight LD due to selection, leading to elevated values of D'^2_{IS}, although pairs of loci between them display more neutral values. In other words, epistatic selection is most likely to be taking place if an elevated D'^2_{IS} is observed for one or a few pairs of loci but not for others in the same region. Such a pattern is extremely unlikely when selection is operating on a single locus with alleles linked to their background, since in this case elevated values of D'^2_{IS} should exist across the entire region.

4.4 Data and Methods

Chicken data

Data were taken from the Synbreed Chicken Diversity Panel (Weigend *et al.* 2014), which represents a wide range of populations of individually genotyped and phenotyped chickens. Chickens included in the panel encompassed wild populations and domesticated breeds of various origin and history. Chickens were genotyped using an Affymetrix® Axiom® HD genotyping array (Kranis *et al.* 2013), for which SNPs were mapped to the Gallus_gallus_4.0 reference genome. The SNPs in this array were selected to have an approximately uniform distribution across the genome in terms of SNPs per centiMorgan, leading to a higher physical density (SNPs per kilobasepairs) on the microchromosomes compared to the

macrochromosomes. Markers that were observed in fewer than 95% of individuals were removed from the dataset. Next, individuals with lower than a 95% call rate for SNPs were removed. After markers and individuals were filtered, only breeds with at least 18 individuals represented were included. This left a total of 72 breeds (Appendix C). After these quality-filtering steps were completed, 1,417 individuals genotyped at 538,298 SNPs remained. The average marker spacing across the entire dataset was one SNP approximately every 1,700 bp.

Computing D'^2_{IS}

With most genotyping strategies, gametic phase cannot be assigned to double heterozygotes, even for loci on the same chromosome. However, a strategy for computing Ohta's variance components of LD that utilizes the Burrows composite measure of LD, Δ_{ij} (Cockerham and Weir 1977), has been previously derived (Black IV and Krafsur 1985). The estimation of D'^2_{IS} from data is particularly relevant, so we reproduce the formula here. Letting T_{ij} measure the approximate frequency with which A_i and B_j appear in the same gamete, as done when calculating Burrows' composite measure of LD (Schaid 2004), D'^2_{IS} may be computed as

$$\boldsymbol{D'^2_{IS}} = \frac{\sum_{k=1}^{s}\left(\sum_i \sum_j \left(T_{ij,k} - \overline{T_{ij}}\right)^2\right)}{s} \quad ,$$

where *s* is the number of subpopulations. T_{ij} is computed on a population bases, so although it is an expectation per population it may be considered a random variable over populations. Notice that T_{ij} is an approximation of the already-discussed random variable $g_{ij,k}$. For this study, we employed R for computation (R. Core Team 2013). To make computations of this magnitude

possible, the resources of the University of Wisconsin, Madison Center for High Throughput Computing were leveraged.

To develop the null distribution, a random sample of 120,000,000 pairs of SNPs was chosen, with the requirement that each locus in a pair involved two loci from a different chromosome. For every pair consisting of two SNPs each with minor allele frequency 0.1 or greater, statistics were calculated. This relatively strict exclusion of loci with low minor allele frequency was imposed to mitigate the high dependency of LD on allele frequencies. Each population was included in the computation only if the minor allele frequency of both SNPs in the pair within that population was 0.05 or greater. Pairs that included SNPs on sex chromosomes were excluded from the null distribution. Since D'^2_{IS} is computed as the mean of a distribution, there is a relationship between the variability of the estimated D'^2_{IS} and the number of populations included in its computation. Specifically, including more populations in the computation of D'^2_{IS} is akin to computing any statistic with a larger sampler size, in which case the sampling error of the estimator will decrease. Therefore, separate critical values were identified for each number of included populations between 15 and 60. Critical thresholds were set as the maximum value of D'^2_{IS} observed in the null distribution for the specified number of populations. Situations when fewer than 15 or more than 60 populations were included in the comparison were too rare for critical values to be reliably drawn. After locus-pairs that did not conform to the filtering criteria were removed from the null distribution, 73,950,705 pairs remained from which the critical thresholds were identified. Since 46 separate critical thresholds were defined (corresponding to comparisons including 15-60 populations), this means that, on average, 73,950,705/46 = 1,607,624 values were used to define critical thresholds for each population number, providing an average p-value per threshold of approximately 6.2×10^{-7}.

The experimental calculations for linked pairs of loci were developed similarly, but the pairs of loci used were not random. Instead, statistics were computed for every pair of markers with no more than 199 markers between them. This 200-marker distance limit was imposed due to computational limitations, as ideally all markers with any level of linkage would be compared. This restriction means that on average, pairwise comparisons between markers up 341 kb distal from one another were computed. The same allele frequency filtering used for the null distribution above was imposed for the inclusion of markers and populations when statistics were computed. To identify significant pairs of loci, D'^2_{IS} values for each pair of SNPs were compared to the critical thresholds for the null distribution that corresponded to the same number of populations as included for that pair of SNPs. After locus-pairs that did not conform to the filtering criteria were removed, 73,413,740 pairs remained, computed from 447,538 unique SNPs. This provides an average of 1,595,951 locus pairs per critical threshold, and therefore an expectation of just under one false positive per threshold, and a total of 45.67 over the entire experiment. This is approximately 2.9% of the number of significant locus pairs that were observed, demonstrating that most identified locus pairs cannot be explained by chance alone.

Using pathway information to investigate potential cases of linked epistatic selection

It may be the case that epistatic genes are sometimes found in close proximity to one another on chromosomes due to epistatic selection. Along these lines, Ohta's original studies of the variance components of LD between and within subpopulations were used to investigate the major histocompatibility complex cluster of genes and the possibility that similarities between species are a result of linked epistatic selection (Ohta 1982a; b). In the spirit of this hypothesis,

pathway information was utilized to evaluate whether evidence of linked genes that are potentially epistatic can be found. This question was addressed in two ways. In the first, all available pathway information for the chicken genome was downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa *et al.* 2014). Groups of genes that are mapped to be in close proximity to one another and predicted to belong to the same pathway were identified. The criteria for forming these groups were that, for a gene to be included, its start position had to be within 500 kb of the start position of another gene in the group, and the genes were required to belong to the same pathway. The "Metabolic pathways" pathway, gga01100, was excluded due to it including a very large number of genes and sub-pathways. When a gene was predicted to belong to multiple pathways, it was permitted to belong to separate groups for each. After such gene groups were identified, they were evaluated for whether or not there was any evidence of significance according to D'^2_{IS} values computed in the whole-genome scan. Significant pairs of loci for which at least one of the loci comprising the pair fell within the physical boundaries of the group were identified. This was taken to putatively suggest that selection, possibly epistatic selection on two or more of the grouped genes, has taken place in the vicinity of the group of genes.

Because the whole-genome scan was limited to markers no more than 200 markers distal from one another, it failed to test potentially epistatic markers corresponding to genes further apart than this. Therefore, genes predicted to belong to the same pathway and reside on the same chromosome were identified. Again, gga01100 was excluded. D'^2_{IS} was computed for all pairwise combinations of markers located within these genes, and the computed values were compared to the critical thresholds calculated from the whole-genome scan.

4.5 Results

Majority of D'^2_{IS} values can be explained by chance

The null distribution of D'^2_{IS}, generated by computing the statistic for pairwise combinations of loci that are unlinked, sets a lower boundary for D'^2_{IS} significance by estimating the degree of extremity that is explainable by chance alone. In practice, the threshold set by this null distribution appears to discriminate putatively selected regions well; among the 73,413,740 values that were computed, approximately 0.00002% were deemed as significant deviations from the null distribution. As mentioned previously, D'^2_{IS} values were dependent on the number of populations included in each locus comparison and, therefore, separate significance thresholds were set for each number of populations. A decreasing trend was observed, such that the more populations were included in a comparison, the lower the D'^2_{IS} significance threshold tended to be (Figure 4.2). When considering all locus-pairs regardless of the number of populations used to compute D'^2_{IS}, however, it was observed that the null and experimental distributions of the statistic were quite similar, with the experimental distribution showing a small amount of inflation (Figure 4.2). For the null distribution, the first quantile, median, and third quantile of the distribution are 0.466, 0.530, and 0.592, respectively, while for the experimental distribution the values are 0.495, 0.566, and 0.634, respectively.

Evidence of selection

At the whole-genome level, 1,553 locus-pairs, encompassing 1,718 unique loci, display values of D'^2_{IS} that exceed the null distribution, suggesting that they may have been under

dispersive selection. Most of these pairs overlap with one another or cluster tightly around potentially interesting regions of the genome. Figure 4.3 depicts the genome-wide distribution of locus pairs that were identified as significant. Such pairs exist on 25 of the 28 chicken chromosomes studied (heterosomes and chromosomes beyond 28 were not tested), with only chromosomes 16, 18, and 25 displaying no significant pairs.

Among the pairs of loci identified as significant, a subset fell into regions that have already been shown to influence traits known to have been under selection in chickens. This supports the efficacy of using D'^2_{IS} and the corresponding null distribution as a genome-wide scan for selection. One of the best-studied of these examples involves the *BCDO2* gene, found on chromosome 24. This region has been shown to control yellow vs. white skin in domestic chickens (Eriksson *et al.* 2008; Lobo *et al.* 2012). The chicken panel studied here included breeds fixed for yellow, white, and black skin, so a sensible hypothesis would be that evidence of selection differentiating these breeds based on skin color would be apparent. This seemed to hold: on chromosome 24, an array of significant D'^2_{IS} values was observed in the immediate vicinity of the *BCDO2* gene (Figure 4.4, Table S1). The lack of many significant SNPs within the *BCDO2* gene itself may result from breeds being fixed for one form of the gene or another, since our test was only computed over populations for which both SNPs in a pair were segregating with a minor allele frequency of at least 0.05.

Another interesting region identified via D'^2_{IS} spanned a large portion of chromosome 7. Chickens harbor a segregating inversion on this chromosome from approximately 14.5 Mb – 21.3 Mb (originally reported as 16.5 Mb – 23.88 Mb in the galGal3 assembly) (Imsland *et al.* 2012). Variants within the inverted region code for comb morphology, among other phenotypes. D'^2_{IS} values indicate an abundance of significant locus-pairs across this region (Figure 4.3).

Similarly to the case with *BCDO2*, as described above, this panel includes several breeds with large phenotypic variability for comb morphology, so it is not surprising that a high divergence between these breeds, measured using D'^2_{IS}, was observed. Additionally, since inversions have large impacts on recombination in a region, therefore affecting LD which the D'^2_{IS} statistic is based on, this suggests that D'^2_{IS} may be useful for detecting selection on structural variation.

Patterns of epistasis and large selective sweeps

Often, regions of the genome harboring pairs of loci with significant values of D'^2_{IS} presented generally elevated values of the statistic across the entire region. This is illustrated in Figure 4.5A, where significant values of D'^2_{IS} are plotted for a region on chromosome 5 that contained several significant pairs. Figure 4.5B shows the complete breakdown of D'^2_{IS} for all locus pairs in this region. Patterns such as this one are consistent with the expectation under a dispersive, large selective sweep between populations, since such an instance is expected to alter gametic frequencies and generate LD across an entire region. However, it cannot be ruled out that epistatic loci on either end of this region and others like it are resulting in the maintenance of LD between locus pairs.

In other instances, the pattern of significant D'^2_{IS} values does not correspond as clearly to the expectation under a large sweep, and it is plausible that epistatic selection has taken place. An example of this is shown in Figures 4.5C and 4.5D, which depict a region on chromosome 6 where significant values of D'^2_{IS} were identified. Importantly, for this region there does not appear to be an overall elevation of D'^2_{IS} values. However, the evidence that this is a case of

epistatic selection is not overwhelming: it may simply reflect a weak selective sweep and therefore less of an abundance of significant signals.

Linked epistatic selection suggested by pathway information

To further evaluate the possibility that some of the locus pairs identified as putatively under selection correspond to cases of linked epistatic selection, pathway information was investigated. The hypothesis is that genes within the same pathway are more likely to be epistatic than are genes in different pathways. Since pathway information is still incomplete, and because there is no guarantee that neighboring genes in the same pathway are epistatic, this analysis should be considered an approximation. We identified 803 groups of genes in chickens that are no more than 500 kb distal from other genes belonging to the same pathway. Of these groups, 67 encompass at least one locus that is part of a pair with a significant value of D'^{2}_{IS}. Appendix D contains information describing each of these 67 gene groups. Since the groups identified pertain to loci with significant D'^{2}_{IS} values, this is evidence that they have been subjected to some sort of dispersive selection. Moreover, because the highlighted groups contain genes previously predicted to correspond to the same pathway, these represent candidates for gene clusters that are epistatic.

Similarly, we computed D'^{2}_{IS} values to test for epistatic selection between pairs of genic SNPs for sets of genes that are both on the same chromosome and predicted to be members of the same pathway. This analysis was not limited by the 200 marker distance limit imposed for the genome scan, yet it was limited to include only markers located within the start and end positions of known gene-models. Three locus-pairs that exceeded the thresholds derived from the

null distribution of D'^2_{IS} were identified (Table 4.1). Interestingly, SNPs for one of the three locus-pairs were approximately 13 Mb apart, with one SNP on each side of the previously mentioned inversion on chromosome 7. To further investigate the possibility of epistatic selection between the two genes implicated from this pair, *ADCY5* and *PDE1A*, pairs of SNPs within 1 MB of the start and stop position of both genes were tested. This analysis incorporated 2,702 SNPs composing 1,393,275 SNP-pairs that passed filtering. Among these tests, the only pair that depicted a significant signal is that portrayed in Table 4.1, which further suggests that these two genes may result from epistatic selection.

4.6 Discussion:

Haplotype-based scan

Methods commonly used to identify selection from dense marker or sequencing data at the whole-genome level can be broadly divided into two classes. First, there are approaches that seek to identify selection based on variability between populations, usually based on some form of the Lewontin and Krakauer test (1973). Secondly are methods tailored for studying variability within a single population, which tend to utilize patterns developed due to genetic hitchhiking (Maynard Smith and Haigh 1974). Another subset of methods is based on a combination of ideas from each of these classes. These include the XP-CLR method, which is based on differences in multi-locus allele frequencies between populations (Chen *et al.* 2010), the XP-EHH method, involving a search for long-range haplotypes that have been generated from differential hard sweeps between populations (Sabeti *et al.* 2007), and more recently the hapFLK approach, using haplotype information and seeking to explicitly incorporate population stratification (Fariello *et*

al. 2013). The use of D'^2_{IS} to detect selection is similar to this latter group of methods given that evaluating gametic frequencies of locus-pairs is a form of haplotype analysis, with haplotypes defined based on two loci. However, the roots of D'^2_{IS} are grounded in its analogy with F_{ST} (Ohta 1982a; b). Therefore, it is comparable to an F_{ST}-based scan as well. The scan may be used to identify pairs of loci depicting correlated patterns of selection while retaining much of the straight-forward interpretability of an F_{ST} approach.

The use of D'^2_{IS} allows constructing a null distribution to be formed from the data. This null distribution captures most of the variability of D'^2_{IS} explainable by chance alone. It has proven useful for defining a lower boundary for significance. This is a substantial improvement over most single-locus tests for selection, especially those that employ F_{ST}, since these tests usually rely on arbitrary outlier thresholds. e.g., the upper 99% quantile of the data-distribution (Akey 2009), guaranteeing false positives if selection has not taken place. The null distribution described here does not eliminate the possibility of drift-generated values that appear to be significant, so it should be considered a hybrid approach between the commonly employed outlier thresholds and "true" statistical significance. This is because our null eliminates most values that are explainable with no selection taking place, yet values above the threshold can potentially be explained without invoking selection as well. Therefore, such locus-pairs should be treated as outliers for which selection is likely, but not necessarily, the explanation. An important advantage of this approach, however, is that unlike other outlier approaches, there is no guarantee of false positives if selection did not take place—if allele frequencies are highly variable due to drift alone it is possible that few or no experimental locus pairs exceed the null.

Epistasis

Although Ohta's introduction of the variance components of LD was for the study of epistatic selection (Ohta 1982a; b), we have demonstrated that D'^2_{IS} is not exclusively limited to that phenomenon. In particular, it is capable of identifying loci that are in LD due to epistasis, or it may identify pairs of loci that are members of a haplotype undergoing a hard sweep. Our results suggest that distinguishing which phenomenon is taking place is not straight-forward. Specifically, when it is known or hypothesized *a priori* that epistasis for a selected trait may exist for a set of loci, using D'^2_{IS} to test this hypothesis seems appropriate and, in such a situation, the null distribution developed here can be used to better establish significance than previous approaches. Using this idea, we isolated and tested pairwise combinations of SNPs within sets of genes on the same chromosome and in the same pathway. We identified three pairs of SNPs suggesting epistatic selection may have taken place, although we cannot rule out the possibility that these genes simply fall within a selected haplotype.

Conversely, in extreme cases where epistatic interactions between linked loci are old enough for LD between nonfunctional locus pairs in the same region to decay, a unique pattern of D'^2_{IS} is expected to form which may indicate linked epistatic selection. However, excluding this pattern (which was rarely seen in this data set), it is probable that D'^2_{IS} generally identifies major sweeps that vary between populations and arise from selection on a single, non-epistatic, locus. As such, an evaluation of the pairwise patterns of significant loci that were observed in this experiment (e.g. Figure 4.5A and B), D'^2_{IS} appears to define the boundaries of selected haplotypes well. In particular, overlap between pairs of loci that are deemed significant may suggest the extent of any sweep that has occurred (a lack of overlap may be indicative of separate sweeps).

