Insights Into Evolution and Adaptation Using Computational Methods and Next Generation Sequencing

BY Alexander G. Shanku

A dissertation submitted to the

Graduate School—New Brunswick

Rutgers, The State University of New Jersey

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

Graduate Program in Computational Biology and Molecular Biophysics

Written under the direction of

Andrew D. Kern

and approved by

New Brunswick, New Jersey

May, 2016

ABSTRACT OF THE DISSERTATION

Insights Into Evolution and Adaptation Using Computational Methods and Next Generation Sequencing

by Alexander G. Shanku

Dissertation Director: Andrew D. Kern

Historically, much of the research in evolutionary biology and population genetics has involved analysis at the level of either a single locus or a few number thereof. However, "Next Generation" sequencing technology has opened the floodgates with respect to both the sheer volume and quality of sequence data that researchers have long needed to address and answer long-standing questions in their fields. Scientists are now, by and large, no longer hampered in their efforts by technological hurdles to obtain data, but are in fact facing the problem of how best to use the vast amount of data that are accumulating at an ever-increasing rate. This is a good problem to have.

The following research described in this dissertation is an attempt to derive answers to questions in the fields of population genetics and evolutionary biology that, until recently, have been either intractable or, at best, extremely

difficult to address. In the first chapter I provide an introduction and a brief historical look at the research efforts that have proceeded my own.

In the second chapter I describe how modern sequencing methods and computational analysis can be used to study, analyze, and answer evolutionary questions about the non-model organism, *Enallagma hageni*, in order to 1) determine this organism's phylogenetic position within Arthropoda, 2) provide answers and insight into the evolutionary history of the protein-encoding genes in the *Enallagma* transcriptome, and 3) give functional annotation to these expressed proteins.

In the third chapter I examine how natural selection acts on the genome and derive a method that can accurately determine the evolutionary cause of nucleotide fixations, having occurred either through positive selection or neutral processes. I then apply the methodology to North American populations of *Drosophila melanogaster*, providing further evidence as to how adaptive evolution proceeds in a newly established population. This is an important question, for though there have been multiple approaches devised to determine the targets and modes of evolution in the genome, to date there has not emerged a definitive method which can determine both the location and type of a selective process, and as a result, the picture of how and where adaptive evolution proceeds in the genome has remained opaque.

In the forth chapter I examine how levels of natural selection within the genome have the potential to inhibit the ability to accurately learn population demographic history. Using a number of modern algorithms and extensive simulations, I first examine whether or not demographic histories that are learned under simple biological assumptions will yield accurate results when the actual data itself does not adhere to these assumptions. Further, I go on to examine more complicated models of demographic history, looking specifically at how positive selection biases inference, which directions these biases occur, and at what levels of selection do inference methods fail to be robust. Finally, I describe potential evolutionary scenarios where these inference methods may be more prone to fail, as well as methods which might mitigate positive selection's effects, thus allowing for more accurate histories to be inferred.

The work contained in this dissertation, at the broadest scale, is an effort to marry state-of-the-art techniques in statistics, computer science, and machine learning algorithms to the technological advances of next generation sequencing; the potent combination of these technologies has provided a means with which to derive answers to multiple, long-standing questions in population genetics and evolutionary biology.

Preface

> The time will come when diligent research over long periods will bring to light things which now lie hidden. A single lifetime, even though entirely devoted to the sky, would not be enough for the investigation of so vast a subject... And so this knowledge will be unfolded only through long successive ages. There will come a time when our descendants will be amazed that we did not know things that are so plain to them... Many discoveries are reserved for ages still to come, when memory of us will have been effaced.
>
> — Seneca, *Natural Questions*

Portions of this dissertation are based on work previously published or submitted for publication by the author [1].

Acknowledgements

This dissertation is based upon the research undertaken as a Ph.D. student while attending Dartmouth College and Rutgers University from 2010 through 2016. These five and a half years have been divided between my research, the classroom, and ultimately, the creation of this document. However, none of these things would have been possible without the help, guidance, and assistance of the many other people that I have had the good fortune of meeting along my academic path.

Accordingly, I must first acknowledge and express my gratitude to my advisor and mentor, Andrew D. Kern. Our first meeting was over a pint while we talked science, and I'm fortunate to say that five years later our meetings often take the same form. The scope of what I've learned under your direction is immense, and for that alone I am grateful. Your love for science, and inquiry in general, permeated the lab and was motivating and inspiring. Your patience and understanding never went unnoticed. I am also happy to say that in addition to the academic relationship we have cultivated, I also am able to call you a friend. Andy, thank you so much for everything.

I would like to thank my Dissertation Committee: Jinchuan Xing, Isaac Edery, and Kevin Chen - your guidance, suggestions, and the wealth of your combined experiences is something that I will benefit from my entire career.

Having the chance to interact with you, be it in journal club, the classroom, or as part of a collaboration has been an extremely valuable experience and I thank each one of you for it.

I have been fortunate to work along side a remarkable colleague and friend, Daniel Schrider. Dan, I certainly owe you quite a bit. You've taught me a great deal and shared so much of what you know, and for that I must thank you. Your energy and passion for science are truly motivating. Along with A.D.K, we've made a point to keep mind and body strong - shouting "ALL MEAT MAN!!" will now only bring stares.

I would certainly not have made it this far, let alone submit this dissertation, if I weren't so fortunate to have known the Associate Director of my graduate program - a confidant, motivator, and friend Gail Ferstandig Arnold. Thank you so much, Gail. I am so appreciative of everything you've done for me.

I couldn't leave the 4th floor without thanking my friend Kathleen McDonald and recalling the many hours we've spent talking in your office. Being able to vent about the "RUt" to each other and about the daily happenings in Nelson was often needed and very cathartic. You certainly helped to keep my spirits up. So, Kathleen, I say thank you for all of your help, your friendship, and great conversations.

This dissertation starts in Tennessee and finishes in New Jersey, with stops in New Hampshire and Brooklyn along the way. During those transitions, serendipity led to a close relationship with the Cline and Cutraro families. Clayton, Allison, Andy, and Erin, your kindness and support mean the world to us. Watching our families grow together during these past five years has been very special.

To the hardest Devil Dogs around - Nathaniel E. Allen, Patrick J. Belmonte, and Craig H. Strohl - Semper Fidelis, and *REPORT*!! I'm especially fortunate for the time I've spent with you, Nat, over the past twenty years. You've kept me engaged with the world, and I treasure every one of our late night conversations, debates, and arguments.

Not everyone is a fortunate as I am to have such wonderful in-laws. Your support, care, and love will always be remembered during this time. Kathy and David Crosslin, and Hal and Kimberly Danson - thank you so very much.

In the same light, I am grateful for my whole family's support and here must specifically acknowledge my sister and brother-in-law, Jennie and Patrick Osman. Your love and support have been crucial, your encouragement has always kept my spirits up, and knowing that you have always been in my corner means a great deal. Thank you both so much. I am excited for the chance to live close to you both again, or for the first time!

My mother, Carol Shanku - your faith in me never faltered, even during those times when my own did. You have given selflessly to me for so long and have supported me in all of my endeavors. I am forever grateful. I could not ask for a better mother, mother-in-law to my wife, and grandmother to my son. I love you very much.

Benjamin, my son, maybe one day you will read this and discover what I was working on the first 5 years of your life. Maybe not though, it's a long read. I am lucky that whenever I will think back about my Ph.D., I'll be thinking about you, too. I love you, I am very proud of you, and I am excited for our next adventure.

Most importantly, I have to acknowledge my wife, Bethany. This has certainly been some adventure. Beth, we've been side by side this entire time, and through all the ups and downs, your love, kindness, and patience - immeasurable as it is - have never wavered. Your endless support has meant so much to me. I could never have made it without you. I love you very much and am so lucky to share my life with you.

Dedication

If we are lucky, we may enjoy the privilege of being surrounded by those whose lives enrich our own. It, too, is a privilege having known those whose memories alone still enrich us and occupy a permanent place in our thoughts.

With happy recollections, love, and admiration, I'm proud to dedicate this dissertation to my step-father and to my father,

Edward J Overstreet
(1938-2002)
George D Shanku
(1943-1993).

Table of Contents

List of Tables

List of Figures

1

Introduction

A.G. SHANKU

> *"Thus, from the war of nature, from famine and death, the most exalted object which we are capable of conceiving, namely, the production of the higher animals, directly follows. There is grandeur in this view of life, with its several powers, having been originally breathed into a few forms or into one; and that, whilst this planet has gone cycling on according to the fixed law of gravity, from so simple a beginning endless forms most beautiful and most wonderful have been, and are being, evolved."*
>
> – Charles Darwin, *The Origin of Species*

From the time that Anaximander first put forth the idea that all animals had their beginnings in the ancient oceans of the planet, mankind has sought to make sense of the natural world and his place therein [7]. In the nearly 2500 years that have passed since Anaximander's time, we find that it is only within the past 150 years that we have made meaningful headway in the answering of these ancient questions. Further, it is within these most recent 150 years that the fields of population genetics and evolutionary biology were born, and theories and observations that would upend the way the modern world viewed all life around it were produced and documented. Moving forward in time, it was within the past one hundred years that the “modern synthesis” began, whereby a host of contemporary biological ideas that had been brewing coalesced into a prevailing framework, able now to examine and predict how life evolves over time [8, and references therein]. Moreover, it was only within the past sixty years that we have had a grasp of what the hereditary material in nature actually was, and that this deoxyribonucleic acid (DNA) that houses the genetic building blocks was formally described [9]. Astonishingly, it was only within the past thirty years that a small number of loci could be sequenced and analyzed. In the last twenty years, however, amazing technological advancements have put us in a position where the combination of next generation sequencing (NGS) can be married to high-powered computing in order to generate data, perform analysis, test hypotheses, and answer questions previously beyond our grasp.

It is now obvious as to how much of an impact next generation sequencing (NGS) has had on the fields of genetics, genomics, and essentially all other areas of study associated therewith. In recent years, next generation based studies have become ubiquitous in science and medicine. Huge amounts of sequence data are being generated, necessitating the need for, and resulting in the development of novel analytical techniques that when paired with computational advances allow researchers a path to derive meaningful understanding of these data. Here, in Chapter 1, I provide a background in the advancements in DNA sequencing technologies, followed by a description of the history of transcriptome analysis in both model and non-model species using next generation sequencing. I go on to detail methods and techniques that have been developed to decipher both where in the genome evolution is occurring and by what route this evolution occurs, and finally, I provide insight on the methods that have been developed and used to make inference on population demographic history.

In the three chapters that follow this introduction I will present my research, composed of a three large projects that seek to utilize the benefits of next generation sequencing and employ novel computational and analytical techniques to address the following distinct questions in evolutionary biology: 1) how can we use next-generation transcriptome sequencing as a means to explore the evolutionary history of a non-model organism, *Enallagma hageni*, 2) devise a framework that incorporates state-of-the-art machine learning techniques in order to determine and describe how natural selection acts on the genome, and 3) to determine how the long reaching dynamics of evolutionary processes at the level of the genome play a roll in, and affect, the ability to make accurate inference about population demographic histories.

1.1 DNA Sequencing

The techniques which first enabled DNA sequence determination have their roots in the early 1970's when Frederick Sanger, a two-time Nobel Prize winner and the forefather of modern sequencing methods, and Alan Coulson developed a DNA polymerase-based sequencing method [10]. Though time consuming and laborious, this "plus and minus" technique was shown to be successful; Sanger and his group sequenced the first full genome in 1977 [11]. This single-stranded genome, bacteriophage ϕX174, consisted of 5375 nucleotides and amongst other notable discoveries, Sanger noted that “Two pairs of genes are coded by the same region of DNA using different reading frames.”, showing for the first time that coding regions of multiple genes may in fact overlap.

Sanger continued to modify his sequencing techniques, later that year developing what is now commonly referred to as “Sanger” sequencing [12]. This technique, the “dideoxy chain termination method”, improved accuracy and read lengths over both the “plus” and “minus” methods, and earned Sanger his aforementioned second Nobel prize in 1980, an award shared along with chemists Walter Gilbert and Paul Berg. Here, a complementary DNA (cDNA) template is generated via natural 2′-deoxy-nucleotides (dNTPs) and terminated via using 2′,3′- dideoxynucleotides (ddNTPs) by DNA polymerase [12, 13]. These resultant fragments are then separated and sorted via gel electrophoresis and further analyzed to determine the genetic sequence.

This technique has been further refined, wherein the ddNTP or primer is labeled with a fluorescent dye, and is known as automated Sanger sequencing. These tagged fragments emit light when hit by a laser and each of these emitted colors correspond specifically with one of the four nucleotide bases in question, allowing for sequence identification (For a full treatment on advances in Sanger-based sequencing methods, the reader is directed to [13]).

1.1.1 Shotgun Sequencing

While Sanger's method was effective in handling reads upwards of 1000bp, it was desired that longer sequences be analyzed. Shotgun sequencing extends the usability of Sanger's chain termination method to much longer sequences. Here, the sequence of interest is randomly fragmented, separated by size, and then cloned in a phage vector [14]. These clones are then sequenced using the Sanger technique to produce "reads". Overlapping ends in multiple reads must then be found, the matching reads joined, thus effectively reassembling the original sequence.

At this time, too, consideration was being given to how the increase in computing power could best be utilized to handle the influx of sequenced regions and the increasing size of these regions. The need for computer programs which would allow analysis of these long sequences led to new software being developed, effectively the precursors to modern genome assembly software. Roger Staden succeeded in producing a series of `FORTRAN` programs that could handle sequences upwards of 20Kb, being able to determine overlaps between sequenced reads (`OVRLAP`), join reads meeting some user-specified criteria (`XMATCH`), and a general sequence file handler (`FILINS`) [15].

Continuing into the early 1980's, shotgun sequencing that employed the chain termination method was used to infer the full sequence of human mitochondrial DNA (mtDNA) [16], and Sanger, along with Coulson and colleagues, sequenced the nearly 50,000 base pair (bp) double-stranded DNA *Enterobacteria phage* λ [17].

1.1.2 Whole-Genome Shotgun Sequencing

Over the subsequent decade technology continued to improve. In 1995, the 1.8 mega base (Mb) genome of *Haemophilus influenzae* Rd was sequenced in an effort led by Craig Venter and Hamilton Smith [18]. This marked the first occurrence that a free-living organism's genome had been sequenced, but perhaps more importantly, this project demonstrated that the shotgun approach to whole genome sequencing was not only a valid tool, but as such, a huge advancement in genome analysis.

Whole genome shotgun sequencing using Sanger's chain termination technique continued to be the de facto sequencing method for the following decade. Another milestone was reached in the year 2000 when it was announced that a team had sequenced and assembled the *Drosophila melanogaster* genome [19]. This was of crucial importance to biology in general, and genetics and evolutionary biology in particular, seeing to this eukaryote's rich 100+ year history as one of the most studied model organisms [20]. This achievement showcased the height to which sequencing technology had risen and marked the first animal genome sequenced using shotgun sequencing.

Quickly following the release of the *Drosophila melanogaster* genome was the publication of the most ambitious project to date that utilized the shotgun sequencing approach: the Human Genome Project (HGP). Proposed in 1985 [21], and initially completed in 2001 [22], this project was one of the most import achievements of the decade as well as the largest scientific collaboration to date [23].

While Sanger sequencing, given its advancements and improvements at the time, was able to produce a finished-grade human genome, it would soon effectively be replaced, especially in whole genome sequencing endeavors, by more modern techniques which were rapidly becoming available. Moreover, it was partially the very difficulties encountered in Humane Genome Project that drove the development of more advanced sequencing technology. Thus, researchers entered the "Next Generation" of sequencing with technology that aimed to usurp Sanger methods in terms of its massively parallel analysis framework, high throughput capabilities, and extremely reduced cost.

1.1.3 Next Generation Sequencing

The Sanger method of sequencing genomes had certainly been shown to be effective throughout the 1990's and early 2000's (Sections 1.1.1 and 1.1.2), however a fundamental shift in sequencing strategies began taking place around 2005. Prior to this period, all Sanger methods can be referred to as "first-generation" sequencing technology, and conversely, all sequencing technologies after this period not using Sanger techniques are called "Next Generation" sequencing (NGS). While no single technical advancement delineates a sequencing protocol being called NGS, the major differences between the older Sanger sequencing paradigm and that of NGS technology is the latter's ability to produce an enormous amount of short-read (Illumina and SOLiD) or long-read (454 and PacBio) data in parallel, at a relatively low cost, without the requirement of plasmid cloning during sequencing [24]. It is further worth noting that the PacBio sequencing technology differs in two main respects from the other NGS sequencers in that it does not require PCR before sequencing and its signal is captured in real time, leading to it occasionally being described as "Third Generation Sequencing" [25].

In the four following sections my aim is to introduce and highlight those next generation sequencing technologies that were the first to follow the Sanger era. Currently, three of these platforms, PacBio, Illumina and AB SOLiD, still remain the most used sequencing technology to this day. At the time of their introduction, all four of these technologies sought to be replacements for Sanger methods in terms of time, cost, and accuracy.

454 Pyrosequencing

The first commercially available NGS sequencer, the GS20, was introduced in 2005. It was created by 454 Life Sciences, a company later acquired by Roche Diagnostics. At its core, this machine made use of a novel technique of large-scale and parallel pyrosequencing [26]. Unique to this technology, a sequence can be inferred by means of a pyrophosphate being released when a complimentary nucleotide is inserted on the template. As a solution containing only one type of nucleotide at a time is applied, then immediately removed, the pattern of luminescent signals emitted and the order with which the nucleotide solutions allows the sequence to be determined.

Currently, the GS FLX Titanium XL+ machine offered by Roche Diagnostics is capable of producing reads of 1000bp, 1×10^6 shotgun reads per run, all at 99.997% consensus accuracy in less than 24 hours. This is likely the last sequencer utilizing this technology that Roche will be producing, as they announced the end of their 454 Life Sciences program in 2013 and discontinuation of product support in mid-2016 [27].

Illumina

Genome Analyzer was marketed by the Solexa company and appeared in 2006, with the capability to sequence 1 gigabase (Gb) in one single run [28]. Solexa was subsequently acquired by Illumina in 2007 and by 2008 they had sequenced a human genome to greater than 30x coverage with 25bp, paired-end reads [29].

Illumina sequencing technology centers around “reversible terminator chemistry”, a method whereby four reversible terminators are tagged with a different removable fluorophore. After each base addition the fluorophore is excited by a laser and the nucleotide type is then recorded via a camera image. Immediately thereafter, the tag is removed and another round of nucleotide addition is performed.

In January of 2014, Illumina released a machine capable of sequencing 18,000 complete human genomes per year, to 30x coverage, for $1000 per genome [30]. This marks the first time in the evolution of sequencing science and development that a $1000 human genome has been made possible, and an absolutely astonishing departure from the $3 billion spent on the Human Genome Project just 13 years hence [22].

AB SOLiD

The last of the original big three NGS technologies was an instrument released in the later part of 2007 by Applied Biosystems. This platform was named “SOLiD” (Sequencing by Oligo Ligation Detection), partially due to its unique method of using oligo adaptor-linked DNA and its utilization of DNA ligase in lieu of polymerase [25, 31, 32]. In this system, amplification is performed by emulsion PCR that is then followed by the ligation-based sequencing. Here, a primer is joined to the adapter thereby facilitating ligation to occur between one of a set of 8mer oligonucleotides that successfully binds to the DNA fragment via ligase. Each base on the template is eventually identifiable due to there being a specific fluorescent tag bound to each of the oligonucleotides in the set.

In the first cycle of sequencing, the 5th base of the 8mer oligonucleotide corresponds to the base that is being identified. This is recorded and the 6,7, and 8 positions of the 8mer are subsequently removed. A new ligation occurs at the end of the cleaved 8mer and a new 8mer - again the 5th base of the 8mer corresponding to the base being identified (spatially the 10th base of the template). This process continues 3 or 5 more times, at an interval that identifies every fifth nucleotide. After the 25th base or 35th base is identified, the newly ligated 8mers are removed and the process starts again, but with the primer being shifted n-1 positions in the 3′ direction of the adapter. A second cycle will therefore identify bases 4, 9, 14, 19, 24, and potentially 29 and 34. A visual graphic of this process may be found at the Applied Biosystems website.

PacBio

Pacific Biosciences brought a sequencer to the fold in 2010 that employed a single molecule real time sequencing (SMRT) paradigm. SMRT makes use of a technique called zero mode waveguide (ZMW). Optical waveguides are used to direct and manipulate electromagnetic waves, a common example being fiber optic cable. What makes the PacBio RS II sequencer unique is that their ZMW framework and XL chemistry kit allows for an average sequence read length of greater than 500bp. Reads of this length are especially useful when non-model organisms are being sequenced, resulting in *de novo* genome assembly [33].

While relatively new technology, researchers have already shown that SMRT sequencing can be used in ways not possible with other sequencing technologies, such as applications in fast and accurate microbial genome sequencing [34], looking for methylation of nucleotides [35], and even sequencing so-called whole “methylomes” of bacteria, that is the whole methylation pattern across the genome [36].

1.2 Transcriptome Analysis

While Section 1.1 has so far described the progression of DNA specific sequencing technologies, another important aspect of next generation sequencing has been the amazing leap forward in the ability to sequence RNA and examine expression levels, even expression levels of all RNA transcripts present in a cell, through the use of RNA sequencing (RNA-seq) [37]. Prior to recent NGS methods - roughly the mid 1990's until the mid 2000's - the main tool available to probe RNA content and expression levels was the microarray.

1.2.1 Microarrays

The driving mechanism behind the microarray is the property of hybridization, whereby a single stranded nucleotide sequence binds to its complementary strand to form a double stranded sequence. Microarrays have their beginnings in DNA mapping [38] and hybridization sequencing studies [39]. In most microarray studies, the goal is to not just identify the DNA or RNA present, but to determine the expression level of many hundreds or thousands of genes in a single experiment. In order to do so, microarray technology has led to "chips" where tens of thousands or even hundreds of thousands of different oligonucleotides are mounted using solid phase and semiconductor techniques. These oligonucleotides are present in millions of copies on each chip, and represent the complimentary strands to the sequences one wishes to probe (i.e. genes). When the target RNA is introduced to the chip, those oligonucleotide probes hybridize with those sequences present in the target sample. By staining with a fluorescent molecule, the intensity of the resultant emitted light can be measured and as the location of each group of probes are known, not only the sequences themselves that are present in the sample, but also the amount of target sequence in the sample can be inferred [40].

The microarray has been used in many studies, both across academia and in the clinical setting, for finding sets of genes with expression level differences (e.g. breast cancer [41], prostate cancer [42], and response to cancer drug therapies [43]). But the microarray is not limited to just to the study expression levels. Recent studies have shown that many desirable biological phenomena can be analyzed such as copy number variation [44] and single nucleotide polymorphism (SNP) detection [45].

There are, however, drawbacks to the use of the microarray method. Specifically, if one were interested in examining the total RNA content present, then the reference sequence of that species must be known a priori. Further, there exists the potential for the target sample and the probes to cross-hybridize or mis-hybridize, resulting in high background signals and potentially leading to spurious results [46, 47]. Additionally, it is problematic to compare across studies due to normalizing methods involved [37]. By harnessing the power of next generation sequencing, however, there now exist techniques that can avoid these aforementioned pitfalls.

1.2.2 RNA Sequencing (RNA-seq)

At the heart of modern transcriptome analysis is a technology called RNA sequencing (RNA-Seq). This process is made practical, if not possible, via next generation sequencing, that in turn allows the experimenter to obtain the profile of all RNA present in a given tissue or sample at the moment when that sample is isolated. In RNA-seq, isolated RNA is converted to cDNA fragments that are then ligated with adapters. Sequencing proceeds as described in Section 1.1.3, thereby producing a collection of reads that may vary in length from a few tens of nucleotides to a few hundred nucleotides (see Section 1.1.3).

Once sequencing is complete, the bioinformatic analysis is dependent on whether there exists a reference genome or reference transcripts for the sampled organism. In the case where a reference exists, reads can be mapped and aligned to said reference. If a reference does not exist, as in the case of many non-model organisms, analysis proceeds differently. In this case, de novo assembly must be performed.

1.2.3 De Novo Transcriptome Assembly

Specifically, assembly in this context describes the process that seeks to use a hierarchical structure to convert the sequenced data back into the original target form. Assembling a transcriptome in this manner begins by grouping sequencing reads that overlap into structures called contigs, followed by the grouping of overlapping contigs into scaffolds. The structure of these scaffolds may take the form of simple path, or a more complex topology such as a network [48].

There are generally two classes of algorithms that may be used during assembly: the overlap/layout/consensus (OLS) approach or the de Bruijn Graph (DBG) approach [49, 50, 51]. The choice of which assembler to use is generally determined by the read length of the sequenced data. For long reads greater than 50bp, the usual choice would be to use OLS, whereas short reads under 50bp are better assembled via DBG methods.

The OLS makes use of an overlap graph consisting of three steps [48, 52]:

1. Overlap - both the forward and reverse orientations of every read is compared in an all-against-all approach.
2. Layout - once the overlap graph is constructed, the aforementioned contigs and scaffolds are determined. In this state, all bases are not represented, allowing large graphs to be computationally manageable.
3. Consensus - here the consensus sequence is determined using all reads in each scaffold

The de Bruijn Graph method has some attractive features in that it doesn't necessitate an all-vs-all comparison like the OLS method and that it better handles redundant nucleotides in the sequence. However, DBGs make use of the real sequences, potentially leading to computational memory issues [48, 51]. The DBG method makes use of k-mers ranging from 25-50bp and a real time hash table lookup to align these k-mers. The consensus sequence is then determined by finding the Eulerian path through the k-mer graph.

Beyond assembly, it is often desirable to functionally annotate the assembled contigs to determine the biological roll the putative proteins take. A common method to achieve this is to use a Basic Local Alignment Search Tool (BLAST) [53] based approach to compare the contigs to a curated database of sequences whose biological sequences are known, such as the NCBI's *nr* (non-redundant) database (`ftp://ftp.ncbi.nlm.nih.gov/blast/db/`).

1.2.4 The Power Of Modern Transcriptomics

Since it's introduction in 2005, next generation sequencing has provided a means for researchers to explore the transcriptomes of hundreds of organisms in thousands of studies (`http://www.ncbi.nlm.nih.gov/Traces/wgs/`). In evolutionary biology, specifically, this has allowed researchers to bring forth multitudes of new findings, both in model and in non-model species. Below, I highlight example cases where new insights in areas of evolutionary biology has been garnered, specifically in convergent evolution, determining rates of gene evolution, and phylogenetic analysis.

Convergent Evolution

In studies concerned with convergent evolution, the phenomena whereby similar phenotypic traits have evolved in multiple species whose last common ancestor did not posses such a trait, transcriptome studies have been extremely valuable. For example, a recent study presents transcriptome-derived phylogenetic evidence which supports two distantly related species of cephalopods have evolved bacterial bioluminescent organs (photophores) in distinct and disparate ways [54]. Strikingly, in that same study, it was found that the gene expression patterns between the photophores themselves were also convergently evolved.

It has been proposed, at least in insects, that convergent evolution of molecular mechanisms lead to convergently evolved traits [55, 56]. Berens et al. (2015) utilized a comparative transcriptome approach to examine eusociality in three species of Hymenoptera [57]. The authors demonstrate that there exist significant overlap in the pathways and functions associated with insect castes, and reaffirm that convergent social behavior phenotypes result from the modulating of networks and pathways that themselves are convergently evolved.

Rates Of Gene Evolution

The use of RNA-seq in transcriptome analysis has had a meaningful impact on studies examining rates of gene evolution, as well. Yang et al. (2012) examined how the schizothoracine fish *Gymnodiptychus pachycheilus*, a species living 4,500m above sea level in the Tibetan Plateau, evolved under such highland conditions [58]. Using zebrafish to determine one-to-one orthologs, they found 350 genes had been lost, while 41 had been gained since their divergence. Using branch models and likelihood ratio tests, they compared their genes with orthologs found in four other fish species and found evidence for accelerated rates of evolution genome wide. Interestingly, they present evidence that nine of their candidate genes evolving at fast rates are directly associated with hypoxia and energy response.

Looking at sex-biased genes in *Drosophila melanogaster* and *Drosophila pseudoobscura*, Assis et al. (2012) examined, amongst other things, whether male-biased or female-biased genes were evolving at rates different from each other as well as unbiased genes [59]. The ratio of nonsynonymous substitutions per non-synonymous site to synonymous substitutions per synonymous site (d_N/d_S) were calculated in the aforementioned three classes of genes. The results show a significant increase in the rate of sex-biased gene evolution relative to unbiased genes. Further, when comparing the male and female sex-biased genes to each other, they found that male genes have evolved faster than female-biased genes, and that the female-biased genes evolve at the same rate as the unbiased genes.

Phylogenetics

Transcriptome studies also have been shown to be valuable in the search for determining evolutionary relationships between species. The large amount of transcript data that is amassed during sequencing increases the ability to find orthologous genes between species, even species separated by vast amounts of evolutionary time. With this, the ability to conduct phylogenetic analysis is enhanced, as the compliment of aligned sequences between species of interest also increases. Using this information, researchers have been able to show, for instance, the ability to resolve the phylogenetic tree between ten non-model organisms (two Annelida, two Arthropoda, two Mollusca, two Nemertea, and two Porifera) [60] and five dipterans including *Drosophila melanogaster* [61]. Additionally, Brawand et al. (2011) created multiple phylogenies across ten species of mammals for each of six tissue types (brain, cerebellum, heart, kidney, liver, testis) [62]. Their results comport with the known mammalian phylogenetic tree, and interestingly, using these trees the authors speculate that changes in gene regulation may accumulate on evolutionary timescales, highlighted by their expression level similarity.

1.2.5 Conclusion

In Section 1.2.4, I have only scratched the surface in demonstrating the usefulness and power that RNA-seq brings to transcriptome analysis. However, the message is clear - the ability to obtain high quality data of whole transcriptptomes has given researchers the ability to perform analysis which heretofore were often improbable or even impossible. In Chapter 2 I make use of a number of the scientific techniques described in Section 1.2 and examine the transcriptome of the damselfly, *Enallagma hageni*. The power of RNA-seq, and the newfound ability to study non-model organisms, has provided a platform with which I determine the phylogenetic history of this damselfly and its place within Arthropoda, determine the rates at which its coding genome is evolving, and provide a detailed functional annotation of the organism's expressed proteins.

1.3 Evolutionary Adaptation and Positive Selection

Natural selection, the phenomena first put forth by Charles Darwin in his seminal work, *On The Origin of Species* [63], describes the mechanism by which physical traits that improve an organisms fitness, or reproductive success, are more likely to be passed on to offspring, and as such, increase in frequency within that population. While this selection for beneficial traits takes place on the phenotypic level, the evolutionary process proceeds at the genotypic level. Positive selection describes this adaptive process whereby one allele is favored over another and, in turn, the favored allele proceeds to increase in frequency, until ultimately reaching frequency 1.

One of the main driving forces behind biological evolution is the action of positive selection. Thus, determining those regions of the genome that are the targets of selection remains a long-standing goal in evolutionary biology and population genetics [64, 65, 66, 67, 68, 69, 70, 71].

If a new mutation is to fix in a population, that is, to reach frequency 1, there are three routes it may take: 1) the mutation may not impact fitness and may drift to fixation by chance (a neutral fixation), 2) the mutation may be beneficial and rapidly fix by the action of natural selection, a process that is sometimes known as a “hard sweep” [72], or 3) the mutation may be initially neutral, or nearly so, and drift in frequency until such a time where that mutation then becomes favorable, perhaps as the result of environmental change, and is then quickly swept to fixation (sometimes known as “soft sweeps”; [e.g. 73]).

The process of selection acting upon a favorable allele will alter the frequencies of its neighboring, linked alleles as well. Those linked neutral alleles, or even potentially slightly deleterious alleles [74, 75], that were present on the genetic background where the beneficial mutation originated will increase in frequency along with the selected allele in a process called the “hitch-hiking effect” [72]. The extent of this hitch-hiking effect is jointly dependent upon the strength of selection and the rate of recombination in the region [76].

Under both the hard and soft selective sweeps modes of evolution, levels of genetic diversity in that region will be reduced, linkage disequilibrium will be increased, and the site frequency spectrum will appear skewed [72, 77, 78]. Each of these routes to fixation should, however, leave a distinct population genetic signature in local variation surrounding the site that has fixed [79].

The search for selection in the genome has been greatly expanded by the advances in next generation sequencing technologies (see Section 1.1.3) as there now exists a way to obtain the genomes of many individuals of the same species. By building on the tools of classic population genetics and applying them to newly obtainable genomic variation data, the possibility now exists to make direct inference on those regions of the genome that have evolved under natural selection.

1.3.1 Classic Tests Aimed At Finding Selection

There are two categories of tests that exist that use summary statistics of genomic variation data to locate regions in the genome that are under natural selection. The first category utilizes patterns of linkage disequilibrium (LD), or the non-random association of alleles at different loci on the same chromosome. In order to use LD to find regions under selection, it must first be understood that in a scenario where the genome is evolving under neutral evolution, that is where the actions of positive selection are not present, a new allelic variant only reaches high frequency via the stochastic process of genetic drift [80]. This process will take considerable time, allowing recombination to break the association between the allele and its surrounding regions, resulting in a decay in LD in that genomic area [81, 82, 83]. Keying in on this, those common, neutral alleles should only have linkage disequilibrium over a short range. Conversely, an allele which is under selection will quickly rise (sweep) to fixation, preventing the effects of recombination from reducing LD. Thus, by looking at an allele's frequency in the population relative to the linkage disequilibrium surrounding it, putative regions under positive selection may be inferred.