Future directions

While D'^2_{IS} can be employed to successfully identify locus pairs that have undergone selection, there are components of this work that would benefit from further exploration. One aspect that could be improved involves our use of the Burrows approximation (Cockerham and Weir 1977) to estimate gametic frequencies. Without family data, utilizing known gametic frequencies was not possible, but if this information were available it would likely improve the accuracy of the test. Additionally, computation is intensive, which poses a challenge. Ideally, D'^2_{IS} values would be computed for all pairs of loci that are potentially linked. Since this was not feasible without increased computational power, our solution was to compute values for sets of loci that are most likely to be linked. This was done by allowing the distance between them to span up to 200 markers, or approximately 341 kb on average. With advancements in computational methods and in high throughput computing, extending the distance between pairs of markers will be possible in the near future. Additionally, an analysis at an even larger scale may become more feasible if coded in a lower-level language than R, as was used here. However, as the number of locus-pairs considered is expanded, so too must be the number of pairs used to develop the null distribution, which adds to the computational challenges.

Finally, we have mentioned the difficulty of distinguishing whether a significant D'^2_{IS} value for a pair of loci is the result of a single-locus sweep or epistatic selection between pairs of loci. We also discussed a unique pattern of D'^2_{IS} that may establish, more conclusively, if epistatic selection is at play. A more formal investigation of the differences between patterns generated by each type of selection may prove fruitful.

4.7 Acknowledgements:

We are indebted to all the breeders who supported sampling from their animals. This research was funded by the German Federal Ministry of Education and Research (BMBF) within the AgroClustEr ‘‘Synbreed - Synergistic plant and animal breeding’’ (FKZ 0315528). Additionally, we used the computational resources and assistance of the University of Wisconsin—Madison Center For High Throughput Computing (CHTC) in the Department of Computer Sciences. The CHTC is supported by UW-Madison and the Wisconsin Alumni Research Foundation, and is an active member of the Open Science Grid, which is supported by the National Science Foundation and the U.S. Department of Energy's Office of Science.

4.8 Figures

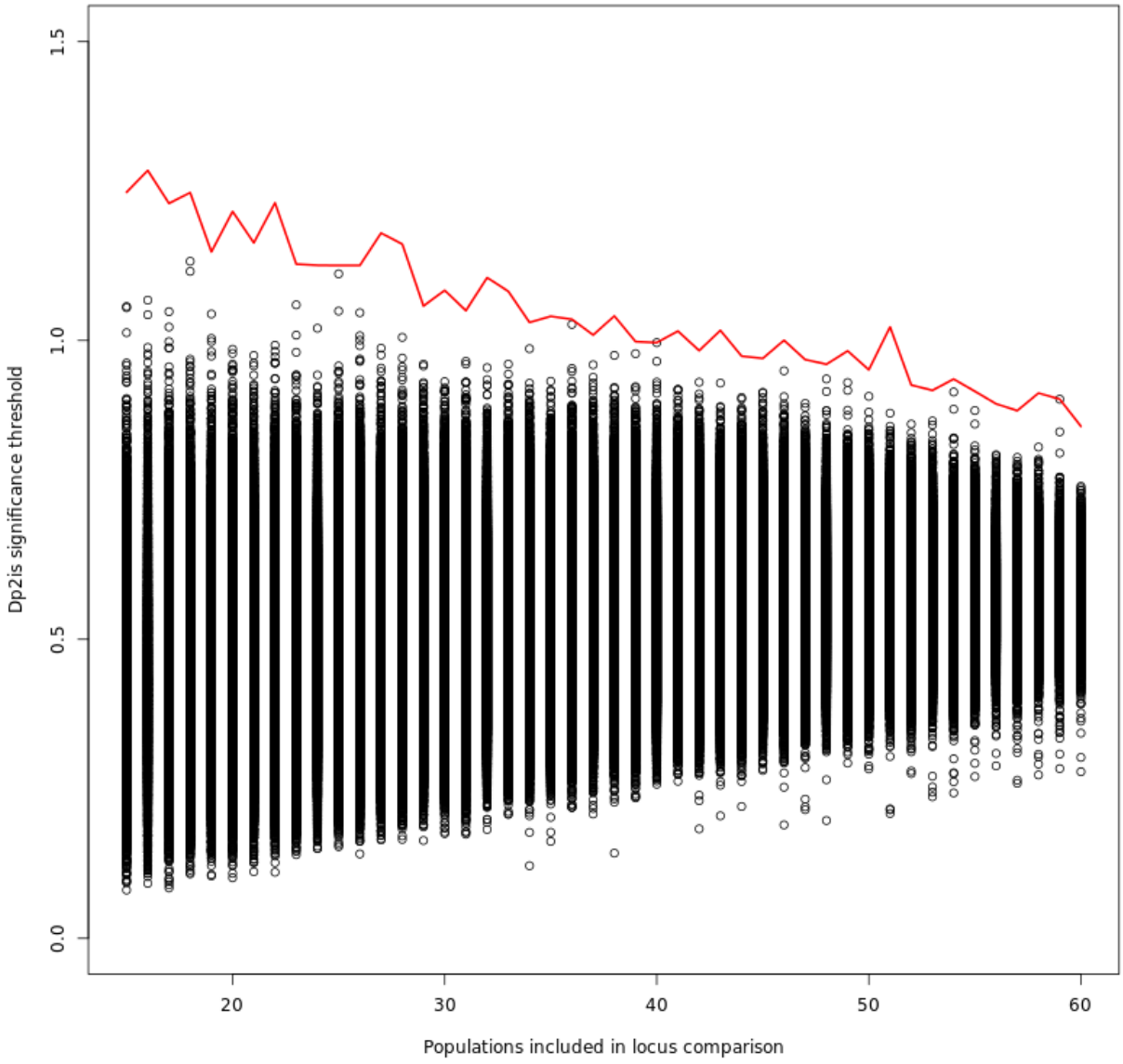


Figure 4.1: Significance thresholds.

Significance threshold for each number of populations included in a locus comparison is shown in red. Black points depict 2 million D'^2_{IS} values randomly chosen from the null distribution of more than 70 million. The null distribution was used to compute a separate significance threshold for each potential number of populations included in a locus-comparison between 15 and 60.

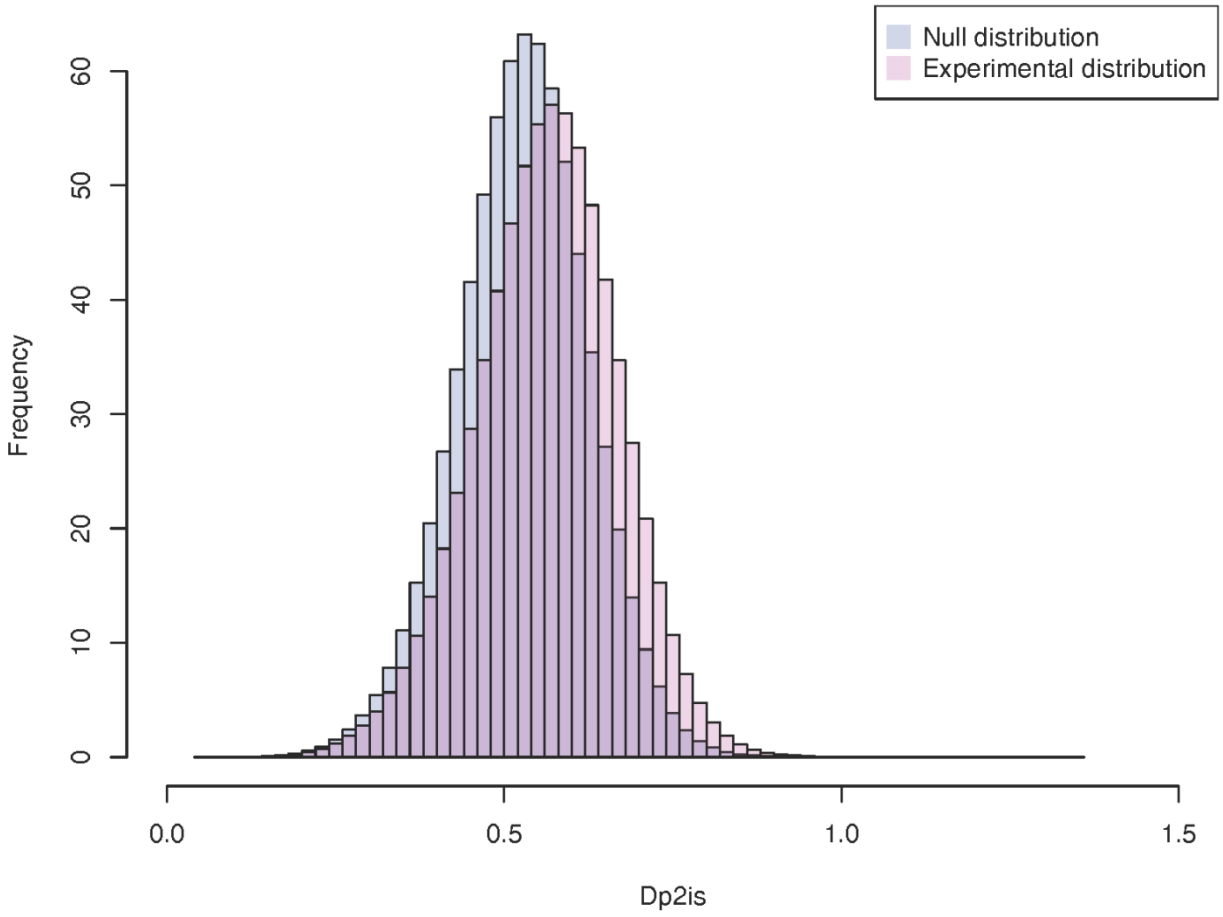


Figure 4.2: Null vs. experimental distributions.

Histograms comparing the null and experimental distribution for D'^2_{IS}. Observe that the distributions are quite similar, but the experimental distribution is slightly inflated.

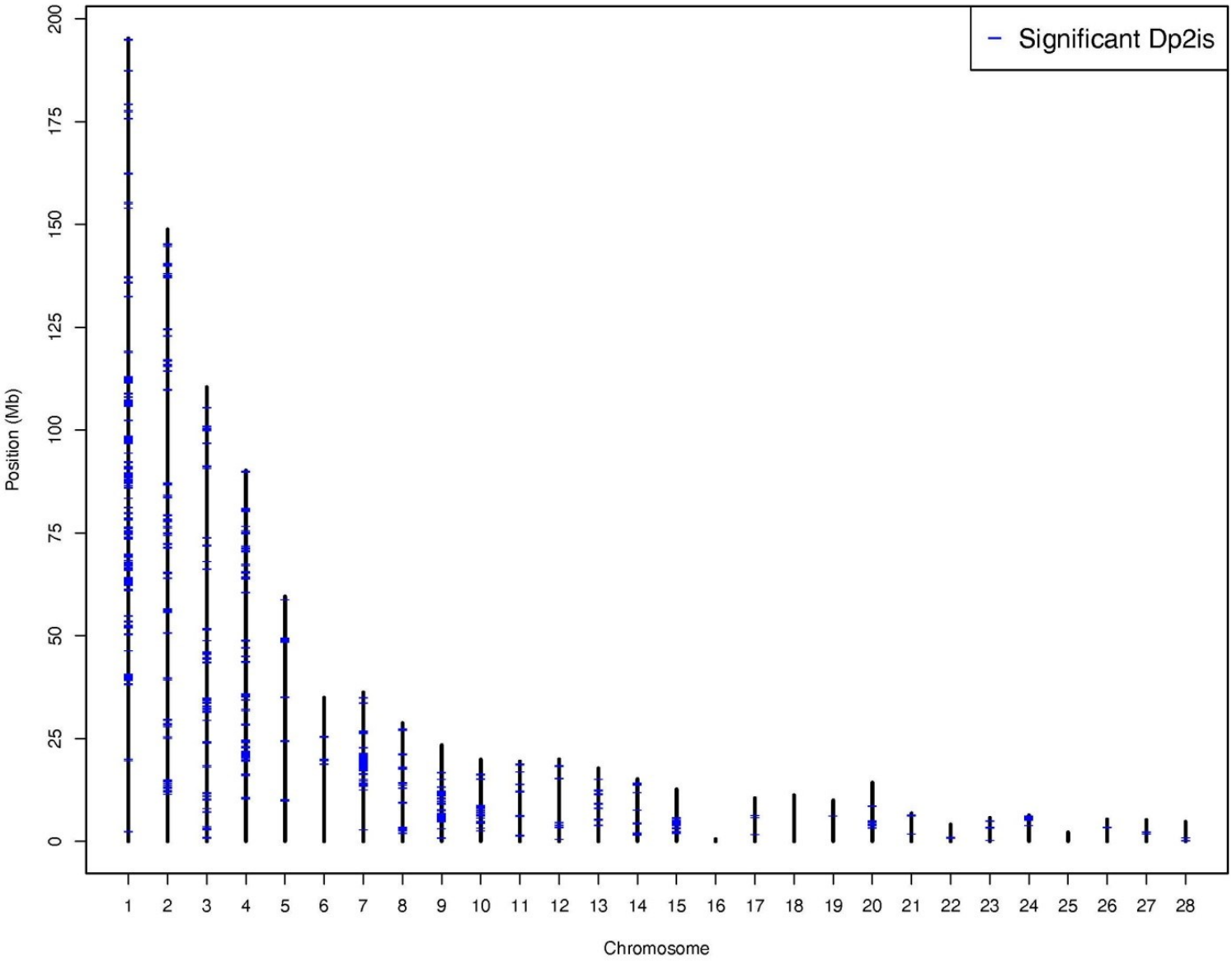


Figure 4.3: Location of significant locus-pairs.

A whole-genome map of loci where significant values of D'^{2}_{IS} were observed in the chicken genome. The position of every marker that was part of a significant pair is represented. Positions are plotted according to the galGal4 assembly. These loci are those putatively under dispersive selective sweeps or dispersive cis-acting epistatic selection.

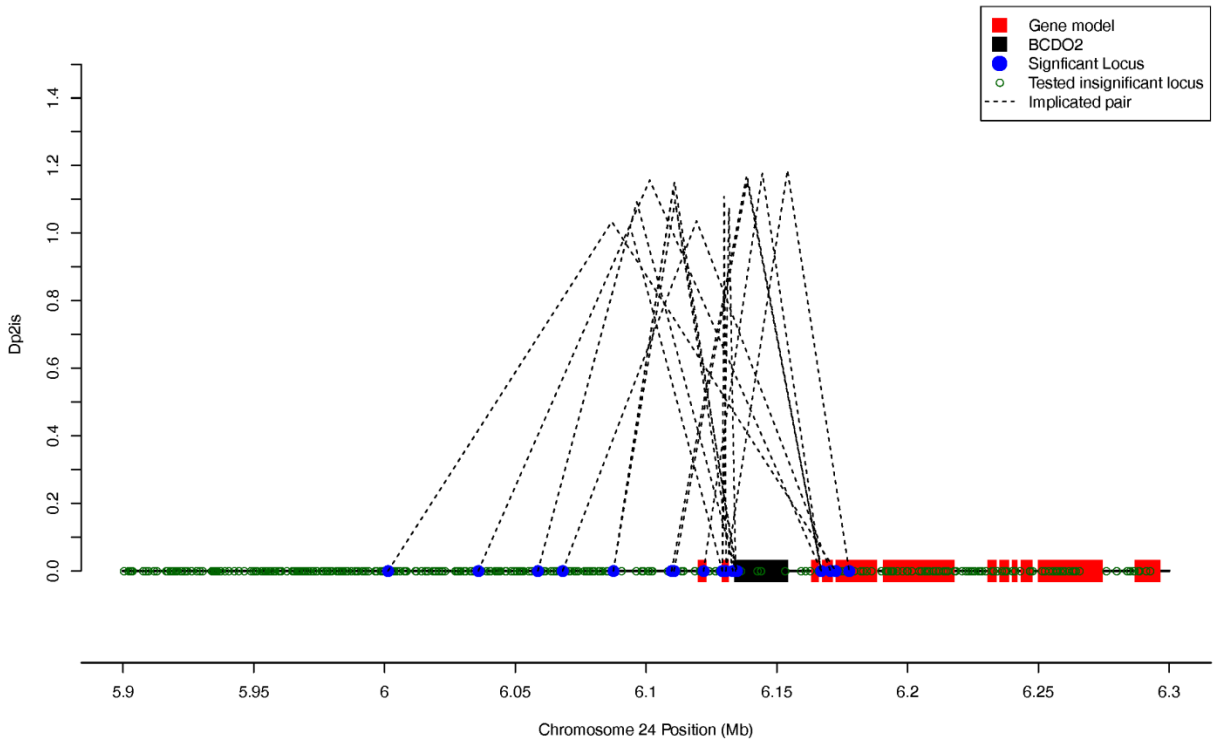


Figure 4.4: The *BCDO2* locus.

Depiction of significant locus pairs identified in the *BCDO2* region, and the D'^2_{IS} value observed for each significant pair. The height of each dashed line depicts the value of D'^2_{IS} observed for that pair.