The second category of tests that can detect signals of positive selection are those making use of the site frequency spectrum (SFS) [84, 85, 86]. The SFS is used to summarize patterns of genetic variation, specifically to describe the distribution of allele frequencies across a set of single nucleotide polymorphisms (SNP). Essentially, the SFS is a histogram where each bin represents an allele frequency and, subsequently, these bins are filled according to the number individuals in a population who posses an allele at such a frequency. It has been shown that the shape of the SFS is altered under non-neutral scenarios, such as selective sweeps, and has the effect of skewing the SFS toward an access of rare alleles [78]. By looking for skew in the SFS in regions across the genome, it may be possible to locate loci having undergone adaptive evolution [87].

1.3.2 Composite Likelihood Ratio Test

More advanced methods to detect selection began to emerge in the early 2000's that built upon the methods described in Section 1.3.1 and showed great promise in the search for selection in the genome. These tests take a parametric approach to explain observed genomic variation as having resulted from either the actions of neutral processes or those of positive selection. Kim and Stephan (2002) developed a composite likelihood ratio test (CLR) that compares the maximized composite likelihood of the observed data under the neutral null hypothesis (L_0) of a randomly mating population of constant size, where evolution proceeds without selection, to a maximized composite likelihood where the observed data is the product of a selective sweep, L_1 [88]. In the neutral case, L_0, all that is required to calculate the likelihood is the population mutation rate, θ, which is assumed known. The parameters needed to calculate L_1 are N, the population size, u, the per-site mutation rate, ρ, the recombination rate, s, the strength of selection, and the location of the selected site, here called X. However, in this test, the only free parameters are X and s; the other parameters are considered known and fixed.

The CLR test conditions on a mutation arising at frequency $1/2N$, drifting to frequency ϵ, at which point the frequency is deterministically changed to $1 - \epsilon$, and that allele subsequently drifts to fixation. The framework for the test consists of some data, D, containing n individual sequences, of chromosome length L. There are a number of observed snps, S. The number of derived alleles at each site, i, are described by y_i, for $i = 1, \ldots, L$. Thus, y can take values of $[0, \ldots, n-1]$, meaning that it can be present in anywhere from zero individuals to $n-1$ individuals. The maximum composite likelihood estimators of X and s, $\hat{X}$ and $\hat{s}$, respectively, are then found via a maximization step (Powell's Method [89] was used in Kim and Stephan (2002)) such that:

$$\left(\hat{X}, \hat{s}\right) = \underset{X,s}{\arg\max}\, L_s\left(X, s | D\right) \tag{1.1}$$

where

$$L_s\left(X, s | D\right) = P\left(D | X, s\right) = \prod_{i=1}^{L} P\left(Y_i = y_i | X, s\right) \tag{1.2}$$

and where

$$P\left(Y_i | X, s\right) \tag{1.3}$$

describes the probability of observing y derived alleles at some site in the sample. This probability is defined as

$$P\left(y\right) = \int_0^1 \binom{n}{y} p^y \left(1-p\right)^{n-y} \phi\left(p\right) dp \tag{1.4}$$

where, in the neutral hypothesis, $\phi\left(p\right)$ is given by

$$\phi_0\left(p\right) dp = \frac{\theta}{p} dp \tag{1.5}$$

as shown by [90], and where, in the selective sweep hypothesis, $\phi\left(p\right)$ is given by

$$\phi_1\left(p\right) = \begin{cases} \frac{\theta}{p} - \frac{\theta}{C}, & \text{when } 0 < p < C \\ \frac{\theta}{C}, & \text{when } 1 - C < p < 1 \end{cases} \tag{1.6}$$

as shown by [76, 88]. C is approximated to be $1 - \epsilon^{r/s}$, where the optimal ϵ was found to be $1/2Ns$ [88].

In the case where each individual is fixed for the derived allele at some site, Equation (1.4) becomes

$$P\left(y = 0\right) = 1 - \left(P\left(y = 1\right) + \ldots + P\left(y = n - 1\right)\right) \tag{1.7}$$

as shown in [88, eq. 5].

The test proceeds by determining the maximum composite likelihood of the data in the sweep model, $L_1\left(\hat{X}, \hat{s}|D\right)$, and comparing it to the composite likelihood of the neutral case, $L_0(D)$. The CLR test statistic is given by

$$\Lambda = \ln\left(\frac{L_1\left(\hat{X}, \hat{s}|D\right)}{L_0(D)}\right) \tag{1.8}$$

The null distribution of Λ can be found by replicating the CLR test on sets of simulated neutral data under a fixed θ (Kim and Stephan (2002) find 200 replicate tests to suffice). The null model can be rejected when the test statistic, Λ, is larger than the $100\left(1 - \gamma\right)$ percentile of the null distribution. A γ of 0.05 was used in [88, 91].

Variations on the CLR method of [88] have been proposed since its inception in 2002. The first was a report by Kim and Nielsen (2004), who modified the CLR with an attempt to incorporate patterns of linkage disequilibrium, along with the original test using spatial patterns of polymorphism (i.e. the SFS), to find regions under selection [92]. While this method gained very little power to detect sweeps, the authors produced a new statistic, ω, based upon the idea that a selective sweep leaves an increase in linkage disequilibrium within the regions adjacent to a fixed or selected site, but that this excess of LD does not extend across the selected site (see Appendix A.6).

Nielsen et al. (2005) continued the parametric approach by building on the original CLR method, this time avoiding the use of the standard neutral model. In this study, instead, they derive a null model from the genomic background signature of the SFS from the data itself [93]. In addition to this advancement, they apply the modified CLR test to simulated sets of data generated under varying demographic models, including a simple growth model as well as a more complex population bottleneck model. They note that the improvements over the original CLR test result in a greater robustness to non-stationary demography as well as to varying rates of recombination.

While it has been shown that the composite likelihood approach appears to perform well in some cases, there exists a number of major drawbacks to this methodology. The Nielsen et al. (2005) modified CLR test made use of the empirical background SFS as a null hypothesis and showed it to be robust under a number of non-stationary demographic histories and also to perform better than using the standard neutral model as a null hypothesis [93], however, it has been shown that this robustness to population structure and demography does not hold true in general [91, 94, 95]. Moreover, the composite likelihood test assumes both that the data being analyzed are independent and that true, neutral regions of the genomes exist. It has been shown, too, that these assumptions are violated in genomic data [e.g. 96]. As a result of these assumptions, the CLR test as described may produce spurious conclusions as to where adaptive evolution is occurring in the genome, even to the point of returning false-positives nearly 90% of the time, under certain demographic scenarios [91, 97, 98].

1.3.3 Approximate Bayes Approach

While the CLR test of Kim and Stephan (2002) and its derivatives, specifically the software package SweepFinder based on Nielsen et al. (2005), are arguably the most popular tools used to scan for sweeps, there exists another methodology that makes use of the coalescent process [99], and software that allow genealogies to be simulated under the coalescent [100], in a non-likelihood based approach to infer the genomic regions subject to selection (See Appendix C).

Simulating chromosomes under the coalescent is fast and efficient, and allows for data to be generated under multiple conditions, such as non equilibrium population demography, varying rates of recombination, and selection [101, 102, 103]. Conceptually, it might seem that an ideal approach to differentiating regions as being under selection from those evolving neutrally would be to simulate data under many models and determine the likelihood of these models given the data. Currently, these likelihood calculations are not tractable and as such, hypothesis testing under this framework has not been implemented. However, an algorithm for estimating likelihoods exists and can be used in the case where either a true likelihood function doesn't exist, or is infeasible to calculate.

The Approximate Bayes computation (ABC) method circumvents the need to calculate an explicit likelihood function through generating numerous simulations under some model and comparing these with observed data [e.g 104]. Briefly, simulations are drawn from a set of known parameter values and compared, usually using a Euclidean distance metric, to the observed data. If a simulation is deemed to be close to the observed data by some heuristic tolerance, it is accepted. These accepted simulations, and the parameters that generate them, are stored. Following this rejection-sampling step, an estimated posterior probability for each model parameter is generated (See Appendix B for a through description of the ABC methodology).

Taking advantage of this technique, a number of studies have applied ABC to genomic variation data to search for signals of selection [104, 105]. For example, Przeworski (2003) uses this method in an attempt to provide support for a candidate selective sweep as well as estimate the time, T, that the beneficial allele fixed in the population [106]. Jensen et al. (2008) used ABC in an attempt to determine the rate of weakly and strongly selected substitutions in an empirical population of *Drosophila melanogaster* and conclude that both common and weak, along with rare and strong positive selection will yield similar average levels of genome variation [107].

Most recently, Garud et al. (2015) uses ABC in an attempt to locate and differentiate between hard and soft selective sweeps on the basis of haplotype information. In order to determine the likelihood of a hard or soft sweep given the haplotype-based summary statistics $H12$ and $H2/H1$, Bayes factors are estimated via ABC across a range of selection coefficients, time of fixations, and frequencies when the allele came under selection [108].

1.3.4 Supervised Learning Approach

The latest approach in the search for selection within the genome are methods making use of machine learning algorithms [3, 109, 110]. Machine learning has its roots in computer science, and broadly described, seeks to recognize patterns in data, learn from these patterns, and subsequently make predictions.

While still in their infancy as applied to evolutionary biology and population genetics, machine learning has a rich history in other fields and applications including character recognition [111], facial recognition and detection [112], social networking applications [113], and analysis of medical data [114].

Pavlidis et al. (2010) used a class of supervised learning algorithms called a Support Vector Machine (SVM) (see Appendix D.1) in an effort to classify a genomic region as either neutral or selected [109]. In this implementation, the authors used two summary statistics as features in their classifier: ω_{MAX} and Λ_{MAX}. The former makes use of linkage disequilibrium patterns [92, Appendix A.6], while the later is based entirely on the patterns contained within the SFS [88, Section 1.3.2]. In addition to the aforementioned statistics, they make use of combinations of ω and Λ, utilizing distances between their peaks with the reasoning that these two statistics should be maximally correlated when a true sweep is present.

Along with a neutral demography, classification accuracy is examined under two models of a population bottleneck which varied in length and severity. In this study the SVM preformed better than ω or Λ alone, albeit still performing poorly (true positive rates dropping to 63% when false positive rates reached 50%) when attempting to detect both young and old sweeps in the presence of a severe bottleneck. However, their method showed marked improvement when the task was to classify neutral region under stationary demography vs. a selected region under the same stationary demography, as well as regions subject to non-equilibrium neutral histories vs. selection in equilibrium populations. This study marks the first application of a machine learning algorithm to the classification problem of differentiating neutral regions from those under selection, and further demonstrates that the potential exists to incorporate such methods into population genetics in the future.

Following Pavlidis et al. (2010), Ronen et al. (2013) produced an analysis similar in scope and proposed an algorithm called SFselect, with naming convention based upon the fact that their SVMs were trained on a binned and scaled site frequency spectrum, exclusively [3]. In order to build their classifiers, training data was generated using forward simulations where strength of selection and fixation times constituted a grid of known values, $s \in (0.005, 0.01, 0.02, 0.04, 0.08)$ and $\tau \in (0, 100, 200, \ldots, 4000)$. Effective population size was fixed at $N_e = 1000$ and each simulated chromosome was 50kb. Per cite mutation rate and recombination rate were fixed at $\mu = 2.4 \times 10^{-7}$ and $r = 3.784 \times 10^{-8}$.

The first set of tests looked to examine how their classifier performs when the model parameters, s and τ, are known. This is essentially a form of model validation and is named SFselect-s to indicate "specific" parameters. They train 200 SVMs, one for each parameter combination, and then compare their power at each parameter combination to that of other popular selection-finding methods, including Tajima's D [86], Fay and Wu's H [76], OmegaPlus [115] - an implementation of the Kim and Nielsen's (2004) ω statistic, and SweeD [116] - an improved implementation of Nielsen et al. (2005) SFS-based Λ statistic. The results show, broadly, that when classifying regions as either neutral or under selection when the parameters are known, the SFselect-s classifier is superior to all of the other methods listed. This is especially true in two cases: 1) when selection is weak, $s = 0.01$, and 2) at time points pre and post fixation, that is, at those times right when the beneficial mutation originates and at times long after the fixation has swept through the population. For example when

$s = 0.08$, SFselect-s has 87% power at 2000 generations post-fixation, vs. 42% for the next best method, Fay and Wu's H. Also, at the time when a fixation occurs, here 1000 generations, looking at a selection coefficient of $s = 0.02$, they have 85% power vs. 57% using Tajima's D.

In the second model, SFselect, an approach is developed to train an SVM that can be used in a more general framework, that is to classify data where the values of s and τ are not known. Using a cosine distance metric to determine similarity between SVM feature weights, two new SVMs are trained by aggregating simulation data across a range of selective strengths and fixation times. Tests of these SVMs show reduced power compared to SFselect-s, however SFselect still has more power to differentiate selection than the alternative methods listed in the preceding paragraph. These performance increases, however, are limited to classification events at the actual time of the sweep's fixation and not at periods prior to or following fixation as was demonstrated in the general model.

Finally, SFselect was applied to human population data obtained from the 1000 Genomes Project [117]. They determined 339 genomic regions where selection has taken place, 217 of these overlapping regions containing known genes. Of these 339 regions 36 others had been described in previous studies as being the putative products of adaptation, including genes associated skin pigmentation the region containing the lactase gene [83, 118, 119, 120]. Novel candidates that were found included a cluster of olfactory receptor genes and two regions potentially associated with various immune response.

1.3.5 Conclusion

The search for selection in the genome has been an important topic in genetics and evolutionary biology for some time now. The methods used in this search have advanced from simple detection of outlier regions, to composite likelihood scans of the genome, to methods utilizing state-of-the-art supervised learning techniques. And while these advancements have helped to shed light on the questions as to how evolution proceeds on the genomic level, there are still a number of issues that plague all of the above mentioned scans.

Using the variation that exists in population data has been extremely valuable in the search for selection, however most methods to date have not made full use of the information contained therein. In the case of the newest tests, Garud et al. (2015) uses only haplotype structure, whereas Ronen et al. (2013) only makes use of the signals present in a scaled site frequency spectrum. Pavlidis et al. (2010) does use both SFS and LD information, albeit using only a single summary statistic for either. While these methods have shown success, the question still remains - does utilizing more of the information present in the data improve the power of selection scans?

Further, current methods have concentrated on differentiating selected regions from those of neutral regions. There is not yet in place a powerful method that can classify those selected genomic regions as having arisen as the result of de novo beneficial mutation, a hard sweep, or as having been selected for from standing variation - a soft sweep [121, 122, 123]. Garud et al. (2015) attempts to tease apart hard and soft sweeps, however the effectiveness of their method radically drops with strength of selection and fixation times occurring the recent past.

The last, and perhaps most major concern is that neutral scenarios such as non-equilibrium demography, population bottlenecks for example, or gene flow from other populations can create population genetic signatures that are nearly identical to those produced by selection and thus confound searches or scans for selective sweeps [124, 125, 126].

Therefore, the driving motivation behind Chapter 3 is to develop a method that makes use of all the available information in the data, and is able to accurately distinguish hard sweeps from soft sweeps - even in the face of non-stationary demographic history.

1.4 Population Demographic Inference

Anatomical, linguistic, and genetic data have long been sought after for the purpose to elucidate potential routes of migration that ancient populations might have undergone or to describe fluctuations in a population's size in order to paint an archaeological picture of some population's past [127]. In recent years there has been an interest in using genetic variation data to make inferences about various populations' demographic histories, and along with the massive influx of data due to next generation sequencing advances, many new techniques have been proffered to accomplish this goal [e.g. 5, 91, 128, 129, 130, 131, 132, 133, 134].

Early attempts made use of the method of moments approach in determining demography, as with Rogers and Harpending [135] who studied how the distribution of pair-wise nucleotide differences (π; see Appendix A) between individuals can be affected by population size expansions and contractions [135, 136]. Modern approaches to learn demographic histories, however, can be split into two broad categories - those using likelihood-based approaches and those using methods in a likelihood-free framework. In the following section I will describe how polymorphism data, both within and between populations, coupled with statistical approaches, has been used in an effort to learn the properties and dynamics of natural populations.

1.4.1 Likelihood Based Methods

In the past three decades researchers have been making use of the increasing amount of available genetic data in an effort to make sense of the complex demographic histories of humans and other species. Much of this work has its roots in population genetics theory that pre-dates modern sequencing methods [e.g. 137, 138]. With more population data now being generated, an ideal approach to make demographic inferences would be through likelihood based methods. Many methods have been developed to this extent, for example, Kuhner et al. [139] use maximum likelihood (ML) to infer population growth rates whereas Nielsen [140] derives a likelihood based method to do the same. In these approaches, however, the assumption in the former is that recombination is not present, or in the later that all SNP loci are independent of each other. Under models which incorporate recombination and linked sites, the likelihood approach becomes more challenging.

Method which use summary statistics in lieu of the full data have also been developed. Pluzhnikov et al. [141] make use of unlinked human autosomal data and use the means and variances of Tajima's D and Fu and Li's D^* to test different demographic models in a number human populations [142]. Weiss and von Haeseler [143] examined the variation in the non-coding hypervariable region I (HVRI) of mitochondrial DNA (mtDNA) in an attempt to infer population growth in three human populations [143].

Composite-likelihood (CL) techniques have been been developed and used to infer rates of recombination as in Frisse et al. [144], who used information present linkage disequilibrium amongst African, European, and Asian human populations. Voight et al. [145] extended this model to incorporate levels of polymorphism data and site frequency spectrum data in those same three human populations. The CL method was used infer ancient human population structure as well as the possibility of admixture events with Neanderthals [146].

Adams and Hudson [129] forgoes the use of summary statistics and uses the site frequency spectrum in a ML approach to infer a model of either growth alone, or population decline followed then by exponential growth, in a number of extant human populations.

Perhaps the most advanced method currently being used in demographic inference is that of Gutenkunst et al. [133]. Unlike techniques described so far, here the authors derive a method that uses diffusion-based approximations of the site frequency spectrum data from multiple populations in order to test hypothesis. Further, this method, called ∂a∂i (Diffusion Approximations for Demographic Inference), allows for testing separate rates of migration between each population, multiple admixture events, changing population sizes and divergence times, and can handle the incorporation of selection. In ∂a∂i, partial differential equations (PDEs) are numerically solved, requiring optimization techniques that might prove to be computationally expensive under complicated demographic models or large sample sizes [133, 147]

1.4.2 Approximate Likelihood Approaches

Often, likelihood functions associated with demographic inference do not offer closed form solutions. In the case where likelihoods can be determined up to a normalizing constant, methods based on Monte Carlo algorithms might be used such as Markov Chain Monte Carlo (MCMC) and importance sampling [148]. However, if a model is too highly parameterized, or if a likelihood function doesn't exist, there exists a method called Approximate Bayes computation (ABC) which effectively side-steps the likelihood function altogether (The reader is strongly encouraged to examine Appendix B which contains a description of the ABC algorithm).

Briefly, in ABC it is first necessary to be able to generate a large number of samples under some model, and by doing so, it is possible then to replace the likelihood function by an approximation to the likelihood. While rejection sampling forms the basis for the simplest methods of ABC, techniques have been developed which expand on this framework, including modifications to the estimated posterior distributions [149], incorporating MCMC into the simulation step [150], and a partial least squares transform of summary statistics [151].

ABC methods have been increasingly used and applied to population demographic problems. In a comprehensive simulation study, Sen Li [152] tested the ability of local linear regression ABC to estimate parameters in three complex demographic models using simulated data consisting of hundreds of thousands of SNPs. Further, they evaluated how the choice of summary statistics, both haplotype and linkage disequilibrium based, affect ABC performance [152].

Thornton and Andolfatto [153] used the means and variances of several standard summarty statistics in an ABC framework to estimate the time, duration, and severity of a recent out-of-Africa population bottleneck in European *Drosophila melanogaster*, concluding that a severe bottleneck occurred $\sim$ 16,000 years ago. Following this, Duchen et al. [134] modeled a joint demographic history of *Drosophila melanogaster* in African, European, and North American populations. They conducted ABC with summary statistics that included the number of segregating sites and haplotypes, S and K, respectively, Watterson's θ estimator, θ_W, π, Tajima's D, and Kelly's ZnS. Using model selection via Bayes factors under the ABC framework, they concluded that of the five models examined, the best supported history was one in which the North American population was a product of a recent admixture event between Europe and Africa [134].

Looking to differentiate between seven demographic models, Shafer et al. [154], used an ABC approach to estimate model parameters and perform model selection. Further, they compared their ABC results to results obtained from ∂a∂i, showing similar power between the two methods. Finally, they applied their method to an empirical data obtained from a population of Atlantic walrus *Odobenusrosmarus rosmarus* [154].

1.4.3 Pairwise Sequential Markovian Coalescent Approach

The last algorithm described is that of Li and Durbin (2011) [155]. This algorithm is a departure from the coalescent algorithms described in Appendix C, which the reader is encouraged to examine, as the PSMC is an algorithm used to make inference on demographic and population parameters given some data, whereas the algorithms described in Appendix C are used to sample genealogies given some genetic data.

Knowing a population's demographic history is crucial in understanding how that population evolved. In the case of human evolution, it has been proposed numerous times that a population bottleneck occurred during an out-of-Africa migration approximately 60,000 years ago [156]. Using the PSMC method, the authors posit that their method might provide clearer insights into this demographic event, as their model requires fewer parameters and assumptions than those of allele frequency and summary statistic based methods [5, 6, 156, 157, 158, 159, 160].

The theory behind PSMC begins with the idea that within a diploid genome there exists hundreds of thousands of independent loci, and that each pair of these loci have different time to most recent common ancestor (TMRCA). The attempt then is to exploit this situation by inferring the TMRCA across the whole genome as a function of how the density of heterozygosity changes along the genome. Specifically, we are looking for genomic regions of constant TMRCA and how those regions are separated by recombinations. The PSMC is a Hidden Markov Model (HMM) along a haplotype, which is described in detail, below.

PSMC Algorithm

The PSMC model is a HMM who's underlying framework is the Sequential Markovian Coalescent (SMC) model of McVean and Cardin [161]. The data consists of a diploid sequence which is partitioned into contiguous 100bp segments. The actual observations in the HMM are labeled "1", "0", or "·". Specifically, if a 100bp window has more than 10 sites that are inferable and contain no heterozygous sites it is labeled a “0”. If there are more than 10 sites that are inferable and at least 1 or more heterozygous site this observation is a “1”. Otherwise, windows with at least 90 or more sites that are either filtered or uncalled are labeled a “$\cdot$”.

Concerned with only inferable data, the emission probabilities from state t, given observation $[1, 0]$ are:

$$e(1|t) = \mathrm{e}^{-\theta t} \tag{1.9}$$

$$e(0|t) = 1 - \mathrm{e}^{-\theta t} \tag{1.10}$$

The transition probability from s to t is:

$$p(t|s) = (1 - \mathrm{e}^{-\rho t})q(t|s) + \mathrm{e}^{-\rho s}\delta(t - s) \tag{1.11}$$

The transition probability conditional on a recombination event is:

$$q(t|s) = \frac{1}{\lambda_{(t)}} \int_0^{min\{s,t\}} \frac{\mathrm{e}^{-\int_\mu^t \frac{\mathrm{d}v}{\lambda_{(v)}}}}{s} \mathrm{d}\mu \tag{1.12}$$

Where

$$\lambda_{(t)} = \frac{Ne_{(t)}}{N_0} \tag{1.13}$$

is the relative population size at coalescent time t, and where the scaling factor,

$$N_0 = \frac{\theta}{4\mu} \tag{1.14}$$

In order to infer parameters from PSMC, real time is discretized into segments that are evenly distributed in log space. Specifically, the set $\{t_i\}_{i=0\ldots n}$ where

$$t_i = 0.1(\mathrm{e}^{\frac{i}{n}\log(1+10T_{max})} - 1) \tag{1.15}$$

and $T_{max} = t_n$ is set so only a few coalescences happen beyond this value.

The authors determined that $T_{max} = 15$ is optimal and used 64 atomized time intervals, t_i. To decrease the search space complexity, the TMRCA was not found for each time interval, t_i, but rather neighboring time intervals were joined and assumed to have the same population size parameter. For example, on the autosome, the pattern for joining neighboring time units was $1{*}4 + 25{*}2 + 1{*}4 + 1{*}6$ (author's notation). That is, 1 population parameter was used for the first 4 of the 64 time intervals, the next 25 parameters were used for sets of 2 intervals, the next (27th) parameter spans 4 time intervals, and the last parameter is used for the final 6 intervals. Therefore, 28 N_e's are learned over the 64 time intervals. In the case of the X chromosome, 60 time intervals are used, $T_{max} = 15$, and the pattern for learning TMRCA was $1{*}6 + 2{*}4 + 1{*}3 + 13{*}2 + 1{*}3 + 2{*}4 + 1{*}6$.

The free parameters in the model are the scaled mutation rate, the rate of recombination, and the vector of population sizes. For the EM estimation, the initial population parameter was set to 1, Watterson's θ was used for mutation rate, and recombination was set to $1/4$ of the mutation rate. For each of the time intervals, t_i, the scaled population size is calculated directly as the inverse of the coalescent rate, $1/\lambda_{(t_i)} = Ne_{(t_i)}/N_0$ (See Equation (1.13)). The learned TMRCA for each time interval is in units of mutation per site. To convert to years the authors assumed a 2.5×10^{-8} neutral mutation rate per site per generation and used a generation time of 25 years.

In order to validate the model, one hundred 30Mb sequences were simulated with a bottleneck demography designed to simulate the human out-of-Africa migration, followed then by a population expansion. The parameters are recovered quite well between 2×10^4 and 3×10^6 years ago, while times occurring before or after this range preform poorly due to lack of recombination events. For further testing, five alternative demographic histories were simulated which included sharp bottlenecks and population splits. Overall, the PSMC model performs well when examining time between 2×10^4 and 3×10^6 years ago, although the PSMC tends to smooth large changes in population size across few time intervals [155].

In applying PSMC to empirical autosomal data, it was demonstrated that six populations examined have very similar population sizes between 1.5×10^5 and 1.5×10^6 years ago. However, the two Yoruban autosomes show greater population sizes than the other four examined autosomes from 1.1×10^5 years ago until 1×10^4. It is interesting to note that at approximately 150,000 years ago effective population size was estimated to be 13,500 for Yoruban, Asian, and European individuals, which then saw a dramatic decrease in European and Asian populations to $N_e = 1200$ at a time of 40,000 to 20,000 years ago. The African populations also showed a reduction in N_e, albeit less severe, decreasing to 5700 at 50,000 years ago. This bottleneck resolved with a Yoruban population increase to $N_e = 8700$ at 20,000 years ago, which the authors point out coincided with the Last Glacial Maximum [155].

It is worth examining that at least two scenarios might be problematic for the PSMC when applied to empirical data. First, PSMC estimated that all populations, including African, had an increased population size during the time we believe modern humans to have arisen [162]. However, when examining how the PSMC performed on a structured population, it was found that the estimated effective population size was larger than the sum of the subpopulations [155]. As coalescences occur less often in structured populations, it appears as though a larger population size exists. Thus, the increase in population size estimated by PSMC on the empirical data may actually be indicative of underlying population substructure. Second, the authors note that at 1,000,000 years there was increased population size across all populations, and that at $\sim 3,000,000$ years ago, a huge spike in population size occurred. While this time frame does coincide with previous estimates of the human-chimp split, it is possible that this ancient population size is due to long genomic sections of high heterozygosity. During simulations, the authors generated chromosomes using a mutation rate 10x higher than the normal simulations, mimicking the effects of both balancing selection and duplication. The effect of these hypermutated regions led to the PSMC estimated extremely large values of ancient N_e. Again, the results thus seen in the empirical data could stem from large duplications in the genome, rather than a split from our ancestral primates [155].

Despite the potential pitfalls of this method listed above, and despite PSMC not taking into account selection across the genome, it appears that this new method offers an convenient way to estimate the ancestral effective population size using single, diploid genomes. This method is tractable for large chromosomes and has been shown more recently to be able to be extended to more than one diploid genome [163].

1.5 Conclusion

Using modern advances in technology coupled with newfound understanding of evolutionary biology as described above, and while being cognizant of their limitations, I present the following research. I apply the power of next generation sequencing and quantitative computer methods to accomplish the following: 1) analyzing and answering evolutionary questions about a non-model organism, *Enallagma hageni*, through the use of transcriptomics 2) devise a framework that incorporates state-of-the-art machine learning techniques in order to determine and describe how natural selection acts on the genome, and 3) to determine how the long reaching dynamics of evolutionary processes at the level of the genome play a roll in, and affect, the ability to make accurate inference about population demographic histories.

2

Evolutionary Analysis Of *Enallagma hageni* Using Next Generation Transcriptome Sequencing

A.G. Shanku, M.A. McPeek, A.D. Kern

Functional Annotation and Comparative Analysis of a Zygopteran Transcriptome. G3, 3(4):763-770, January 2013

> *"A living organism must be studied from two distinct aspects. One of these is the causal-analytic aspect which is so fruitfully applicable to ontogeny. The other is the historical descriptive aspect which is unraveling lines of phylogeny with ever-increasing precision. Each of these aspects may make suggestions concerning the possible significance of events seen under the other, but does not explain or translate them into simpler terms."*
>
> – Sir Gavin de Beer

2.1 Introduction

Enallagma damselflies are aquatic invertebrates belonging to the order Odonata. Included in this group are dragonflies (suborder Anisoptera) and other damselflies (suborder Zygoptera), which together represent one of the most ancient branches of the winged insects (Pterygota) and furthermore represent a basal group within the division Palaeoptera [164]. The damselfly has a rich history as an organism used in evolutionary and ecological studies, spanning research in speciation [165, 166], species distribution [165], selection [167, 168], population diversity [169], and predator-prey interactions [170, 171, 172].

Despite the fact that this organism is an ideal candidate for many types of biological studies, there has been relatively little examination of the genetic makeup of damselflies on a large scale [173, 174, 175]. For example, most of the sequence data used to determine phylogenetic relationships among *Enallagma* species, as well as to infer *Enallagma* phylogenetic relationships within Odonata, has been in the form of mtDNA [176, 177] or ribosomal nuclear DNA [178]. Therefore, in this study, we attempted to investigate the nuclear, protein-encoding gene profile of the damselfly *Enallagma hageni* by using next-generation sequencing technology and, by doing so, (1) give further resolution and support to this organism's phylogenetic position within Arthropoda, (2) determine the evolutionary rates of the protein-encoding genes in the *Enallagma* transcriptome, and (3) give functional annotation to the proteins expressed in our dataset.

2.2 Results

2.2.1 Transcriptome Assembly

After assembly we obtained 31,662 contigs comprised of 13,191,394 nucleotides. Of these contigs, 1656 were singletons (5.23%). Median coverage was 25 reads/contig (mean = 179.71 reads/contig, SD=746.27) and median contig length was 355 bases (mean 416.6, SD=429.7). With singletons excluded, the dataset was reduced to 29,996 contigs. Of these, median coverage was 26 reads/contig (mean coverage = 173.73 reads/contig, SD=677.99) and median contig length was 406 bases/contig (mean contig length = 439.7 / SD = 429.9). The largest contig in the dataset was composed of 3036 nucleotides. The assembled transcriptome contained an AT bias at 59.86%, GC at 40.13% and 0.01% were labeled N. CpG sites occurred at 2.69% of the transcriptome. (Figure F.2 and Table E.3 for assembly details.)

2.2.2 Translated Proteins

Translation of the *Enallagma* contigs yielded 14,813 individual open reading frames comprised of 1,621,208 amino acids (singletons not included). Mean length was 109 amino acids. Shortest and longest protein sequences were comprised of 19 amino acids and 735 amino acids, respectively Figure F.3.

We have deposited our raw sequences and assembled transcriptome at the National Center for Biotechnology Information (NCBI). The *Enallagma hageni* Bioproject (Accession: PRJNA185185 ID: 185185) contains links and access to all data, including insect sampling data: BioSample (SAMN01881995), raw sequencing data: SRA (SRR649536), and transcriptome data: (SUB156504).