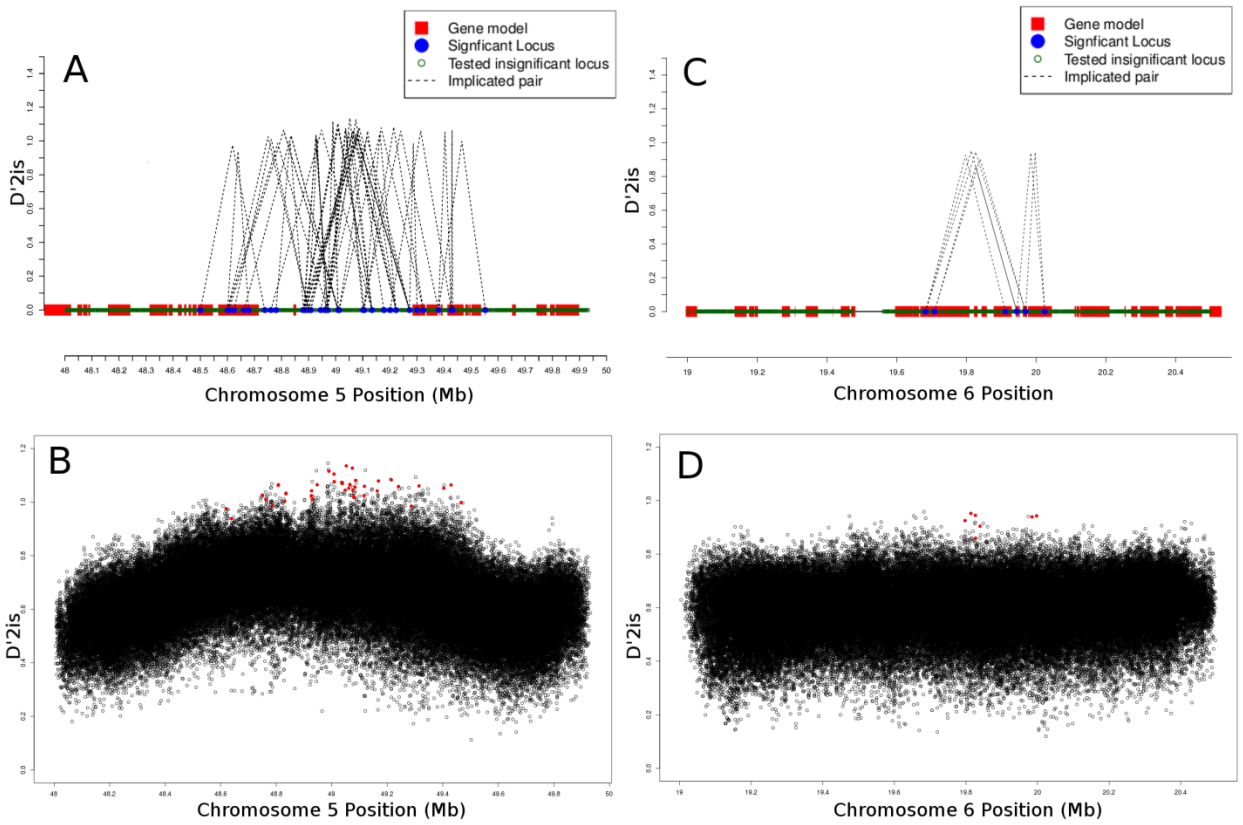


Figure 4.5: Sweeps vs. potential epistatic selection.

D'^2_{IS} values for two example regions. Panels A and B demonstrate region on chromosome 5 showing patterns consistent with a large selective sweep. Panels C and D are of a region on chromosome 6, where significance was not clear except for specific locus pairs, suggestive of epistasis. A) Significant locus pairs identified across a region of chromosome 5; B) D'^2_{IS} for every locus pair across the same region of chromosome 5. Values are plotted at the midpoint of the locus pair. Significant pairs are highlighted in red; C) Significant locus pairs identified across a region of chromosome 6; B) D'^2_{IS} for every locus pair across the same region of chromosome 6.

4.9 Tables

Table 4.1: Three locus pairs displaying putative evidence of epistatic selection.
SNP pairs shown here correspond to pairs of distinct genes predicted to be in the same pathway (KANEHISA *et al.* 2014). The corresponding D'^2_{IS} values exceed the critical thresholds computed from the null distribution.

Chr.	Marker_1	Marker_2	Marker_1 Position	Marker_2 Position	Number Pops	D'^2_{IS}	Pathway	Gene_1	Gene_2
4	**AX-76622925**	**AX-76623883**	**20,836,864**	**213,39,719**	**56**	**0.91451**	**gga04080**	**RXFP1**	**GLRB**
7	**AX-80925777**	**AX-77016330**	**13,465,414**	**26,503,508**	**39**	**1.00719**	**gga00230**	**ADCY5**	**PDE1A**
9	**AX-77173228**	**AX-77173770**	**4,913,635**	**5,113,445**	**17**	**1.268478**	**gga04150**	**PIK3CB**	**CAB39**

4.9 References

Akey J. M., 2009 Constructing genomic maps of positive selection in humans: Where do we go from here? Genome Res. **19**: 711–722.

Akey J. M., Zhang G., Zhang K., Jin L., Shriver M. D., 2002 Interrogating a High-Density SNP Map for Signatures of Natural Selection. Genome Res. **12**: 1805–1814.

Beissinger T. M., Hirsch C. N., Vaillancourt B., Deshpande S., Barry K., Buell C. R., Kaeppler S. M., Gianola D., Leon N. de, 2013 A Genome-Wide Scan for Evidence of Selection in a Maize Population Under Long-Term Artificial Selection for Ear Number. Genetics: genetics.113.160655.

Black IV W. C., Krafsur E. S., 1985 A FORTRAN program for the calculation and analysis of two-locus linkage disequilibrium coefficients. Theor. Appl. Genet. **70**: 491–496.

Chen H., Patterson N., Reich D., 2010 Population differentiation as a test for selective sweeps. Genome Res. **20**: 393–402.

Cockerham C. C., Weir B. S., 1977 Digenic descent measures for finite populations. Genet. Res. **30**: 121–147.

Eriksson J., Larson G., Gunnarsson U., Bed'hom B., Tixier-Boichard M., Stromstedt L., Wright D., Jungerius A., Vereijken A., Randi E., Jensen P., Andersson L., 2008 Identification of the Yellow Skin Gene Reveals a Hybrid Origin of the Domestic Chicken. PLoS Genet. **4**.

Fariello M. I., Boitard S., Naya H., SanCristobal M., Servin B., 2013 Detecting Signatures of Selection Through Haplotype Differentiation Among Hierarchically Structured Populations. Genetics **193**: 929–941.

Garnier-Gere P., Dillmann C., 1992 A Computer Program for Testing Pairwise Linkage Disequilibria in Subdivided Populations. J. Hered. **83**: 239–239.

Imsland F., Feng C., Boije H., Bed'hom B., Fillon V., Dorshorst B., Rubin C.-J., Liu R., Gao Y., Gu X., Wang Y., Gourichon D., Zody M. C., Zecchin W., Vieaud A., Tixier-Boichard M., Hu X., Hallböök F., Li N., Andersson L., 2012 The Rose-comb Mutation in Chickens Constitutes a Structural Rearrangement Causing Both Altered Comb Morphology and Defective Sperm Motility. PLoS Genet **8**: e1002775.

Kanehisa M., Goto S., Sato Y., Kawashima M., Furumichi M., Tanabe M., 2014 Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Res. **42**: D199–205.

Kelly J. K., 1997 A test of neutrality based on interlocus associations. Genetics **146**: 1197–1206.

Kimura M., 1965 Attainment of Quasi Linkage Equilibrium When Gene Frequencies Are Changing by Natural Selection. Genetics **52**: 875–890.

Kranis A., Gheyas A. A., Boschiero C., Turner F., Yu L., Smith S., Talbot R., Pirani A., Brew F., Kaiser P., Hocking P. M., Fife M., Salmon N., Fulton J., Strom T. M., Haberer G., Weigend S., Preisinger R., Gholami M., Qanbari S., Simianer H., Watson K. A., Woolliams J. A., Burt D. W., 2013 Development of a high density 600K SNP genotyping array for chicken. BMC Genomics **14**: 59.

Lewontin R. C., Krakauer J., 1973 Distribution of Gene Frequency as a Test of the Theory of the Selective Neutrality of Polymorphisms. Genetics **74**: 175–195.

Lobo G. P., Isken A., Hoff S., Babino D., Lintig J. von, 2012 BCDO2 acts as a carotenoid scavenger and gatekeeper for the mitochondrial apoptotic pathway. Dev. Camb. Engl. **139**: 2966–2977.

Ma X.-F., Hall D., Onge K. R. S., Jansson S., Ingvarsson P. K., 2010 Genetic Differentiation, Clinal Variation and Phenotypic Associations With Growth Cessation Across the Populus tremula Photoperiodic Pathway. Genetics **186**: 1033–1044.

Maynard Smith J., Haigh J., 1974 The hitch-hiking effect of a favourable gene. Genet. Res. **23**: 23–35.

Miyashita N. T., Aguadé M., Langley C. H., 1993 Linkage disequilibrium in the white locus region of Drosophila melanogaster. Genet. Res. **62**: 101–109.

Nagylaki T., 1993 The evolution of multilocus systems under weak selection. Genetics **134**: 627–647.

Ohta T., 1982a Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. **79**: 1940–1944.

Ohta T., 1982b Linkage Disequilibrium with the Island Model. Genetics **101**: 139–155.

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

R. Core Team, 2013 *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sabeti P. C., Reich D. E., Higgins J. M., Levine H. Z. P., Richter D. J., Schaffner S. F., Gabriel S. B., Platko J. V., Patterson N. J., McDonald G. J., Ackerman H. C., Campbell S. J., Altshuler D., Cooper R., Kwiatkowski D., Ward R., Lander E. S., 2002 Detecting recent positive selection in the human genome from haplotype structure. Nature **419**: 832–837.

Sabeti P. C., Varilly P., Fry B., Lohmueller J., Hostetter E., Cotsapas C., et al., 2007 Genome-wide detection and characterization of positive selection in human populations. Nature **449**: 913–918.

Schaid D. J., 2004 Linkage disequilibrium testing when linkage phase is unknown. Genetics **166**: 505–512.

Song B.-H., Windsor A. J., Schmid K. J., Ramos-Onsins S., Schranz M. E., Heidel A. J., Mitchell-Olds T., 2009 Multilocus patterns of nucleotide diversity, population structure and linkage disequilibrium in Boechera stricta, a wild relative of Arabidopsis. Genetics **181**: 1021–1033.

Storz J. F., Kelly J. K., 2008 Effects of Spatially Varying Selection on Nucleotide Diversity and Linkage Disequilibrium: Insights From Deer Mouse Globin Genes. Genetics **180**: 367–379.

Tang K., Thornton K. R., Stoneking M., 2007 A New Approach for Using Genome Scans to Detect Recent Positive Selection in the Human Genome. PLoS Biol **5**: e171.

Turner T. L., Stewart A. D., Fields A. T., Rice W. R., Tarone A. M., 2011 Population-Based Resequencing of Experimentally Evolved Populations Reveals the Genetic Basis of Body Size Variation in Drosophila melanogaster. PLoS Genet **7**: e1001336.

Vitalis R., Riba M., Colas B., Grillas P., Olivieri I., 2002 Multilocus genetic structure at contrasted spatial scales of the endangered water fern Marsilea strigosa Willd. (Marsileaceae, Pteridophyta). Am. J. Bot. **89**: 1142–1155.

Voight B. F., Kudaravalli S., Wen X., Pritchard J. K., 2006 A Map of Recent Positive Selection in the Human Genome. PLoS Biol **4**: e72.

Weigend S., Janssen-Tapken U., Erbe M., Ober U., Weigend A., Preisinger R., H S., 2014 Biodiversität beim Huhn – Potenziale für die Praxis. Züchtungskunde **86**: 25–41.

Wisser R. J., Murray S. C., Kolkman J. M., Ceballos H., Nelson R. J., 2008 Selection Mapping of Loci for Quantitative Disease Resistance in a Diverse Maize Population. Genetics **180**: 583–599.

Wright S., 1949 The Genetical Structure of Populations. Ann. Eugen. **15**: 323–354.

Chapter 5 General Discussion

Author: Timothy M. Beissinger

Discussion of methods for assessing the effect of selection on genetic variability

It has been recognized for well over a century that selection, both natural and artificial, systematically modifies allele frequencies at both advantageous and detrimental loci (Darwin, 1859). In its simplest form, Darwin's theory of natural selection, applicable to natural populations, crop species, domestic livestock, and experimental populations, states that individuals who are better able to survive in their environment will have a greater opportunity of passing on their heritable characteristics to the following generation. Over time, this will lead to the expectation of an increase in the frequency of favorable alleles at the expense of those that are deleterious or neutral. Darwin's work did not involve genes or DNA, but rather was based on the characteristics of individuals and their observable phenotypic traits. It was only a short time, however, before scientists such as Mendel (1865) began to perform experiments allowing the study of inheritance at something that represented the gene level. Still, the structure of DNA was not revealed until almost a century after Mendel's work (Watson and Crick, 1953). In recent decades technological advancements in molecular methods allowed scientists to explore variation within and among individuals at the level of DNA. Early genotyping work involved techniques such as microsatellites (Queller *et al.* 1993), amplified fragment length polymorphisms (AFLPs; vos *et al.* 1995), and restriction fragment length polymporhisms (RFLPs; Botstein *et al.* 1980), among others. These technologies allowed geneticists and breeders to make great strides toward understanding and utilizing genetic variation, but their high cost and limited density across genomes led to the development of single nucleotide polymorphism (SNP) markers as a more ideal system for assessing genomic variation. SNPs are abundant in nearly all genomes and can be scored relatively inexpensively using a variety of techniques. These techniques include array-based technologies (Fan *et al.* 2006) and sequencing-

based approaches such as genotyping-by-sequencing (GBS; Elshire *et al.* 2011). An analysis of, and suggestions for, the implementation of a GBS protocol was provided in Chapter One of this dissertation. With the abundance of genotyping tools now available, it is not surprising that scientists are devoting enormous efforts, with great success, towards the discovery of variation that has been subjected to selection over generations.

The approach taken to identify such selected variants is to define a genetic pattern expected to result from selection, and then screen or scan the set of available markers to isolate those displaying that pattern. Several such patterns are known to exist. For example, it follows from Darwin's theory that the frequency of a favorable variant, and consequently the frequency of a marker in linkage disequilibrium (LD) with that variant, will increase over generations as a result of selection. Therefore, identifying variants that are more common after several generations of selection than they were before selection began (e.g. Wisser *et al.* 2008; Kelly *et al.* 2013) is one mechanism for identifying selection.

The aforementioned example is just one out of a multitude of signals that may result from selection. Overall, these signals and the genomic scans for selection that seek to identify them can be broken down into three main categories: 1) Assessment of allele frequencies; 2) Identifying LD relationships across segments of DNA; and 3) Measuring the site-frequency-spectrum (SFS), or the distribution of alleles at particular frequencies, across potentially functional sites such as genes or genomic windows. Variations and combinations of these categories are often employed as well. Ultimately, the outcome from a scan for selection consists of a set of genomic regions that putatively are involved in the control of the trait or set of traits that were selected, along with insight into the genetic and biological mechanisms controlling

those traits. Because these three categories collectively represent all indicators known to arise from selection, each warrants further explanation.

Assessment of allele frequencies: Perhaps the most intuitive technique to scan for signatures of selection involves comparing allele frequencies in a selected population to those of the population before selection, and identifying loci at which frequencies have deviated more than what is expected from random process such as genetic drift. Wisser *et al*. (2008) called this method selection mapping. Markers or regions with allele frequency changes that exceed the expected amount of genetic drift are deemed to have potentially been under selection, likely for the trait(s) that the population was selected for.

This method is based on basic population and quantitative genetics. Specifically, it depends on the frequencies of favorable alleles at a locus under selection increasing at the expense of deleterious or neutral alleles, as well as on the frequencies of all alleles at unlinked, and neutral loci remaining approximately stable. An excellent depiction of how selection increases the frequency of favorable alleles in a population can be found in chapter five of Crow and Kimura (1970). The relative stability of allele frequencies at neutral loci is often given less attention, even though it too can display a complex and difficult to quantify pattern of variability even in populations with relatively simple mating schemes. If, for a single locus with initial frequency p, an equal number, N/2, of male and females contribute gametes at each generation, the expected variance of the allele frequency in the t^{th} generation, without selection, is given by:

$$V^{(t)} = p(1-p)\left[1-\left(1-\frac{1}{2N}\right)^{t}\right], \qquad (5.1)$$

and the expected mean allele frequency does not change (Crow and Kimura, 1970). However, this formula is applicable only to a single diallelic locus for which equal numbers of males and females are contributing gametes to subsequent generations. For more realistic scenarios, which are inherently more complex, computer simulations of drift are a commonly employed approach (e.g. Wisser *et al.* 2008). Furthermore, when the loci being considered have more than two alleles, any quantification of drift must include the covariance between alleles. Waples (1989) characterized some frequently improper measures used to assess variability in allele frequency over multiple generations and developed a test of a drift-only null hypothesis that is applicable to multiple sampling schemes and multiple alleles. The test proposed is an extension of the standard chi-square test, and is similar but more general than tests previously found in the literature (Fisher and Ford, 1947; Meuller *et al.* 1985; Schaffer *et al.* 1977; Gibson *et al.* 1979; Wilson, 1980; Watterson, 1982). However, the Waples test still requires a relatively simple population mating scheme to be applicable without bias, so a simulation approach may still be preferred.