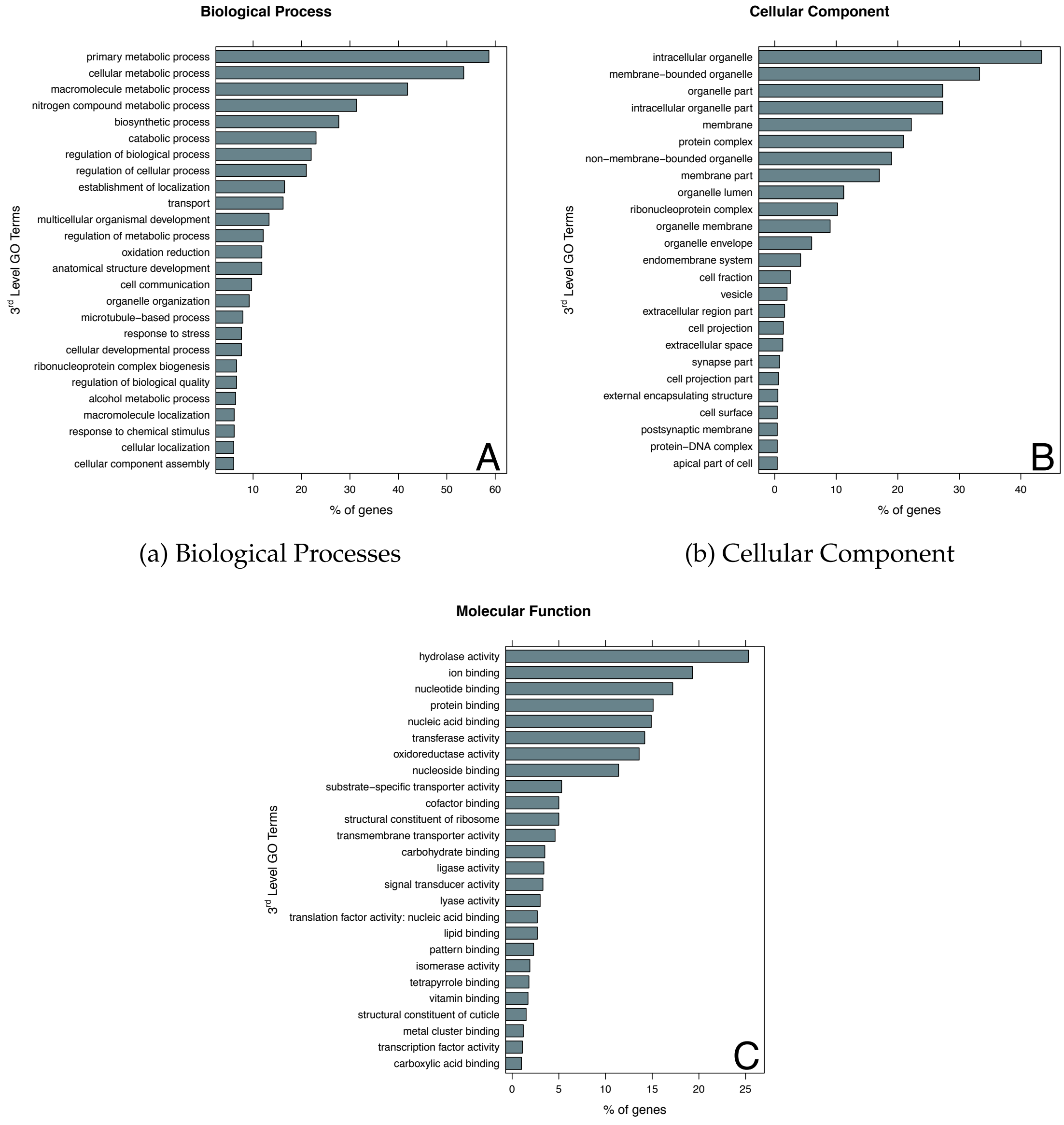
2.2.3 Orthologs

The one-to-one, reciprocal best method of elucidating orthologous proteins generated 634 orthologs across the 11 species in the study. The *Enallagma* orthologs, themselves, contained 108,866 amino acids with a mean length of 171 amino acids, and shortest and longest sequence length of 46 amino acids and 413 amino acids, respectively Table E.4.

2.2.4 Gene Ontology Annotation

Our annotation methodology mapped 3998 *Enallagma* genes to at least one GO term, using Blast2GO and the NCBI nr database. There were 24,439 total GO terms mapped to those 3,998 genes, with 3,812 of the GO terms being unique. The mean mapping was 6.1 GO terms/gene with a minimum and maximum mapping of 1 and 78 GO terms per gene, respectively. Using 3rd and 4th level GO term distributions, we mapped our dataset to 404 GO terms across 3 ontologies for 3rd level terms (Cellular Component, Biological Process and Molecular Function) and 1,463 terms across 3 ontologies for 4th level terms (Figures 2.1 and 2.2). At the 3rd level of the hierarchy the top GO terms represented are 1) Biological Processes: 58.7% of the genes were mapped to "primary metabolic processes" (GO:0044238), 53.5% of genes to "cellular metabolic processes" (GO:0044237), and 41.9% to "macromolecule metabolic processes" (GO:0043170). 2) Cellular Components: 43.4% to "intracellular organelles" (GO:0043229), 33.3% to "membrane-bound organelles" (GO:0043227), and 27.3% to "organelle parts" (GO:0044422). 3) Molecular Function: 25.3% to "hydrolase activity" (GO:0016787), 19.3% to "ion binding" (GO:0043167), and 17.2% to "nucleotide binding" (GO:0000166). See Figure 2.1

for 3rd level distribution. See Figure F.4 for 4th level distributions.



(a) Biological Processes

(b) Cellular Component

(c) Molecular Function

Figure 2.1: **3rd Level Go Term Distributions For Annotated *Enallagma hageni* Genes.** GO term distributions were plotted for each of the three 1st level categories. The full dataset mapped to 404 unique GO terms at the 3rd level. Shown are the top 25 terms in each of the broadest, 1st level categories: (a) Biological Process, (b) Cellular Component, and (c) Molecular Function.

To look for enriched or diminished GO terms, we then compared the *Enallagma hageni* GO annotations to the *Drosophila melanogaster* GO annotations. We queried 3,986 annotated *Enallagma* genes against 13,127 annotated Drosophila genes and found that 1,080 unique (1089 total) *Enallagma* GO terms were enriched or diminished. Described in terms of the GO hierarchy, we discovered 33 2nd level GO terms and 161 3rd level GO terms.

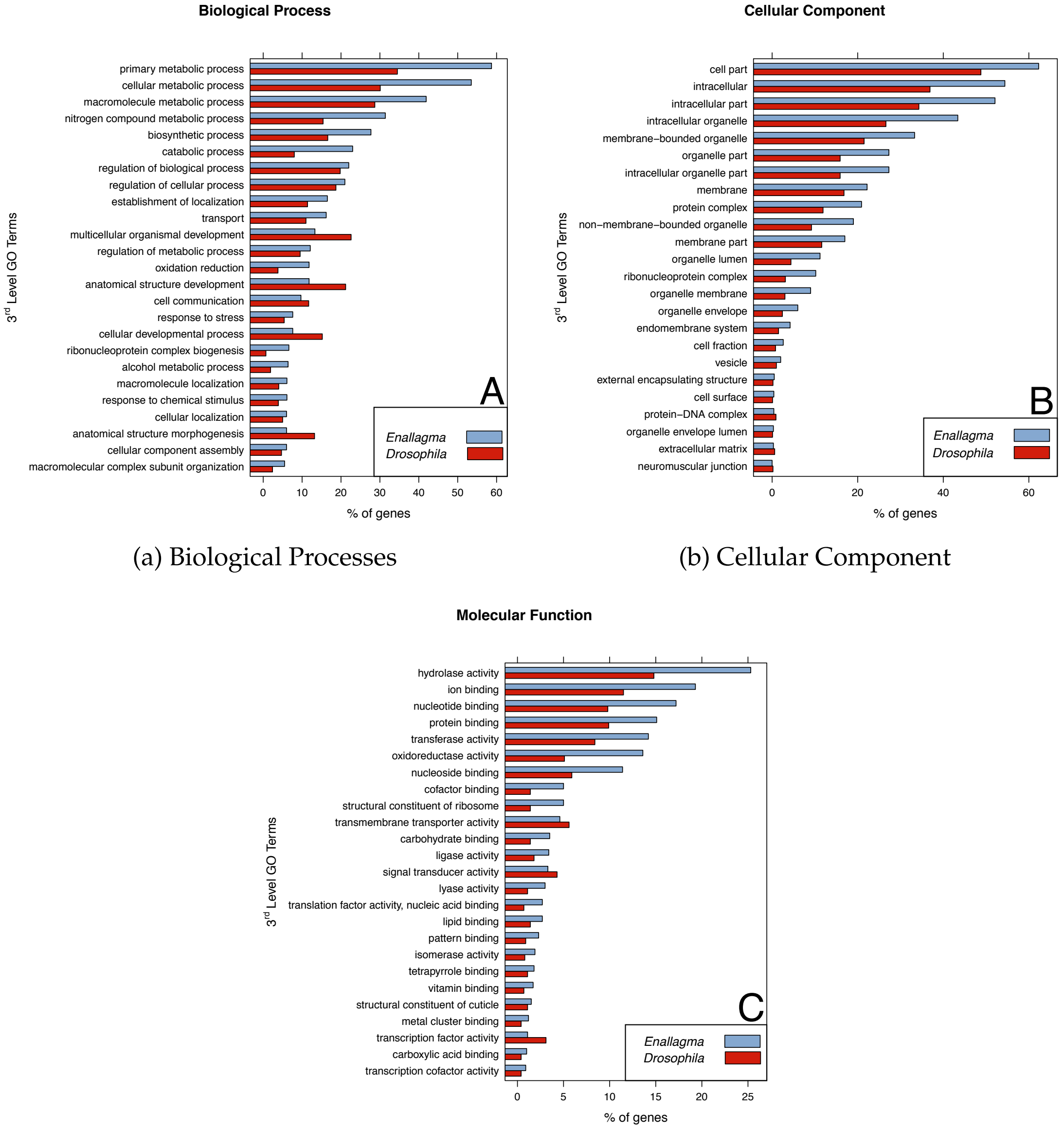
Some of these enriched 3rd level GO annotations include: hydrolase activity (GO:16787), ion and nucleotide binding (GO:43167 and GO:0000166), and primary metabolic processes (GO:44238). Examples of diminished GO terms include: anatomical structural development (GO:48856) and protein-dna complex (GO:32993).

Additionally, we mapped 488 genes within the orthologous protein-coding set to 1669 GO IDs, 691 of these GO IDs being unique. For the gene ID, GO ID, and gene product and function, see Table E.5.

2.2.5 Phylogenetics

After concatenating the 634 orthologous genes, the resulting multi-way alignment contained 182,478 amino acid positions. This alignment was then filtered with Gblocks, using the default parameters that does not allow for gaps at any position in the matrix, resulting in an un-gapped alignment of 27,594 amino acids positions (15.1% of the original data). This ungapped matrix was then analyzed using Mr. Bayes with settings described in Section 2.4.8.

We removed 50 samples of burn-in after each MCMC run, therefore sampling from the posterior 2,952 times for each of the two runs. Each of the two MCMC analyses took 224,340 seconds (62.3 hours) and 227,756 seconds (63.3



(a) Biological Processes

(b) Cellular Component

(c) Molecular Function

Figure 2.2: **Enrichment Or Reduction Of *Enallagma hageni* GO Terms Relative To Annotated *Drosophila melanogaster* Genes.** Using *D. melanogaster* as a background set, hypergeometric distribution tests were performed to identify Enallagma genes that were enriched or diminished. The background set consisted of 13,127 *D. melanogaster* annotated genes and was queried by 3986 Enallagma genes. We discovered 1080 unique enriched or diminished terms. Shown in (a) Biological Process, (b) Cellular Component, and (c) Molecular Function are the top 25 most significant results.

hours) to complete, respectively. The plotted phylogram, based on the consensus tree data of the MCMC runs, is shown in Figure 2.3.

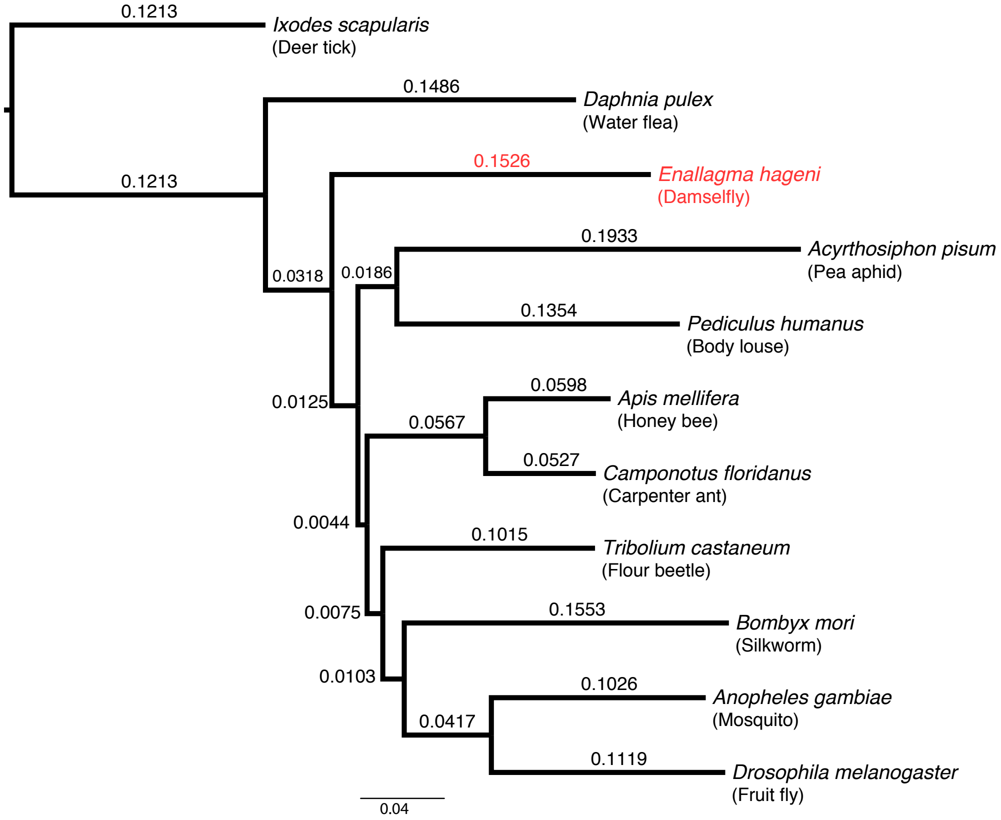


Figure 2.3: **Arthropod Phylogram** 11 taxa and 23,679 amino acids positions were used in the analysis. Branch lengths are labeled, and posterior probabilities at each branching node are 1.0.

Ixodes scapularis (Class: Arachnida) was chosen as the out-group and the tree was rooted upon it. The posterior probability for each node in the tree was 1.0. Trace plots of the MCMC analysis and Gelman convergence plots can be found in Figures F.5 and F.6.

2.2.6 Rate Testing

The branch length test indicated that 439 of the 634 (69.2%) orthologs fit a local clock model better and were therefore deduced to be evolving at a rate that varied relative to that gene's orthologs ($p < 0.05$). However, a Bonferroni correction for multiple tests, ($p < 0.05/634 = 0.0000788$), reduced that set and yielded 169 genes which were shown to be evolving at significantly different rates in Enallagma. Of these 169 genes, 29 genes were shown to be evolving at an accelerated rate, while the remaining 140 were determined to be evolving at a reduced rate. We successfully mapped 37 of these genes to at least one GO term. In the accelerated case, 4 of the 29 genes were mapped to 17 GO terms, while in the decreased case 33 of the 140 genes mapped to 105 GO terms. Of those 37 genes we were able to annotate, no significant enrichment was noted using the hypergeometric test ($p < 0.05$), relative to the background set of all Enallagma GO annotations. Table E.1 shows the four accelerated genes and their gene products. These include *Nol10* a nucleolar protein, *Art7* a protein arginine N-methyltransferase, *Rrp45* a protein involved in RNA processing, and *Uba3* an ubiquitin-like protein (Figures F.7 and F.8 and Table E.2.)

Contributions

In this preceding section, I was responsible for all steps in the analysis from Section 2.2.2 through Section 2.2.6. I generated all figures, tables, and wrote the first draft of the manuscript. M.A.M collected the data in Section 2.4.1 and A.D.K performed the assembly in Section 2.4.2. M.A.M and A.D.K contributed to the writing of the manuscript.

2.3 Discussion

At the level of resolution we examined (other species within Arthropoda which had assembled transcriptomes), our phylogenetic analysis of *Enallagma* and the compared Arthropods appears congruent to that of other current studies and reviews [179, 180, 181].

Our hypergeometric tests of the accelerated and decreased rates of proteins GO annotations, relative to the background set of all genes we were able to annotate, indicated no significant enrichments ($p < 0.05$ raw, FDR corrections). Nevertheless, the GO term distributions of the altered rate genes were shown to similarly represent the distributions of the overall dataset. For example, the top three GO terms represented by both the Biological Processes and Cellular Component 3rd level domains were the same. In the case of Biological Processes, we saw the terms "primary metabolic process", "cellular metabolic process", and "macromolecule metabolic process" encompassing the top three positions, while the top three terms in the domain of Cellular Component were "intracellular organelle", "membrane-bounded organelle", and "intracellular organelle part". However, there were some deviations from that, especially in the Molecular Function domain. For example, the top two GO terms represented in the decelerated genes category, in the "Molecular Function" domain, were shown to be "nucleotide binding" and "nucleic acid binding", whereas in the full set, the top two expressed GO terms for that same domain were "hydrolase activity" and "ion binding".

One of the interesting ecological and evolutionary scenarios involving *Enallagma* is that various *Enallagma* lineages have adapted to living with predators by increasing their burst swimming speeds to increase their probability of escape during predator attacks [182, 183, 184]. In agreement with this, we annotated genes involved in muscle mass increase and differentiation (GO:0003012) and genes with roles in Arginine Kinase (GO:0004054), and Arginine methylation (accelerated; see Table E.1) (GO:0019918) which has been shown to partially responsible for the observed rapid movements of the damselflies [183, 184].

Another issue worth noting is that analysis by short read sequencing in transcriptome assembly relies on the use of reads typically 35-250bp in length [31, 185]. Our annotation methodology mapped 3998 *Enallagma* genes to at least one associated GO term. While this number represents less than 30% of the genes in our dataset associating with a GO term, it should be noted that small contigs, like those generated in 454 sequencing, can be difficult to successfully map to GO terms and that mapping success increases successively with read size [186, 187].

In summary, we have generated a draft functional annotation of nearly 4,000 genes in *Enallagma hageni's* transcriptome, which to our knowledge is the first examined and annotated transcriptome of any palaeopteran in the literature. We examined the rate at which *Enallagma hageni's* proteins are evolving and found 169 genes which fit better the hypothesis of having an altered evolutionary history, relative to other genes in its transcriptome. We examined the distributions of GO terms for each of three classes of our data: the whole annotated transcriptome, the transcriptome with *Drosophila melanogaster* as a background, and the set of altered genes with all *Enallagma* genes as a background. Of those, we additionally deduced which annotations are enriched or diminished through the use of hypergeometric distribution testing. Finally, we have produced a strongly supported phylogenetic analysis that in turn further strengthens support for Odonata's place in the Arthropoda tree.

2.4 Methods

2.4.1 Insect Capture and RNA preparation

Individuals across the entire life cycle were included in the sample from which RNA was extracted. Some *Enallagma* larvae are difficult to identify to species, with *E. hageni* being one of these. *E. hageni* larvae are largely indistinguishable from four other species that are all derived from a very recent radiation [166]. To ensure that we were unambiguously collecting *E. hageni* larvae, we collected larvae from Martin's Pond, Green Bay, VT, a lake where we have only found *E. hageni* and none of these other species as adults in previous years (McPeek, pers. obs.). Embryos were obtained by allowing females to oviposit in the laboratory, then allowing two weeks for development prior to RNA extraction. Aquatic larvae from across the entire range of the larval period and adults were collected, and immediately placed in RNAlater (Ambion Inc.) until RNA isolation. Total RNA was isolated from the pooled material of roughly 50 embryos, 150 larvae, and 25 adults by first flash freezing the insects in liquid nitrogen, and then processing the frozen material using Qiagen RNAeasy protocols. From our isolations we collected roughly 100 mg of total RNA.

2.4.2 Transcriptome Sequencing and Assembly

mRNA isolation, library construction, and 454 sequencing were contracted out to Beckman Coulter Genomics using 1mg of total RNA as starting material. All sequencing was of unnormalized cDNA libraries using standard 454 protocols on the 454GS instrument. This produced 976,767 reads (see results for details on the sequencing output).

To perform de novo transcriptome assembly on our reads we used the Newbler assembler (v2.3) using parameter settings specifically for mRNA assembly (Table E.3).

2.4.3 Protein Translation

To compile a dataset of proteins which would form the basis of our analysis, assembled contigs were translated using Virtual Ribosome [188]. Each of six open reading frames (ORF) was translated `--readingframe=all` and the longest resulting translated read was kept, provided it initiated with a start codon `--orf=any`. To account for contigs that may have had their upstream start codon truncated during assembly, we again translated over six ORFs all contigs that did not posses a start codon, but terminated with a stop codon `--orf=none`. Of these two sets of putative proteins, the longest read that possessed both a start and stop codon was determined to be the translated protein for a given contig unless a fragment not initiated by a start codon, but terminating with a stop codon, was greater in length. Contigs composed of fewer than 10 nucleotides were excluded from translation and removed from further analysis.

2.4.4 Arthropod Proteins

Comparative analysis of phylogenetic relationships necessitates the alignment of homologous sequences amongst individuals being compared. To compile the data for such an analysis, we began by conducting a search aimed at identifying orthology across expressed proteins in a group of selected Arthropods. To build this set of putative orthologous proteins, we obtained transcriptome data from ten arthropod species housed on public databases (See Table 2.1 and Figure F.1).

Binomial Name	Common Name	Class / Order	Public Database
Acyrthosiphon pisum	Pea aphid	Insecta / Hemiptera	NCBI
Anopheles gambiae	Mosquito	Insecta / Diptera	Vectorbase
Apis mellifera	Honey bee	Insecta / Hymenoptera	NCBI
Bombyx mori	Silkworm	Insecta / Lepidoptera	Silkworm Gen. DB
Camponotus floridanus	Carpenter ant	Insecta / Hymenoptera	Hymenoptera Gen. DB
Daphnia Pulex	Water flea	Branchiopoda / Cladocera	wFleaBase (Daphnia Gen. Proj.)
Drosophila melanogaster	Fruit fly	Insecta / Diptera	Flybase
Ixodes scapularis	Deer tick	Arachnida / Ixodida	Vectorbase
Pediculus humanus	Body louse	Insecta / Phthiraptera	Vectorbase
Tribolium castaneum	Red flour beetle	Insecta / Coleoptera	NCBI

Table 2.1: **Arthropod Species Used In Transcriptome Analysis.** These ten species transcriptomes were obtained from publicly accessible databases. Included in the dataset are one Arachnid (*Ixodes scapularis*), one Branchiopod (*Daphnia Pulex*), and 8 Insects. All data were downloaded from their respective database in January 2011.

2.4.5 Ortholog Detection

To construct a working set of orthologous proteins we utilized a method of one-to-one reciprocal best BLAST hits [189, 190], rather than attempting to use ortholog clustering methods (e.g. OrthoMCL; [191]). We performed a BLAST search between protein-coding genes in each species transcriptome and those in *D. melanogaster*, and conversely, the *D. melanogaster* transcriptome was BLASTed against all protein-coding genes present in each of the species transcriptomes in the dataset. The best hit was determined using the `-K 1` and `-b 1` BLAST parameters, which limit output, in this case the `-m 8` tabulated output format, to the best scoring hit of each BLAST query. Following this methodology, and using mpiBLAST, an open-source, parallelized version of BLAST [192], we constructed a set of reciprocal-best, one-to-one orthologs. To expedite computational processing time, each species database file was partitioned into 94 fragments `nfrags=94`, and parameter settings `--output-search-stats --use-parallel-write --use-virtual-frags --removedb` were used for each job. Using custom scripts, individual orthologs that were present across all 11 arthropod species were grouped together into individual `.fasta` files. Following this search and grouping method, the protein sequences within each file were aligned using clustalw2, using the flags `-OUTPUT=FASTA` and `-OUTORDER=INPUT`, the latter being necessary to later allow for concatenation of all aligned orthologs when conducting phylogenetic analysis [193].

2.4.6 Phylogenetics

Each orthologous gene alignment was concatenated into a "super-gene" [194], that is, we took individual `.fasta` files and joined them into one singular, interleaved `.nexus` file using a custom Ruby script. If an amino acid position in the concatenated alignment contained a gap at a position in any of the species, or in multiple species, that position was removed prior to analysis using Gblocks 0.91b [195], as we did not use a model of sequence evolution that allowed for insertions or deletions.

2.4.7 Model Selection

To determine the optimal model of protein evolution for phylogenetic analysis of our dataset, ProtTest 2.4 was utilized for model selection [196, 197]. All amino acid evolutionary rate models available in ProtTest were examined, as were the +I, +G, and +F parameters (Dayhoff [198], JTT [199], WAG [200], mtREV [201], MtMam [202], VT [203], CpREV [204], RtREV [205], MtArt [206], HIVb/HIVw [207], LG [208], and Blosum62 [209]).

Ideally, we would optimize tree topology, branch lengths and the parameters of the model, for each model investigated. This is inefficient in our case, as the dataset is too large to realistically attempt topology optimization for each model and each additional model parameter associated with that model. Instead, we allow a neighbor-joining tree to be constructed given our data, fix the topology, and from that topology, optimize branch lengths and select model parameters [210].

2.4.8 Bayesian Phylogenetic Inference

Once the optimum model was selected, we searched topology space of the 11 arthropod species in our dataset using a Bayesian MCMC approach via Mr.Bayes v3.1.2 [211, 212, 213].

The following settings were used in our MCMC analysis: Two runs, 750,000 generations, number of chains=240, sample frequency=250. 240 processors were utilized in parallel. The evolutionary model used was the WAG model that allows for 20 states. Rates were set to `Invgamma`, with the gamma shape parameter $\sim U(0.00, 200.00)$. The proportion of invariable sites was $\sim U(0.00, 1.00)$. All topologies were equally probable and branch lengths were unconstrained.

2.4.9 Rate Testing

To address the question of whether certain orthologous protein-coding genes present in *Enallagma* were evolving at different rates relative to the other arthropods, branch length rate tests were conducted on each *Enallagma* gene in our dataset. Using PAML [214], two models for each protein were generated: one which assumed a global clock across all species, and the other which fixed the rate of evolution of each *Enallagma* protein to a local clock, while keeping the rest of the species evolutionary rates confined to a global clock. In this manner, we generate two likelihood estimates (one for each model) for these proposed modes of evolution of a particular protein. To that extent, a likelihood ratio test was performed between the null model (global clock) and alternative model (local clock).

$$D = -2(\ln L_G - \ln L_L) \tag{2.1}$$

Where D is the test statistic, $\ln L_G$ is the log likelihood of the global clock model and $\ln L_L$ is the log likelihood of the local clock model. The probability distribution of the test statistic, D, can be approximated by the chi-squared distribution, where the degrees of freedom of the distribution is equal to the number of free parameters of the global model minus the number of free parameters of the local model, which, for our purposes will be 1. (Parameters required in local model = 11 while the parameters required in the global model = 10.) Once a raw probability for each likelihood ratio was calculated, we performed Bonferroni corrections to determine significance [215].

2.4.10 Gene Ontology Annotation

The complete set of all *Enallagma hageni* proteins was queried against a local NCBI Non-redundant (nr) protein database (10/14/2011) using MPIBlast. The output was saved in `.xls` format (`-m 7 --output-search-stats`), which was then analyzed using Blast2Go without graphical interface (B2G4PIPE) and a local B2G Database [216].

We examined GO term distributions for three partitions of our dataset. First, we derived the distributions of 3rd and 4th level GO term hierarchies for the complete dataset of *Enallagma* proteins. The hierarchical system of the Gene Ontology is represented as a directed acyclic graph in which parent-child relationships describe specific GO terms. That is, parent terms are less specific in their description of a biological function than is their respective child terms. This leads to "levels" within the Gene Ontology structure, with the 1st level containing the broadest categories: Biological Processes, Cellular Components, and Molecular Function. An individual gene may then have many parents and many levels of categorization before reaching the 1st level [217]. Secondly, using *Drosophila melanogaster* as a background dataset, we determined those *Enallagma* genes that were enriched by a hypergeometric distribution test and correcting for multiple tests with FDR under dependency [218, 219]. Finally, we evaluated those *Enallagma* genes that were shown to have undergone either accelerated or reduced rates of evolution, per the branch length rate tests. These genes were examined for their overall GO 3rd and 4th level profile as well as analyzed to determine if any gene was enriched. Enrichment was determined by setting all *Enallagma* genes as a background and using the hypergeometric test with FDR correction mentioned above.

We constructed a hash table for each of the 3 partitions using the annotations from the Blast2GO pipeline. Each gene and that gene’s associated GO accession terms made up the “key:value” relationship, which was then imported into the WeGO web-based program in order to sort the data by GO term hierarchy [220].

3

Machine Learning Methods To Classify The Evolutionary Cause Of Nucleotide Fixations

A.G. Shanku, D.R. Schrider, A.D. Kern

> *"To be ignorant of what occurred before you were born is to remain always a child. For what is the worth of human life, unless it is woven into the life of our ancestors by the records of history?"*
>
> – Marcus Tullius Cicero, *Orator Ad M. Brutum*

3.1 Introduction

Characterizing the genome-wide targets of natural selection is one of the major goals of comparative and population genomics [64, 65, 66, 67, 68, 69, 70, 71]. Determining which genomic changes between species or populations are the result of adaptive versus neutral evolution will improve our understanding of both the targets of natural selection in the genome and modes by which adaptation proceeds in natural populations. At the molecular level, adaptation results from the fixation of beneficial alleles. If one had a reliable method to identify such beneficial fixations one could, in principle, catalog the history of adaptive genetic change in a given species.

Evolutionary genetics, as a field, has been schizophrenic in its characterization of adaptation at the molecular level. On one hand, quantitative geneticists have amply demonstrated that breeding populations respond to selection in proportion to levels of standing genetic variation [221]. On the other hand, the dominant paradigm of adaptation for population geneticists has been the Maynard-Smith and Haigh model of a de novo beneficial mutation rising quickly to fixation [72]. Recently this disconnect has been given more theoretical attention [e.g. 79, 222, 223], however, to date there has been relatively limited efforts to determine what the dominant mode of adaptation might be from genome sequence data [108, 123, 224, 225].

For an individual mutation within a population there are at least three routes to fixation (i.e. its sojourn to frequency one): 1) the mutation may not impact fitness and may drift to fixation by chance (a neutral fixation), 2) the mutation may be beneficial and rapidly fix by the action of natural selection, a process that is sometimes known as a "hard sweep" [e.g. 72], or 3) the mutation may be initially neutral, or nearly so, and drift in frequency until such a time that the environment changes and the mutation then becomes favorable and is quickly swept to fixation (sometimes known as "soft sweeps" [e.g. 73]). Each of these modes of evolution should leave a distinct population genetic signature in local variation surrounding the site that has fixed [79], yet it remains extremely challenging to statistically differentiate between these competing models [e.g. 123].

As a beneficial mutation sweeps through a population, it carries with it its linked genetic background. This "genetic hitchhiking" leads to a local reduction in polymorphism and a skew in the site frequency spectrum (SFS) at linked neutral sites [72, 77]. Accordingly, many population genetic tests for hard sweeps that have been proffered look for an aberrant SFS (e.g. Tajima's *D*; [86]) or use model-based approaches for the expected SFS following a hard sweep [88, 93]. However, many neutral scenarios with non-equilibrium demography, such as population bottlenecks, can create population genetic signatures that are nearly identical to those produced by selection and thus confound searches or scans for selective sweeps [124, 125].

A possible way around the confounding effects of demographic changes in the search for selective sweeps might be to use machine learning methods capable of integrating multiple population genetic summaries. For instance, Pavlidis et al. (2010) used a supervised machine learning approach (Support Vector Machines; SVMs) to integrate a combination of two population genetic summary statistics in an attempt to classify genomic regions as selected or neutral, given simulations from a demographic model used for training [109]. In that study, the authors demonstrated that the SVM approach may have power to detect hard selective sweeps even in the face of non-equilibrium demography. More recently, Ronen et al. (2013) demonstrated that an SVM approach using the entire site frequency spectrum (SFS) of variation at a locus could also successfully identify regions of the genome that had recently experienced hard sweeps. Indeed, both of these SVM-based approaches achieve greater power than more standard population genetic tests currently in use [3, 109].

Here we introduce an SVM framework aimed at classifying individual nucleotide fixations as being the result of either drift, a soft sweep from standing variation, or a hard sweep, and apply our model to whole genome sequencing data from *Drosophila melanogaster* [226, 227, 228]. Our SVM approach is similar in spirit to the approaches of Pavlidis et al. [109] and Ronen et al. [3] in that it uses as training data simulated datasets drawn from demographic models approximating what is believed to be the true demographic model for *Drosophila*. However, rather than utilizing either the SFS or a limited set of population genetic summary statistics for a given focal window of the genome, we use a large set summary statistics that has good power to distinguish among our alternative routes to fixation. We apply our SVM to the well known case of environmental adaptation in the North American colonization of *Drosophila melanogaster* [71, 229] and use recent, deep samples from North Carolina [226] and Africa [227, 228] to polarize changes and identify a set of candidate fixations in the North American lineage. Our classifications allow us to directly examine the frequency of soft versus hard sweeps during this putative bout of adaptation and identify individual candidate loci underlying adaptation in a lineage specific manner.