It was only shortly after genotyping with molecular markers became possible in the 1960's (e.g. Hubby and Lewontin, 1966), that studies attempting to quantify selection based on allele frequency change in a selected population began to appear. One early study in maize looked at three isozyme loci and found a significant correlation between increase in yield and allele frequency change at one of the loci (Stuber and Moll, 1972). However, this study did not consider genetic drift as a possible explanation for the correlation. Another study identified highly significant isozyme allele frequency differences in a natural population of *Dacus oleae* based on a chi-square test. When the authors estimated N_e and employed equation 5.1, they concluded that genetic drift could not be ruled out as the sole force contributing to these differences (Krimbas and Tsakas, 1971).

Over time, the relevance of this type of research has not diminished, especially because of the constantly improving genotyping technologies. Generally, selection mapping studies have been performed with model species due to the requirement of a genotyped base population before selection and of a quantifiable selection scheme; model species tend to have short generation intervals, allowing experiments to be performed for many generations in months or a few years. For instance, a recent experiment identified 21 QTL for heat stress tolerance in yeast by sequencing populations of tens of millions of individuals and subjecting them to selection for twelve generations (Parts *et al.* 2011). The authors reported that nine of the 21 QTL intervals identified contained two or fewer genes, providing substantially higher resolution than previous linkage-based studies (i.e. QTL mapping type approaches). Because of the short generational intervals in yeast, the selection program took only twelve days. Similarly, Turner *et al*. (2011) conducted an experiment based on selection for body size in replicated populations of *Drosophila melanogaster*. Their selection persisted for over 100 generations, and ultimately they found that between 304 and 1,236 regions may have been under selection. Their results suggest that body size in *Drosophila melanogaster* meets the classical definition of a polygenic trait, with many loci of small effects determining a continuously variable trait.

Unlike *Drosophila* and yeast, which have generation intervals of approximately one and 14 days, respectively (Parts *et al.* 2011; Frankham and Loebel, 1992), crop species have generation intervals on the order of months or years. For example, maize landraces reach maturity after two to eleven months (Kuleshov, 1933). As such, over the course of time required to complete a multi-generation selection experiment with an agronomic species, genotyping technologies change enough to make genotypes obtained from the base population nearly irrelevant. In the case of grain-producing plants, however, this problem can be alleviated if

remnant seed of the base population is still available for genotyping at the end of the experiment; the base population can then be genotyped with methods identical to those used on the selected population. Wisser *et al*. (2008) performed a selection mapping experiment in maize to identify QTL for northern leaf blight resistance. Although selection began in 1991, the authors genotyped both the base and selected populations using 151 simple sequence repeat (SSR) loci more than a decade later, including an initial set of 87 loci and a later step where additional loci were added. The authors concluded that 25 SSR loci showed evidence of selection.

When samples from the base population, before selection, are not available, but instead samples representing multiple distinctly selected populations exist, the pattern of allele frequency variability between and within these populations can be used to identify putatively selected sites. As an overview, loci that display abnormally divergent (or abnormally similar) frequencies between populations may indicate that selection at these loci has taken place. Such approaches for finding selection have a rich history in population genetics, beginning with work by Sewall Wright in the first half of the twentieth century (Wright, 1951). These studies are typically based on a statistic known as F_{ST}, which merits a somewhat in-depth discussion here.

Beginning in the 1920's, Sewall Wright began working on dividing the genetic variability of species into within and between population components. Wright termed the statistics used for this division '*F'* statistics, and introduced F_{IS}, F_{IT}, and F_{ST}, which distinguish variation between individuals relative to their sub-population, variation between individuals relative to the total population, and variation between sub-populations relative to the total population, respectively (Wright, 1951). These statistics have been widely applied in statistical, quantitative, and population genetics (Holsinger and Weir, 2009). Various equivalent definitions of Wright's *F*-statistics abound, and an intuitive set is given below (Holsinger and Weir, 2009):

$$F_{IS} = \frac{\sigma_P^2 + \sigma_l^2}{\sigma_P^2 + \sigma_l^2 + \sigma_G^2} \quad (5.2)$$

$$F_{ST} = \frac{\sigma_P^2}{\sigma_P^2 + \sigma_l^2 + \sigma_G^2} \quad (5.3)$$

$$F_{IT} = \frac{\sigma_l^2}{\sigma_l^2 + \sigma_G^2} \quad (5.4)$$

In these definitions, σ_G^2 denotes the variance among genotypes, σ_P^2 denotes the variance among populations, and σ_l^2 denotes the variance among individuals within populations. Additionally, an excellent description of estimators of the various *F* statistics is provided by Weir and Cockerham (1984). A simple estimator for F_{ST} that is applicable for a large sample size is $\hat{F}_{ST} = \frac{s^2}{\bar{p}(1-\bar{p})+s^2/r}$, where s^2 is the sample variance of allele frequency between populations, $\bar{p}$ is the mean allele frequency across populations, and *r* is the number of populations. When *r* is large, the s^2/r term may be dropped from the estimator. Also, it should be noted that the three F-statistics are related through the following equality:

$$F_{ST} = \frac{F_{IT} - F_{IS}}{1 - F_{IS}} \quad (5.5)$$

Cockerham (1969) developed an alternative approach to defining F-statistics, and this was the first attempt to include the random effects of genetic sampling into the calculation. It is surprising that approximately forty years elapsed between Wright's first partitions of genetic

variation into between and within sub-population components and Cockerham's extension because, in other studies, error attributable to genetic sampling has been acknowledged as 'genetic drift' since Wright himself introduced the term in 1931 (Wright, 1931).

Now, we may shift the focus to using these statistics to identify signatures of historical between populations. This class of method is often referred to as the Lewontin and Krakauer (1973) test. The idea behind this application of *F*-statistics is that an excess of genomic differentiation between populations, at a particular locus, is indicative of divergent selection. Alternatively, a dearth of differentiation at a locus could suggest some sort of balancing selection – selection against differentiation. Akey *et al*. (2002) performed an F_{ST} analysis to scan for selection in a human data set of individuals from three populations typed at over 25,000 single nucleotide polymorphisms (SNPs). The authors compared single-locus estimates of F_{ST} to the empirical genome-wide distribution of the statistic. Based on this analysis, 174 genes were identified as corresponding to historical selection. In another study, Beaumont and Balding (2004) performed simulations to determine the power of F_{ST} to identify signatures of selection. It was found that both Bayesian and Frequentist methods can detect divergent selection if the selection coefficient is at least five times the migration rate. However, the authors also found that none of the methods they tested was capable of identifying balancing selection.

Identifying selection by comparing molecular markers between sub-populations is a potentially powerful approach, but like any method, it is not without shortcomings. Perhaps the most obvious of these involves the cohorts on which the scans are conducted. To perform this type of scan, genotypic information is required on many individuals from multiple populations. With more populations and more individuals, the scan has increasing power. The number of individuals genotyped is limited by resources, but it is not difficult to imagine scenarios where

identifying and genotyping several populations is impossible. For instance, in a species where the boundaries of migration are unclear, it may be difficult to group individuals into respective sub-populations. Another drawback can be found in the situation where an allele is favored in each of the typed sub-populations. In this case, the frequency of the favorable allele in every sub-population is expected to be high and thus the variation between sub-populations will be small (Holsinger and Weir, 2009).

LD relationships across segments of DNA: A second class of signature that may become apparent occurs when an allele at a locus under selection rapidly increases in frequency, and as it does so it carries along the frequency of alleles at linked sites (Maynard Smith and Haigh, 1974). The signature created is that of increased linkage disequilibrium (LD) between the selected allele and those within the haplotype where arose. When selected allele has risen to fixation it is expected that fixation will also be observed across the haplotype in which that allele appeared. This type of signature is most pronounced when the selected allele began at extremely low frequency or as a novel mutation, when the strength of selection is strong, and when the signature is identified during selection or a small number of generations after selection ceased. Such a selective sweep has been deemed a 'hard sweep'. Conversely, the situation in which selection operates on a variant that was already segregating in the population before selection began has been referred to as a 'soft sweep' (Hermisson and Pennings, 2005). If the selected allele is not rare, it will exist in a multitude of haplotypes and therefore no increase in LD is to be expected. Also, under weak selection the allele will rise in frequency no faster than recombination breaks its LD with surrounding alleles. Lastly, as generations elapse, recombination is expected to break down LD regardless of the selection strength. As soon as the

selected allele reaches fixation, any LD will no longer be identifiable, thus making identification of the selected locus difficult (Akey *et al.* 2009). One major advantage of scanning for this type of selection is that only one population of individuals is required for the signature to be identified. Drawbacks, however, include the criteria for the signature's creation, as described above, along with the fact that extremely dense marker sets are necessary for increased LD over a region to be identified.

One of the first studies that presented methods to search for this type of selection signature was Sabeti *et al.* (2002). The authors first identified 'core haplotypes' based on SNPs of sufficient density such that recombination within the core haplotypes was extremely rare. Therefore, individuals with a shared core haplotype are almost certainly identical by descent at that haplotype. Next, LD decay from the core haplotype to surrounding SNPs was measured, based on a statistic called extended haplotype homozygosity (EHH). EHH is defined as the probability that two chromosomes, chosen at random from all individuals carrying the core haplotype, are identical by descent up to a specified distance from the core haplotype. Thus, EHH measures the size of a selective sweep around a specified core haplotype. Finally, EHH scores were simulated under a constant-sized population, expansion, bottleneck, and population structure to obtain p-values for selection under each of these population-genetic scenarios. The authors used this approach to show that two loci conveying resistance to malaria, *G6PD* and *CD40*, have been under selection in the human genome. They also showed that a variety of other more classical tests for selection were insufficient to implicate these loci.

The work by Sabeti *et al*. (2002) was groundbreaking in that it provided a framework to identify recent selection based on genomic signals that had previously been mostly overlooked. However, the methods developed were optimized to test candidate loci for selection and their

extension for performing a genome-wide scan was not immediately clear. Later, Voight *et al*. (2006) generalized the EHH method to be fully applicable to genome-wide data and termed their statistic the integrated haplotype score (iHS). The iHS is based on the EHH from a SNP, rather than a core haplotype, and it also differs in that it is based on the integral of the EHH as the distance from the core SNP increases. The iHS is defined as:

$$iHS = \frac{ln\left(\frac{iHH_A}{iHH_D}\right) - E_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]}{SD_p\left[ln\left(\frac{iHH_A}{iHH_D}\right)\right]} \quad (4.6)$$

In the above equation, iHH_A and iHH_D represent the integrated EHH computed from the ancestral and derived alleles, respectively. The expectation and standard deviations are empirically determined based on all SNPs with allele frequency p. The authors did not offer a technique for computing precise p-values. Instead they suggested comparing iHS values to the genome-wide distribution of iHS, as is often done for scans of selection based on other statistics, such as F_{ST}.

Recent applications of this type of scan are numerous in studies of selection in humans, where dense genotypic information has been available for many years. One group used the iHS test, along with related tests including some that were created to incorporate differences between populations, to identify over 300 regions in the human genome that have been under selection in recent years (Sabeti *et al.* 2007). In another example, a genome-wide association study was conducted to look for genomic regions associated with blood pressure in humans. After associated SNPs were identified, a test for selection on one of the associated SNPs with a particularly strong signal was conducted based on iHS, and evidence for selection on the allele

conferring lower blood pressure was apparent (Newton-Chech *et al.* 2009). This type of scan has not been restricted to SNP studies; Conrad *et al.* (2010) used both the EHH and iHS tests to explore the patterns of selection exhibited for copy number variants in the human genome and found many that have been under selection.

After the success of scans for selective sweeps in humans, animal geneticists have begun applying these tests in agricultural species. For instance, Qanbari *et al*. (2011) employed both the iHS test as well as a Bayesian F_{ST} approach (Gianola *et al.* 2010) across 13 cattle populations to identify an array of regions that have been under recent selection. Some of the most interesting regions identified exhibited overlap between the two tests as well as with previously identified QTL with strong effects. However, their test employed fewer markers than many of the human studies and therefore it is possible that important regions were missed. In plants, method-improvement work has been performed utilizing *Arabidopsis thaliana* as a model (Günther and Schmid, 2011). Also, Olsen *et al*. (2006) utilized the EHH test to show evidence of a strong selective sweep in rice corresponding to the absence of amylose in certain rice varieties, a trait that is prevalent in Northeast Asian cuisine.

This class of test has the most power for studying hard sweeps, where a new variant is immediately favorable upon a arising from a mutation. Soft sweeps, however, for which a variant may exist at intermediate frequency before becoming selectively advantageous due to, for instance, a change in selective pressures, are more difficult to identify via LD-based methods. Much of the research that has discriminated between hard and soft selective sweeps has been performed using plants, because the process of crop domestication sets up a situation in which soft sweeps are likely to arise. Innan and Kim (2004) modeled selection during a domestication event and found that, unlike in the case of a hard sweep, genetic variability surrounding a

selected locus during domestication may not be diminished even if the selected locus has reached fixation. This results from the likelihood that loci selected during domestication are segregating in the population before domestication, so there is a high possibility that the selected locus is not in LD with a long-extending haplotype. The authors then compared their findings to what is known about the domestication of maize and concluded that, with the exception of a known domestication gene, *tb1* (Wang *et al.* 1999; Clark *et al.* 2004), the evidence for selection at other candidate domestication genes (Whitt *et al.* 2002) is consistent with their model of selection on standing variation. Artificial selection on existing crops mimics the pattern of selection during a domestication event, i.e. new selective forces allow variants that may have formerly been neutral to become advantageous. Therefore, selection experiments involving existing varieties are likely to show patterns of soft sweeps. This was demonstrated by Raquin *et al.* (2008), who showed that selection for plant height in wheat successfully modified the frequency of a known, large effect gene, *Rht-B1*. However, the authors found no evidence of a decrease in diversity at loci surrounding *Rht-B1*, which suggests the occurrence of a soft sweep.

The site frequency spectrum

A final class of signature that may indicate selection is an abnormality in the SFS across a region. In a sample of n sequences, each segregating site will be found on at least one and no more than $n-1$ sequences. The SFS refers to the number of sites observed on exactly k sequences, where $1 \leq k \leq n-1$. In other words, the SFS may be thought of as the distribution of alleles at particular frequencies. Under a neutral model, the SFS is expected to display a pattern that is determined by population demography alone. However, Tajima (1989) showed that selection on

a beneficial mutation is expected to shift the distribution so that there is an excess of sequences with few mutations. This results from the fact that, as an allele is selected for or against, the level of genetic variability surrounding the respective locus is expected to decrease. A statistic denoted as 'D' was therefore proposed to test the neutrality of sites. Tajima's D statistic is defined as

$$D = \frac{\pi_n - S/a_n}{\sqrt{\mathrm{Var}(\pi_n - \frac{S}{a_n})}}, \text{ with } a_n = \sum_{i=1}^{n-1} \frac{1}{i} \quad , \tag{5.7}$$

where *S* is the total number of segregating sites in a DNA sample and π_n represents the average number of differences between any two sequences chosen from the total sample of n sequences. Under neutrality, it is expected that $\pi_n = 4Ne\mu$, where *Ne* is the effective population size and μ is the mutation rate (Tajima, 1983), and that $S/a_n = 4Ne$ (Waterson, 1975). Therefore *D* represents difference between these two estimates standardized by their variability. Furthermore, under the assumption of neutrality Tajima's *D* is expected to follow a scaled beta distribution with mean zero and variance one, and thus *D* can be used as a test for selection. Scaling is performed so that the support of the distribution is [D_{min}, D_{max}] (Simonsen *et al.* 1995). Significantly positive values suggest a selective sweep, while significantly negative values suggest balancing selection.

A related set of statistics that can be used to test for selection was proposed by Fu and Li (1993). These statistics are often referred to as Fu and Li's *D** and *F**. The Fu and Li tests are both based on the same principle as Tajima's test, in that they compare estimators of $4Ne\mu$ that are equivalent under neutrality but differ under selection, but the tests proposed utilize estimators determined by the number of singletons in a sample. The D* test compares $4Ne\mu$ based on the number of segregating sites to that based on singletons, while the F* test compares $4Ne\mu$ based on pairwise differences to that based on singletons.

Simonsen's *et al.* (1995) showed through simulations that the Tajima test tends to be the most powerful for identifying selective sweeps. However, the authors also showed that the beta distribution offered by Tajima may be overly conservative, and they offered an alternative specification of critical values. Moreover, regarding the Fu and Li tests, the alternative method for critical value determination given by Simonsen *et al.* (1995) also generalizes some of the previous requirements (i.e. that the true $4N_e\mu$ falls between two and 20). A variety of additional tests based on the SFS have been proposed. These include Fay and Wu's H test (Fay and Wu, 2000) and Zeng *et al.*'s E test (Zeng *et al.* 2006). Both of these tests involve an exploration of low, intermediate, and high frequency alleles, which makes the tests powerful for identifying selection. But, they have the added complication of requiring information pertaining to ancestral and derived alleles, which may be inferred from an outgroup.

Statistically separating selection from other forces

As described in the previous section, a wide variety of signals may be generated as a result of selection. However, all of these signals are affected by additional population forces as well. Mutation, migration, recombination, and drift are all phenomena that impact genetic variability. In many instances, these forces may partially or completely mask the effects of selection. Identifying selected polymorphisms is therefore not simply a matter of identifying those loci that display a pattern expected to result from selection, but instead it is necessary to determine those displaying a pattern consistent with no forces other than selection. Moreover, these additional forces are rarely precisely quantifiable in natural populations, and even in well controlled experimental populations they are often difficult to predict accurately.

This was apparent in the artificial selection experiment described in Chapter Two. Here, a maize population of a known size and with a controlled number of mating individuals was subjected to strong selection for 30 generations for an increase in ear number. Migration was not permitted during the experiment, and the total number of individuals and generations was small enough that mutation was not expected to be of consequence. Individuals representing the population before and after selection were sampled and sequenced, and therefore identifying selected sites was theoretically a simple matter of determining which loci depicted allele frequency changes that were greater than the expectation due to drift alone. Equation 5.1 gives the variability of allele frequencies expected due to drift, but this equation was not applicable due to the different number of mating males and females, and because in this experiment drift and sampling error occurred at three levels: during the mating process, during sampling of individuals to be sequenced, and during sequencing itself.

Instead, simulations were employed to estimate the maximum magnitude of drift expected for individual loci and, after correcting for multiple testing, it was apparent that even the largest changes in allele frequency could rarely be distinguished from the combined effects of drift and sampling error. However, much of this uncertainty resulted not from genetic drift, but instead from sampling and sequencing a limited number of individuals (Figure 5.1). The most straightforward way to address this complication would be to perform increased sequencing on individuals rather than pools, but this was prohibited by cost and still would be for all but very well-funded experiments. As an alternative solution, allele frequencies were compared by averaging F_{ST} estimates from individual SNPs over windows of 25 adjacent SNPs. This increased the coverage per observation from a minimum of 20 to a minimum of 500. Figure 5.2 demonstrates the dramatic reduction in sampling error that this approach enabled.

Even with this adjustment to the analysis protocol, complications remained. By utilizing F_{ST} over windows instead of for individual SNPs, comparable simulations measuring genetic drift to determine significance also needed to be based on averages over windows, and to perform these appropriately, knowledge of the levels of linkage disequilibrium at the onset of the selection protocol was required. Since linkage disequilibrium estimates depend on the relationships between variants within individuals, there is not an accurate method to estimate these from pooled sequencing data with short read-lengths, as were generated for this experiment.

Therefore, even though this experiment involved an experimental population with a known reproductive scheme, after sampling error was reduced to a manageable level it was impossible to establish statistically sound probabilities for a locus having been impacted by selection. Because the population was subjected to strong artificial selection, though, it was reasonable to assert that the most outlying windows putatively corresponded to selection, and therefore 99.9% and 99.99% outlier thresholds were used to isolate the most likely selection candidates. For natural and artificial populations that are less well-documented, such an outlier approach is often the best that can be achieved (Akey, 2009). Even though outlier thresholds such as these may successfully identify selected sites, there is no guarantee that they do so. In populations minimally or not at all effected by selection, outlier thresholds guarantee false positives, since for any distribution a top 0.1% must exist. Similarly, when selection has effected a larger portion of the genome, an outlier threshold may fail to identify many of the important candidates. In the current era of massive amounts of data and tremendous computational tools, it is unfortunate that no mechanism yet exists to determine statistical significance for these types of studies.

Strides toward the goal of more appropriately characterizing and distinguishing outliers from these types of studies are continuously being made. In Chapter Three, a method was proposed that is capable of using the semi-continuous nature of LD between adjacent markers to separate blocks of the genome that may be functionally related. This approach may allow a better separation of outliers and non-outliers. In brief, the method involves fitting a cubic smoothing spline to individual marker estimates of F_{ST} and identifying the inflection points of the fitted spline. The inflection points are then used to define separate windows over which F_{ST} means may be computed and subsequently compared.

Although this method does not require LD information, the roughness of the spline is dependent upon the relatedness of estimates from individual markers, and therefore LD is approximately accounted for. Specifically, when windows of an arbitrarily defined and uniform size are used to analyze data, the breaks between adjacent windows have little meaning and outlying regions do not necessarily stand out. Graphically, this can be observed in Appendix A, which corresponds to a sliding-window analysis of F_{ST} data from Chapter Two, where it is clear that every outlying window is surrounded by others that are just below the outlier level. However, Figure 3.2 demonstrates that by defining non-overlapping windows according to spline inflection points, the separation between outliers and the genomic background may become more pronounced. If selection had not occurred, the expectation for figure 3.2 (bottom) would be of one with no isolated points.

In addition, because the inflection points of the fitted spline are dependent upon LD, future research should be conducted to formally evaluate this relationship. As discussed previously, the only way to identify significance thresholds for selection is to account for the non-selective forces that also impact genetic variability. This means that realistic simulations

must be conducted, and if a statistic is computed over windows of markers the simulations must also be computed for windows of markers. If it is determined that spline inflection points can be used to estimate blocks of genetic material that have been inherited as a unit, these blocks may be used for the purpose of simulating drift in the context of linkage and linkage disequilibrium.

While simulations are useful for establishing the significance of selection, another approach was utilized in Chapter Four. In this experiment, 72 distinct breeds of chicken were analyzed for evidence of selection that operated jointly on pairs of linked loci. Although the recent demographic history of some of these breeds is approximately known, for the majority of them it is not. Many of the breeds studied have been maintained by breeders around the world for unique traits, utilizing highly variable and constantly fluctuating population sizes and migration schemes. This precluded the implementation of simulations to assess the effects of drift and migration. However, because LD between pairs of loci was utilized for the test, a different approach that was a hybrid between demographic history and outlier thresholds was possible.

For single-locus selection that affects two loci within a haplotype block, or for epistatic selection between a pair of loci that are genetically linked, the level of LD between the two loci is expected to increase over generations. This is because the frequency of individuals with the selected haplotype or epistatic pair will increase at a faster rate than recombination will break apart the relationship. However, for loci that are not genetically linked, recombination is expected to break the LD relationship between the two loci at a faster rate than selection can increase it (Kimura, 1965; Nagylaki, 1993). This is true for all but the most extreme cases of epistatic selection between unlinked loci, for which selection must be so strong that fixation of the epistatic relationship will occur almost instantly. Therefore, if D'^{2}_{IS}, the statistic described in Chapter Four, is computed for pairs of loci that are not linked (e.g. loci on different

chromosomes), the value of this statistic is effected by the background levels of genetic variability that are not due to selection on a haplotype or epistatic selection on a locus pair. This background variability captures most of the demographic history of the population but not joint selection on a pair. Conversely, when two loci are linked the value of D'^2_{IS} computed between them represents demographic history as well as selection. Therefore, the maximum values of D'^2_{IS} that are observed for unlinked loci may be used to set critical thresholds for values of D'^2_{IS} that do not include selection. These thresholds account for the non-selective forces that affect genetic variability, except for the situation where two non-selected loci drift together do to linkage between them. Because of this situation, this approach represents a compromise between an outlier test and one that determines true significance. In essence, a lower limit for significance is identified, whereby all values of D'^2_{IS} that may be explained without invoking selection are removed, and those pairs of loci remaining are the most promising candidates for selection, although they do not certainly correspond to it.

This approach removed the overwhelming majority of locus-pairs as explainable without selection, leaving a set of pairs that appear to correspond to selection, and one to three that seem to show cases of selection on an epistatic relationship. At its root this is still an outlier-based technique, but unlike typical outlier approaches, the strategy does not ensure a set number of outlying sites. When there is a large amount of the genome affected by selection, an increased number of pairs is expected to be identified, and when there is little selection it is possible that no pairs will be identified. Still, rather than asserting that all of the significant sites correspond to selection, there is the possibility that some of those that were identified correspond to pairs that have drifted by an abnormally large amount. Overall, this provides further evidence that even when complex statistics are computed and unique thresholds are derived, the biological

processes in play during the evolution of populations are extraordinarily intricate and the identification of selection is no trivial matter.

In fact, in studies such as the one in Chapter Four, where a diverse array of populations was utilized, even those loci that clearly correspond to selection may not have functional roles that are immediately interpretable. This is because unlike genome wide association studies or quantitative trait locus mapping, scans for selection are blind to phenotypes and therefore the selected variants could correspond to any characteristic that is relevant to fitness. For instance, among the selected sites in the chicken genome that were identified, only two could be linked to definitive biological functions. These two regions had previously been studied in detail using other techniques, and these earlier studies are the reason that biological interpretations were possible. A third example region, found on chromosome 5 of the chicken genome (Figure 4.4A,B), clearl depicts a region with almost certain evidence of dispersive selection on a haplotype, and yet identifying the function of this region among the 72 breeds studied would require an entire additional study of its own.

Even though complications for determining the significance of selected loci in natural and experimental populations abound, there is one remaining technique that, when it can be employed, represents a ‘gold standard’ for determining true significance. This approach involves performing selection on experimental populations with replications, controls, or preferably both. By utilizing controls that mimic the mating scheme of the selected population or populations, boundaries for changes in genomic variability due to non-selective forces can be identified. In many ways, such controls correspond to biological implementations of the simulations described above. Replicated selection populations can also provide additional evidence that loci demonstrating patterns of selection do reproducibly. It is necessary that all of the control and

selection populations are generated from the same base population for this method to be employed. This is because even for very similar populations, levels and blocks of LD can be unique, leading to different patterns of drift and selection.

Turner and various colleagues (2011, 2012) employed such an approach to investigate body size and courtship song traits using experimental populations of *Drosophila*. In both of these studies, the authors utilized only six populations: Two selected in one direction, two in the other direction, and two controls. Even with this limited number of replications, they were able to convincingly identify a large number of loci displaying patterns of selection that were discernible from drift, based on directly assessing changes in allele frequency at SNP markers. By utilizing control populations, boundaries for drift could be identified, and the replicated selection populations in diverging directions further confirmed that the observed selected sites were not spurious.

However, *Drosophila* is unique compared to most species of interest because of its short generation intervals and the ease with which mating can be controlled. In species with longer generation intervals or that cannot be as well-controlled, comparable levels of replication rarely exist or are not possible. For example, a comparable study in chickens (Johansson *et al.* 2010) utilized divergent populations that had been artificially selected for 50 generations for body weight. Nevertheless, generating these populations was a multi-decade task, with the selection protocol beginning in 1957. Since the initiators of this selection protocol did generate control populations, it was not possible to definitively rule out drift as the source of extreme allelic variation between the populations. Still, divergently selected populations represent a type of replication, and the authors were able to utilize this to identify selected loci.

These *Drosophila* and Chicken experiments are comparable to that employed in Chapter Two, except that the maize population described there did not include any type of control or replication. As discussed, no known statistical analysis can reconstitute the demographic history of such a population or the levels and blocks of LD present within it, so definitive statements about selected loci cannot be made.

Conclusions

Considering all that has been discussed here, it would be wise for newly generated experiments that seek to invoke selection on a population for any goal to include biological controls, regardless of the immediate goals of the experiment. These experiments often involve years of work, generating valuable resources that may be analyzed in a manner that was not feasible when the experiment began. When replication, and especially controls, were included from the start, it is clear that the utility of the population today is dramatically increased.

Still, in naturally occurring populations in the wild, or for human populations, opportunities for replication will never exist. In these instances identifying outliers, or conducting an array of simulations that together represent the range of demographics that is realistic, will likely remain the best option. Therefore, an identification of loci that display patterns most consistent with selection should not be considered a final step. Instead, these studies must be coupled with downstream experiments that incorporate phenotypic information, and only then can the role of loci be validated. Chapters Two and Four of this dissertation have described experiments that will ideally lead to such later experiments by providing promising regions of the maize and chicken genome for further study, that are likely to contribute to

important traits. In Chapter 3, a method was proposed to refine the identification of such regions and provide increased power to detect them coupled with a more reasonable isolation of their location. Ultimately, if quantitative trait locus mapping or genome wide association studies are used to validate the putatively selected regions, Chapter One provided an analysis of tools for generating and utilizing the markers that may be used in such studies. Together, these chapters provide information that will help to better understand and make use of populations undergoing selection.

Figures

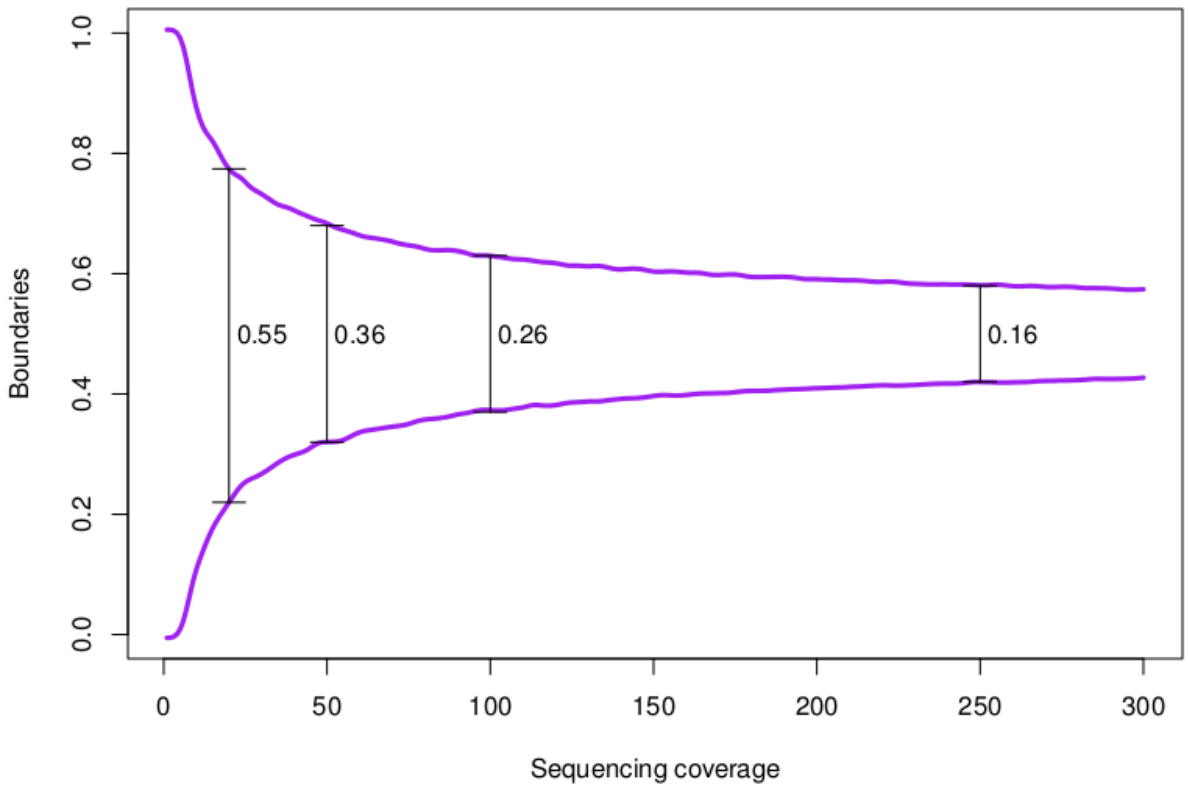


Figure 5.1: Error due to pooled sequencing.

Error intervals arising from the effects of pooled sequencing alone. If pooled sequencing is performed, allele frequencies at each locus will be estimated with error. This plot depicts a 99% interval for the allele frequency based on pooled sequencing of a diallelic locus with true allele frequency in the pool of p=0.5. Purple lines depict the expected maximum and minimum estimates for a specified coverage that will hold 99.5% of the time, based on a binomial approximation. Vertical bars depict the 99% interval for sequencing to a coverage of 20X, 50X, 100X, and 250X, from left to right, respectively.

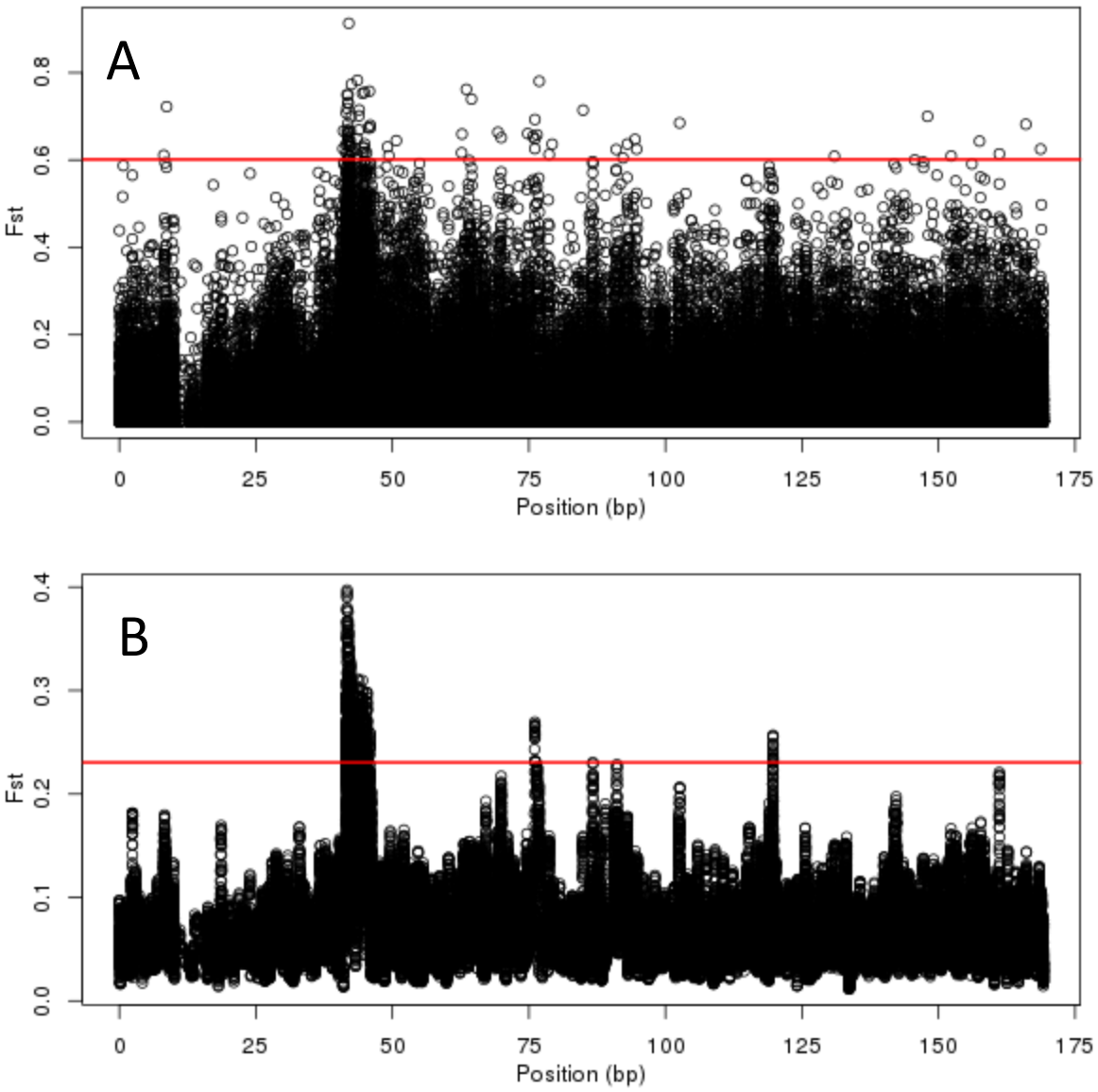


Figure 5.2: Increase in power due to window approaches.