3.2 Results

Our supervised machine learning strategy for classifying nucleotide fixations is illustrated in Figure 3.1. Briefly, we first seek reasonable demographic parameter estimates assuming a model of population divergence and admixture to describe the history of a North American *Drosophila melanogaster* dataset. For parameter estimation we utilize site frequency spectrum information from North American and African populations and do inference using the software package, ∂a∂i [133]. With those parameters in hand, we then train SVMs on coalescent simulations that condition on either neutral fixations, soft sweeps, or hard sweeps, each in the context of one of three demographic scenarios: 1) constant population size, 2) a strong bottleneck, or 3) a more complex model of population divergence with admixture. Given recent evidence that North American populations might have undergone recent changes in population size, and may represent an admixed sample with African ancestry [134], we focus our attention here on the third demographic scenario, but provide simulation results showing the power of our SVM approach in each case. Finally, we apply our trained SVMs to a large set of nucleotide fixations that we have identified from alignments of population genomic samples from North American, European, and African populations. From this we obtain a single classification for each fixation in the genome - either a neutral fixation, a hard sweep, or a soft sweep. Below, we first assess the power of our method on simulated datasets used for training and testing and then apply this classifier to empirical data.

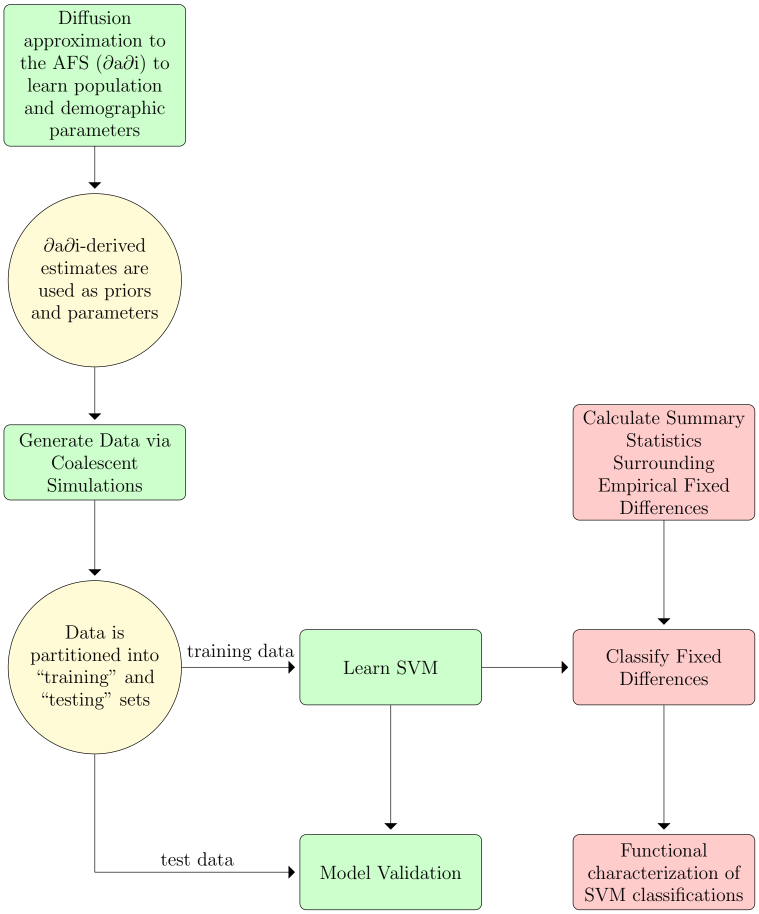

Diffusion approximation to the AFS (∂a∂i) to learn population and demographic parameters
∂a∂i-derived estimates are used as priors and parameters
Generate Data via Coalescent Simulations
Data is partitioned into "training" and "testing" sets
training data
test data
Learn SVM
Model Validation
Calculate Summary Statistics Surrounding Empirical Fixed Differences
Classify Fixed Differences
Functional characterization of SVM classifications

Figure 3.1: **Flowchart Describing The Strategy For Classifying Nucleotide Fixations.** We first learn demographic parameters via diffusion approximations to the allele frequency spectrum (afs) using ∂a∂i. These parameters then, in turn, are used to generate coalescent simulations that serve as training and testing data for our SVMs and are further used to characterize the classifier's performance. We then apply the learned SVM to population genomic data from a North American *D. melanogaster* population.

3.2.1 Demographic History Of *Drosophila melanogaster* Population Samples

We assess the power and accuracy of our classifier under a range of models, including simple equilibrium population size and a strong population bottleneck. However, ultimately we are interested in applying the method to data from recent sequencing surveys of *Drosophila melanogaster*. To this end, we fit two population demographic models using ∂a∂i [133], utilizing joint SFS information from North America and the recent deep sequencing of a population sample from Zambia [226, 228]. As noted above, the population history of N. American *D. melanogaster* is thought to be characterized by both a history of population bottlenecks and recovery as this lineage first left the ancestral range in sub-Saharan Africa and then later upon colonization of the new world, as well as later admixture with African flies [134, 230]. Thus, we focused attention on four demographic models that generally had this structure and compared between them using the Aikake Information Criterion (AIC [231]; see Section 3.4 and Section 3.4.4 for model details).

Generally, all models optimized quite well with multiple runs converging to similar solutions, even those that did not include population admixture between N. American and African lineages. Table E.6 gives parameter estimates, likelihood values, and AIC values for 3 independent optimizations for each of the four models tested. Model choice via AIC supports a model with admixture from Africa into the N. American population as well as constant, bidirectional gene flow between both populations. Our best supported model is in general agreement with [134] who used a different source of African variation, one that potentially has been secondarily admixed with cosmopolitan variation [227], however there are some notable differences. First, unlike parameter estimates from [134], our results support a smaller population size in N. America than in Africa currently. This is well supported by levels of population genetic variation that have been observed for many years (e.g. [232]). Secondly, our best supported model includes constant gene flow since the split of N. American and African populations in both directions, with a slight bias in migration from N. America to Africa. Finally, the estimated admixture event from Africa into N. America is older in our estimates ($\sim$ 2900 ya) than what was assumed by Duchen et al. (2013) and stronger - we estimate the proportion of African ancestry during the admixture to be $prop_{AF \rightarrow NA} = 0.376$. As we fit our demographic model using the joint SFS alone, we were interested in seeing how well our parameter estimates fit other summaries of the data. To this end, we performed posterior predictive simulations using coalescent simulations and examined a set of commonly used population genetic summaries. As can be seen in Figure F.10, our posterior predictive simulations show excellent agreement to the N. American data.

3.2.2 Power Analysis and SVM Accuracy

With representative demographic model parameterizations in hand, we next turned attention to building and training our support vector machine. An SVM is a binary classification algorithm, assigning a label to each example in two classes of data. As such, the multiclass problem that we are faced with here (“hard”, “soft”, and “neutral”) requires that we either, 1) generate multiple one-vs-one, and one-vs-all models, or 2) use some other method of multiclass SVM, such as DAG-SVM [233]. Though we tested both methods with similar results (DAG-SVM not shown), we opted to proceed with a combination of

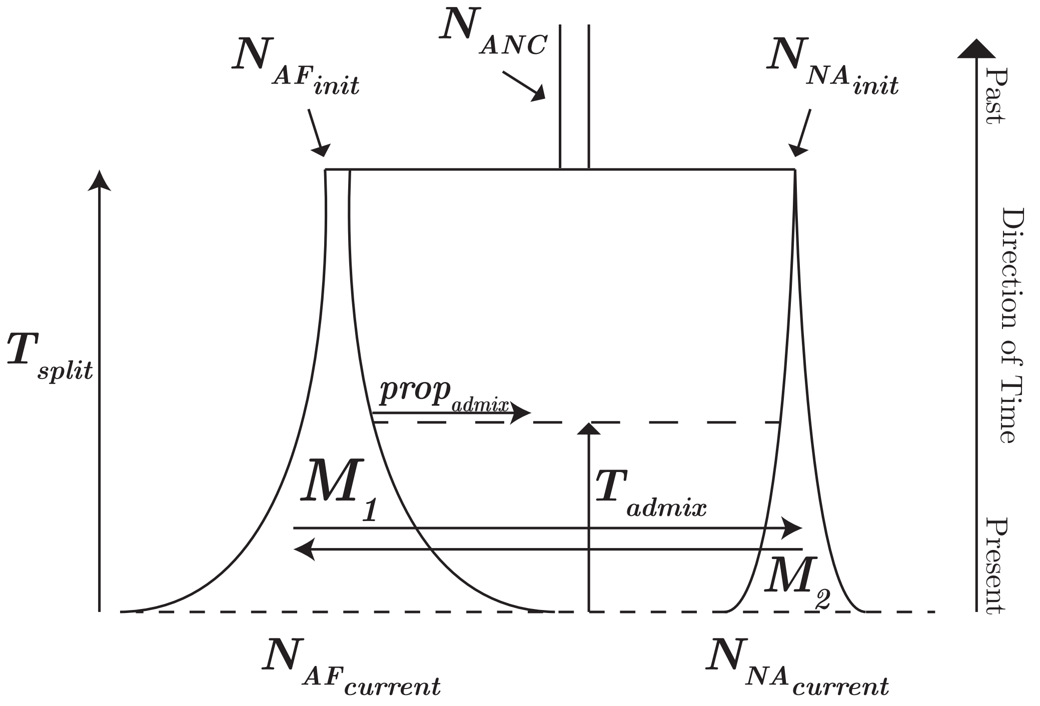


Figure 3.2: **North American *D. melanogaster* Demographic History Described By A Two Population Admixed Model.** We model North American demography using, bottlenecks, exponential growth, admixture, and migration between an ancestral African population and an out-of-Africa cosmopolitan population. Key parameters of the admixture model are shown, but see Table 3.1.

Coalescent Simulation Priors and Parameters	
$\theta = 75.04$	$\rho \sim U(150.0, 450.0)$
$\alpha \sim U(1000, 10000)$	$\tau \sim U(0.0, 0.1)$
$M_{NA \rightarrow AF} = 1.50$	$M_{AF \rightarrow NA} = 1.15$
$T_{NA\_AF_{split}} = 0.242$	$N_{NA_{init}} = 0.082$
$T_{admix} = 0.11$	$prop_{AF \rightarrow NA} = 0.376$
$N_{NA_{current}} = 1.68$	$N_{AF_{current}} = 11.28$
Additional Soft Sweep Prior	
$f_0 \sim U(N^{-1}, 0.2)$	

Table 3.1: **Prior Distributions And Parameters Of Coalescent Simulations For SVM Training**. We used a combination of prior distributions and point estimates to parameterize coalescent simulations which themselves were used in SVM training. Here θ is the mutation rate, ρ is the recombination rate, α is the strength of selection, τ is the time of nucleotide fixation, and f_0 is the frequency of the neutral allele in the population before it underwent selection. $M_{NA \rightarrow AF}$ and $M_{AF \rightarrow NA}$ represent the rates of migration of North America into Africa and Africa into North America, respectively. For the admixture model examined, we used ∂a∂i estimations for each parameter considered. Explanation of the demographic parameters is given visually in Figure 3.2.

one-vs-one SVMs to classify fixed positions in the *Drosophila* data. In particular, for the empirical analysis we used a combination of two classifiers: first, sites were classified as selected vs. neutral using a classifier of Hard & Soft sweeps vs. Neutral fixations, and then subsequently those sites classified as selected were divided into hard and soft using the Hard vs. Soft Sweep SVM (also see Section 3.4.8).

In the next sections we present the results of the following one-vs-one SVM tests: Hard sweeps vs. Neutral fixations, Hard vs. Soft, Soft vs. Neutral, Hard & Soft vs. Neutral, and Hard vs. Soft & Neutral. We demonstrate the power of our classifiers under multiple conditions by generating simulations while varying the strength of selection ($\alpha = 2Ns$) and time of allele fixation (τ) for each of the five binary classifiers. We thus varied our simulations both over a range of fixation times, $\tau = [0.0, 0.05, 0.1]$ (in units of $2N$ generations), and the strength of selection, $\alpha = [100, 500, 1000, 5000, 10000]$. The complete parameterization of our simulations used for SVM training and testing under the admixture model can be found in Table 3.1. For these power analyses we utilize training data sets with 10,000 examples per parameter combination and test our model on test data consisting of 10,000 examples per parameter combination (See Section 3.4.6). For SVM training and testing we summarized patterns of polymorphism in the sample using π, segregating sites (SS), θ_H, Tajima's D, Fay & Wu's H, the number of haplotypes, Kelly's ZnS, ω_c, and windowed-π [76, 86, 92, 234]. ω_c is a modified version of Kim and Nielsen's (2004) statistic, where rather than look for the position that maximizes the value of ω as in the original paper, we only calculate ω centered upon our candidate fixation (thus we call it ω_c; see Section 3.4.3 and Appendix A for further details). In machine learning terms this set of population genetic summaries represents our feature space. Accordingly, this feature space was also used in our empirical analysis (see Section 3.4.6).

3.2.3 Accurate Discrimination Between Modes Of Fixation

In Figure F.11 we show classifier accuracies under constant population size. Once selection is strong, say $\alpha = 1000$, our Hard vs. Neutral classifier for mutations that have fixed in the previous generation ($\tau = 0.0$) is 99.8% accurate and only decays to an accuracy of 80.5% at time $\tau = 0.75$. Further, at $\alpha = 1000$ our ability to distinguish selected fixations (the combination of hard and soft sweeps) from neutral fixations varies from 90.06% ($\tau = 0.0$) to 77.16% ($\tau =$

0.75), and accuracy for the Hard vs. Soft comparison in the same time span rages from 94.19% to 86.5%. Under this demographic history, we have very good power to accurately classify fixations, even reaching back into the past. The specific decays vary according to the strength of selection, however across all but the weakest selection considered our accuracy remains quite good.

Next we considered the accuracy of our one-vs-one classifiers when a strong population bottleneck has occurred in the recent past (Figure F.9), a history known to confound many scan statistics for selective sweeps [124, 125]. Figure F.12 shows the accuracies of our one-vs-one classifiers in the context of a bottleneck. Across the board our accuracies are a bit less than what is seen in the equilibrium population size case (Figure F.11), though even in the presence of a bottleneck we have very good power to distinguish among fixations classes when selection is strong. For instance, at $\tau = 0.0$ and $\alpha = 1000$ we have 99.1% accuracy in the Hard vs. Neutral case, 90.06% accuracy for Hard & Soft vs. Neutral, and 94.19% accuracy for Hard vs. Soft. One thing worth noting is that in the bottleneck case the timing of the fixation matters quite a bit, as those fixations that have occurred while population size was small are much harder to classify accurately. In general however, even with a bottleneck our classifiers are accurate over a wide range of fixation times and strengths of selection.

Finally, in Figure 3.3 we show classifier accuracies under our admixture model (Figure 3.2), which we will apply to *Drosophila* data, below. Again, for stronger and more recent selection, our SVM classifier works quite well. For instance, at $\alpha = 1000$ our ability to distinguish selected fixations (the combination of hard and soft sweeps) from neutral fixations varies from 98.8% ($\tau = 0.0$) to 92.1% ($\tau = 0.05$), only dropping to 64.5% when fixations occur in the past

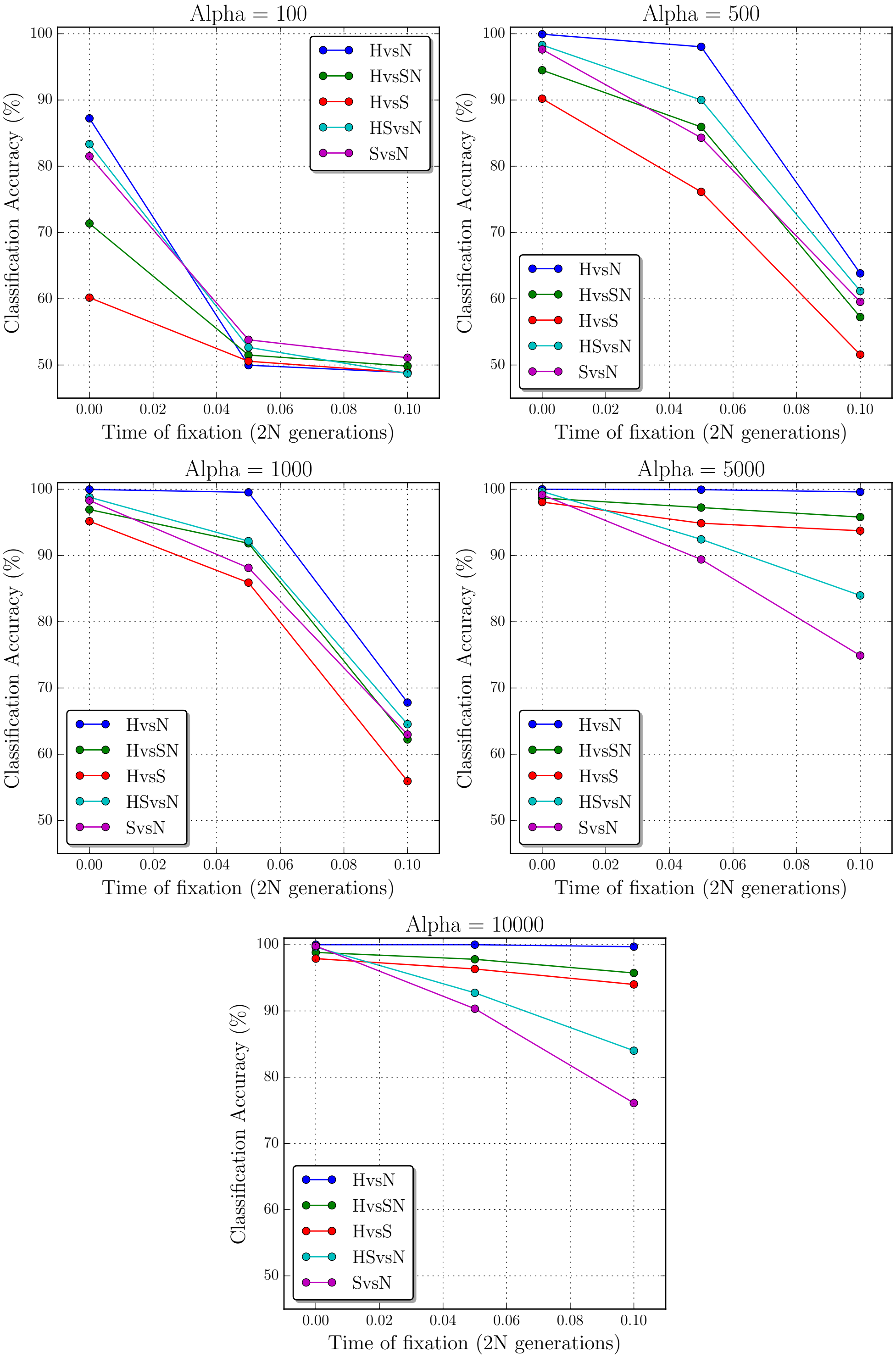


Figure 3.3: **SVM Performance As A Function Of Age Of Fixation.** Here we show accuracy of our set of binary classifiers while varying fixation times and strengths of selection in the admixture model (Figure 3.2).

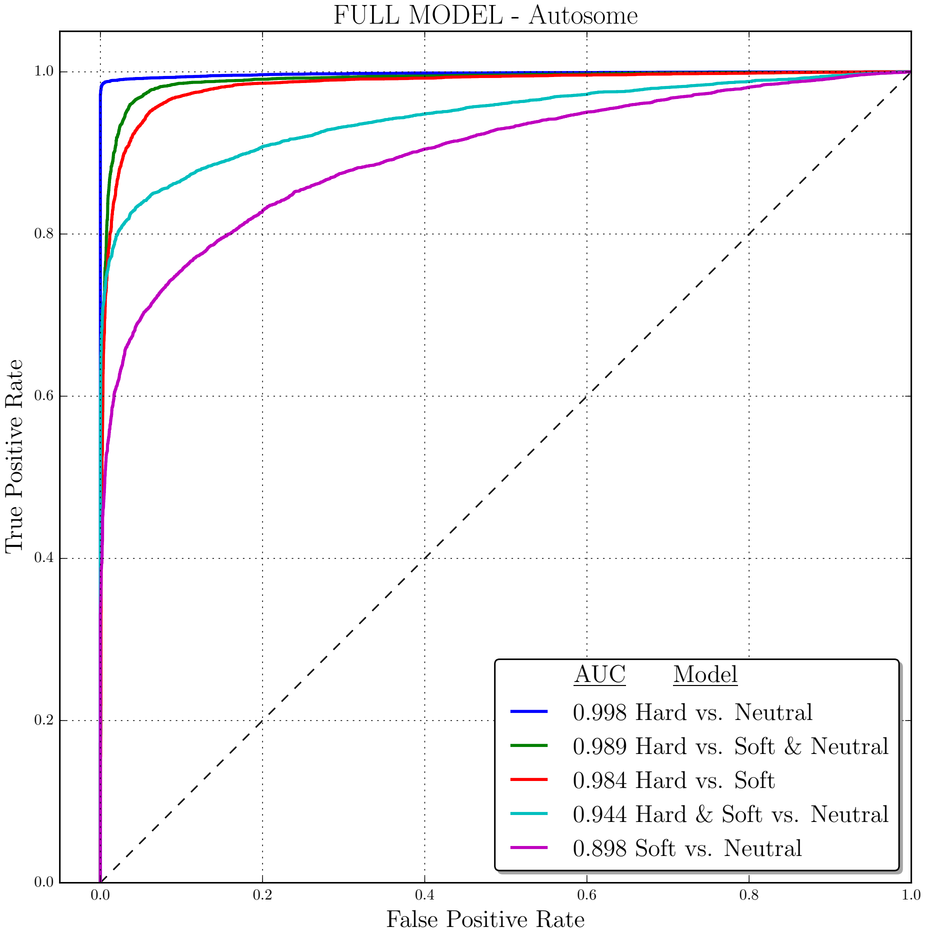


Figure 3.4: **Performance Of The Autosome Classifier.** Here we consider the classification performance of SVM$_{auto}$ under the more realistic scenario where parameters underlying fixations are drawn from distributions (Table 3.1).

at $\tau = 0.1$. As strength of selection increases to $\alpha = 5000$, accuracy for the Hard vs. Soft comparison in the same time span rages from 98.0% to 93.7%. This is an encouraging result, as even under a complex model of demography we should have excellent power to find strongly beneficial mutations, even those that have fixed at some considerable period of time in the past, and to infer whether they were the result of hard sweeps or soft sweeps. Again, the specific decays vary according to the strength of selection, and in all but the weakest levels of selection our classifier's performance remains quite accurate.

There are a few generalities to draw from these tests: 1) our power to differentiate between models is quite good if selection is strong and recent, and decays with increasing time since the sweep or weaker selection; 2) rank order in accuracy among classifiers changes primarily with strength of selection, as seen with the Hard & Soft vs. Neutral relative performance decreasing with α, and, conversely, Hard vs. Soft, and Hard vs. Soft & Neutral increasing with selection; 3) We have very good power to distinguish among fixation classes in the context of a population bottleneck or admixture. (ROC curves and AUC values for all one-vs-one tests for the admixture model can found in Figures F.13 to F.15)

We then considered the accuracy of our SVM under the more realistic scenario in which the parameters underlying a given fixation are drawn from a distribution. For instance, under the soft sweep model we simulate using distributions on the strength of selection, α, the frequency at which the nucleotide comes under selection, f_0, and the time of fixation, τ, and then consider the accuracy of our trained SVM (See Methods). We designate two models of classifiers, SVM$_{auto}$ and SVM$_X$. The former is designed to classify fixations on

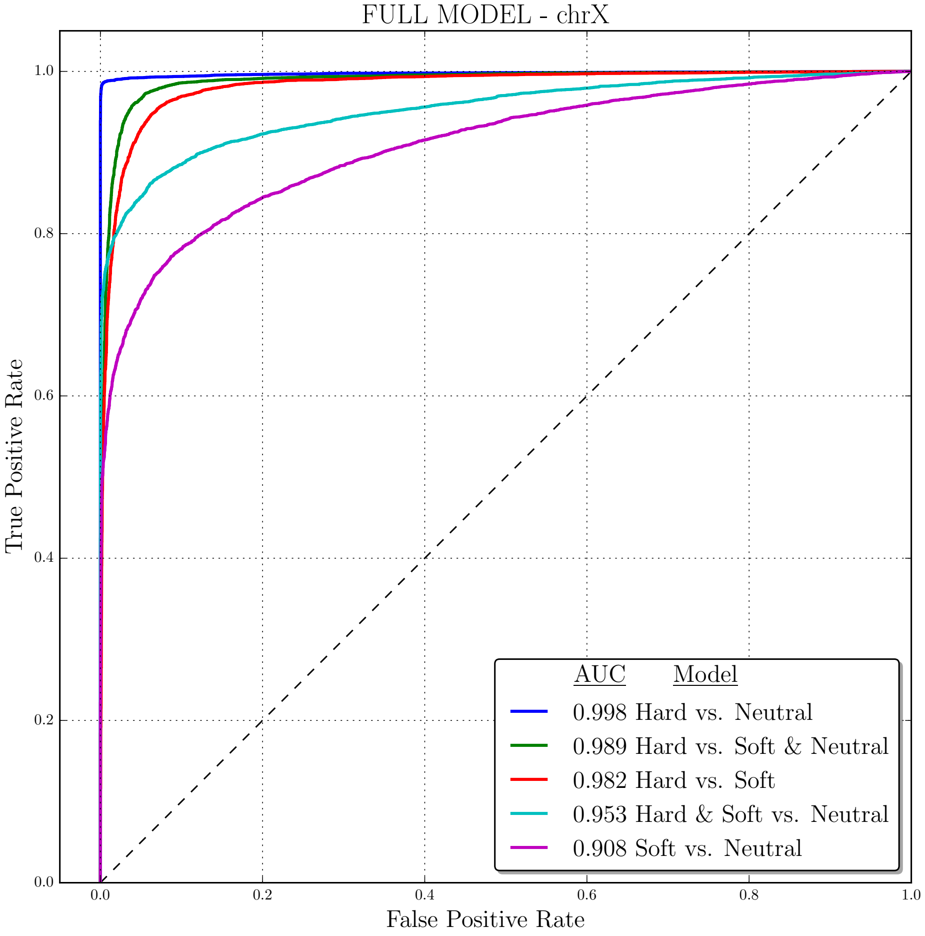


Figure 3.5: **Performance Of The X Chromosome Classifier.** Again, we consider the ROC curve and AUC value as a measure of performance for our classifier, SVM$_X$.

the autosome, explicitly, and is trained on simulations generated using the parameters shown in Table 3.1. SVM$_X$ operates on the X chromosome and was trained by scaling θ, ρ, and both migration rates, M_1 and M_2, by 0.75 (Assuming an equal sex ratio; Table 3.1).

In both classifiers, fixations are allowed to have occurred at a time such that $\tau \sim U(0.0, 0.1)$. We observe that both of these classifiers performed quite well, with very little difference between their performance. Both classification accuracy and ROC/AUC values demonstrate that our SVMs do an excellent job at identifying the correct class across each data set (Figures 3.4 and 3.5). Specifically, the Hard vs. Neutral SVMs were the most accurate classifiers tested, with a classification accuracy of 99.1% and an AUC of 0.998 for SVM$_{auto}$, and 99.0% and an AUC of 0.998 for SVM$_X$ classifier. The SVM$_{auto}$ performed as follows: the Hard vs. Soft & Neutral classifier had an accuracy of 96.1% and an AUC of 0.989. Hard vs. Soft achieved 94.3% classification accuracy and an area under the curve of 0.984. Hard & Soft vs. Neutral was able to classify 89.4% of sites correctly and had an AUC of 0.944. Soft vs. Neutral performed at 82.8% with an AUC of 0.898. The SVM$_X$ examined fixations specific to the X chromosome with the Hard vs. Soft & Neutral classifier having an accuracy of 95.8% and an AUC of .989. Hard vs. Soft performed at 94.2% and and AUC of 0.982. Hard & Soft vs. Neutral achieved 89.6% at an AUC of 0.953. Lastly, the classifier with the most challenging task of separating soft sweeps from neutrality (Soft vs. Neutral) performed at 84.1% with an AUC of 0.908. Again, while we present the power of all five classifiers for completeness, the analysis of empirical *Drosophila* data relies only upon the Hard vs. Soft classifier and the Hard and Soft vs. Neutral classifier.

3.2.4 SVM Comparisons

The features used to train the SVM consist of a collection of summary statistics, both of the SFS and the haplotype structure We sought to compare our SVM's performance (specifically the SVM_{auto} classifier) to 1) a classifier which utilizes only those statistics that summarize the number and frequency of derived alleles as features (SVM_{sfs}), and 2) a classifier based on the Ronen et al. [3] linear transformation of the SFS which here are used as features (SVM_{Ronen}; [3]). We find that our SVM performs at least as well, and at times better, than the two tested alternatives (Table E.7). In the Hard vs. Neutral test, the SVM_{auto} achieves an accuracy and AUC of 99.2% and 0.998, respectively. The Ronen and SVM_{sfs} classifiers perform similarly, with accuracies of 99.0% and 99.0% and AUC values of 0.998 and 0.997, respectively (Figure F.16). All three classifiers perform well in the Hard vs. Soft & Neutral test; SVM_{auto} obtains an accuracy and AUC of 96.1% and 0.990, respectively, versus SVM_{Ronen}'s 96.1% and 0.990, and the SVM_{sfs}'s 95.9% and 0.988 (Figure F.16). Again, we find that the three classifiers perform similarly in the Hard vs. Soft case. Here, our SVM classifies 94.4% of the test data correctly and has an AUC of 0.983. SVM_{Ronen} performs well, with an accuracy of 94.3% and an AUC of 0.983. Performance of the SVM_{sfs} classifier is nearly equal, classifying 94.0% of test points correctly and achieving an AUC of 0.981 (Figure F.16). We see that our classifier begins to perform appreciably better than the alternatives when we test Hard & Soft sweeps vs. Neutral fixations. In this test, the SVM_{auto} classifier is correct in 89.2% of classification tasks and has an AUC of 0.945. SVM_{Ronen} has an accuracy of 86.6% and an AUC of 0.928, while the SVM_{sfs} classifies 87.5% correct and has an AUC of 0.935 (Figure F.16). In our last test, Soft vs. Neutral, we see the most noticeable difference amongst the classifiers. Our SVM$_{auto}$ classifier performs best with an accuracy of 83.2% and an AUC of 0.901, compared to SVM$_{Ronen}$ with 78.6% correct classifications and an AUC of 0.863, and SVM$_{sfs}$ with an accuracy of 80.3% and an AUC of 0.876 (Figure F.16).

3.2.5 Misclassification Due To Linked Selection

A serious concern with classification of genomic windows centered upon fixations is that misclassification could occur as a result of linked selection. Indeed, we have found elsewhere that population genetic statistics aimed at identifying soft or partial sweeps can produce spurious signals of selection when linked hard sweeps have occurred [4]. To examine to what degree this would be a problem for our SVM classifier, we used simulations of chromosomal regions linked to a hard sweep at various recombination distances (Section 3.4.9) and examined how our trained SVMs would classify neutral regions subject to linked selection. Figure 3.6 shows misclassification rates of the SVM$_{auto}$ classifier as a function of increasing recombination distance from the site of a hard sweep (given in units of r/s). We can see that at shorter genetic distances, $r/s = 0.1$, neutral loci are getting misclassified as a selective sweep 89.7% of the time. Of these incorrect classifications at $r/s = 0.1$, 68.1% are called hard sweeps. In the next window of r/s, and in all subsequent windows, soft sweeps represent the great majority of misclassifications. Soft sweeps represent $> 75\%$ of the misclassified neutral sites for $r/s = 0.2$, and when $r/s > 0.3$, more than 97% of misclassifications are due to erroneous soft sweep calls. Thus, our model is likely to misclassify neutral sites linked to hard sweeps as soft selective sweeps - an example of the "soft shoulder" effect [4].