Power is increased by using windows instead of estimates from individual SNPs. A: A plot of F_{ST} along chromosome 2 from the maize experiment described in Chapter 2, for estimates based on individual SNP markers. B: A plot of F_{ST} along chromosome 2 from the same data, but using estimates based on averages over windows of 25 adjacent SNPs. Observe that when estimates are made from individual markers (A), identifying differences between signal and noise is difficult. When estimates from markers are averaged over windows (B), sampling error due to sequencing is reduced and signal becomes distinguishable from noise.

References

Akey, J. M., Zhang, G., Zhang, K., Jin, L., and Shriver, M. D. (2002). Interrogating a high-density SNP map for signatures of natural selection. *Genome Research*, 12.

Akey, J. M. (2009) Constructing genomic maps of positive selection in humans: Where do we go from here? *Genome Res.* 19**(5)**: 711-722.

Beaumont, M. A. and Balding, D. J. (2004). Identifying adaptive genetic divergence among populations from genome scans. *Molecular Ecology*, 13.

Botstein, D., R. L. White, M. H. Skolnick, and R. W. Davis. (1980). Construction of a genetic map in man using restriction fragment length polymorphisms. Am. J. Hum. Genet. 32: 314-331.

Clark, R. M., E. Linton, J. Messing, and J. F. Doebly. (2004). Pattern of diversity in the genomic region near the maize domestication gene *tb1*. Proc. Natl. Acad. Sci. 101(3): 700-707.

Cockerham, C. C. (1969). Variance of gene frequencies. *Evolution*, 23.

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., Aerts, J., Andrews, T. D., Barnes, C., Campbell, P., et al. (2010). Origins and functional impact of copy number variation in the human genome. *Nature*, 464.

Crow, J. F. and Kimura, M. (1970). *An introduction to population genetic theory*. Harper and Row Publishers.

Darwin, C. (1859). *On the origin of species by means of natural selection*. John Murray.

Elshire, R. J., Glaubitx, J. C., Sun, Q., Poland, J. A., Kawamoto, K., et al. (2011). A robust, simple genotyping-by-sequencing (gbs) approach for high diversity species. *PLoS ONE*, 6.

Fan, J.-B., Chee, M. S., and Gunderson, K. L. (2006). Highly parallel genomic assays. *Nature Reviews Genetics*, 7.

Fay, J. C. and Wu, C. (2000). Hitchhiking under positive Darwinian Selection. Genetics 155:1405-1413.

Fisher, R. A. and Ford, E. (1947). The spread of a gene in natural conditions in a colony of the moth *Panaxia dominula* l. *Heredity*, 1.

Frankham, R. and Loebel, D. A. (1992). Modeling problems in conservation genetics using captive *Drosophila* populations: Rapid genetic adaptation to captivity. *Zoo Biology*, 11.

Fu, Y., and W. Li. (1993). Statistical tests of neutrality of mutations. Genetics 133: 693-709.

Gianola, D., Simianer, H., and Qanbari, S. (2010). A two-step method for detecting selection signatures using genetic markers. *Genetics Research*, 92.

Gibson, J., Lewis, N., Adena, M., and Wilson, S. (1979). Selection for ethanol tolerance in two populations o f*Drosophila melanogaster* segregating alcohol dehydrogenase allozymes. *Aust. J. Biol. Sci.*, 32.

Günther, T. and Schmid, K. J. (2011). Improved haplotype-based detection of ongoing selective sweeps toward an application in arabidopsis thaliana. *BMC Res Notes*, 4.

Hermisson, J. and P. S. Pennings. (2005). Soft sweeps: Molecular population genetics of adaptation from standing variation. Genetics 169:2335-2352.

Holsinger, K. E. and Weir, B. S. (2009). Genetics in geographically structured populations: defining, estimating and interpreting F_{ST}. *Nature reviews genetics*, 10.

Hubby, J. and Lewontin, R. (1966). A molecular approach to the study of genic heterozygosity in natural populations i. the number of alleles at different loci in *Drosophila pseudoobscura*. *Genetics*, 54.

Kelly, J. K., B. Koseva, and J. P. Mojica. (2013). The genomic signal of partial sweeps in *Mimulus guttatus*. *Genome Biol Evol* 5(8): 1457-1469.

Kimura, M. (1965). Attainment of quasi linkage equilibrium when gene frequencies are changing by natural selection. *Genetics*, 52.

Krimbas, C. B. and Tsakas, S. (1971). The genetics of *Dacus oleae* v. changes of esterase polymorphism in a natural population following insecticide control–selection or drift? *Evolution*, 25.

Kuleshov, N. (1933). World's diversity of phenotypes of maize. *Journal of the American Society of Agronomy*, 25.

Lewontin, R. C. and Krakauer, J. (1973). Distribution of gene frequency as a test of the theory of the selective neutrality of polymorphisms. *Genetics* 74: 175-195.

Maynard Smith, J. and Haigh, J. (1974). The hitch-hiking effect of a favorable gene. *Genetical Research*, 23.

Mendel, G. (1865). Versuche über pflanzenhybriden. In *Verhandlungen des naturforschenden Vereines*, Brün.

Meuller, L., Wilcox, B., Ehrlich, P., Heckel, D., and Murphy, D. (1985). A direct assessment of the role of genetic drift in determining allele frequency variation in populations of *Euphydryas editha*. *Genetics*, 110.

Nagylaki, T. (1993). The evolution of multilocus systems under weak selection. *Genetics*, 134.

Newton-Chech, C., Johnson, T., Gateva, V., Tobin, M. D., Najjar, S. S., Zhao, J. H., Heath, S. C., Eyheramendy, S., et al. (2009). Genome-wide association study identifies eight loci associated with blood pressure. *Nature Genetics*, 41.

Olsen, K. M., Caicedo, A. L., Polato, N., McClung, A., McCouch, S., and Purugganan, M. D. (2006). Selection under domestication: Evidence for a sweep in the rice *waxy* genomic region. *Genetics*, 173.

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

Qanbari, S., Gianola, D., Hayes, B., Schenkel, F., Miller, S., Moore, S., Thaller, G., and Simianer, H. (2011). Application of site and haplotype-frequency based approaches for detecting selection signatures in cattle. *BMC Genomics*, 12.

Queller, D. C., J. E. Strassmann, and C. R. Huges. (1993). Microsatellites and kinship. Trends in Ecology and Evolution 8: 285-288.

Raquin, A. L., P. Brabant, B. Rhone, F. Balfourier, P. Leroy, and I. Goldringer. (2008). Soft selective sweep near a gene that increases plant height in wheat. Molecular Ecology 17: 741-756.

Sabeti, P. C., Reich, D. E., Higgins, J. M., Levine, H. Z. P., Richter, D. J., Schaffner, S. F., Gabriel, S. B., Platko, J. V., Patterson, N. J., McDonald, G. J., Ackerman, H. C., Campbell, S. J., Altshuler, D., Cooper, R., Kwiatkowski, D., Ward, R., and Lander, E. S. (2002). Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419.

Sabeti, P. C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E. H., McCarroll, S. A., Gaudet, R., Schaffner, S. F., Lander, E. S., and The International HapMap Consortium. (2007). Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449.

Schaffer, H., Yardley, D., and Anderson, W. (1977). Drift or selection: a statistical test of gene frequency change over generations. *Genetics*, 87.

Simonsen, K. L., G. A. Churchill, and C. F. Aquadro. (1995). Properties of statistical tests of neutrality for DNA polymorphism Data. Genetics 141: 413-429.

Stuber, C. W. and Moll, R. (1972). Frequency changes of isozyme alleles in a selection experiment for grain yield in maize (*Zea mays* l.). *Crop Science*, 12.

Tajima, F. (1983). Evolutionary relationship of DNA sequences in finite populations. Genetics 105: 437-460.

Tajima, F. (1989). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

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

Turner, T. L., Miller, P. M. (2012). Investigating natural variation in *Drosophila* courtship song by the evolve and resequence approach. *Genetics*, 191: 633-642.

Voight, B. F., Kudaravalli, S., Wen, X., and Pritchard, J. K. (2006). A map of recent positive selection in the human genome. *PLoS Biology*, 4.

Vos, P., R. Hogers, M. Bleeker, M. Rijans, T. Van de Lee, *et al.* (1995). AFLP: A new technique for DNA fingerprinting. Nucleic Acids Research 23: 4407-4414.

Wang, R., A. Stec, J. Hey, L. Lukens, and J. Doebly. (1999). The limits of selection during maize domestication. Nature 398: 236-239.

Waples, R. S. (1989). Temporal variation in allele frequencies: testing the right hypothesis. *Evolution*, 43.

Watson, J. D. and Crick, F. H. C. (1953). A structure for deoxyribose nucleic acid. *Nature*, 171.

Watterson, G. (1982). Testing selection at a single locus. *Biometrics*, 38.

Weir, B. S. and C. C. Cockerham. (1984). Estimating F-statistics for the analysis of population structure. Evolution 38(6): 1358-1370.

Whitt, S. R., L. M. Wilson, M. I. Tenaillon, B. S. Gaut, and E. S. Buckler. (2002). Genetic diversity and selection in the maize starch pathway. Proc. Natl. Acad. Sci. 99(20): 12959-12962.

Wilson, S. (1980). Analyzing gene-frequency data when effective population size is finite. *Genetics*, 95.

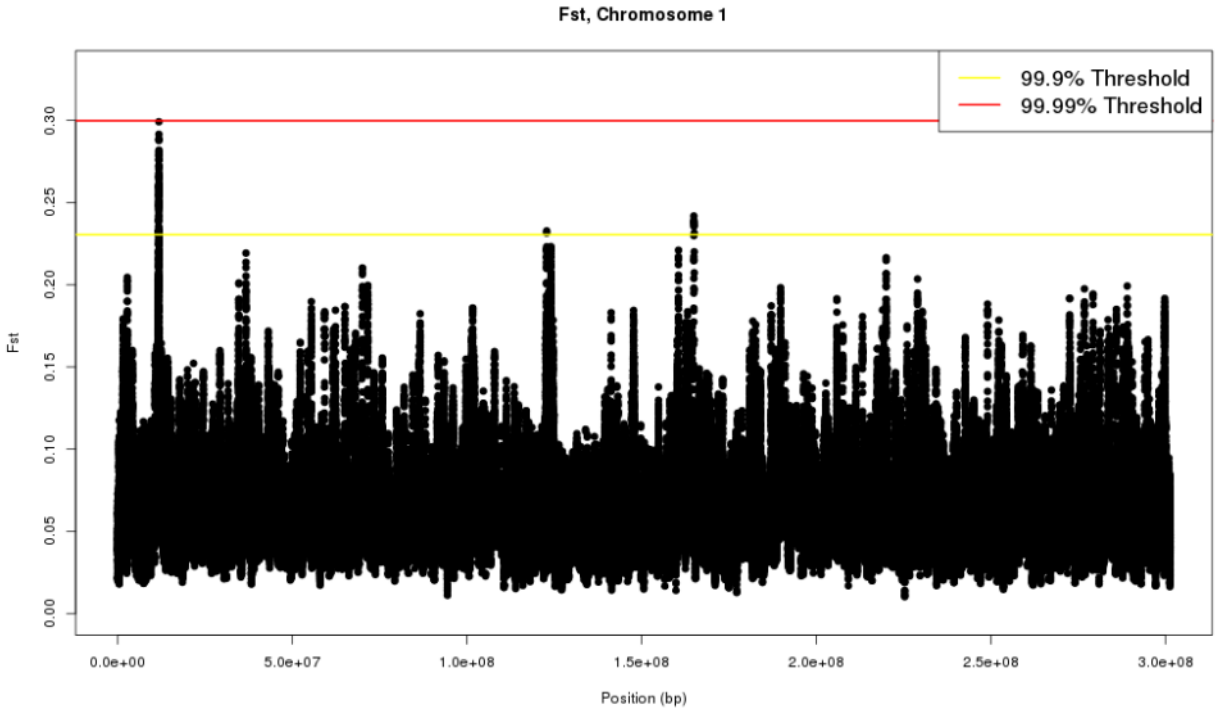
Wisser, R. J., Murray, S. C., Kolkman, J. M., Ceballos, H., and Nelson, R. J. (2008). Selection mapping of loci for quantitative disease resistance in a diverse maize population. *Genetics*, 180.

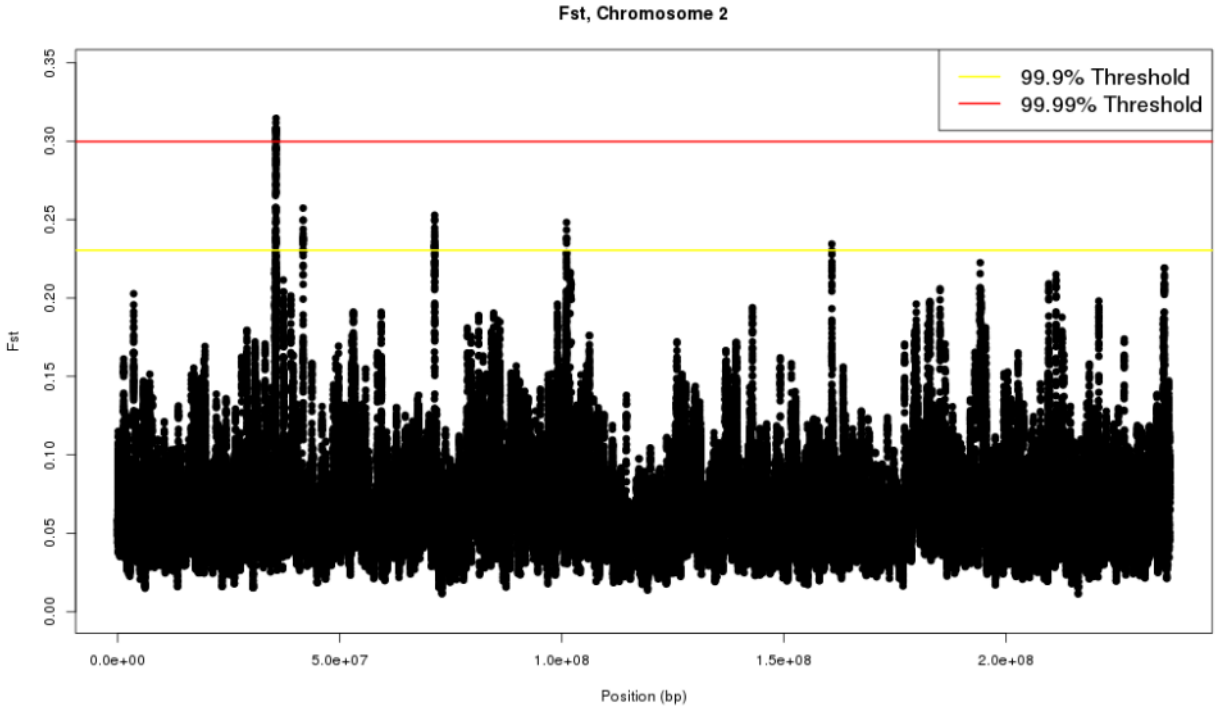
Wright, S. (1931). Evolution in mendelian populations. *Genetics*, 16.