To ameliorate this effect in our analysis of empirical data, we could limit

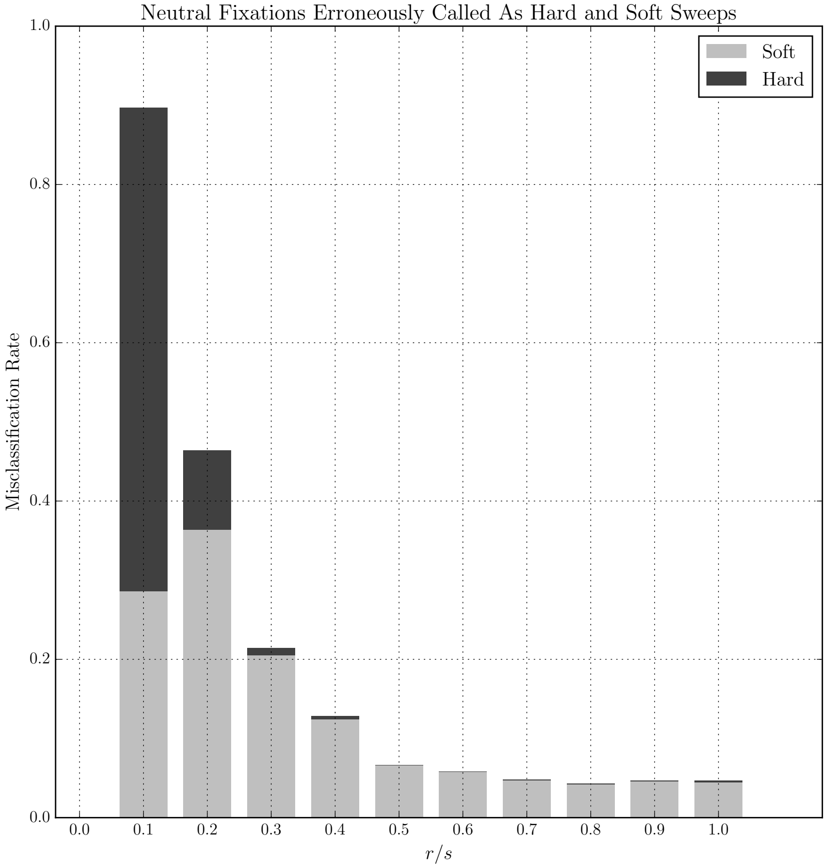


Figure 3.6: **Erroneous Soft Sweeps**. To determine how neutral fixations that neighbor hard sweeps are misclassified, we generated 10kb windows with r/s ranging from 0.1 to 1.0 that contained only neutral fixations. These windows were classified using our standard SVM method (Section 3.4.9). Bins with $r/s \geq 0.1$ again contain only neutral fixations, and rates of misclassification vary with r/s. It is apparent that at all ratios of r/s, save for $r/s = 0.1$, the majority of misclassifications are called as soft sweeps. Allowing for 5% misclassification, we determined a cut off value of $r/s \geq 0.482$. Using these values and the recombination rates across the *Drosophila melanogaster* genome given by Comeron et al. 2012, we removed sites which were classified as soft sweeps if they fell within a determined distance from neighboring hard sweeps (See Sections 3.2.5 and 3.4.9).

our calls of soft sweeps to only those that occur beyond a given genetic distance from the nearest hard sweep, having assumed something about the strength of selection. For instance when $r/s = 0.482$, 95% of windows were correctly classified. Assuming our estimated $N_{NA_{current}}$ =341,745 and an average strength of selection $\alpha = 2000$, this corresponds to a genetic distance of 0.141cM. We used this genetic distance cutoff in conjunction with recent fine-scaled estimates of the recombination rate along the *Drosophila melanogaster* genome [235] to filter soft sweep calls neighboring hard sweeps. Thus, the physical distance of the "buffer zone" between a called hard sweep and the next considered soft sweep will change across the genome, reflecting local variation in the recombination rate.

3.2.6 Application To *Drosophila* Data

To examine adaptation in the North American *Drosophila melanogaster* population, we partitioned derived fixations, relative to African *Drosophila melanogaster*, into two sets "filtered" and "unfiltered" (Section 3.4.2). First, we looked for sites that shared the derived fixation in the N. American sample [226], and were either fixed for the ancestral allele or segregating in the African *Drosophila. melanogaster* population at some frequency ("unfiltered"). We further constrained the set of fixations such that the derived allele frequency in the African population was present at $<= 0.5$ ("filtered"). We found 115,257 such fixed positions present in the unfiltered set (33,685 and 81,572 on the X and autosomes, respectively), and 3413 fixed positions specific to the filtered set (2720 and 693 on the X and autosomes, respectively). These sets of fixations do not include sites found in non-recombining regions [235] or where positions were located within areas of missing data (those windows where more than 35% of the bases in the alignment are "N").

Using Fisher's Exact Test, we asked whether certain annotation categories were enriched for fixations in either the filtered or unfiltered fixations. Starting with the unfiltered set, we found significant enrichment ($p < 0.005$) of introns, mRNAs, genes, transcription factor binding sites, and ncRNAs (Figure F.17). We found that the X chromosome was enriched for the presence of fixations, relative to the whole genome (fold enrichment = 1.55, $p \leq 0.005$), while the autosomes were depleted for fixations (fold enrichment = 0.87, $p \leq 0.005$). The filtered set of fixations was also enriched ($p < 0.005$) for genes, mRNAs, introns and transcription factor binding sites (Figure F.18). As is the case with the unfiltered set, the X chromosome houses a significant number of fixed positions in the filtered set as well (fold enrichment = 4.23, $p \leq 0.005$).

For each fixation in both the unfiltered and filtered sets, we calculated our selection of summary statistics (Section 3.4.3) in 10kb windows centered around the focal site for use as input to our classifiers.

3.2.7 Recent Adaption Of North American *D. melanogaster* Was Driven Primarily By Soft Sweeps

We next used our two trained classifiers (SVM_{auto} and SVM_X) to classify 10KB windows surrounding fixations of each respective type. In the unfiltered set of fixations, our SVM_{auto} classified 13,338 fixations on the autosome as putative sweeps (1014 hard sweeps and 12,324 soft sweeps), and SVM_X called 9234 fixations on the X chromosome as sweeps (3844 hard and 5390 soft sweeps) after filtering putative misclassifications due to the soft shoulder effect (Table 3.2). This amounts to 80.42% of fixed positions being identified as neutral, 4.21%

as hard sweeps, and 15.37% as soft sweeps. In the set filtered by African derived allele frequency, we classified 311 fixations on the autosomes as putative sweeps - (39 hard and 272 soft sweeps), and 1275 fixations called as sweeps by SVM_X - (656 hard sweeps and 619 soft, again, after filtering for shoulder regions) (Table 3.2). Here 53.54% of our called fixations are identified as neutral, 20.36% as hard sweeps, and 26.10% as soft sweeps.

We were first interested in seeing if our predictions recovered previously identified targets of sweeps. In the unfiltered set, we found numerous genes overlapping our classified hard sweeps that had been previously reported, most notably a massive hard sweep at *Cyp6g1*, a gene associated with pesticide resistance [236, 237, 238, 239], and a closely linked methyltransferase gene *pimet (Hen1)* [238, 239]. Further, we find many other putative targets of hard sweeps, such as: *unc-119*, which is associated with adaptation in both European [240] and North American flies [239]; *kirre* [241]; *ras*, a protooncogene involved in the synthesis of guanine nucleotides and signal transduction by cell surface receptors [239, 242, 243]; and *par-1* [239].

We also recover examples of putative soft sweeps that have previously been noted in the literature, including: *acxC*, a gene important to spermatogenesis [227, 244], and *dally*, which is associated with life span and fecundity [245].

Looking for previously identified sweeps in the set of fixed differences filtered by derived allele frequency, we find that many of the sweeps listed above are absent. However, called sweeps that were not removed by filtration include *Cyp6g1* [236, 237, 238, 239], *unc-119* [239, 240], and *kirre* [241]. In addition to successfully recovering known sweeps, we also identify a rich set of novel candidates. See Tables E.15 to E.18 for complete lists of genes classified as having undergone hard and soft sweeps.

Fixed Difference Class	Sweep Type	Location	Sweep Count	Fixed Difference Count	Proportion of Fixed Differences Called as Sweeps
Filtered	Hard	Autosome	39	693	0.0562
		ChrX	656	2720	0.2411
	Soft	Autosome	272	693	0.3924
		ChrX	619	2720	0.2275
		Total	1586	3413	0.4646
Unfiltered	Hard	Autosome	1014	81572	0.0124
		ChrX	3844	33685	0.1141
	Soft	Autosome	12324	81572	0.1510
		ChrX	5390	33685	0.1600
		Total	22572	115257	0.1958

Table 3.2: **Classified Hard And Soft Sweeps** Here we show the distributions of hard and soft sweeps as called by their respective classifiers. The right column shows the fraction of fixed differences, either filtered or not, and either occurring on the autosome or X chromosome, called as sweeps.

3.2.8 Genomic Distribution Of Hard And Soft Sweeps

We next asked if hard and soft sweeps were distributed uniformly across the genome. In the case of those called sweeps determined in the unfiltered set of fixed differences, we discovered that the X chromosome was highly enriched for the presence of hard sweeps (fold enrichment = 4.2, $p \leq 0.005$) and enriched for soft sweeps (fold enrichment = 1.65, $p \leq 0.005$) Further, conditioning on being a fixed position, we tested if there existed a preference for hard or soft sweeps at those locations. We found that fixed positions on the X chromosome were enriched for both hard sweeps (fold enrichment = 2.71, $p \leq 0.005$) and for soft sweeps (fold enrichment = 1.04, $p \leq 0.005$).

The genomic distribution of sweeps in the filtered set followed a similar pattern to those in the unfiltered set, above. However, we found that fixations on the autosomes were enriched for soft sweeps (fold enrichment = 1.5, $p \leq$

0.005), whereas fixations on the X chromosome were enriched for hard sweeps (fold enrichment = 1.18, $p \leq 0.005$) These results echo prior research showing that the X chromosome may play host to an excess of selective sweeps [246, 247, 248].

We next examined how fixations that were classified as either hard or soft were distributed across genetic elements, and utilized a permutation test (See Section 3.4.10) to determine if any genetic elements were enriched for these sweeps. We find that a number of elements were enriched for either the presence of hard sweeps or soft sweeps. In the set of unfiltered fixations we find six elements enriched for hard sweeps: CDS ($p < 0.005$), Exon ($p < 0.001$), transcription factor binding sites ($p < 0.001$), Gene ($p < 0.001$), 3′ UTR ($p < 0.001$), and 5′ UTR ($p < 0.001$). Those same elements were also enriched for the presence of soft sweeps, and further included: mRNA ($p = 0.025$) and snoRNA ($p = 0.036$). In the set of filtered fixations, we found that two elements were enriched for hard sweeps: Exon ($p = 0.028$) and transcription factor binding sites ($p < 0.001$). No elements were enriched for the presence of soft sweeps in the filtered set. Inasmuch, our sweep calls are highly enriched for annotated functional regions of the genome as would be expected a priori under most models of adaptation.

We then considered if adaptive fixations showed any clustering with respect to function by using the DAVID annotation tool [249, 250]. We found that unfiltered hard sweeps were enriched for many annotation clusters, including a differentiation and morphogenesis GO cluster (enrichment score $(e.s.) = 6.2$), with neuron differentiation, neuron development, axonogenesis, and others being significantly enriched ($p < 0.001$, Benjamini corrected). We also found a second GO cluster associated with development and morphogenesis ($e.s = 5.1$)

that possessed multiple enriched terms, including post-embryonic organ development and appendage development ($p < 0.005$).

Unfiltered soft sweeps were enriched for multiple clusters as well, including a morphogenesis and development cluster ($e.s. = 4.6$) where the GO terms post-embryonic organ development, wing disc morphogenesis, and wing disc morphogenesis were enriched ($p < 0.005$, Benjamini corrected). We also discovered a "binding" cluster ($e.s. = 4.2$) where nucleotide binding and atp binding were enriched ($p < 0.005$).

Interaction	Sweep Class	#Interactors	Mean # Permuted Interactors	Enrichment	p-value
Flybase genetic interactions	Both	659	427.82	1.54	p = 0.0
	Hard	61	47.19	1.29	p = 0.13
	Soft	344	210.95	1.63	p = 0.0
RNA-gene interactions	Both	42	122.49	0.34	p = 1.0
	Hard	0	0.926	0.0	p = 1.0
	Soft	37	105.9	0.35	p = 1.0
TF-gene interactions	Both	814	616.24	1.32	p = 0.033
	Hard	168	99.38	1.69	p = 0.069
	Soft	514	399.6	1.29	p = 0.11
Flybase other physical interactions	Both	412	213.89	1.93	p = 0.0
	Hard	35	17.41	2.01	p = 0.007
	Soft	221	118.67	1.86	p = 0.0
yeast two-hybrid	Both	299	165.15	1.81	p = 0.0
	Hard	6	8.89	0.67	p = 0.86
	Soft	213	98.16	2.17	p = 0.0
co-affinity purification	Both	1447	627.52	2.35	p = 0.0
	Hard	71	33.69	2.11	p = 0.005
	Soft	1008	391.99	2.57	p = 0.0

Table 3.3: **Interactions In Selective Sweeps Called From The Unfiltered Set Of Fixed Differences.** Here we test for an excess of interacting pairs of genes both experiencing selective sweeps (Section 3.4.10)

We examined the filtered hard and soft sweep calls, as well. While we found no significant hard clusters, filtered soft sweeps were enriched for reproductive developmental and formation cluster (enrichment score ($e.s.$) = 2.8). Within this cluster, GO terms including reproductive developmental processes, female gamete generation, and oogenesis were significantly enriched

($p < 0.05$, Benjamini corrected). See Tables E.11 to E.14 for complete enrichment results from the DAVID analysis.

We examined classified sweeps that occurred in coding regions of the genome in an effort to determine the prevalence of synonymous and nonsynonymous adaptive fixations. Using all fixed positions in coding regions as a background, we used a Fisher's Exact test to first discern if sweeps were enriched for either synonymous or nonsynonymous substitutions. We found that the set of unfiltered hard sweeps was enriched for synonymous substitutions (fold enrichment = 1.026, $p \leq 0.05$). There were no enrichments for either hard or soft sweeps in the filtered set.

We also examined whether nonsynonymous adaptive fixations were enriched for either radical or conservative amino acid changes. Radical changes involve the transition between amino acids that either possess a different charge or different polarity, and as such might significantly alter the structure and function of its resulting protein. This implies that these changes might be subject to stronger purifying selection and less prevalent than conservative nonsynonymous changes. However, those radical changes that do occur may be indicative of positive selection [251, 252]. Here we find that neither hard sweeps nor soft sweeps are enriched for radical amino acid fixations in either the filtered or unfiltered set.

Finally, we looked for the presence of a change from unpreferred codons to preferred codons at synonymous sites in our called sweeps. We found that in the set unfiltered hard sweeps, 66.81% of synonymous substitutions were from an unpreferred to a preferred codon, a significant enrichment of sweeps when using the set of synonymous fixations as a background (fold enrichment = 1.06, $p \leq 0.01$). Further, we found that only 9.52% of synonymous substitutions in the unfiltered hard sweeps were from preferred to unpreferred codons, relative to all synonymous fixations, (fold enrichment = 0.728, $p \leq 0.0005$). Conversely, we found the opposite pattern in the set of unfiltered soft sweeps. Here, only 61.94% of synonymous substitutions were of the unpreferred to preferred type, a significant depletion relative to the background of all synonymous fixations (fold enrichment = 0.98, $p \leq 0.05$). Again, opposite to what we observed in the hard sweeps, we find soft sweeps are enriched for preferred to unpreferred codon usage (fold enrichment = 1.06, $p \leq 0.05$). When looking at the filtered set of called sweeps, neither the hard or soft sweep sets were significantly enriched or depleted.

3.2.9 Selective Sweeps Disproportionately Affect Genes Interacting With One Another

In order to investigate whether positive selection has recurrently acted on the same pathways or multiprotein complexes, we also asked whether interacting gene pairs were enriched for selective sweeps. In particular, we counted the number of pairs of genes that interact with one another and that each contained at least one classified sweep. We then compared the number of observed pairs of interacting genes both experiencing a sweep in our true classifications to the corresponding numbers observed in permuted data sets (Section 3.4.10). We performed this test on several types of interactions for which data are available in the Drosophila Interactions Database (http://www.droidb.org; [253, 254]), including genetic interactions from FlyBase [255], physical interactions from FlyBase [255], transcription factor-gene interactions [256, 257], microRNA-gene interactions [257, 258, 259, 260], protein-protein interactions from a yeast two-hybrid experiment [261], and protein-protein interactions from a co-affinity and purification and mass spectrometry experiment [262].

Strikingly, in the unfiltered set of fixations we observed a significant excess of interacting pairs of genes that both experienced sweeps for each type of interaction, except for miRNA-gene interactions (Table 3.3). Hard sweeps alone showed a significant excess two types of interactions, whereas soft sweeps alone showed excess in four categories. With regards to hard sweeps having only two significant categories, this may be due in part to lower power as we made fewer hard sweep classifications. In any case, this result suggests that genes that have experienced recent selective sweeps are more likely to have interacting partners that have also recently acquired adaptive fixations. (See Table E.8 for filtered fixation interactions.)

3.2.10 Novel Candidate Genes In Sweeps

In addition to those recovered sweeps discussed above, we find many heretofore unknown instances of genes with hard sweeps in our unfiltered set, including *ken*, a gene involved in genital formation [263], *sh (Shaker)*, a gene involved in sex pheromone discrimination [264], *Prp8*, involved in mRNA binding and splicing [265], *Neto*, associated with flight, hatching behavior, and locomotion [266], and *yippee*, a gene associated with *Drosophila melanogaster* wing shape [267]. See Table E.15 for a complete list of hard sweeps.

We found a number of putative soft sweeps that to our knowledge have not been reported as such in the literature and show the full results of called soft sweep genes in Table E.16 . A sample of these genes include: *unc79* which has been shown to critical in circadian locomotor rhythms [268], *her*, a gene known to play a role in sex determination [269], and *ihog (interference hedgehog)*, with roles in *hedgehog* family protein binding [270], compound eye development, and eye photoreceptor cell differentiation [271].

While the filtered set contains fewer hard and soft sweeps than found above, amongst others we retain *sh (Shaker)*, *yippee*, *Neto*, and *her*. See Tables E.17 and E.18 for a complete list of hard and soft sweeps found in the filtered set of fixations.

Contributions

In this preceding section, I was responsible for all steps in the analysis with the exceptions of demographic parameter estimation in Section 3.2.1, which was performed by A.D.K, and the permutation tests found in Section 3.2.8 and Section 3.2.9, which were carried out by D.R.S. I generated all figures (with the exception of Figure F.10, A.D.K), tables, and contributed in writing the draft of the manuscript, along with D.R.S and A.D.K.

3.3 Discussion

Understanding the genetic basis of adaptation in natural populations is a central goal for evolutionary and population genetics. However, methods to detect and localize the targets of natural selection in the genome still lag in their ability to deal with non-stationary demographic processes [109, 124, 125]. Here we use a combination of population genomic data and machine learning methods to find regions of the genome that have experienced recent sweeps in a manner that is demonstrably robust to demographic history. Our SVM method classifies nucleotide fixations into one of three groups on the basis of surrounding nucleotide variation: those that are neutral, those that are due to a soft sweep, and those that are due to a hard sweep. Application of our method to North American *D. melanogaster* genomic data thus allows us to ask which mode of adaptation has been more frequent in the history of this population.

Our method of determining a fixation's history was tested under multiple scenarios designed to examine the model's power as 1) the strength of selection varied, 2) as the time of the fixation varied, and 3) as the demographic history of the population varied. When we examined how these three factors contributed to our misclassification rate we found that, as a whole, performance was moderately affected by the time at which the fixation occurred, but was accurate for the more recent times associated with North American colonization by *D. melanogaster*. Accuracy was more affected by the strength of selection (Figure 3.3) such that power to detect weakly beneficial fixations is quite low across all demographic scenarios. Nevertheless, we determined that our trained SVMs, in which we include an explicit demographic history, and integrate over ranges of selection coefficients and fixation times, were powerful and robust. Indeed, we demonstrate that our method has very good power to detect recent hard and soft sweeps (Figures 3.4 and 3.5).

Additionally, we found that our classifier performed better than both an SVM using only site frequency spectrum summary statistics as features, as well as Ronen's SVM utilizing a linearly transformed SFS [3] (Table E.7). This suggests that our classifier, which incorporates not only SFS data, but also utilizes linkage disequilibrium-based summary statistics, captures more information contained within the population sample, resulting in better performance, especially in cases where the classification tasks were more difficult (i.e. - Hard & Soft vs. Neutral, and Soft vs. Neutral; Table E.7). It is worth noting that both Ronen et al.'s methods and our own are quite effective at detecting selection even in the face of demography, thus demonstrating the unique power of the supervised machine learning approach.

In applying our method to North American samples of *D. melanogaster*, we are able to study adaptation associated with the recent colonization of a new environment [229, 272]. This analysis yields several conclusions: at the broadest scale, we find that the vast majority of fixations (80.42% and 53.54%) in the unfiltered and filtered sets, respectively, are classified as neutral fixations. Among those sites that we determined to be adaptive fixations, we find that soft sweeps represent the majority in both comparisons. Among North American adaptive fixations in the unfiltered set, soft sweeps are ~ 4 times as frequent as hard sweeps across the genome (Table 3.2). In the filtered set we observe that soft sweeps are ~ 1.25 times more numerous than hard sweeps. Given the very recent timing of North American colonization by *D. melanogaster*, it is perhaps to be expected that selection from standing variation would be the dominant mode of adaptation, as there has been little time for new beneficial mutations to appear in the nascent population. On theoretical grounds there is good reason to believe that in the first phase after an environmental shift selection from standing variation should dominate selection on de novo mutations [73, 79]. Only later, after all such newly beneficial standing variation has been exhausted should de novo beneficial mutations contribute greatly to adaptation. Our observations concerning the percentage of adaptation from soft sweeps would be consistent with this idea.

Estimates of the proportion of adaptive fixations in Drosophila using comparisons of polymorphism and divergence have revealed abundant positive selection in both the protein-coding portion of the genome and the non-coding portion (e.g. [273, 274, 275]). For instance, Keightley and Eyre-Walker (2012) use an extension of the McDonald-Kreitman test to show that a full 57% of amino acid replacements are adaptive. Similarly, Andolfatto (2005) used contrasts of polymorphism and divergence to show that a large proportion of intergenic, UTR, and intronic fixations are adaptive (20-70%). Our estimates for North American fixations are well in line with this observation, where we find that at least 19.58% of fixations are adaptive genome-wide. Comparing across genomic annotations (Tables E.9 and E.10), we see that an even higher percentage of amino acid fixations are adaptive (30.8% and 55.5%). As our estimates are based only on patterns of within population variation, our results should be taken as independent confirmation that natural selection drives a large percentage of nucleotide substitution. Here our estimates may be conservative, in that we have reduced power to detect sweeps going back in time, so we will be missing many sweeps that were associated with the out-of-Africa migration.

However, despite the accuracy and power of our classifier, our estimates of the numbers of soft or hard sweeps are undoubtedly only approximate. There are two reasons for this: the first is that when multiple fixations occur in a window classified as a sweep (hard or soft), often more than one fixation will share the class designation, as a result of linked selection broadly influencing patterns of genomic variation. A second, but related problem is that neutral sites closely linked to hard sweeps often are misclassified as soft sweeps [4]. Using coalescent simulations we have shown that at small genetic distances this problem can be quite dramatic but at greater genetic distances misclassification due to linked selection is no longer a problem. This "soft shoulder" effect is not just a problem for the present SVM approach but also other summary statistics aimed at detecting soft or partial sweeps [4]. We see this effect in our empirical analysis as before filtering out soft sweep calls near hard sweeps in the unfiltered and filtered Drosophila genome we find that 86.8% and 88.8%, respectively, of soft sweeps occur within 10 kb of hard sweeps, as expected from our simulations. We thus only consider those soft sweep calls that are at a considerable genetic distance away; we selected a genetic distance that would allow for a 5% nominal misclassification rate assuming moderately strong selection $\alpha = 2000$ (See results). Using the cutoff of $r/s = 0.482$ we recover 17,714 soft sweeps in the unfiltered set and 891 soft sweeps in the set filtered by derived allele frequency (Methods). Of course our assumption about the average strength of selection of sweeps is purely speculative. If we instead assume that the strength of selection is weaker, say $\alpha = 500$, and an $r/s = 0.3$, we end up with 32,871 soft sweep calls. If we assume selection is much stronger on average, say $\alpha = 15000$, and use a distance cutoff of $r/s = 0.6$, we would call only 2421 soft sweeps (Figure F.19). With the true strength of selection across the genome unknown, we are undoubtedly classifying some fixations as soft sweeps as a result of linked selection, however our results strongly support selection on standing variation as being the primary mode of adaptation even if the average selection coefficient associated with hard sweeps is very strong.

While we have modeled soft sweeps throughout this report as selection from standing variation, an unknown percentage of our called soft sweep regions may be the result of recurrent mutation of a beneficial allele [c.f. 121]. Indeed, there is evidence from the *Ace* locus that population sizes in *Drosophila melanogaster* might be large enough to support such a phenomenon [276]. Using similar SVM classifiers to what we have used here, we previously have shown that selection on standing variation closely mimics selection on recurrent mutation, thus our set of called soft sweeps may include sweeps of both types. While this may be so, it seems reasonable to think that the series of population bottlenecks associated with North American colonization history may limit the frequency of recurrent mutation.

A complete understanding of adaptation would include the genomic regions subject to selection as well as the mode of selection. By focusing on fixed differences among populations, we focus our search for adaptive changes on those regions that should be the most likely candidates a priori [277, 278, 279]. We found statistical over-representation of both hard and soft sweeps in coding sequences, 3′ UTRs, exons, and transcription factor binding sites. Thus, both coding and non-coding portions of the genome are putatively involved in the adaptive response to the colonization of North America [96, 239, 275]. In addition, we found evidence for an enrichment of synonymous substitutions classified as hard sweeps in a manner that might be suggestive of adaptation at the level of transcriptional or translational activity, which is consistent with recent observations of local adaptation in levels of transcription [280].

We observed a much higher rate of adaptive evolution on the X chromosome than the autosomes in our sets of unfiltered and filtered out-of-Africa fixations. These findings are consistent with population-genetics theory, which predicts a “fast-X” effect when adaptation occurs primarily through new mutations which are at least partially recessive [281]. Although numerous studies have produced evidence in favor of a “fast-X” effect in *Drosophila* [70, 246, 282, 283, 284, 285, 286], these results have remained controversial due to a lack of observed “fast-X” in other studies [287, 288, 289]. Our own findings strongly support the notion of a “fast-X” effect as the X chromosome is enriched above autosomes for hard sweeps, although not soft sweeps, in our filtered data. Theory would predict that hard sweeps should be differentially aided by the dominance effects of the X if soft sweeps often act on segregating recessive mildly deleterious variants [222], which should be underrepresented on the X. Indeed, we see such a pattern in our data as the X contains 79% of our identified hard sweeps but only 30% of soft sweeps.

While genetic adaptation is often caricatured as allelic replacement at a single locus, a century of quantitative genetics has shown us that most traits are highly polygenic and influenced by alleles segregating at many loci. Natural selection over the short term may thus effect standing variation at a constellation of loci which underlie an adaptation [290]. In this analysis we found strong evidence of coordinated soft sweeps occurring at loci that interact with one another but weaker evidence for interacting hard sweeps. If indeed polygenic adaptation from standing variation were frequent in the colonization of North America such a pattern would be the expectation, although it is clear that in many cases polygenic adaptation may not proceed via fixation of alleles rather than coordinated changes in allele frequency at many loci. Thus, we may be observing a distinct but related phenomenon: where genes linked to fitness in pathway-like networks may undergo cascades of selective sweeps as a population moves towards a fitness optimum.

3.4 Methods

3.4.1 Genome Sequence Data and Alignments

We made use of North American (North Carolina), European (France), and African (Siavonga, Zambia) population genomic datasets of *Drosophila melanogaster* that had been collected previously by the Drosophila Genetic Reference Panel (DGRP) consortium and the Drosophila Population Genomics Project (DPGP) [226, 227, 228]. We obtained aligned datasets from the Drosophila

Genome Nexus resource (v1.0; [228]), and subsequently filtered from those alignments regions that showed strong identity-by-decent (IBD) among lines using the `ibd_mask_seq.pl` script provided with the alignments. Note that we did not filter regions flagged as potentially admixed between cosmopolitan and African lines, as we were interested in inferring a demographic model that included admixture. This dataset thus consists of two deeply sampled populations, yielding sample sizes of $n = 197$ genomes from African lines and $n = 205$ genomes from N. American lines, and a much more shallowly sampled population from Europe, $n = 7$.

3.4.2 Candidate Fixed Positions

Using an ancestral sequence reconstruction described earlier [238], we generated two sets of candidate fixations - one set in which fixations were filtered for sites where the derived allele frequency was > 0.5 in the African population (filtered set), and a second set that contained all fixations, regardless of derived frequency (unfiltered set). Specifically, for the unfiltered set, we searched for positions within our data which met the following criteria: 1) The position was fixed (monomorphic) in North America and Europe. 2) The African position, if monomorphic, must contain the ancestral sequence, or, if polymorphic, the ancestral allele must be segregating at that position. 3) The allele at the North American and European position must differ from the ancestral allele at that same position (i.e. a derived fixation). 4) The ancestral base must be inferable (i.e., not an "N"). 5) Positions that occurred in non-recombining regions of the genome were removed from the list of candidate fixations [235]. This same strategy was used to call the set of filtered fixed differences, with the only difference being the requirement that the derived allele frequency in the African population be < 0.5.

3.4.3 Fixed Position Summary Statistics

Our SVM classifies fixations on the basis of summary statistics calculated from windows throughout the genome. For each fixation, a 10kb window was constructed around each site (5kb on either side of the fixed position). A combination of summary statistics were calculated for each of these 10kb windows which included: π, segregating sites (SS), θ_H, Tajima's D, Fay & Wu's H, the number of haplotypes, Kelly's ZnS, ω_c, and windowed-π. The ω_c (omega center) statistic is a modification of a statistic developed by Kim and Nielson [92, 109]. Their original ω statistic is based upon the idea that a selective sweep leaves an increase in linkage disequilibrium within the regions adjacent to a fixed or selected site, but that this excess of LD does not extend *across* the selected site [92]. Their ω statistic is calculated thusly: if there are S polymorphic sites in the data, they are divided into two groups - one group from the first to the l-th polymorphic site measured from the left and the other group from the $(l-1)$th to the last site $(l = 2, ..., S-2)$.

$$\omega = \frac{\left(\binom{l}{2} + \binom{S-l}{2}\right)^{-1} \left(\sum_{i,j \in L} r_{ij}^2 + \sum_{i,j \in R} r_{ij}^2\right)}{(1/l(S-l)) \sum_{i \in L, j \in R} r_{ij}^2} \tag{3.1}$$

The value of l that maximizes ω, ω_{max}, is then used as a test statistic. In our application, we are conditioning on the fixation being centered on the 10kb region, thus we simply calculate ω at our center site and term it ω_c. The windowed-π statistic consists of 9 windows of equal size per 10kb region, normalized such that we only consider the relative values of π among windows. The calculated summary statistics for each fixed position become a feature vector used by our SVM classifier (See Section 3.4.8).

3.4.4 Demographic Inference

The training of our classifier requires the simulation of hard sweeps, soft sweeps, and neutral fixations - each in the context of an appropriate demographic model. While previous authors have estimated demographic models for *D. melanogaster* populations [e.g. 134], here we seek to estimate new parameters that would be appropriate for the population samples we are using (DGRP, DPGP3). To this end we used *∂a∂i* [133] to estimate a two population model to describe the joint history of our N. American and African population samples.

The joint site frequency spectrum for our *∂a∂i* analysis was constructed from a selection of regions of the genome picked in order to minimize the confounding effects of linked selection. In particular, we excluded genomic intervals that overlapped any of the following: CDS, exons, introns, UTRs, simple repeats, repeat masked regions, annotated transcription factor binding sites, annotated regulatory elements (i.e. Oreganno elements; [291]), and all bases ± 5000bp from genes. This yielded 5530 regions of the genome with an total length of 4.43Mb. These regions in total contain 396,135 SNPs, which we then use for input in our demographic inference. SNPs were rooted using the *D. simulans* reference genome from [70], however all demographic inference was done including a mis-orientation parameter to account directly for rooting error. All coordinates and annotation use FlyBase release 5 [292].

Four demographic models were explored and their relative fits compared using the Aikake information criterion (AIC) [231]. These models were: 1) a simple two population isolation model in which an ancestral population splits into two daughters (N. America and Africa) and each subsequent daughter population experiences growth, 2) an isolation-with-migration (IM) model as above but with asymmetric migration rates between N. America and Africa, 3) an isolation with admixture model that is the same as model 1 but with the addition of a single burst of admixture from Africa into N. America, and 4) an IM model (as in model 2) that adds admixture from Africa to N. America. Each of the four models were optimized for the SFS described above three separate times, each from different initial starting conditions. We found that for the models and data we were using, the supplied optimization functions in the *∂a∂i* package were often unreliable, so we implemented a two-step optimization approach. We first performed a coarse optimization using the Augmented Lagrangian Particle Swarm Optimizer [293] and then refined this solution using Sequential Least Squares Quadratic Programming [294]. We used the implementations of these algorithms available in the `pyOpt` package (version 1.2.0) for optimization in Python [295]. To convert population size scaled parameter estimates to generations we assumed a mutation rate of $\mu = 5.49 \times 10^{-9}$ per gamete per generation [296] and a generation time of 15 generations per year. Python scripts for our optimization are available on `https://github.com/kern-lab`. To assess the fit of our estimated models, posterior predictive simulations were performed using coalescent simulations (Figure F.10).

3.4.5 Coalescent Simulations for Machine Learning

There are three distinct fixation scenarios that we are interested in distinguishing among: 1) fixations due to drift (i.e. neutral fixations), 2) fixations due to a beneficial mutation arising de novo and rapidly fixing under directional selection (i.e. hard sweeps), and 3) fixations of previously neutral mutations that become beneficial after an environmental change (i.e. soft sweeps). To perform coalescent simulations under these evolutionary histories, we use the now conventional technique of altering the genealogy of a sample to be conditional upon the trajectory of an allele moving through the population to eventual fixation (forward in time) [78, 88].

To simulate neutral fixations, the trajectory of the neutral mutant destined to fix is simulated backward in time from frequency 1 to absorption at frequency 0. The time of fixation is assumed to be τ coalescent time units back in history. Conditional on absorption, the frequency of the neutral fixation, p, can be modeled as a jump process of the frequency between small time steps δt. The frequency of the neutral mutation in the next step p' is given by:

$$p' = \begin{cases} p + \mu(p)\delta t + \sqrt{p(1-p)\delta t} & \text{with probability 0.5} \\ p + \mu(p)\delta t - \sqrt{p(1-p)\delta t} & \text{with probability 0.5} \end{cases} \tag{3.2}$$

with $\mu(p) {=} {-}p$ [297]. This trajectory routine was tested for accuracy by calculating the expected time to fixation and by comparing simulation results to those from Tajima (1990) [298] (not shown). To model hard sweeps we use stochastic trajectories rather than deterministic trajectories as in Przeworski *et al.* (2005) [223]. Again the frequency of the selected fixation p is modeled as a jump process as above, however in this case $\mu(p)$ is given by:

$$\mu(p) = \frac{\alpha p(1-p)}{tanh(\alpha p)} \tag{3.3}$$

$$\text{where} \quad \alpha = 2Ns \tag{3.4}$$

The accuracy of these simulations was checked first using deterministic trajectories against software available from Y. Kim [88], and under stochastic trajectories against software provided to us by K. Thornton (pers. comm.). A number of results from these routines are also given in [299].

To simulate soft sweeps we introduce an additional parameter to the model, f_0, the frequency at which the allele came under directional selection. To generate trajectories from this model, we simulate a stochastic selection trajectory back in time until the frequency f_0 is reached, and then switch over to a neutral fixation trajectory until absorption as in Przeworski et al. [223].

Our coalescent software, DISCOAL_MULTIPOP, is available for download from our website `kernlab.rutgers.edu`, and features the ability to generate simulations with multiple populations, population splits, admixture events, instantaneous population size changes, hard and soft sweeps, along with recombination and gene conversion.

3.4.6 Support Vector Machine

There are a wide variety of binary classification algorithms. In the case where one has a substantial amount of data (features, examples, or both) and where the classes of those data are known, a powerful and efficient algorithm is the support vector machine [111, 300, 301]. Support vector machines are a supervised machine learning algorithm in which the user trains the machine on data in which the class of interest is known, then assesses the model's performance on the "test" data, where the class label is also known. This trained machine can then be used to classify new data in which the class label is unknown.

The learning or training phase of an SVM consists of an optimization routine where the goal is to find the optimum hyperplane that separates the two classes of data with a maximum margin between those two classes. In the case where the data are not linearly separable, this algorithm maps the data to a higher dimension through the use of a kernel function, to where linear separability exists. However, if the data, even in their transformed, high dimensional feature space still can not be separated, a slack variable is introduced, which measures the error of misclassification. This formulation is known as the "soft-margin" SVM and allows separation of the data as best as is possible by maximizing the margin between those data points that *can* be linearly separated, and minimizing the penalty incurred for those points that cannot [300].

We create two classifiers, SVM$_{auto}$ and SVM$_X$. Each classifier uses a Gaussian radial basis function (RBF) as a kernel function, it being a good choice for this application as it is numerically stable and requires the tuning only one parameter, γ. In addition to the γ parameter, we also must set the cost parameter of the error term, C, for a total of two parameters to optimize in our model, $[C, \gamma]$. We conduct 100 randomized searches over the parameter space, $C \in \{1, 100\}$ and $\gamma \in \{0.0001, 0.01\}$, using 5-fold cross validation on our training data to determine these optimal parameters. Once these parameters were learned, we tested the classifiers' accuracy on a set of test data.

Prior to training, we transform the data for each classifier, scaling each column of the feature matrix such that each element of that column $\in \{-1, 1\}$. To implement the SVM algorithm we used the open-source package, scikit-learn v0.16 [302].

Power Analysis

In order to examine the classifiers' performance under differing evolutionary scenarios, we simulate data across a range of fixation times and selection strengths for each of our five binary classifiers. Specifically, we jointly model fixations occurring at $[0.0, 0.05, 0.1]$ units of $2N$ generations into the past, and strength of selection $\alpha = [100, 500, 1000, 5000, 10000]$. In these power tests we use the same demographic parameters as those we learned from the ∂a∂i analysis (Section 3.4.4).

We simulated 20,000 examples for each of the 75 combinations of the model parameters: 1.5×10^6 examples total. We utilized the same SVM work-flow as described above and calculated classification accuracy, generated ROC plots, and determined AUC values for each scenario.

Training And Testing Data

We build two separate SVM classifiers, one to classify fixations on the autosomes (SVM_{auto}), and one that classifies those fixations that have occurred on the X chromosome (SVM_X). For each of these two classifiers, we used the estimated parameters from the ∂a∂i analysis and generated 1×10^5, 10kb simulations for each class of fixation (hard, soft, & neutral). Simulations that were used to train and test the SVM_X classifier were generated using the same parameters used the SVM_{auto} classifier, but here ρ, θ, and both migration rates, *M1* and *M2*, were scaled by a factor of 0.75 (assuming an equal sex ratio).

For the first three classification tasks (Hard vs. Neutral, Hard vs. Soft, and Soft vs. Neutral) we concatenated the simulated data for each respective fixation class, then partitioned it into 20% training and 80% testing, it such that 40,000 samples were used for training and 160,000 samples were used for testing. Both the training and test set were composed of equal numbers of the respective sweep type. For Hard vs. Soft & Neutral and Hard & Soft vs. Neutral, we again split the data and created a training and test set as above, however for Hard & Soft vs. Neutral we used a random sample of 50,000 hard and 50,000 soft simulations, and 100,000 neutral examples to comprise each full data set. For Hard vs. Soft & Neutral, training and testing data was created using 100,000 hard fixations, 50,000 soft, and 50,000 neutral examples for each set.

We utilize 100 rounds of random search optimization for the C and γ parameters, at each iteration sampling 20% of the training data. We conducted a parameter search for each SVM and trained our machines using the optimal learned combination of parameters.

We then assessed the performance of SVM$_{auto}$ and SVM$_X$ by classifying the test data (1.6×10^5 examples per comparison) (Figures 3.4 and 3.5). It is important to point out that these test data were not used in any phase of the SVM training. Since they were generated under the same model as the training data, these test data are therefore ideal for assessing our classifier's predictive ability.

3.4.7 Performance Metrics

We assessed the abilities of the classifier using three metrics: Classification accuracy, Receiver Operating Characteristic (ROC), and area under curve (AUC). The accuracy of the classifier for each test is simply the percentage of correct classifications relative to all classifications. The ROC plot graphically shows the sensitivity vs. the specificity of the classifier's performance [303]. More specifically, the ROC curve plots the fraction of true positives out of all positive predictions (true positive rate) on the y-axis, and the fraction of false positives out of all negative predictions (false positive rate) on the x-axis. A curve which starts at the origin and rapidly rises along the y-axis to a value near 1 before moving along along the x-axis is the most desirable. This signifies that the classifier is has a high true positive rate, and a very low fraction of the classifications are false positives. Conversely, a diagonal line from the origin to the top right corner of the plot, ($y = x$), represents a classifier performing no better than random in its ability to predict the class to which the data belong. Lastly, we examine the area under curve (AUC). The area under the ROC curve can be thought of as a single scalar representation of the ROC curve itself. Since this value represents part of the area of a 1×1 square, the AUC $\in [0, 1]$. The AUC of a classifier has the property of being equivalent to the probability that the classifier will rank a randomly chosen positive data point higher than a randomly chosen negative data point [303].

3.4.8 Classifying Fixed Positions

Having trained and evaluated our classifiers, we then used these SVMs to classify both the filtered and unfiltered sets of fixed differences. We scaled these fixed difference datasets using the same parameters we used to scale the simulated data, then classified all of the fixed positions using two of the SVMs: we first applied the Hard & Soft vs. Neutral SVM to all fixations; for those classified as a sweep, we then used the Hard vs. Soft SVM to determine the mode of selection.

3.4.9 Testing For Misclassified Soft Sweeps

To investigate the prevalence of spurious soft sweeps located near called hard sweeps we used simulations along with our trained SVM. We generated 11 sets of 1000 coalescent simulations, each with sample size of 141 and 10kb in length. We fixed $\alpha = 2000$, $\rho = 100$, and the fixation time, $\tau = 0.05$. Each of these 10 sets of simulations incorporates a hard sweep effectively having occurred at some increasing distance away from the 10kb window being evaluated. This results in the first 10kb window of these simulated chromosomes being identical to our classification windows used earlier, i.e. a sweep in the center (Section 3.4.6). The central position of the following window is effectively 10kb away from the hard sweep and has $r/s = 0.1$. The center of the subsequent window is therefore 20kb away from from the hard sweep and has $r/s = 0.2$. In effect, r/s increases 0.1 units in each successive window such that the last window, located 100kb from the hard sweep, has an $r/s = 1.0$.

Using the same SVM methodology used to call a fixation as either a putative hard or soft sweep in the empirical fly data, we calculated summary statistics and classified each example in these 11 10kb windows as either hard, soft, or neutral. It is important to reiterate that all windows, save for the first window where the hard sweep was simulated, contain only neutral mutations. We determined how many examples in each of the 10kb windows were incorrectly classified as a either a hard or soft sweep. We then fit a third order polynomial to the number of misclassified sites across all r/s values and determined the r/s value in which 95% of sites were correctly classified. Next, using this r/s cutoff and assuming N_e =341,745, $\alpha = 2N_e s = 2000$, we obtained a recombination rate cutoff. Using this recombination rate along with Haldane's mapping function [304] we computed the genetic distance. Finally we use the *Drosophila melanogaster* binned recombination rates reported by Comeron et al. [235] to convert to physical distances local to each hard sweep in the genome.

3.4.10 Selective Sweep Enrichment Tests

We sought to determine if genetic elements were enriched for the presence of classified fixed positions and looked for both hard and soft sweep enrichment in the following elements: CDS, gene, exon, intron, 5′ UTR, 3′ UTR, miRNA, mRNA, ncRNA, pre-miRNA, snoRNA, snRNA, tRNA, transcription factor binding sites, and transposable elements. Positions of each element were taken from Flybase Dmel version r5.49 (`ftp://ftp.flybase.net/genomes/Drosophila-melanogaster/dmel_r5.49_FB2013_01/gff`).

We used a permutation test to ask whether sweeps were enriched in particular annotation categories. Rather than randomly permuting the coordinates of hard and soft selective sweeps across the entire genome, we only shuffled the classifications among our sets of shared and North American-specific fixations. The motivation for this approach was to ensure that any bias with respect to the locations of these fixations in the genome would affect both our true set of classifications and our permuted data sets equally, and thus not produce any erroneous signal of enrichment among classified sweeps simply because they are a subset of fixations. Also, we reasoned that nearby fixations were likely to receive the same classification from the SVM, and upon visual inspection of our classification results we noticed many runs of consecutive fixations with the same class label. We therefore chose to shuffle runs of consecutive identical classifications rather than shuffling classifications independently of one another. Because of this, our permutation should be robust to spatial clustering of functionally similar features in the genome, as multiple consecutive fixations within such a cluster were no less likely to receive the same classification in the permuted data set as in the real classification.

Briefly, our algorithm was as follows: 1) Advance to the next, or initially, the first fixation on the chromosome arm, in ascending order. 2) Select a classification (hard, soft, or neutral), y, according to the number of remaining runs of consecutive fixations receiving this classification. Whenever possible y is constrained to differ from the previously selected classification. 3) Draw a run length, l, without replacement from the chromosome arm's length distribution of runs of consecutive fixations classified with label y. 4) Label the next l fixations on the permuted chromosome arm as class y. 5) Repeat from step 1 until the final fixation in the permuted chromosome arm is reached and given a label. At this point all fixation run lengths will have been drawn and the permutation is complete. We performed this permutation separately on each chromosome arm, using that chromosome arm's numbers and length distributions of runs of each class.

We used this algorithm to generate 1,000 permuted data sets. We then searched for annotation features overlapping each randomly classified fixation in each of these sets. When testing for enrichment of adaptive fixations in certain annotation categories, or among interacting pairs of genes, we used the 1,000 permutations to calculate one-sided p-values.

Gene Ontology Analysis

We used the classified fixed positions to determine the presence of any gene ontology (GO) enrichment. From these data we extracted gene names and used "DAVID Bioinformatics Resources 6.7" web-based service to determine individual GO term enrichment, as well as GO cluster enrichment [249, 250]. These genes were used as "gene lists" in DAVID and tested against the whole *Drosophila melanogaster* genome as a background. We considered DAVID clusters with enrichment scores (e.s.) > 2.0 for further examination, and therein looked for individual elements that were shown to be significantly enriched $(p \leq 0.05)$ when corrected for multiple testing [215].

Testing For An Excess Of Sweeps Among Interacting Genes

In order to test for an excess of interacting pairs of genes both experiencing selective sweeps, we counted the number of such pairs observed in our true classification set and compared this number to the numbers observed in each permuted set (Section 3.4.10). Although we attempted to control for the effect of spatial clustering of functionally related genes when performing our permutations, here we took the extra step of not counting interacting gene pairs where the two sweeps were within one Megabase of each other. Moreover, in cases where a gene A had a sweep and interacted with genes B and C, each of which had sweeps, we only counted one of the interactions if the sweeps in B and C were located within 1 Mbp of one another. We used this procedure to obtain a one-sided p-value for each interaction network tested.

Synonymous and Nonsynonymous Analysis

We counted the numbers of synonymous and nonsynonymous substitutions at fixed positions that were classified as either being a hard or soft sweep. Separately for hard and soft sweeps, we then utilized Fisher's exact test to compare the ratio of nonsynonymous to synonymous fixations to that at neutral fixed positions.

We also asked if nonsynonymous sweeps were more likely to be either a radical or conservative amino acid change. We again used Fisher's exact test to check for enrichment of radical substitutions within hard, soft, and both hard and soft sweeps, within coding regions.

Finally, we looked for synonymous substitutions that resulted in unpreferred codons in the African populations changing to preferred codons in the

North American flies. We again utilized Fisher's Exact test, examining the number of preferred codons found in our sweep calls and the number of preferred codons in all fixations, versus the number of synonymous substitutions in our sweeps and the number of synonymous substitutions in all fixations.

4

Effects Of Linked Selective Sweeps On Demographic Inference And Model Selection

D.R. SCHRIDER, A.G. SHANKU, A.D. KERN

> *"Pay no attention to the man behind the curtain!"*
>
> – L. Frank Baum, *The Wonderful Wizard of Oz*

4.1 Introduction

The widespread availability of population genomic data has spurred a new generation of studies aimed at understanding the histories of natural populations from a host of model and non-model organisms alike. In particular, genome-scale variation data allows for inference of demographic factors such as population size changes, the timing and ordering of population splits, migration rates between populations, and the founding of admixed populations [305, 306, 307] (See Sections 1.1 and 1.4). Such efforts can refine our picture of demographic events inferred from the archaeological record [e.g. 156], or reveal such events in species where no archaeological data are available, and can aid conservation efforts by complementing census data [e.g. 308, 309].

Population genomic data is well-suited for this task, simply because demographic changes leave their mark on patterns of genetic variation. Recent population growth, for example, will result in an excess of rare variation compared to equilibrium expectations [310], while population contraction will result in an excess of intermediate frequency alleles [311]. In recent years, researchers have devised a variety of methods that seek to detect the population genetic signatures of these demographic events. These include Approximate Bayesian computation (ABC), where simulation is used to approximate the posterior probability distributions of a demographic models parameters without specification of an explicit likelihood function (See Sections 1.3.3 and 1.4.2, Appendix B). Other approaches, such as $\partial a \partial i$, use the probability density of the site frequency spectrum (SFS) under a given demographic model and parameterization to calculate the likelihood of the observed SFS (Section 1.4.1 and [133]), thereby allowing for optimization of model parameters. More recently, methods based on the sequentially Markovian coalescent (SMC; [161, 312] and Appendix C) have been devised (Sections 1.4.3 and 1.4.3, Appendix C, and [155]), to infer how a populations size has changed over time through the description of patterns of genetic variation along a recombining chromosome.

Demographic inference from population genomic data in its various forms has proven to be successful technique. However, a unifying assumption of these various inference methods (ABC, SFS-based, and SMC-based approaches) is that the genetic data in question are strictly neutral and free from the effects of linked selection in the genome. While this is an important simplifying assumption, it may be the case that in many populations a sizeable fraction of the genome is influenced by natural selection [e.g. 96, 313]. Indeed, natural selection can produce skews in patterns of genetic variation that are quite similar to those generated by certain non-equilibrium demographic histories (Section 1.3.5). For example, positive selection driving a mutation to fixation (i.e. a selective sweep; [72]) may resemble a population bottleneck [124]. Moreover, many demographic perturbations are well known to cause unacceptably high rates of false positives for many classical tests for selection [87, 91, 314]). Thus, if natural selection were to have a substantial impact on genome-wide patterns of variation, then many demographic parameter estimates could be biased [315]. Indeed this has been shown to be the case for at least some scenarios of background selection, where purifying selection reduces levels of neutral polymorphism at linked sites [316].

Here, we examine the potential impact of linked positive selection on three of the most widely used methods for demographic inference: ABC [105, 149], *∂a∂i* [133], and PSMC [155]. We demonstrate that selection can substantially bias parameter estimates, often leading to overestimates of the severity of population bottlenecks and/or the rate of population growth. Moreover, we show that the presence of selective sweeps can result in the selection of the incorrect demographic model; if a reasonably small fraction of loci used for inference are linked to a selective sweep, one may incorrectly infer that a constant-size population experienced a bottleneck. Finally, we discuss the implications of our results when inferences is made in humans and *Drosophila*, and recommend steps that could partially mitigate the bias caused by selection.

4.2 Results

4.2.1 Demographic Parameter Estimates Are Biased By Positive Selection

We sought to quantify the impact of positive selection on demographic parameter estimation under our bottleneck, growth, and contraction-then-growth models (Figure F.20). First, we simulated population samples experiencing no selection and asked how well we could recover the true parameters of the model using diffusion approximations to the SFS via the $\partial a \partial i$ software package (Diffusion Approximations for Demographic Inference; [133]), or with a set of commonly used summary statistics via ABC [317, 318]. Briefly, we used both of these inference methods to fit the focal demographic model to data sampled from 500 unlinked simulated loci, and repeated this process on 100 replicate "genomes" (Section 4.4.1). We then gradually increased the value of f, the fraction of these sampled loci linked to hard selective sweeps (within a distance of $c/s \leq 1.0$; Section 4.4.1). At values of $f = \{0.1, 0.2, \ldots, 1.0\}$, we repeated parameter estimation to assess the extent to which a given amount of selection biases our inference.

Population Bottleneck

When using $\partial a \partial i$ to infer the optimal set of parameters of a population having undergone a bottleneck and experiencing no positive selection (Figure F.20), our estimates were quite accurate (Figure 4.1): the average parameter estimate for the ancestral effective population size, NeA, was 10,060 individuals (a 0.6% deviation from the true parameter value); our mean estimate for the time of recovery from the bottleneck, T_R, was 3,120 generations ago (4.0% deviation); our estimated effective population size during the bottleneck, NeB, was 1,999 individuals (0.05% deviation) on average; and our mean estimate of the present-day effective population size, $Ne0$, was 20,465 (2.3% deviation). Moreover, our inferences were fairly consistent, with most parameter estimates being fairly close to the true value (Figure 4.1). However, while repeating this analysis with increasing numbers of loci linked to a selective sweep, our parameter estimates became increasingly biased. Even a small value of f produced significant underestimates of the effective population sizes $Ne0$ and NeB. For example, the mean inferred value of NeB decreases to 1,764 when $f = 0.2$ (an 11.8% underestimate), to 1,402 when $f = 0.5$ (29.9% underestimate), and to 717 when $f = 1.0$ (64.2% underestimate). A more subtle, but consistent downward bias of $Ne0$ also appears with increasing f. In this case, $Ne0$ is estimated at 19,650 at $f = 0.2$ (1.8% underestimate), 18,371 when $f = 0.5$ (8.1% underestimate), and 15,561 when $f = 1.0$ (22.2% underestimate). By contrast, estimates of the ancestral population size, NeA, and the time since the recovery, T_R, are largely unaffected unless f is fairly high (≥ 0.8), in which case the values of these two parameters are somewhat overestimated.

Like inference using the SFS (i.e. *∂a∂i*), our ABC procedure was able to infer the true parameters with minimal bias when run on simulated population samples experiencing no positive selection: the mean estimates were 10,104 for NeA (1.4% difference from true value), 2,917 for T_R (2.7% difference), 2,311 for NeB (15.6% difference), and 20,078 for $Ne0$ (0.4% difference). However, we note that the NeB estimate was fairly inconsistent (with the middle 50% of estimates ranging from 1447 to 3373), while other parameter estimates exhibited much lower variance. When positive selection is introduced, we obtain significantly biased estimates of all parameters when $f \geq 0.2$, and all but NeA are significantly biased when f is only 0.1. These biases are in the same direction as observed using *∂a∂i* (underestimates for $Ne0$ and NeB, and overestimates for NeA and T_R), but almost always substantially larger. Indeed, for $f \geq 0.2$, our estimates of NeB and T_R are at the boundaries of our prior parameter ranges, respectively (the upper bound of 3,500 for T_R and the lower bound of 100 for NeB). Also note that when f increases to ≥ 0.2, estimates of T_R are also at the upper bound of our prior: we are inferring a very short but extreme bottleneck. Thus, for our bottleneck model, ABC based on our selection of summary statistics appears to be more sensitive to selection than *∂a∂i*. Overall, the presence of positive selection seems to cause both methods to overestimate the extent of population contraction, and underestimate the degree of recovery from the bottleneck. For the simulated datasets used in these analyses, the distance from the sweep to the linked locus is measured in terms of c/s, and is drawn uniformly between 0 and 1. We also repeated these analyses when fixing the value of c/s, and in Figure F.21 we show our distribution of parameter estimates obtained using both *∂a∂i* and ABC on 111 different combinations of f and c/s.

This figure demonstrates how, for a given fraction of neutral loci linked to a selective sweep, increasing the proximity to the sweep increases bias.

Note that for our ABC inference we examined only the means of several population genetic summary statistics (Section 4.4.3). Including the variances caused estimates to behave non-monotonically, because whenever f is not equal to one or zero the distribution of summary statistic values is a mixture of two models, resulting in less accurate parameter estimation. We also show our parameter estimates when including variances in Figure F.21.

Population Growth

Next, we examined the impact of positive selection on parameter estimates for our model of population growth Figure F.20. When our simulated genomes experienced no recent selective sweeps, we again achieved good accuracy when using $\partial a \partial i$ (Figure 4.2): our mean estimates of NeA, T_G, and $Ne0$ were 1,040 (0.8% difference from true value), 955 (3.8% difference), and 36,610 (2.0% difference), respectively. Increasing f again biases our estimates, but the effect is more subtle than for the bottleneck case. This is probably a consequence of the reduced scale of the impact of positive selection on flanking variation under this model relative to the bottleneck model. The most notable pattern that we observe for this model is that T_G decreases with increasing f, while the population size estimates are largely unaffected: when $f = 0.5$ our average estimate is 905 (1.6% difference from true value), versus 872 when $f = 0.8$ (5.3% difference), and 855 when $f = 1.0$ (7.1% difference). In other words, widespread selective sweeps will cause one to infer slightly more recent but more pronounced exponential growth. When c/s is relatively small, our error rates are substantially higher (Figure F.22). Thus, stronger positive selection

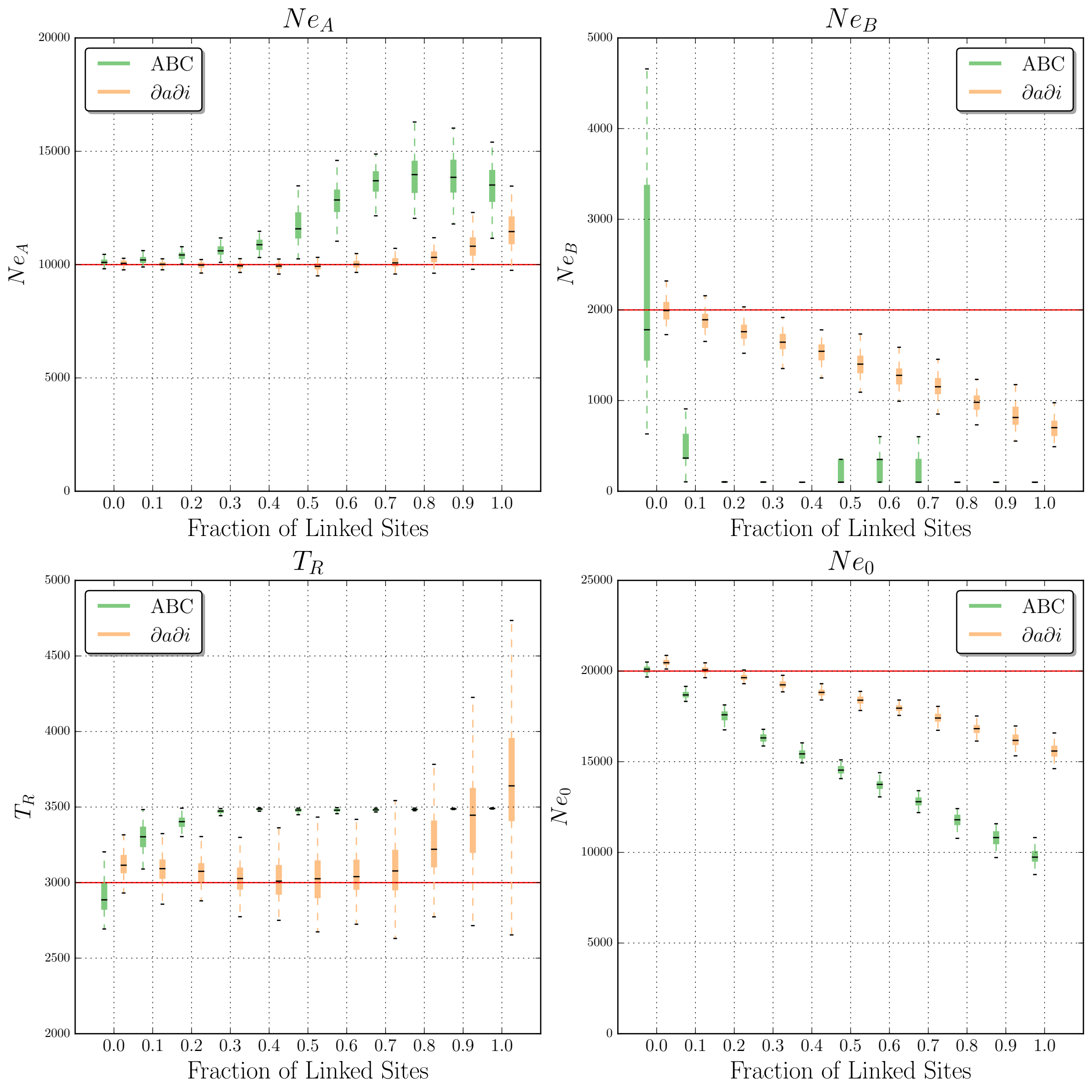


Figure 4.1: **Learning Demographic Parameters In The Bottleneck Model.** Bottleneck model parameter estimates from *∂a∂i* and ABC. Parameter estimation was performed on simulated data sets either evolving neutrally, or with some fraction of loci used for inference linked to a selective sweep. Each box plot summarizes estimates from 100 replicates for each scenario. Note that T_B, the bottleneck onset time, is absent from this figure because it was fixed it to the true value (Sections 4.4.2 and 4.4.3).

could still seriously impair *∂a∂i*'s demographic inferences under this population size history.

We then used ABC to perform parameter estimation under the growth model. In the neutral case, our estimated parameters were largely concordant with the true values, with the exception of some bias observed for T_G (mean estimate of 801, which is 13% below the true value). Our estimates of NeA were also far more dispersed than those obtained from $\partial a \partial i$. Further, unlike our estimates with $\partial a \partial i$, increasing the value of f substantially biases our ABC estimates. For example, $Ne0$ is 39,497 when $f = 0$ (10% greater than the true value), but increases to 42,851 when $f = 0.5$ (an overestimate of 19%), 49,030 when $f = 0.8$ (an overestimate of 37%), and 62,889 when $f = 1.0$ (an overestimate of 75% plus a dramatic increase in variance). The degree to which T_G is underestimated also increases with f: the average estimate is 769 at $f = 0.2$ (16% below the true value), 717 at $f = 0.5$ (22% bias), 667 at $f = 0.8$ (27.5% bias), and 623 at $f = 1.0$ (32.2% bias). Again, we demonstrate the effect of varying the distance c/s of sampled loci from the selective sweep, as well as the effect of performing ABC on the variances of summary statistics in addition to their means, in Figure F.22. Overall, we observe that positive selection under the growth model will cause inferences of more recent, faster population growth, with this effect being far more subtle when using $\partial a \partial i$ than ABC with our set of summary statistics.

Population Contraction Followed By Growth

Finally, we assessed our ability to recover the parameters of our contraction-then-growth model with increasing amounts of positive selection Figure F.20. Without selection, $\partial a \partial i$ estimates NeA, T_G, and $Ne0$ with reasonably high accuracy (Figure 4.3): 14,773 on average for NeA (2.1% over the true value), 862 for T_G (6.3% under the true value), and 37,999 for $Ne0$ (5.8% over the true value). However, T_C and NeC are substantially overestimated at 2,530 (24.0% over the

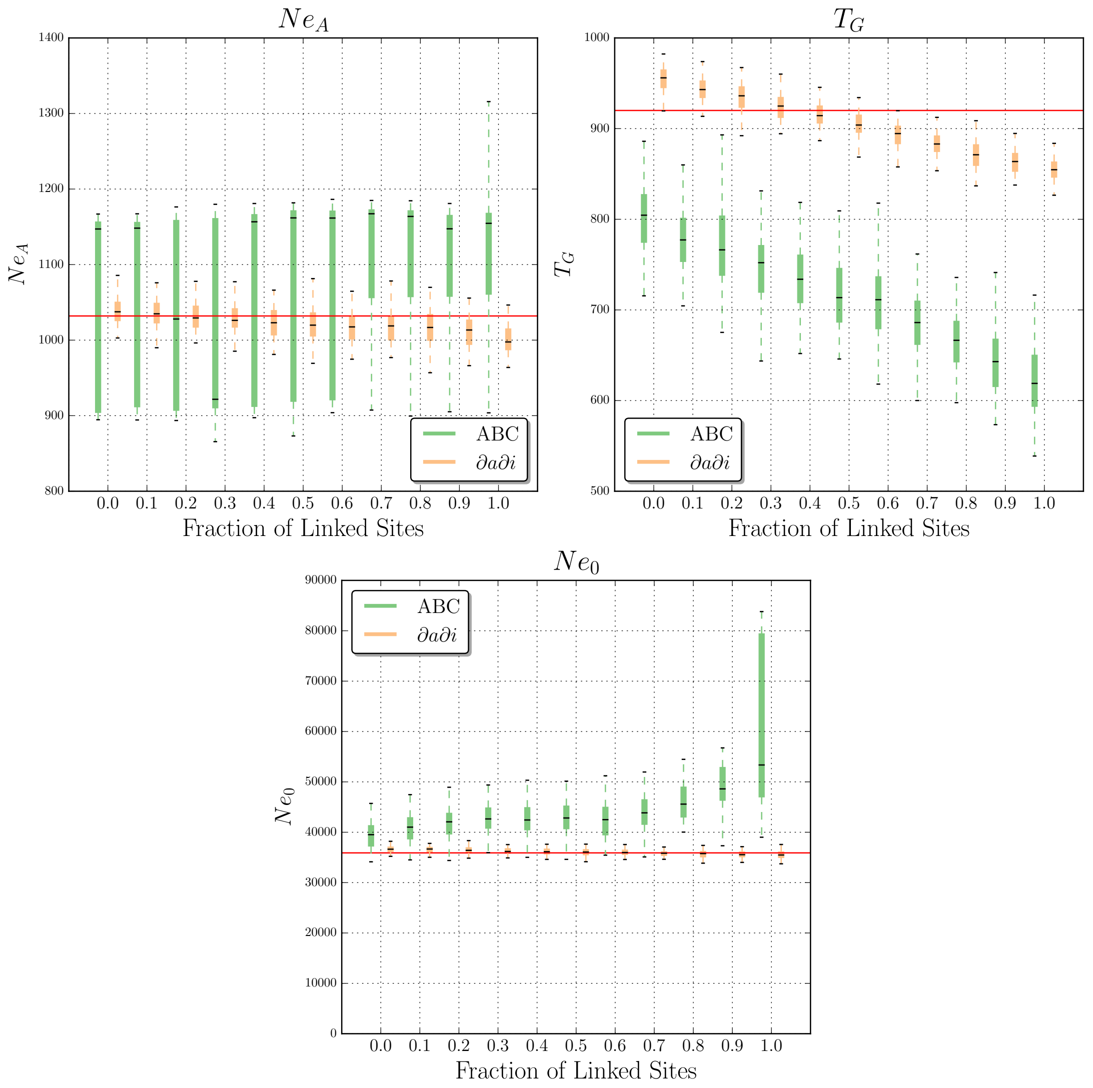


Figure 4.2: **Learning Demographic Parameters In The Growth Model.** Exponential growth model parameter estimates from $\partial a \partial i$ and ABC. Parameter estimation was performed on simulated data sets either evolving neutrally, or with some fraction of loci used for inference linked to a selective sweep. Each box plot summarizes estimates from 100 replicates for each scenario.

true value) and 1,350 on average (30.8% over the true value). As we increase f, our estimates of NeA, T_C, and NeC are inflated, T_G is increasingly underestimated, and $Ne0$ is largely unaffected. The effect on T_C is the largest, resulting in a seemingly linear increase with f: our estimate is 2,886 when $f = 0.2$ (an overestimate of 41.47%), 3,712 when $f = 0.5$ (overestimate of 82.0%), 5,329 when $f = 0.8$ (overestimate of 161.2%), and 7,082 when $f = 1.0$ (an overestimate of 247.2%). NeA and NeC increase more slowly: to 29,069 (an overestimate of 100.8%) and 2183 (an overestimate of 111.6%) when $f = 1.0$, respectively, while T_G on the other hand decreases to 642 (an underestimate of 30.2%) when $f = 1.0$. Thus, positive selection typically results in our $\partial a \partial i$-estimated demographic model to have more protracted population contraction, with larger initial and contracted population sizes. Results for varying values of c/s are shown in Figures F.23 and F.24.