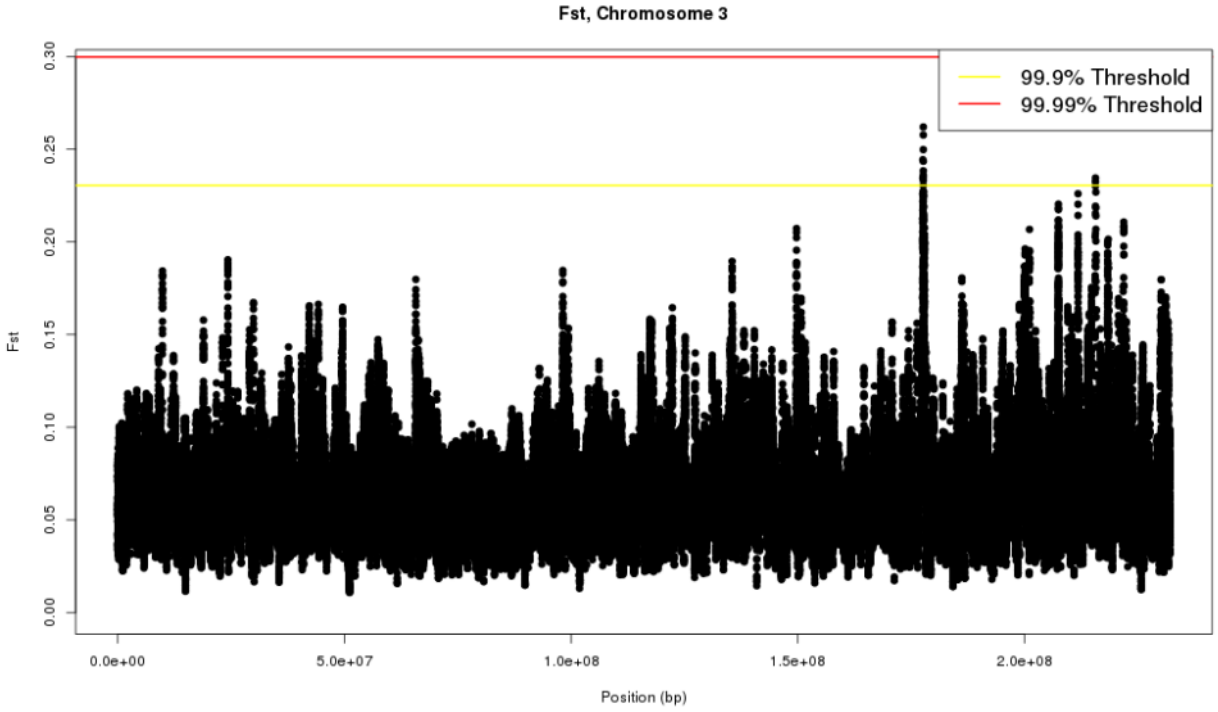
Wright, S. (1951). The genetical structure of populations. *Ann. Eugen.*, 15.

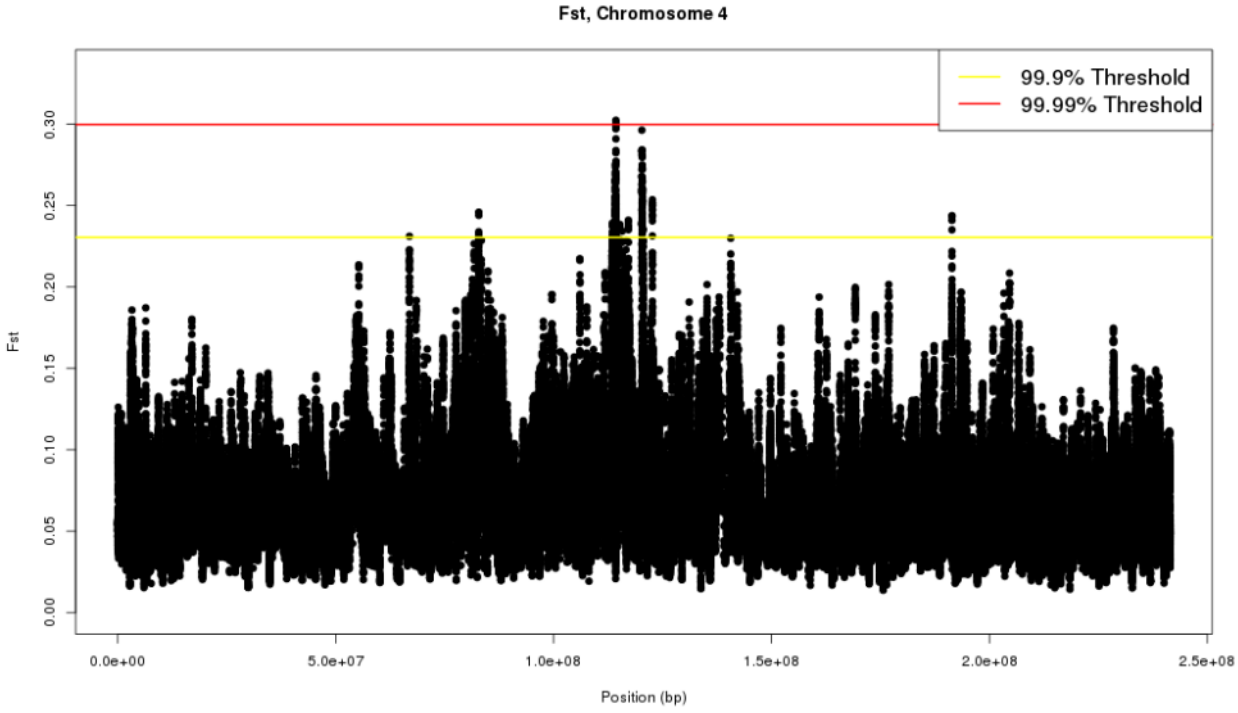
APPENDIX A

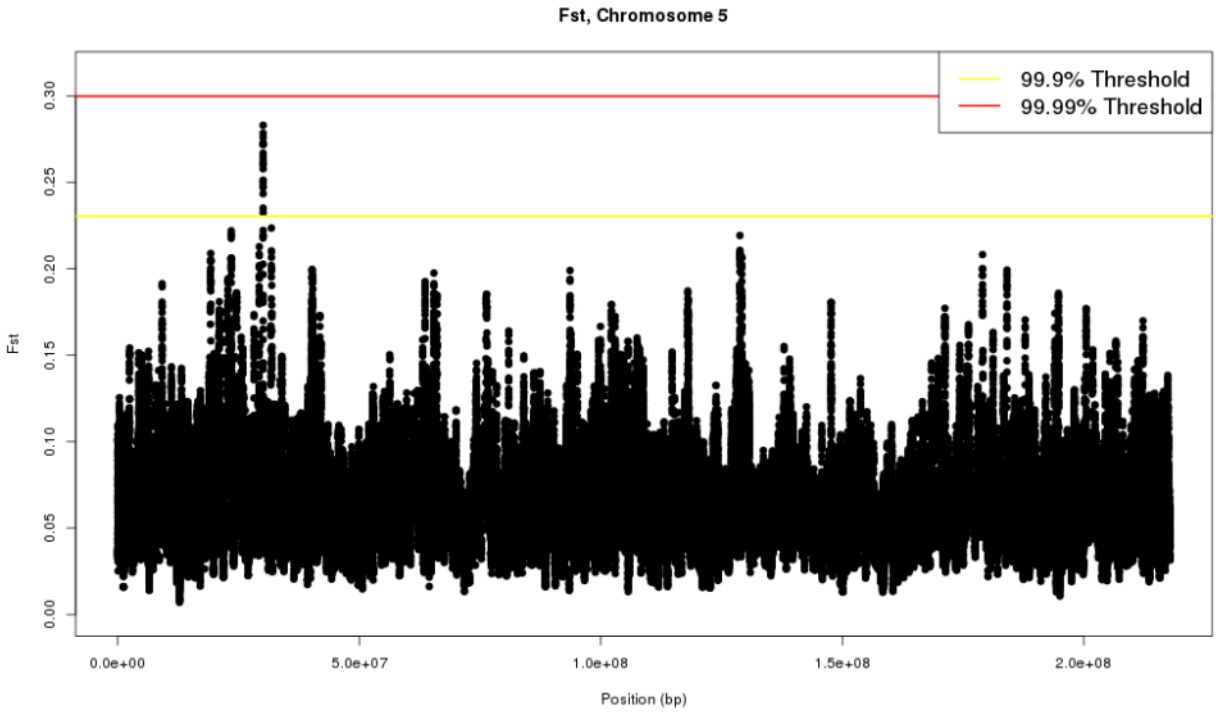
The following figures provide sliding window F_{ST} values, computed by comparing cycle 0 and cycle 30 of the Golden Glow maize population, for each SNP on all chromosomes.

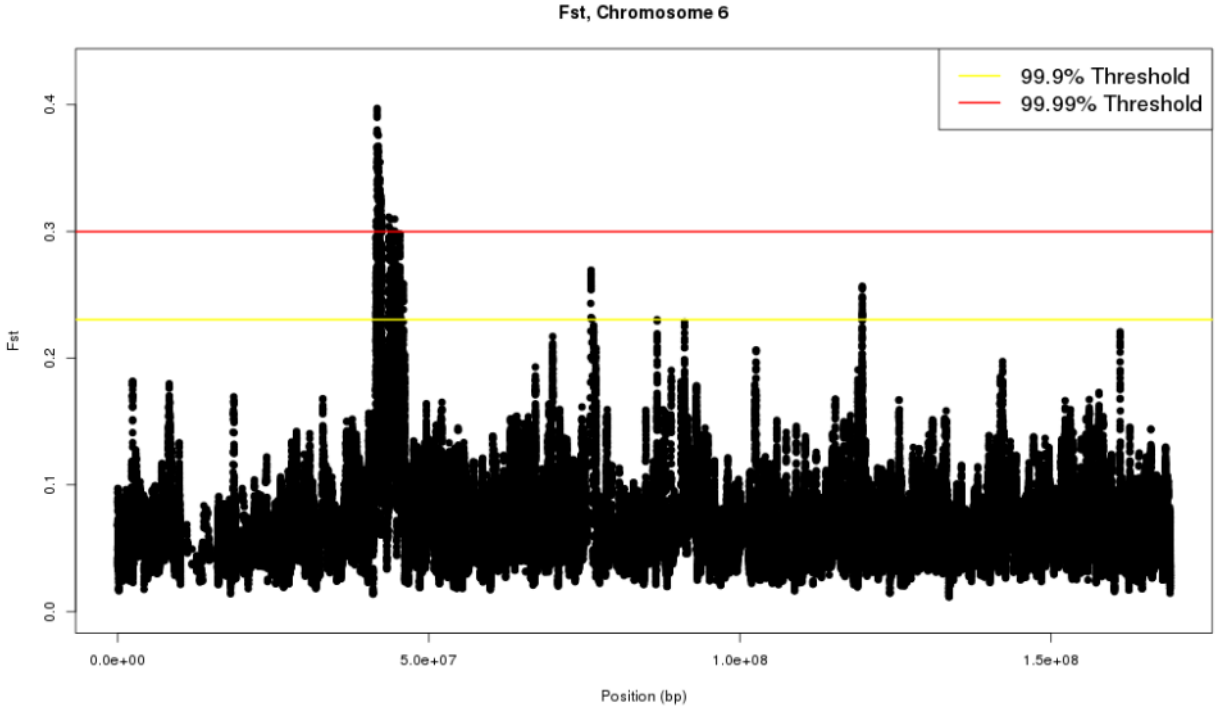


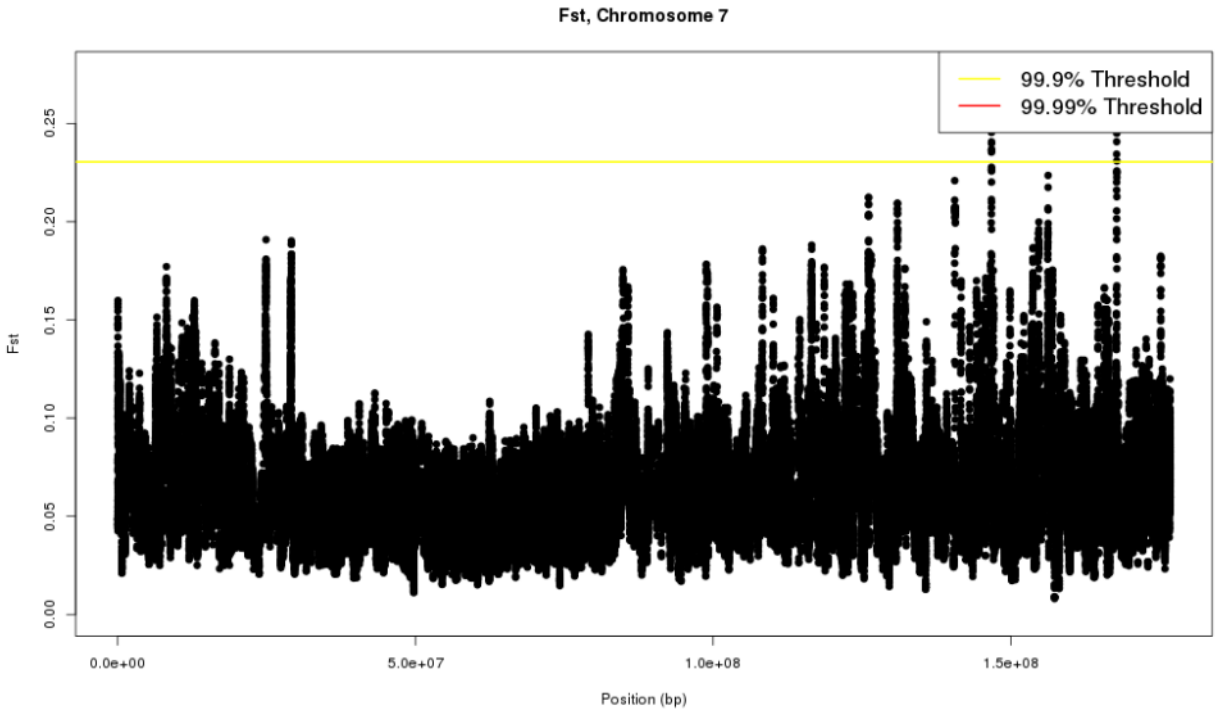


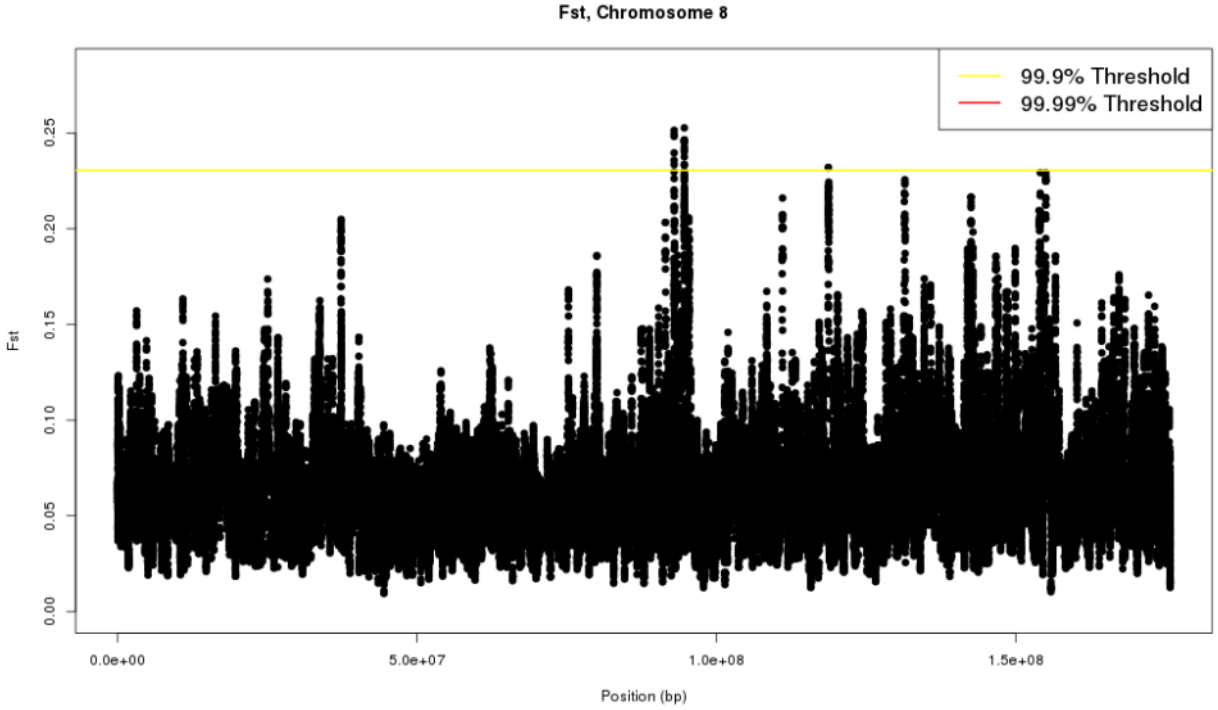

Fst, Chromosome 3
99.9% Threshold
99.99% Threshold
Fst
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
Position (bp)