When repeating these analyses using ABC given our set of summary statistics, we find that under neutrality T_G is grossly underestimated, NeC is slightly overestimated, and NeA and $Ne0$ are estimated with greater accuracy (14,435 or 0.27% under the true value, and 34,373, or 4.3% under the true value, respectively). Thus, we infer a more protracted but slightly less severe population contraction than the true population size history. Even so, we proceed to characterize what the effect of linked selection on parameter estimates using ABC as before. Indeed, our estimates become more biased as we add increasing amounts of positive selection. Most notably, $Ne0$ exhibits a substantial downward bias as we increase f, and is estimated at 31,970 when $f = 0.2$ (11% underestimate), 28,105 when $f = 0.5$ (21.7% underestimate), 25,829 when $f = 0.8$ (28% underestimate), and 24,735 when $f = 1.0$ (31.1% underestimate). Also, as f becomes large NeC shifts from being slightly overestimated to significantly underestimated, and estimates of NeA become slightly upwardly biased. Thus, under this model we find that positive selection again biases parameter estimates, though not in the same manner for $\partial a \partial i$ and ABC; while $\partial a \partial i$

infers a longer phase of reduced population size (as well as larger ancestral and reduced sizes), ABC, we find, infers a more severe contraction followed by a less complete recovery. We show our inference results on the full grid of c/s and f values, as well as when including variances of summary statistics, in Figure F.23.

4.2.2 Effect Of Positive Selection On Population Size History Inference Using PSMC

The pairwise sequential Markovian coalescent (PSMC) is a widely used method that infers a discretized history of population size changes from a single recombining diploid genome [155]. Such inference is possible because coalescence times between the two allelic copies in a diploid, which are governed by the effective population size, will change at the breakpoints of historical recombination events (See Section 1.4.3 and Appendix C). The distribution of coalescence times across the genome thus contains information about population size history, however, this necessitates sampling a large stretch of a recombining chromosome. In order to test the impact of positive selection on inferences from PSMC, we simulated constant-size populations of 10,000 individuals, sampling a 15 Mb chromosomal region from two haploid individuals. We performed 100 replicates of this simulation for each of four scenarios (Section 4.4.4): the standard neutral model, a population experiencing one fairly recent sweep somewhere in this region (reaching fixation $0.2Ne$ generations ago), a population experiencing three recurrent sweeps (fixed 0, 0.2, and $0.4Ne$ generations ago), and a population experiencing five sweeps (0, 0.1, 0.2, 0.3, and $0.4Ne$ generations ago). We find that under neutrality very little population

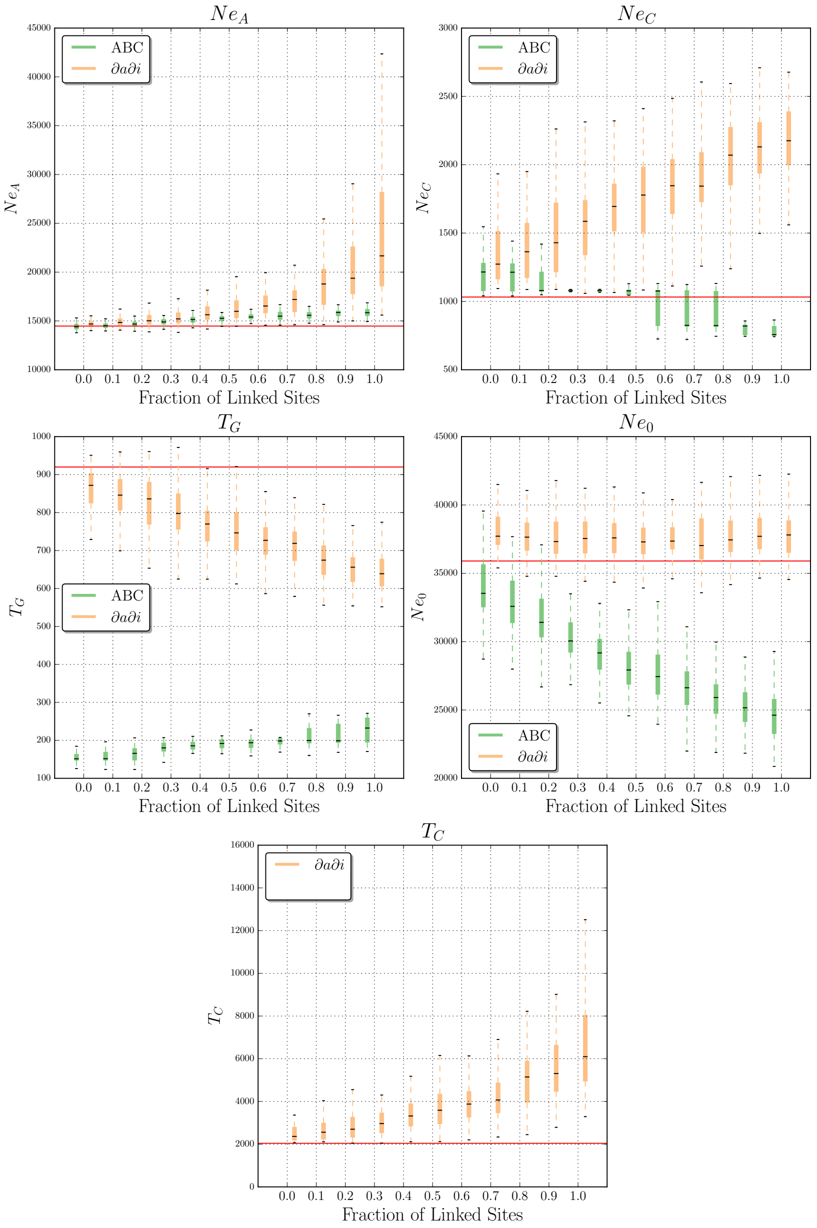


Figure 4.3: **Learning Demographic Parameters In The Contraction-Then-Growth Model.** Contraction-then-growth model parameter estimates from *∂a∂i* and ABC. Parameter estimation was performed on simulated data sets either evolving neutrally, or with some fraction of loci used for inference linked to a selective sweep. Each box plot summarizes estimates from 100 replicates for each scenario. Note that when performing ABC, time of population contraction, T_C, was fixed to the true value and therefore this parameter is only shown for *∂a∂i*.

size change is inferred on average (though there is a fair bit of variance; Figure 4.4). However, when there has been only a single selective sweep, a population bottleneck near the time of the sweep, in which the population contracts to approximately two-thirds of its original size before recovering, is typically inferred (Figure 4.4). When there have been three or five recurrent selective sweeps, the inferred population contraction becomes increasingly severe (Figure 4.4). We observe that this contraction is approximately one-fourth of the original size in the five-sweep case, often with no subsequent recovery. We speculate that this effect may result from scenarios that include a very recent sweep (Section 4.4.4). Thus, we find that positive selection can dramatically skew population size histories deduced by PSMC.

4.2.3 Positive Selection Present In A Stationary Demography Leads To Spurious Support Of Non-Equilibrium History

Demographic inference methods are often used not only to infer parameters of a model, but increasingly to select the best fitting among several competing models. For example, Duchen et al. [134] recently used ABC to infer that a model where the North American *Drosophila melanogaster* population is founded via admixture between European and African flies is a better fit to the data than models without admixture. To ask whether positive selection might affect the outcome of demographic model selection, we simulated genomes with constant population size, again sampling loci for which some fraction, f, is located within $c/s \leq 1$ of a selective sweep. We then performed model selection among our four demographic histories using both $\partial a \partial i$ and ABC (Section 4.4.2 and Section 4.4.3).

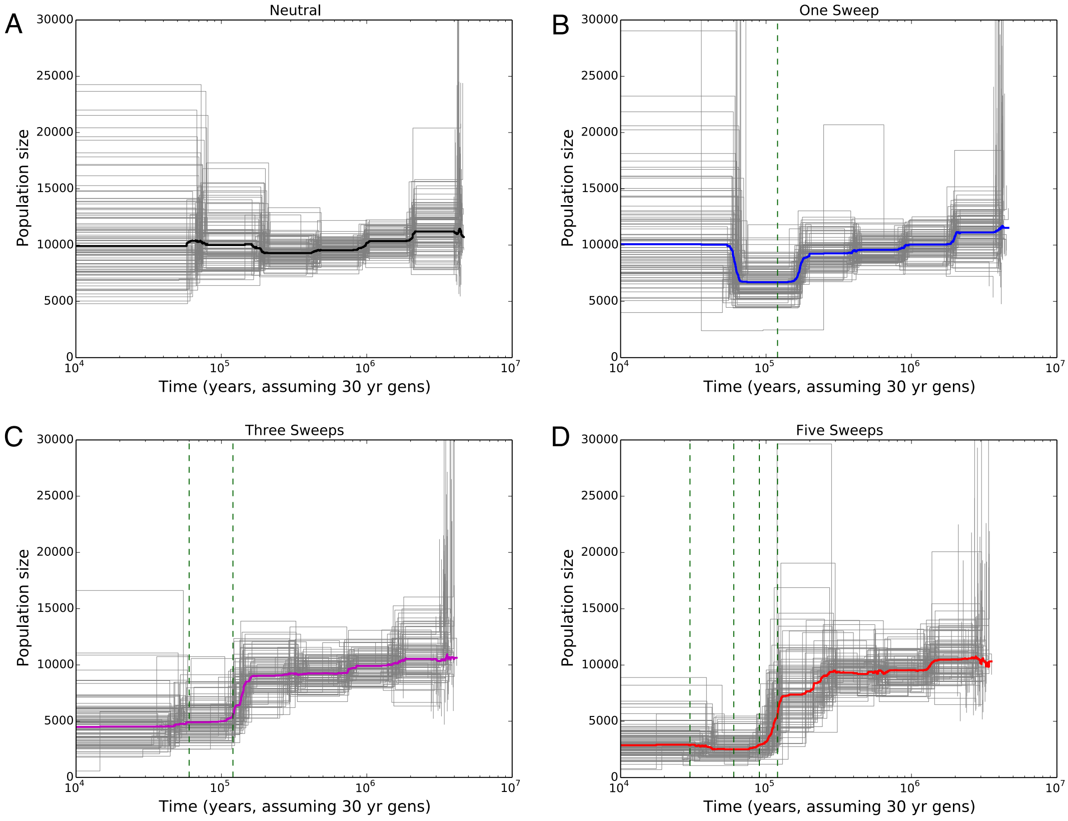


Figure 4.4: In all cases the simulated population size was constant throughout. (A) Population size histories inferred from neutral simulations. (B) Inferences from simulations with one selective sweep, for which the fixation time is shown as a dashed green vertical line. (C) Inferences from simulations with three recurrent selective sweeps. Fixation times for the two older sweeps are shown as dashed green vertical lines, while the most recent sweep fixed immediately prior to sampling. (D) Five recurrent selective sweeps, with fixation times for the four oldest shown as dashed vertical lines; again, the most recent sweep fixed immediately prior to sampling.

Prior to performing model selection with $\partial a \partial i$, we first examined the degree of support for each model when fit to each dataset using the AIC. Examining the differences in AIC between models, we found that even a moderate number of selective sweeps will cause non-equilibrium demographic scenarios to have far stronger support than the true equilibrium history (Figure F.25). This is especially so for the bottleneck and contraction-then-growth models, which achieve better support than the equilibrium model even at small values of f. For example, when $f = 0.2$ the bottleneck model receives an AIC lower than the equilibrium model in 90% of cases , and the contraction-then-growth model has a lower AIC 72% of the time. By contrast, the pure growth model is supported to a lesser extent (a lower AIC in 54% of cases), and occasionally failed to optimize properly, settling on a very low-likelihood parameterization - an indication of a poor fitting model. We may be achieving a better fit of the bottleneck and contraction-then-growth models because they more accurately model the genealogy of a region experiencing a selective sweep: much of the ancestral variation flanking the selected site is removed during the sweep (analogous to contraction), as is replaced by the subset of alleles within the rapidly expanding class of individuals containing the selected mutation (analogous to expansion).

We conducted formal model selection as described in Section 4.4.2, conservatively selecting the equilibrium model unless one of the other models had an AIC at least 50 units higher. We note that it would be preferable to perform parametric bootstraps from competing models to compare the distributions of AIC values, but in the interest of computational efficiency we instead choose this heuristic. Even with this conservative cutoff, we selected a non-equilibrium model for 15% of simulated data sets with $f = 0.2$, for 47%

of datasets with $f = 0.3$, and for 91% of datasets when $f = 0.6$ (Table 4.1). Thus even if a minority of loci are linked to a recent selective sweep then SFS-likelihood based approaches may prefer the wrong demographic model. Interestingly, in every case where a non-equilibrium model was the unambiguous best fit to the data, this model choice was the bottleneck scenario.

Fraction of		Contraction			
Linked Loci	Bottleneck	& Growth	Growth	Ambiguous	Equilibrium
0.0	0	0	0	0	100
0.1	1	0	0	1	98
0.2	5	0	0	10	85
0.3	14	0	0	33	53
0.4	41	0	0	55	4
0.5	77	0	0	23	0
0.6	91	0	0	9	0
0.7	100	0	0	0	0
0.8	99	0	0	1	0
0.9	100	0	0	0	0
1.0	100	0	0	0	0

Table 4.1: **Model Selection Using $\partial a \partial i$.** The column labeled "Ambiguous" indicates the number of tests for which no one model fit better than any other (AIC $>$ 50).

Next, we performed model selection on our constant-size population samples using ABC (Section 4.4.3). For each of these datasets, we estimated Bayes factors for each pairwise comparison of demographic models. Again, we find that non-equilibrium demographic models receive stronger support than the constant-size model when a sizable fraction of loci are linked to selective sweeps. For example, when $f = 0.4$, the bottleneck model has nominally stronger support (Bayes factor > 1) than the equilibrium model for 55% of datasets, the growth model has stronger support than equilibrium in 9% of datasets, and the contraction-then-growth model has stronger support in 4% of datasets. When f is increased to 0.8, we see even stronger support for the

Fraction of Linked Loci	Bottleneck	Contraction & Growth	Growth	Ambiguous	Equilibrium
0.0	0	0	0	0	100
0.1	0	0	0	1	99
0.2	0	0	0	6	94
0.3	2	0	0	32	66
0.4	25	0	0	45	30
0.5	88	0	0	2	10
0.6	99	0	0	0	1
0.7	100	0	0	0	0
0.8	100	0	0	0	0
0.9	100	0	0	0	0
1.0	100	0	0	0	0

Table 4.2: **Model Selection Using Appoximate Bayes Computation**. The column labeled "Ambiguous" indicates the number of observations for which no one model fit better than any other (Bayes Factor > 20).

non-equilibrium models, with 100% of the bottlenecks datasets, 79% of the contraction-the-growth datasets, and 26% of the growth model datasets having a BF > 1. (Figure F.27).

We used these Bayes factors to perform model selection in a manner similar to our analysis with *∂a∂i*, conservatively selecting the equilibrium model if there was no alternative model that was a significantly better fit to the data (i.e. having a Bayes factor relative to the equilibrium model of ≥ 20 [319]). Again, we find that even if a minority of loci are linked to a sweep, then there is a substantial probability that the constant-size model will not be selected: for 6% of datasets we select a non-equilibrium model when $f = 0.2$, for 34% of datasets with $f = 0.3$, and for 99% of datasets when $f = 0.6$; (Table 4.2). As with *∂a∂i*-optimized models, we found that in every instance where we were able to unambiguously select a single non-equilibrium demographic history as the best fit we chose the bottleneck model. When we include the variances of our set of summary statistics in our ABC procedure, we find that non-equilibrium models are strongly supported in an even higher proportion of simulated data sets, though in this case we typically select the contraction-then-growth model rather than the bottleneck model (Table E.20).

Contributions

In this preceding section, I was solely responsible for all ABC analysis, estimation, and model selection in Section 4.2.1. D.R.S was responsible for ∂a∂i analysis. I contributed to the PSMC analysis in Section 4.2.2, along with D.R.S. I generated all figures (with the exception of Figure 4.4, D.R.S), tables, and contributed in writing the draft of the manuscript, along with D.R.S and A.D.K.

4.3 Discussion

It is well known that natural selection profoundly affects genealogies and therefore patterns of genetic polymorphism [77, 320], thus it is reasonable to expect that linked selection will bias demographic inference that assumes strict neutrality of population genomic data. Indeed, background selection has recently been shown to skew demographic inferences using the site frequency spectrum [316]. Here, we show through extensive simulation that positive selection can severely impair demographic model selection and parameter estimation based on the SFS, summary statistics of variation, and reduced approximations of the ancestral recombination graph (i.e. PSMC). The extent to which this is so depends on the fraction of genetic loci examined during inference affected by a recent sweep, and the ratio of the genetic distance between the locus and the target of selection to the selection coefficient, c/s.

When the fraction of loci affected by linked selection is low, we have shown that point estimates of population parameters estimated under the correct demographic scenario are reasonably accurate using both SFS-based inference and ABC with summary statistics (Figures 4.1 to 4.3); the exact fraction, however, depends on the model in question. Unless f is quite low, our results indicate that when model selection is applied using either SFS-based or ABC inference, linked selection can bias model choice (Tables 4.1 and 4.2). In many of our simulated datasets we have assumed that loci linked to sweeps are on average a distance of $c/s = 0.5$ away from the sweep (i.e. drawn uniformly from between 0 and 1). In real genomes this may correspond to quite a large physical distance. For instance, if we assume a selection coefficient of 0.05 (i.e. selection as strong as in our simulations) and a crossover rate of 2 cM/Mb (similar to estimates in Drosophila; [e.g. 235]) this corresponds to a physical distance of 1.25 Mb. If instead we assume a crossover rate of 1 cM/Mb (similar to estimates from humans; [e.g. 321]), this corresponds to a physical distance of 2.5 Mb. While we have assumed a fairly high value of s that may not be representative of all selective sweeps, known sweeps in humans may often have selection coefficients fairly close to 0.05 [322].

Thus, even if there are a small number of recent selective sweeps, the majority of the genome may nonetheless be sufficiently impacted by linked selection to produce biased demographic inferences. For example, if a human population experienced 1,200 recent sweeps fairly evenly spaced across the genome (the equivalent of one recent sweep in $\sim$ 5% of genes), every site in the genome would be within a c/s distance of 0.5 from the nearest sweep (i.e. $f = 1.0$). In *Drosophila*, this would be the case if there were 120 evenly spaced recent sweeps. This is a very small number indeed, equivalent to a single recent sweep affecting $< 1\%$ of all genes. Indeed, in *Drosophila* the fraction of loci affected by recent positive selection may be quite large [239]. Numerous studies have estimated that the fraction of adaptive amino acid substitutions in *D. melanogaster* is considerable, with estimates ranging from 10-50% [226, 239, 323, 324]. Positive selection may therefore be particularly troublesome for demographic inference in Drosophila and other organisms were adaptive natural selection is similarly pervasive. In humans, where positive selection is perhaps less common [325], this may be less of a thorny issue. However, some have argued that selection may be pervasive in the human genome as well [326, 327], and certainly humans show many adaptations to local environments [e.g. 328, 329, 330, 331, 332]. Nonetheless, the 10-fold increase in the number of sweeps (under our parameterization) required to produce the same level of bias suggests that the confounding effect of demography in humans may be less considerable than in *Drosophila*. However, given uncertainty in the number, location, strength, and type of selective sweeps, we are unable to quantify the extent to which demographic inferences in either species are skewed by adaptation.

A new and promising class of methods for inferring demographic histories rely on estimating approximations to the ancestral recombination graph (ARG) using a sequentially Markovian coalescent (See Section 1.4.3 and [155, 163]). Our findings suggest that natural selection may alter the shape of, and inflate the degree of change in, these inferred histories. Indeed, because of the specific way in which a sweep perturbs the ARG locally during the coalescent history of a chromosome, PSMC inference on regions that have experienced one or more sweeps in the past may lead to erroneous estimation of a population bottleneck. If sweeps continue until the present day, PSMC inference might appear to support a population contraction rather than a bottleneck, though this may very well be a result of PSMC having lower power for very recent population dynamics [155]. Under a truly recurrent sweep model [e.g. 333], it is unclear what the behavior of inference using PSMC might be. Note that with PSMC we infer population size changes from a large simulated chromosomal segment (corresponding to $\sim$ 15 Mb in the human genome) experiencing only a single selective sweep $0.2Ne$ generations ago. This is equivalent a total of $\sim$ 200 fairly recent sweeps across the human genome.

Our results are broadly concordant with those of Messer and Petrov's (2013), examination of the McDonaldKreitman tests ability to infer the fraction of substitutions that were adaptive (α) under a simulated recurrent hitchhiking scenario with constant population size [334, 335]. Their study found that Eyre-Walker and Keightleys DFE-alpha method [336], which simultaneously estimates α, the distribution of fitness effects, and a two-epoch population size history, incorrectly inferred the presence of population size changes [334]. It is therefore reasonable to assume that positive selection could have a substantial confounding effect on a variety of population genomic methods for demographic inference in practice, beyond those considered here. In the empirical literature, numerous recent studies of demographic history have found support for contractions and recent expansions of natural populations [6, 134, 153, 156, 337]. While such population size changes are probably common and our results do not call the major findings of these studies into question, they do suggest that natural selection exaggerates the inferred intensity of these changes.

This study has examined only a single model of adaptive natural selection, and therefore has several limitations. Throughout we have assumed that positive selection occurs only through completed hard selective sweeps. Indeed soft sweeps [79, 108, 122, 338] and partial sweeps [82, 339, 340], may be widespread, and differ in their effects on linked polymorphism [4, 222, 223, 341]. Polygenic selection, in which alleles at several different loci underlying a trait under selection will experience a change in frequency, is also thought to be widespread [342, 343]. Such polygenic adaptation is known to leave its own unique signature on patterns of population genetic variation [343]. These alternative modes of positive selection could skew demographic inferences in a different manner than what we have observed in this study. Positive selection may also affect estimation of multi-population demographic scenarios: though we did not examine this here, Mathew and Jensen [344] recently showed that selective sweeps will impair parameter estimates for a two-population isolation-with-migration model. Thus our results, combined with those of Ewing and Jensen [316], Messer and Petrov [334], Mathew and Jensen [344], strongly suggest that the problem of natural selection skewing demographic inference is a general one.

The observations we have made here also suggest some steps that can be taken to mitigate the impact of positive selection. First, we note that in general $\partial a \partial i$ (i.e. SFS-based inference) appears to be somewhat more robust to the effects of selection than does our ABC approach based on summary statistics. Perhaps this is because $\partial a \partial i$ uses an SFS summed across loci, such that more polymorphic regions will have a greater weight on the shape of the SFS. Thus, the extent to which regions most affected by sweeps contribute to the SFS is diminished implicitly, as these regions will exhibit less variation. Relying on the SFS rather than summaries of variation that, to a greater extent, depend on the number of segregating sites may therefore reduce selection's confounding effect on inferred relative population size changes via estimates of $4N_u$ and therefore the absolute population size may be biased. We also found that including variances of summary statistics when performing ABC can inflate error when an intermediate number of loci are linked to sweeps, perhaps because this mixture of two evolutionary models (neutrality and positive selection) inflates the variance. Omitting variances may therefore reduce the confounding effect of selection.

Finally, we have shown convincingly that the proximity of selective sweeps to genomic regions used for inference (as measured by c/s) has a large effect on the magnitude of bias (Figures F.21 to F.23). It is of paramount importance, therefore, to select regions located as far away in genetic distance as possible from genes and other functional DNA elements [315]. While this is so, it may not be possible to move far enough away from potential targets of selection to completely eliminate any bias (as discussed above). Moreover, it is essential to omit regions with lowered recombination rates, where the impact of linked selection will be strongest [345]. Our results also motivate the challenging task of simultaneous estimation of parameters related to natural selection and demographic history [336, 346]. Until an approach to obtain accurate estimates of demographic parameters in the face of natural selection is devised, population size histories inferred from population genetic datasets could remain significantly biased.

4.4 Methods

4.4.1 Simulating Demographic And Selective Histories To Test Inference Methods

To test the robustness of ∂a∂i and ABC to positive selection, we generated coalescent simulations from four different demographic scenarios: 1) a constant population size model; 2) a three-epoch population bottleneck (the European model from Marth et al. [5]); 3) a model of recent exponential population growth; 4) and a three-epoch model with a population contraction followed by stasis and then recent exponential population growth (the European model from Gravel et al. [6]). These models, their prior distributions, and parameters are shown in Figure F.20 and Table E.19.

For each demographic model, we simulated 100 observed genomes experiencing no natural selection, each of which was summarized by a collection of 500 unlinked loci sequenced in a sample size of 200 individual sequences. We then repeated these simulations while stipulating that a specified fraction of loci (f) were linked to a recent selective sweep where the selected mutation reached fixation immediately prior to sampling. The selection coefficient, s, for this mutation was always set to 0.05, with a completely additive fitness effect ($h = 0.5$). For each simulation with a selective sweep, we specified the genetic distance of the sweep from the sampled locus by the ratio c/s, where c is the crossover rate per base pair multiplied by the physical distance to the sweep, and s is again the selection coefficient. We examined values of f that were multiples of 0.1 between 0.1 (10% of loci linked to a sweep) and 1.0 (100% of loci linked to a sweep). Values of c/s examined were multiples of 0.1, raging from

0.0 (the sweep occurred immediately adjacent to the locus being used for inference) to 1.0 (4.17 Mb given our value of s and our recombination rate). We generated sets of simulations with a given value of f by combining the appropriate numbers of neutral simulations and simulated loci linked to a sweep. For each combination of f and c/s (110 combinations in total), we generated 100 sets of 500 unlinked loci.

We also simulated large chromosomal regions to which we applied PSMC. These simulations were of 15 Mb regions with a constant-size population ($Ne = 10,000$), from which two individuals were sampled. These 15 Mb regions either experienced no selective sweeps, one selective sweep fixing 0.4N generations ago, three selective sweeps (fixing 0.4N generations ago, $0.2N$ generations ago, and immediately prior to sampling) or five selective sweeps ($0.4N$,$0.3N$, $0.2N$, $0.1N$, or 0 generations prior to sampling). The location of each sweep was thrown down randomly along the chromosome. For each scenario, 100 replicate simulations were generated.

For all simulations we used parameters relevant to human populations: a recombination rate of 1.0×10^{-8} (approximately equal to the sex-averaged rate from Kong et al. 2010 [321]), and a mutation rate of 1.2×10^{-8} (from Kong et al. 2012 [347]).

4.4.2 Parameter Estimation and Model Selection Using $\partial a \partial i$

We downloaded version 1.6.3 of $\partial a \partial i$ [133], which we programmed to optimize the parameters of the bottleneck, growth, and contraction-then-growth models. For each model we used a two-step constrained optimization procedure to find the combination of demographic parameters that have the highest likelihood given the site frequency spectrum measured across all 500 unlinked loci in the simulated genome. First, we performed a coarse optimization using the Augmented Lagrangian Particle Swarm Optimizer [293], and then refined this solution using Sequential Least Squares Quadratic Programming [294]. Both of these techniques are implemented in the pyOpt package (version 1.2.0) for optimization [295].

To asses the accuracy of point estimation of parameters in the face of varying amounts of and genetic distances to selective sweeps, we optimized the parameters of each demographic model against each data set simulated under that model, comparing estimated values to the true values. As shown in Section 4.2, this approach was quite successful recovering the true parameter values of each demographic model when applied to data simulated under neutrality. However, one exception was the bottleneck model, for which the optimal solution was typically a shorter but more severe bottleneck than the one we had simulated. We therefore fixed the bottleneck duration to the true value of 500 generations, after which $\partial a \partial i$ was able to estimate the remaining parameter values with acceptable accuracy.

To assess the support for a given demographic model, we obtained the likelihood for a simulated data set of each demographic model under the optimal parameters estimated by $\partial a \partial i$, and then from this likelihood and the number of parameters of the model calculated the AIC [231]. For the constant population size model, there are no optimized parameters, so the AIC is simply $-2 \ln L_M$ of the model, M. Model selection was performed for each data set simulated with constant population size, with or without selection. For each model with variable population size, the python script we used to perform parameter optimization and obtain the likelihood of the optimal parameterization has been deposited at `https://github.com/kern-lab/demogPosSelDadiScripts`, as has the script used to obtain the likelihood under the constant population size model.

To perform formal model selection, we asked for a given simulated data set whether any non-equilibrium model had an AIC at least 50 units greater than that of the equilibrium model. If so, we asked whether any of our three non-equilibrium models had an AIC at least 50 units than the other two, in which we selected that model; otherwise we classified the simulated data set as "ambiguous but non-equilibrium". If no non-equilibrium model had an AIC at least 50 units greater than the equilibrium model, then we conservatively classified the simulated data set as "equilibrium".

4.4.3 Parameter Estimation and Model Selection Using Approximate Bayes Computation

To estimate the parameters of each of our four demographic models under an Approximate Bayes framework, we create two datasets: an "observed" dataset and a "sample" dataset. Using the coalescent simulations described in Section 4.4.1, as the basis for these observed data, we first summarize their SFS and haplotype structure by calculating the means and variances of π, the number of segregating sites, Tajima's D, $\hat{\theta}_H$, and haplotype count. As our goal is to observe how robust ABC is to the effects of increasing levels of linked selection when estimating parameters, we create 10 observed datasets by combining a number of unlinked loci with a varying fraction of loci, f, linked to a sweep, where $f \in \{0.1, 0.2, \ldots, 1.0\}$. In order to model varying genetic distances, c/s, of these linked loci from the associated selective sweep, each individual linked locus is randomly sampled from loci with $c/s \in \{0.0, 0.1, \ldots, 1.0\}$. Thus, for each demographic model, we create 10 observed data sets; one set composed entirely of unlinked neutral loci, and nine sets possessing unlinked loci combined with varying fractions of linked sites.

We next turn our attention to creating a sample dataset for each of our four demographic models. The sample data contains 5.0×10^5 examples, each example representing coalescent simulations of 500 unlinked loci, each of sample size 200. Here, the simulation parameters are drawn from prior distributions, those distributions conditional on the demographic model of interest (Figure F.20 and Table E.19). As with the observed data, these coalescent simulations are summarized using π, the number of segregating sites, Tajima's D, $\hat{\theta}_H$, and haplotype count.

We utilize the `ABCreg` software package to perform parameter estimation for each of our four demographic models [317]. We opted to apply the conventional tangent transformation procedure to the parameters sampled from our prior distributions via passing the `T` flag and set the tolerance parameter, `-t = 0.001`, thus retaining 0.1% of our sample data for use in estimating the posterior parameter distributions [348]. After a rejection sampling step using the indicated tolerance, the now-standard practice of weighted linear regression is applied [149]. From each of the resultant estimated posterior parameter distributions we calculate the maximum a posteriori (MAP; posterior mode) and retain these as our parameter point estimators.

For the bottleneck and contraction-then-growth models, our initial efforts to estimate parameters under neutrality failed to approximate the true parameterizations. We therefore fixed the values of the times of population contraction parameters of these models (referred to as T_B and T_C respectively) during both parameter estimation and model selection. These parameters were always set to the true values during sampling simulations and their values were not estimated. After this change, we were able to estimate the parameters of each model under neutrality with reasonable accuracy.

For each of the 100 simulated genomes described in Section 4.4.1, we repeat the above procedures: ten new observed datasets are constructed, one for each fraction of unlinked loci replaced with loci linked to a selected sweep, and the ABC analysis is again conducted. We summarize the demographic parameter estimates obtained via ABC using boxplots shown in Figures 4.1 to 4.3.

To determine how model choice under the ABC framework may be affected selection, we simulated neutral, unlinked loci under a stationary demographic history as described in Section 4.4.1. These simulations, which serve as the "observed" data, were again summarized, as in Section 4.4.3, using the means and variances of π, segregating sites, Tajima's D, $\hat{\theta}_H$, and the number of haplotypes. For each of the four demographic models being examined we created "sample" datasets composed of 5.0×10^5 examples, each example representing coalescent simulations of 500 unlinked loci of sample size 200.