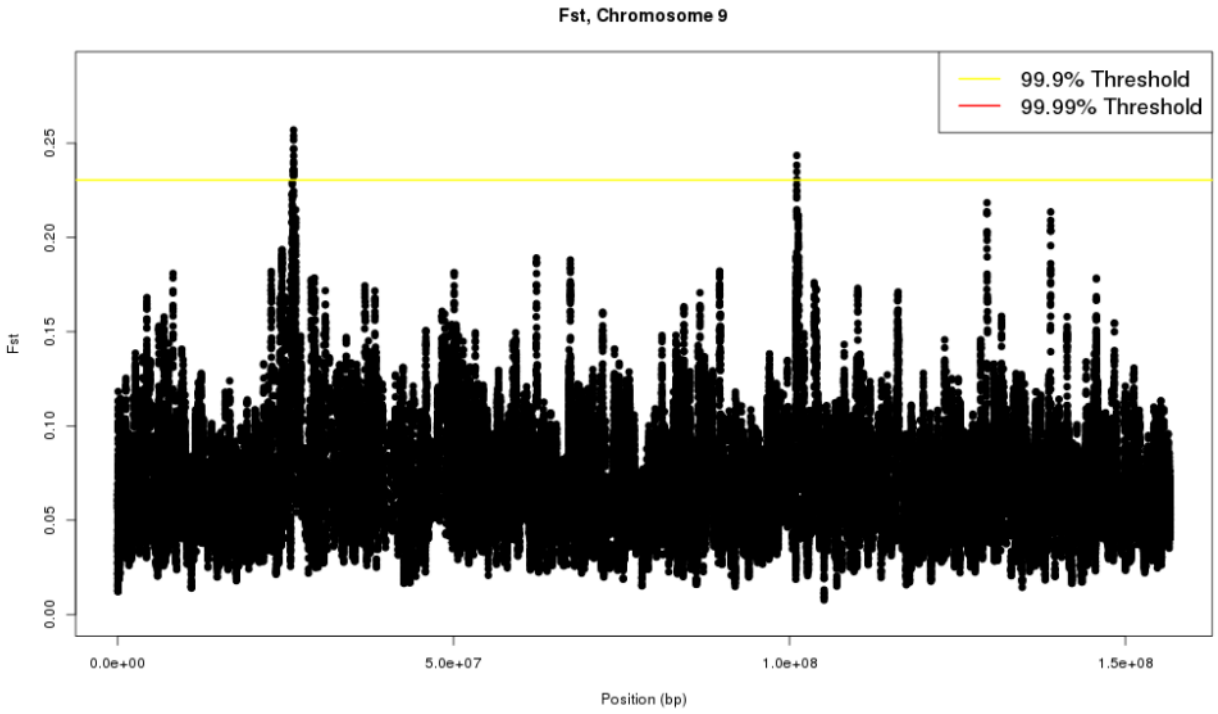

Fst, Chromosome 4
99.9% Threshold
99.99% Threshold
Fst
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
2.5e+08
Position (bp)

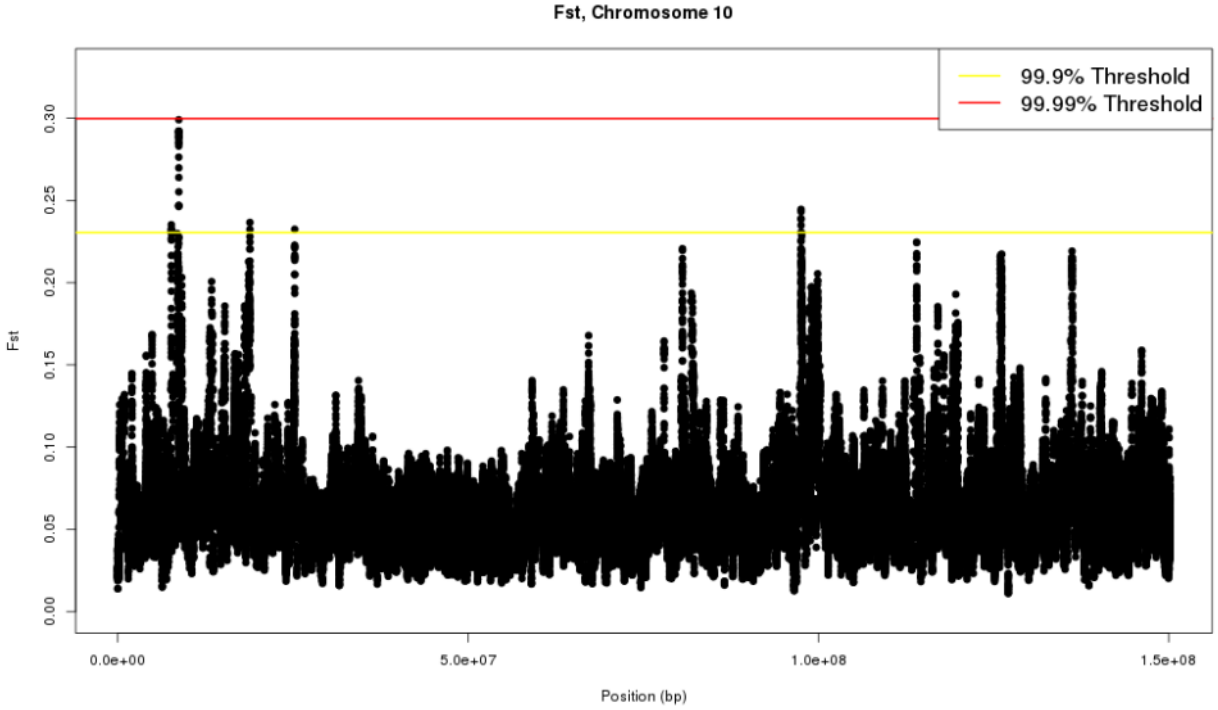

Fst, Chromosome 5
99.9% Threshold
99.99% Threshold
Fst
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0e+00
5.0e+07
1.0e+08
1.5e+08
2.0e+08
Position (bp)


Fst, Chromosome 6
99.9% Threshold
99.99% Threshold
Fst
0.0
0.1
0.2
0.3
0.4
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position (bp)


Fst, Chromosome 7
99.9% Threshold
99.99% Threshold
Fst
0.00
0.05
0.10
0.15
0.20
0.25
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position (bp)


Fst, Chromosome 8
99.9% Threshold
99.99% Threshold
Fst
0.00
0.05
0.10
0.15
0.20
0.25
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position (bp)


Fst, Chromosome 9
99.9% Threshold
99.99% Threshold
Fst
0.00
0.05
0.10
0.15
0.20
0.25
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position (bp)


Fst, Chromosome 10
99.9% Threshold
99.99% Threshold
Fst
0.00
0.05
0.10
0.15
0.20
0.25
0.30
0.0e+00
5.0e+07
1.0e+08
1.5e+08
Position (bp)

APPENDIX B

The following R code defines a function for evaluating genomic data over spline-based windows, as described in Chapter 3. This function will be made available on CRAN.

```
#######################################################################
### This is a functions that can be called for spline                 ###
### analysis of sliding window data                                   ###
#######################################################################

### Inputs:
  # Fst: a vector of Fst values at each marker
  # map: a vector of positions for each marker with a corresponding
  #      Fst value
  # smoothness: The level of resolution (in base pairs) for computing
  #             the spline
  #             and its derivatives
  # plotFst: Whether or not to include a plot of raw Fst values, with
  #          the computed spline
  # plotWindows: Whether or not to include a plot of Fst values over
  #              the computed windows
  # method: The method for controlling amount of smoothing: 1, 2, 3, \
  #         or 4. See documentation of smooth.Pspline for description.
  #         Usual choices are either 3 for generalized cross
  #         validation or 4 for ordinary cross validation.

### Outputs: The output is a list. Its elements are:
  # rawSpline: The fitted spline object
  # breaks: The spline-suggested window breaks
  # distinctFst: A full table of mean Fst values computed over spline-
  #              suggested windows.

### Begin function
GenWin <- function(Fst,map,smoothness,s2=NA,mean=NA,plotFst=FALSE,
                   plotWindows=FALSE, method=3){
  library(pspline)
  ### Function to find roots
  roots <- function(data){
   data1 <- c(NA,data[1:{length(data)-1}])
   data2 <- data
   posneg <- which(data1>0 & data2<0) - 0.5
   negpos <- which(data1<0 & data2>0) -0.5
   zero <- which(data == 0)
   roots <- sort(c(posneg,negpos,zero))
   return(roots)
  }
  #set up data sets
  rawData <- data.frame(Pos=map,Fst=Fst)
  data <- rawData[which(is.na(rawData$Fst)==F),]
  #compute spline and derivatives
  pspline <- smooth.Pspline(data[,1],data[,2],norder=2,method=method)
```

```
  predict <- predict(pspline,seq(0,max(pspline$x),by=smoothness))
  psplinederiv <- predict(pspline,seq(0,max(pspline$x),by=smoothness),
                          nderiv=2)
  psplineInflection <- roots(psplinederiv)*smoothness
  # Print number of windows
  print(paste("Total number of windows = ",
              length(psplineInflection) + 1))
  #create window table
  print(" ---- Computing window statistics ----")
  if(is.na(s2)) s2 <- var(Fst)
  if(is.na(mean)) mean <- mean(Fst,na.rm=T)
  cat(1, "of", length(psplineInflection)+1, "\r") # print progress
  Distinct <- data.frame(WindowStart=rep(NA,
                         length(psplineInflection)+1),WindowStop=NA,
                         SNPcount=NA,MeanFst=NA, Wstat=NA)
  Distinct$WindowStart[1] <- 0
  Distinct$WindowStop[1] <- psplineInflection[1]
  Distinct$SNPcount[1] <- length(which(data[,1] <=
                                 psplineInflection[1]))
  Distinct$MeanFst[1] <- mean(data[which(data[,1] <=
                                   psplineInflection[1]),2],na.rm=T)
  Distinct$Wstat[1] <- {mean(data[which(data[,1] <=
                           psplineInflection[1]),2],na.rm=T)-mean}/
                           sqrt(s2/length(which(data[,1] <=
psplineInflection[1])))
  for(i in 2:length(psplineInflection)){
    cat(i, "of", length(psplineInflection)+1, "\r")
    Distinct$WindowStart[i] <- psplineInflection[i-1]
    Distinct$WindowStop[i] <- psplineInflection[i]
    Distinct$SNPcount[i] <- length(which(data[,1] >=
                                   psplineInflection[i-1] & data[,1]
                                   <= psplineInflection[i]))
    Distinct$MeanFst[i] <- mean(data[which(data[,1] >=
                                 psplineInflection[i-1] & data[,1]
                                 <= psplineInflection[i]),2],na.rm=T)
    Distinct$Wstat[i] <- {mean(data[which(data[,1] >=
                               psplineInflection[i-1] & data[,1]
                               <= psplineInflection[i]),2],na.rm=T) -
                               mean}/ sqrt(s2/length(which(data[,1]
                               >= psplineInflection[i-1] & data[,1]
                               <= psplineInflection[i])))
  }
  ### Fill out final window
  i <- i+1
  cat(i, "of", length(psplineInflection)+1, "\r")
  Distinct$WindowStart[i] <- psplineInflection[i-1]
  Distinct$WindowStop[i] <- max(Pos)
  Distinct$SNPcount[i] <- length(which(data[,1] >=
                                 psplineInflection[i-1] ))
  Distinct$MeanFst[i] <- mean(data[which(data[,1] >=
                              psplineInflection[i-1]),2],na.rm=T)
  Distinct$Wstat[i] <- {mean(data[which(data[,1] >=
```

```
                        psplineInflection[i-1]),2],na.rm=T) - mean}/
                        sqrt(s2/length(which(data[,1] >=
                        psplineInflection[i-1])))

  print(" ---- done ---- ")
  #make plots if requested
  if(plotFst==T & plotWindows==T) par(mfrow=c(2,1))
  if(plotFst==T){
    plot(data,xlab="Position (bp)",ylab="Fst")
    lines(seq(0,max(pspline$x),by=smoothness),predict,col="red")
  }
  if(plotWindows==T){
    plot((Distinct$WindowStop-Distinct$WindowStart)/2+
          Distinct$WindowStart,Distinct$Wstat,xlab="Position (bp)",
          ylab="Spline Wstat",pch=19)
  }
  return(list(rawSpline=pspline,breaks=psplineInflection,
              distinctFst=Distinct))
}
```

APPENDIX C

A list of the 72 chicken breeds included in this study and the number of individuals genotyped from each breed.

Breed	Sample Size
Albanische Kräher	19
Altenglische Kämpfer	19
Antwerpener Bartzwerge wachtelfarbig	20
Appenzeller Spitzhaube silber-schwarzgetupft	18
Araucanas wildfarbig	20
Asil rotbunt	19
baier	18
Bantam schwarz	20
Bergische Kräher schwarz-goldbraungedobbelt	29
Bergische Schlotterkämme schwarz	18
Brahma rebhuhnfarbig-gebändert	20
Brahma weiß-schwarzcolumbia	20
Broiler dam line B	18
Broiler sire line B	19
Chabo gelb mit schwarzem Schwanz	20
Chabo schwarz mit weißen Tupfen	27
chahua	19
Cochin schwarz	20
Denizli	18
Deutsche Lachshühner lachsfarbig	19
Deutsche Sperber gesperbert	19
Deutsche Zwerghühner goldhalsig	20
Federfüßige Zwerghühner gold-porzellanfarbig	20
Federfüßige Zwerghühner schwarz	18
Gallus Gallus Gallus	18
Italiener rebhuhnhalsig	19
Italiener schwarz	19
Kastilianer schwarz	18
Ko Shamo gold-weizenfarbig	18
Krüper schwarz	19
Kuchi	20
Lakenfelder	20
langshan	20
Leghorn weiß	20
Malaien gold-weizenfarbig	19
Marans schwarz-kupfer	20

Minorka schwarz	20
Morogoro Medium	19
New Hampshire goldbraun	19
Ohiki goldhalsig	18
Ohiki silberhalsig	20
Orloff rotbunt	19
Orpington gelb	20
Ostfriesische Möwen silber-schwarzgeflockt	20
Paduaner	21
Phönix	21
Plymouth Rock gestreift	19
Prat-Hühner	18
Rheinländer rebhuhnhalsig	20
Rheinländer schwarz	20
Rhodeländer dunkelrot	20
Sebright gold-schwarzgesäumt	19
Sebright silber-schwarzgesäumt	20
Seidenhühner weiß	19
SH	19
Shamo schwarz	19
SP	26
Sumatra schwarz	19
Sundheimer weiß-schwarzcolumbia	20
Tau Vang	19
Totenko goldhalsig	21
Vorwerk	20
Vorwerk Erhaltung	20
Wannan Three-yellow	19
Westfälische Totleger silber	19
wugu	20
Wyandotten silber-schwarzgesäumt	20
Wyandotten weiß	19
xiaoshan	19
Yokohama weiß-rotgezeichnet	20
Zwerg-Cochin schwarz	20
Zwerg-Cochin weiß	19

APPENDIX D

Information describing groups of neighboring genes that were both predicted to be members of the same pathway according to KEGG (KANEHISA *et al.* 2014), and which displayed locus pairs with D'^{2}_{IS} values that indicate selection putatively took place. Because genes are often annotated as belonging to multiple pathways, the same region or similar regions are often depicted in duplicate rows of this table, corresponding to the different pathways that may be involved.

Pathway	Chr	Group_start	Group_end	Genes_contained	Number_Locus_Pairs
path:gga04010	1	50264535	50594411	4	5
path:gga04910	1	67195294	67359692	2	1
path:gga04540	1	67337860	68044295	2	22
path:gga04912	1	67337860	68044295	2	22
path:gga00010	1	76434647	76776197	3	2
path:gga01200	1	76434647	76776197	3	2
path:gga01230	1	76434647	76776197	3	2
path:gga04310	1	79451031	79960263	2	1
path:gga04514	1	90875216	91642963	3	5
path:gga00260	1	1.09E+08	1.09E+08	2	1
path:gga00270	1	1.09E+08	1.09E+08	2	1
path:gga01230	1	1.09E+08	1.09E+08	2	1
path:gga03013	1	1.36E+08	1.36E+08	2	1
path:gga04110	3	3037097	3122490	2	3
path:gga04114	3	3037097	3122490	2	3
path:gga04914	3	3037097	3122490	2	3
path:gga04080	3	44679918	44839972	2	1
path:gga04916	3	1.05E+08	1.06E+08	3	1
path:gga04080	4	10622204	11126899	3	2
path:gga04120	4	16354478	16512992	2	1
path:gga04540	4	20224667	20768790	2	1
path:gga04080	4	20823798	21376386	3	10
path:gga04115	4	34042665	34467515	2	1
path:gga04120	4	34042665	34542198	2	1
path:gga04620	4	60395342	60939133	2	1
path:gga05164	4	60395342	60939133	2	1
path:gga05168	4	60395342	60939133	2	1
path:gga00230	4	64341542	64393109	2	2
path:gga04060	6	18559508	18812561	2	1
path:gga01200	7	2653627	2886932	2	1
path:gga01230	7	2653627	2886932	2	1

path:gga04810	7	13314358	13852740	2	13
path:gga04622	7	20543411	20975797	2	17
path:gga04145	7	21327981	21707280	2	8
path:gga04270	7	26451247	26859021	2	2
path:gga00230	8	26764668	27426443	3	6
path:gga04630	8	27103615	27564574	4	6
path:gga04060	8	27238562	27564574	3	2
path:gga04150	9	4894537	5175814	2	53
path:gga04120	9	9533376	9749897	2	4
path:gga03013	9	15087461	15678579	4	2
path:gga04120	10	2999037	3528542	2	1
path:gga03018	11	907974	1453089	2	1
path:gga04080	13	11646223	12648860	4	2
path:gga04141	13	15039101	15237378	3	1
path:gga04810	14	4169150	4366882	3	4
path:gga05132	14	4169150	4366882	3	4
path:gga05164	14	4169150	4502734	3	7
path:gga02010	14	7628434	7703987	2	1
path:gga04145	15	4508303	4924155	2	6
path:gga04142	15	4909039	5219304	2	10
path:gga04020	15	5297891	5584055	4	18
path:gga04110	17	1475849	1792711	2	1
path:gga04114	17	1475849	1792711	2	1
path:gga04120	17	1475849	1792711	2	1
path:gga04914	17	1475849	1792711	2	1
path:gga04080	20	8120601	8620192	2	1
path:gga04080	22	827828	1253785	2	4
path:gga04060	24	5424232	5595172	2	1
path:gga03018	24	5556681	5587498	2	1
path:gga00020	24	6167375	6188311	2	3
path:gga01200	24	6167375	6188311	2	3
path:gga04916	26	3436365	3854232	2	1
path:gga00600	28	99392	549556	2	2
path:gga03010	28	676818	888920	2	1
path:gga00190	28	794175	890297	2	1
path:gga04144	28	832292	1032575	2	1