We used the R package `abc` to conduct model choice, which allows for logistic regression-based estimation of the posterior probabilities of a model [318], given a minimum of two models to compare. We examined six pair-wise model selection scenarios, comparing each demographic model against the others (equilibrium demography & growth demography, equilibrium demography & contraction/growth demography, equilibrium demography & bottleneck demography, bottleneck demography & contraction/growth demography, bottleneck demography & growth demography, and contraction/growth demography & growth demography). We first determine if any of the non-equilibrium models have stronger support than the equilibrium model by a

Bayes Factor $\geq$ 20. If they do, we then determine if this particular model has greater support than all other models by a Bayes Factor $\geq$ 20. If so, we then call that model the supported model. If they do not, and the equilibrium models Bayes Factors are all $\geq$ 20, we declare the equilibrium model unambiguously the best fit. However, if a non-equilibrium model fits the data better than the equilibrium model, but does not have the strongest support across all other non-equilibrium models (BF $\geq$ 20), we declare that "ambiguous but non-equilibrium".

Each of the model comparisons were applied to ten sets of the observed data that, as described above, are composed of unlinked loci simulated under a stationary demographic history, combined with successively increasing amounts of linked loci, resulting in the overall fraction of linked loci in a given observed dataset to range from $\{0.0, 0.1, \ldots, 1.0\}$. In addition to Tables 4.1 and 4.2 that show the results of the model selection analysis, we further summarize the results of the ABC model choice for each non-equilibrium vs. equilibrium pair-wise comparison by plotting the estimated posterior probability of the alternative model at each fraction of linked loci in the observed data in Figure F.26.

4.4.4 Inferring Population Size Histories With PSMC

We ran PSMC in order to infer the history of population size changes of our 15 Mb simulations from which two individuals were sampled (see above). Briefly, we converted our simulation output to the same format generated by running msHOT-lite (`https://github.com/lh3/foreign/tree/master/msHOT-lite`) with the `-l` flag. We then ran PSMCs `ms2psmcfa.pl` script with default parameters to generate input for PSMC, which we ran with default parameters.

Finally, we ran PSMCs `psmc2history.pl` script with default parameters to output the inferred population size history. We then rescaled the output from units of Ne to years (after rescaling to generations and assuming a 30 year generation time) and numbers of individuals using the estimated value of θ. For each selective scenario, we ran PSMC separately on all 100 simulated population samples. Finally, for the purposes of visualization we obtained a median estimate of population size across time by examining a large number of time points (one every 100 years) across the entire period examined, and at each time point taking the median population size estimate from all 100 simulations.

Appendix A

Population genetics summary statistics

A.1 Nucleotide Diversity

A common measure of genetic variation between genomic samples is the average number of nucleotide differences between sites or sequences, often denoted as π [349].

$$\pi = \sum_{ij} x_i x_j \pi_{ij} = 2 \sum_{i=2}^{n} \sum_{j=1}^{i-1} x_i x_j \pi_{ij} \tag{A.1}$$

where n is the number of sequences examined, x_i is the ith sequence in the population, and π_{ij} is the number of nucleotide differences per nucleotide site between the ith and jth sequences.

A.2 Segregating Sites

Segregating sites are defined as the number of nucleotide positions in an alignment of genomic sequence data that exhibit a polymorphism.

A.3 Tajima's *D* Statistic

The *D* statistic was developed to differentiate between a DNA sequence that is evolving under neutral or non-stochastic processes [86].

$$D = \frac{d}{\sqrt{\widehat{V}(d)}} = \frac{\widehat{k} - \frac{S}{a_1}}{\sqrt{[e_1 S + e_2 S(S-1)]}} \tag{A.2}$$

where,

$$e_1 = \frac{c_1}{a_1} \tag{A.3}$$

$$e_2 = \frac{c_2}{a_1^2 + a_2} \tag{A.4}$$

$$c_1 = b_1 - \frac{1}{a_1} \tag{A.5}$$

$$c_2 = b_2 - \frac{n+2}{a_1 n} + \frac{a_2}{a_1^2} \tag{A.6}$$

$$b_1 = \frac{n+1}{3(n-1)} \tag{A.7}$$

$$b_2 = \frac{2(n^2+n+3)}{9n(n-1)} \tag{A.8}$$

$$a_1 = \sum_{i=1}^{n-1} \frac{1}{i} \tag{A.9}$$

$$a_2 = \sum_{i=1}^{n-1} \frac{1}{i^2} \tag{A.10}$$

Here, $\widehat{k}$, is the measure of nucleotide diversity, π, and S is the number of segregating sites.

A.4 Fay and Wu's H Statistic

In the same vein as Appendix A.3, the H statistic also aims to draw a distinction between evolution resulting from neutral processes or non-random processes. Here, however, H seeks to specifically allow for differentiating neutrality from positive adaptation [76]. To calculate H, we first need to calculate $\widehat{\theta_H}$ and then simply subtract from this value the value of π.

$$\widehat{\theta}_H = \sum_{i=1}^{n-1} \frac{2S_i i^2}{n(n-1)} \tag{A.11}$$

where n is the sample size, S_i is the number of derived variants, and i is the number of times those variants are found. H is then the difference between π and $\widehat{\theta}_H$. An outgroup is needed to infer the derived and ancestral allele states. While π is affected by intermediate frequency variants, $\widehat{\theta}_H$ is influenced greatest by variants of high frequency [76]. Tajima's D (Appendix A.3) and Fay and Wu's H both look to reject neutrality in favor of an alternative hypothesis, D will only reject neutrality when an excess of low frequency variants are present, the corollary holds that H will reject neutrality when an over-abundance of high frequency variants are present [76].

A.5 Kelly's Z_nS Statistic

Kelly's Z_nS statistic, [234], is used to measure linkage disequilibrium in a sample of polymorphic sites, S. Specifically, Z_nS is the average pairwise LD across all polymorphic sites, S, within a sample of size n and is defined as

$$Z_nS = \frac{2}{S(S-1)} \sum_{i=1}^{S-1} \sum_{j=i+1}^{S} \delta_{ij} \tag{A.12}$$

where δ_{ij} is a standardized measure of linkage disequilibrium. This squared correlation between locus i and locus j is defined as

$$\delta_{ij} = \frac{D_{ij}^2}{p_i(1-p_i)p_j(1-p_j)} \tag{A.13}$$

where p_i and p_j are the frequency of the derived allele at the ith and jth loci, respectively, and D_{ij} is the measure of linkage disequilibrium between i and j. It is defined as

$$D_{ij} = p_{ij} - p_i p_j \tag{A.14}$$

and p_{ij} is the frequency when derived alleles are present at both i and j.

A.6 Kim and Nielsen's ω Statistic

Their original ω statistic is based upon the idea that a selective sweep leaves an increase in linkage disequilibrium within the regions adjacent to a fixed or selected site, but that this excess of LD does not extend *across* the selected site [92]. Their ω statistic is calculated thusly: if there are S polymorphic sites in the data, they are divided into two groups - one group from the first to the l-th polymorphic site measured from the left and the other group from the $(l-1)$th to the last site $(l = 2, ..., S-2)$.

$$\omega = \frac{\left(\binom{l}{2} + \binom{S-l}{2}\right)^{-1} \left(\sum_{i,j \in L} r_{ij}^2 + \sum_{i,j \in R} r_{ij}^2\right)}{(1/l(S-l)) \sum_{i \in L, j \in R} r_{ij}^2} \tag{A.15}$$

The value of l that maximizes ω, ω_{max}, is then used as a test statistic.

Appendix B

Approximate Bayes computation (ABC)

Bayes theory relates the conditional probability of a parameter, θ, given some data, D, to the probability of D given θ, by

$$p\left(\theta|D\right) = \frac{p\left(D|\theta\right)p\left(\theta\right)}{p\left(D\right)} \tag{B.1}$$

where $p\left(\theta|D\right)$ is known as the posterior probability, $p\left(D|\theta\right)$ is the likelihood, $p\left(\theta\right)$ is the prior probability, and $p\left(D\right)$ is known as the marginal likelihood and is often ignored in the explicit calculation of $p\left(\theta|D\right)$ as it is a normalizing constant. This leads to the expression

$$p\left(\theta|D\right) \propto p\left(D|\theta\right)p\left(\theta\right) \tag{B.2}$$

Even without having to calculate the evidence (marginal likelihood) in eq. (B.1), it may be that solving $p\left(\theta|D\right)$ remains intractable. In that case, the

ABC approach has been shown to be effective at estimating the posterior distribution of θ [149, 153].

In ABC, simulations $D^{\star}$ are generated under a model M using known parameter values θ. $D^{\star}$ is "accepted" if it resembles the true data, D. More formally,

$$p\left(D^{\star}, D\right) \leq \varepsilon \tag{B.3}$$

where $\varepsilon \geq 0$.

Typically, a simple distance metric such as Euclidean distance is used to measure $p(D^{\star}, D)$. And this is the where the approximation of the posterior distribution occurs, namely, that instead of sampling from $p\left(\theta|D\right)$, samples are actually being drawn from $p\left(\theta|p(D^{\star}, D\right) \leq \varepsilon)$. However, if the tolerance, ε is small, then $p\left(\theta|p(D^{\star}, D\right) \leq \varepsilon)$ should be a close approximation to the true posterior.

It should be noted, however, that in population genetic analysis summary statistics of the data, both SFS and LD based, are frequently used in ABC studies [105]. In the case of ABC analysis this means that not only are approximating the true posterior distribution, you are also losing whatever information is not contained when mapping the data to its summarized form. If summary statistics were truly sufficient statistics of the data, this would not be a problem, however it is not known if there exist sufficient statistics in the population genetics to describe linkage structure or patterns of polymorphism, and as such, information is inevitably lost. The use of summary statistics, $S(D)$, leads to eq. (B.3) being rewritten as

$$p\left(S(D^{\star}), S(D)\right) \leq \varepsilon \tag{B.4}$$

Appendix C

The Coalescent

C.1 Basic Coalescent

The coalescent, in its most basic form, is a stochastic process that describes the distribution of genealogies of a sample of genes or alleles. In this framework, as time moves backwards from the present into the past, alleles are allowed to coalesce with each other until there is only one allele (lineage) remaining, which is termed the "most recent common ancestor" (MRCA) of the gene or allele. The theory was first formalized in 1982 by Kingman [99] and from this work a number of applications have been developed, as well as advancements in the theory itself.

C.1.1 Discrete Time Coalescent

We begin with the Wright-Fisher model of reproduction [350, 351] and operate under the following assumptions: there exists discrete generations that do not overlap, we deal with either haploid individuals or two subpopulations of males and females, a constant population size, no fitness differences, and no recombination.

Working under a Wright-Fisher model we might wish to know how long it took 2 genes in a sample of $2N$ genes to find their MRCA, and we could approach it likewise: If these two genes were to find their MRCA in one generation, the probability would be $1/2N$. That is, the first gene can choose any ancestor as a parent, but the other gene must choose the same parent, and that probability depends on the number of genes in the sample, in this case $2N$. Therefore, the probability of finding the MRCA in one generation is $1/2N$, and it follows that the probability of *not* finding the MRCA in one generation is $1 - 1/2N$. To find the probability that these two genes find a common ancestor in j generations,

$$P\left(T = j\right) = \left(1 - \frac{1}{2N}\right)^{j-1} \frac{1}{2N} \tag{C.1}$$

This follows from the fact that for $j - 1$ generations the genes do not pick the same ancestor, then at generation j they choose the "correct" (same) ancestor. Notice that the expression in Equation (C.1) is actually the probability mass function for the geometric distribution.

$$P\left(T = j\right) = \left(1 - p\right)^{j-1} p \tag{C.2}$$

In this case, time to coalescence (in discrete generations), T, is geometrically distributed with parameter $p = 1/2N$. The expected time to coalesce is then, $\mathbf{E}(T) = 1/p = (1/2N)^{-1} = 2N$.

We might also wish to know the probability of k genes in a sample of n genes coalescing in T generations. In this case, we may make use of the binomial coefficient and Equation (C.2).

$$P\left(T = j\right) \sim \left[1 - \binom{k}{2}\frac{1}{2N}\right]^{j-1} \binom{k}{2}\frac{1}{2N} \tag{C.3}$$

C.1.2 Continuous Time Coalescent

In the above examples we use discrete units of time, but we are conditioning on population size in our calculations. By using the mean time for two genes to find a common ancestor, $2N$ generations, we can scale time to continuous units, only needing population size if we wish to convert time back into generations. More specifically, we set time $t = j/2N$, where j is generations, we can say that the time for k genes to find $k-1$ ancestors is an exponential variable, $T \sim \mathbf{Exp}\big(\binom{k}{2}\big)$:

$$\mathbf{P}(T \leq t) = 1 - e^{-\binom{k}{2}t} \tag{C.4}$$

C.1.3 Tree Height and Total Branch Length

With the continuous time coalescent we are now able to calculate two quantities of interest: The height of the coalescent tree, and the total branch length of the tree. The latter distribution is obtained via a convolution of exponential variables [352]. We can calculate the mean height of the tree,

$$\mathbf{E}(H) = \sum_{j=2}^{n} \mathbf{E}(T_j) = 2\sum_{j=2}^{n} \frac{1}{j(j-1)} = 2\left(1 - \frac{1}{n}\right) \tag{C.5}$$

where $n =$ number of genes in the sample and $j = n, n-1, \ldots, 2$ ancestors.

Using properties of the exponential distribution, we can also calculate both the distribution of total branch lengths and the expected total branch length [352],

$$\mathbf{P}(L \leq t) = \left(1 - e^{-t/2}\right)^{n-1} \tag{C.6}$$

$$\mathbf{E}(L) = \sum_{j=2}^{n} j\mathbf{E}(T_j) = 2\sum_{j=1}^{n-1} \frac{1}{j} \tag{C.7}$$

Worth noting is that the right expression in Equation (C.7) is proportional to the log of n,

$$\sum_{j=1}^{n-1} \frac{1}{j} \propto \log(n) \tag{C.8}$$

C.1.4 Coalescent with recombination

The coalescent as described above assumes no recombination. A coalescent model featuring recombination was reported by Hudson shortly after Kingman's seminal work was published [100]. Recombination is problematic in the coalescent for a number of reasons. First, it allows different parts of a chromosome to have different tree topologies. Secondly, the relationship between sequences now takes the form of a graph, not a simple tree. Actually, at each position in the sequence there exists a "local tree", and the whole genealogy is comprised of all local trees, one tree for each position in the sequence.

Coalescent events and recombination events "compete" to create a genealogy. In the same way that the waiting time to a coalescence was shown in Equation (C.4), we also would like to know the waiting time and the probability of a recombination event. Using a scaled rate of recombination, $\rho = 4Nr$, where r is the crossing over rate per generation, and by approximating the geometric distribution with an exponential distribution, the waiting time until a recombination event occurs within a sequence is an exponential variable with parameter $\rho/2$. More specifically,

$$\mathbf{P}(T \leq t) = 1 - (1-r)^j = 1 - \left(1 - \frac{2Nr}{2N}\right)^{2Nt} \approx 1 - e^{-\rho t/2} \tag{C.9}$$

where $j = 2Nt$. Note the second expression from the left is the geometric cdf. Given k ancestral samples, the waiting time to recombination is also an exponential variable, parameterized by $k\rho/2$.

From Equation (C.4) the waiting time to a coalescence is $T \sim \mathbf{Exp}\big(\binom{k}{2}\big) = T \sim \mathbf{Exp}\big(k(k-1)/2\big)$ and from Equation (C.9) the waiting time until a recombination event is $T \sim \mathbf{Exp}\big(\rho k/2\big)$. These independent distributions allow us to write the combined rate parameter as

$$\frac{k\,(k-1)}{2} + \frac{\rho k}{2} \tag{C.10}$$

The probability that the first event is a coalescence is an exponential variable with parameter:

$$\frac{\lambda_{coal}}{\lambda_{coal} + \lambda_{rec}} = \frac{\big(k(k-1)/2\big)}{\big(k(k-1)/2\big) + \big(\rho k/2\big)} = \frac{k-1}{k-1-\rho} \tag{C.11}$$

and the probability that the event is a recombination is an exponential variable with parameter:

$$\frac{\lambda_{rec}}{\lambda_{coal} + \lambda_{rec}} = \frac{\big(\rho k/2\big)}{\big(k(k-1)/2\big) + \big(\rho k/2\big)} = \frac{\rho}{k-1+\rho} \tag{C.12}$$

Given waiting times for either a recombination or coalescence we can now simulate a genealogical process: If $k = n$ number of genes, time to the next event is and exponential variable who's rate parameter is given in Equation (C.10). With probability given in Equation (C.11), that event is a coalescence, otherwise that event is a recombination. This process continues until $k = 1$ (See Algorithm 1).

The algorithm in the preceding section results in a graph being generated, deemed the *ancestral recombination graph* (ARG) [353]. An alternative to this algorithm is presented in the next section.

Algorithm 1 Coalescent With Recombination

```
$k = n$ number of genes
procedure TIME_TO_EVENT($combined\_rate\_param$)    ▷ Given in Equation (C.10)
    $T = \mathbf{Exp}(combined\_rate\_param)$
end procedure
procedure COAL_OR_RECOMB($coal\_rate, recomb\_rate$, k)    ▷ From Equations (C.11) and (C.12)
    while $k \neq 1$ do
        $\mathbf{P}_{coal} = \mathbf{Exp}\ (coal\_rate)$
        if $\min(\mathbf{U}(0,1), \mathbf{P}_{coal})$ then    ▷ The first event is a coalescence
            $k = k - 1$
        else    ▷ The first event is a recombination
            $k = k + 1$
        end if
    end while
end procedure
```

C.2 Spatial Coalescent

In contrast to the coalescent process described in the previous section, Wiuf and Hein introduced an alternative algorithm in 1999 which moves along the sequence [354]. They called this implementation a "spatial" coalescent, as the genealogical history is modified not as going back in time, but as recombinations occur as one moves left to right along the sequence. The algorithm works in the following manner: Starting at the first position in the sequence (left), simulate a normal genealogy. Next, find the first recombination "break point" as you move left to right along the sequence. Once the break point is found, choose a branch to undergo recombination and modify the genealogy.

Using the example given in [352] pictured here in Figure C.1, an example of the spatial algorithm follows: Starting at the left-most position, and assuming the total branch length is 1.8, draw an exponential variable with parameter 1.8.

In Figure C.1, this variable is 0.87. Choose a random point on the tree for a recombination event to occur at, such that all positions from 0.0 to 0.87 have the same local tree and positions greater than 0.87 have different trees. Now, assuming that the total branch length for the tree less than or equal to 0.87 is 3.3, we repeat these steps again. Draw an exponential variable with parameter 3.3 (in Figure C.1 this value is 1.05) and add it to the previous breakpoint, $0.87 + 1.05$, which gives the location of the next recombination break point, 1.92. Again, choose a random location on the branches and coalesce it at 1.92. Continue this process until the complete history has been created, that is, until the drawn recombination distance is past the right end of the sequence.

The spatial coalescent builds up a graph as one moves left to right along the sequence. This graph is actually embedded in the ARG, as are all subgraphs generated in the spatial algorithm (SAG - spatial algorithm graph). It is important to note that while the coalescent with recombination is a Markovian process since we can infer generation $t + 1$ from generation t moving into the past, the spatial coalescent is not truly Markovian. This is due to the fact that in order to determine the genealogical history at position $q > p$ along a sequence, it requires that sequences non-ancestral to p be utilized. This means that p itself doesn't contain enough information to move spatially - that is, knowing the local tree at p is not sufficient to infer the tree at q [352, 354]. An approximation to this algorithm that retains a true Markovian structure is presented in the next section.

C.3 Sequential Markov Coalescent

As mentioned in the previous section, a potentially undesirable feature of the spatial coalescent is that when a recombination occurs, these newly created sequences can coalesce back to branches consisting of non-ancestral material. This leads to a huge graph, and as such, the need to track this whole, ever-expanding ARG. This is the point of departure between the spatial algorithm of Wiuf and Hein, and the work pioneered by McVean and Cardin, which the later have termed the Sequentially Markov Coalescent (SMC) [161]. In their model, an approximation to the coalescent with recombination, the algorithm progresses just like the spatial coalescent, however, lineages that share no interval of ancestral material are not allowed to coalesce. That is, when a recombination breakpoint is posited, one line in the ARG detaches from the graph and the portion of this line above is deleted from the graph - just like in the spatial coalescent. However, in SMC, this line now has to reattach to the local tree (See Figure C.2). In this way, there is no need to track the full ARG. This serves as the distinction between the spatial coalescent and the SMC, and which further differs in the following three ways: 1) The ARG state space is greatly reduced, 2) the SMC algorithm greatly reduces the number of recombinations in the genealogical history [355], and 3) a true Markovian structure now exists in each sequential genealogy along the sequence [161].

The sequential Markovian coalescent algorithm serves as the basis for the *psmc* which is described in the next section. It should be noted that while the spatial coalescent of Wiuf and Hein results in identical genealogies as the standard coalescent, this algorithm is quite inefficient for large chromosomes. The SMC algorithm of McVean and Cardin has been shown to do quite well in approximating the standard coalescent, and, importantly, is computationally efficient and allows for the genealogies of large chromosomes to be sampled [355].

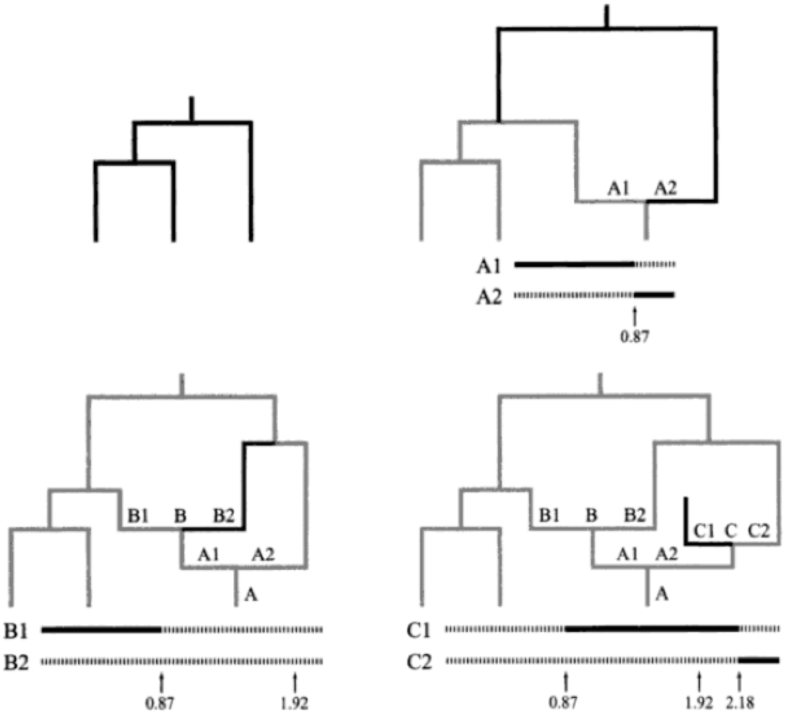


Figure C.1: Taken from Hein, et al. 2005 [352].

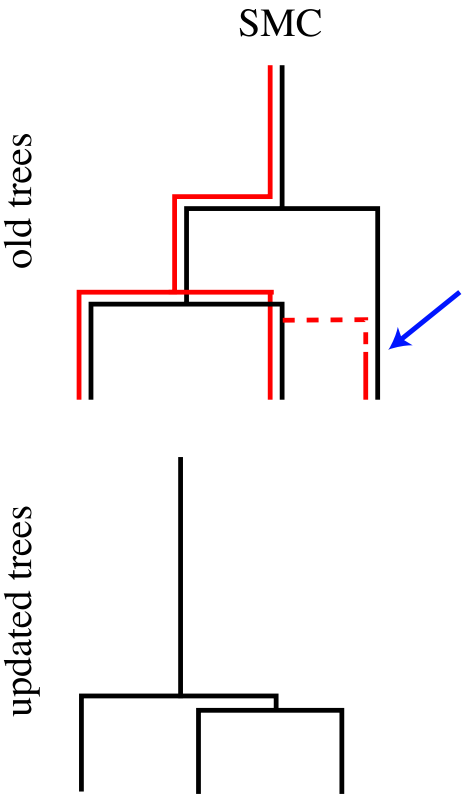


Figure C.2: Taken from Eriksson et al., 2009 [355]

.

Appendix D

Support Vector Machine

D.1 Support Vector Machine

The learning or training phase of an SVM consists of an optimization routine of which the goal is to find the optimum hyperplane that separates the two classes of data with a maximum margin between those two classes (Figures D.1 and D.2). In the case where the data are not linearly separable, this algorithm allows the user to map the data to a higher dimension through the use of a kernel function, to where linear separability exists. However, if the data, now in their transformed, high dimensional feature space still can not be separated, we then introduce a slack variable, ξ_i, which measures the error of misclassification of $\mathbf{x}_i$. This formulation is known as the "soft-margin" SVM and allows one to still separate the data as best as possible by maximizing the margin between those data points which *can* be linearly separated, and minimizing the penalty incurred for those points that cannot [300].

More specifically, we have a set of training data $(\mathbf{x}_i, y_i)$ $i = 1, ..., l$, where $\mathbf{x}_i \in$

R^n and $\mathbf{y} \in \{1, -1\}$. The data, $\mathbf{x}_i$, are row vectors of summary statistics and the class labels, y_i, are either 1 or -1. We therefore wish to solve the optimization problem:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l} \xi_i \tag{D.1}$$

$$= \min_{\mathbf{w},b,\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l} \xi_i \tag{D.2}$$

$$\text{where} \quad y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i,\ \forall i, \tag{D.3}$$

$$\xi_i \geq 0 \tag{D.4}$$

Here, $\mathbf{w}$ is the normal vector to the hyperplane, C is the cost parameter of the error term and the radial basis function $\phi(\mathbf{x}_i)$ maps the original data to a higher dimensional feature space. In this report, we use one of the most popular kernels, known as the Gaussian radial basis function (RBF) to transform our data. This kernel is a good choice, as it is numerically stable and requires tuning only one parameter, γ. Further, it can be shown that if one uses model selection, there is no need to attempt to use a linear kernel [356]. Additionally, the sigmoid kernel is shown to be similar in performance to the RBF, but the potential to become unstable exists as the kernel matrix may not be positive definite [357]. The RBF kernel function that maps pairs of two points to higher dimension:

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2) \tag{D.5}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j) \tag{D.6}$$

$$\text{where} \quad \gamma > 0 \tag{D.7}$$

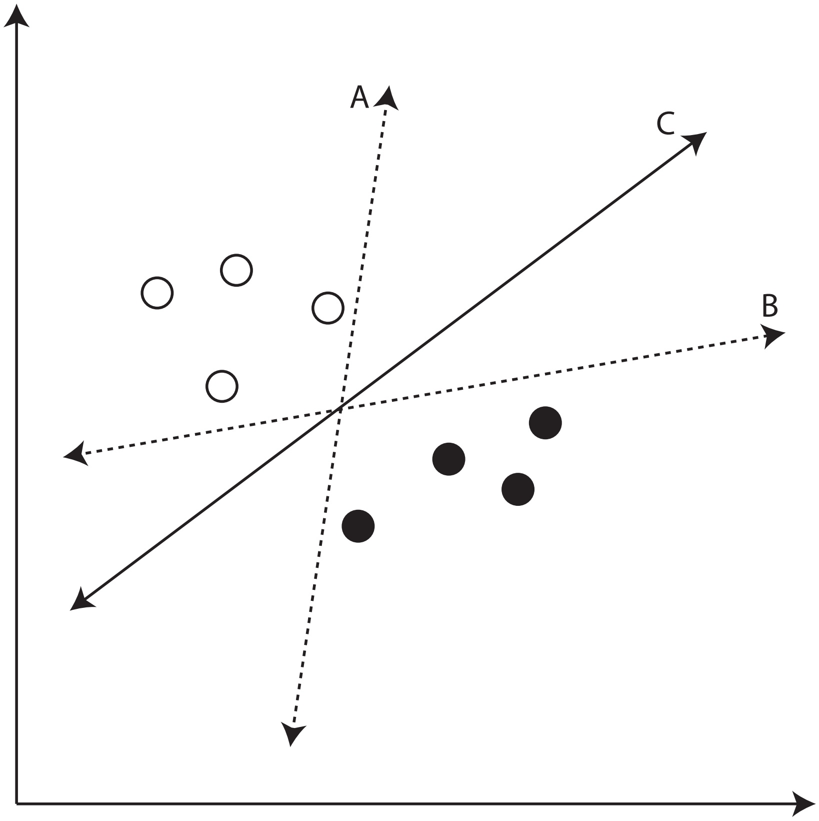


Figure D.1: **Support Vector Machine Maximum Margin.** Two classes of 2-dimensional data are represented by black and white circles. While all three lines, $\{A, B, C\}$, shatter, or separate these data, only one line, C, separates with a maximum margin between the two classes.

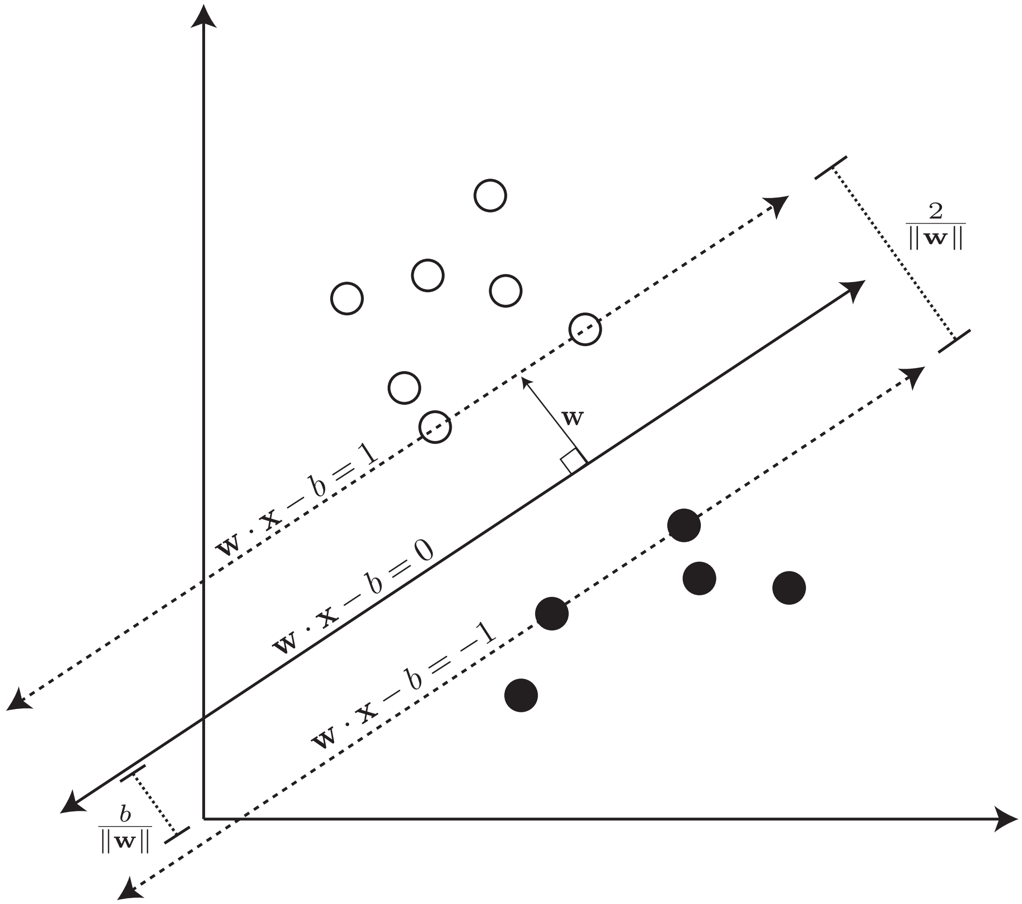

$\frac{2}{\|\mathbf{w}\|}$
$\mathbf{w}$
$\mathbf{w}\cdot\mathbf{x}-b=1$
$\mathbf{w}\cdot\mathbf{x}-b=0$
$\mathbf{w}\cdot\mathbf{x}-b=-1$
$\frac{b}{\|\mathbf{w}\|}$

Figure D.2: **The Support Vectors Learned By The Support Vector Machine.** The bold, black line separating the two data clusters provides the maximum margin between the classes. Data points with dashed lines passing through them are the support vectors.

Appendix E

Supplemental Tables

E.1 Chapter 2

ORF	Gene	Gene Product	GO ID's			
contig12757	Nol10	Nucleolar Protein 10	GO:0005730			
contig13640	Art7	Protein arginine N-methyltransferase 7	GO:0005737	GO:0019918	GO:0035243	
contig03660	Uba3	Ubiquitin-like modifier activating enzyme 3	GO:0016881	GO:0008641	GO:0005524	GO:0045116
contig12629	Rrp45	mRNA processing	GO:0000178	GO:0005730	GO:0005829	GO:0051252
			GO:0004532	GO:0017091	GO:0006364	GO:0005515
			GO:0043928			

Table E.1: **Accelerated *Enallagma* Genes and Their Gene Products.** In the *Enallagma* transcriptome, 29 genes were shown to be evolving at an accelerated rate. Of these, four could be annotated. The *Enallagma* ORF, associated gene, gene product, and GO IDs are shown.

contig01776	GO:0003677				
contig01831	GO:0008486				
contig06005	GO:0000502				
contig08606	GO:0005840				
contig11500	GO:0005488				
contig13110	GO:0000139				
contig14049	GO:0006904				
contig00167	GO:0009306	GO:0005788			
contig06160	GO:0008380	GO:0005681			
contig09479	GO:0005634	GO:0003677			
contig11985	GO:0016740	GO:0008270			
contig14077	GO:0005737	GO:0006915			
contig16813	GO:0016747	GO:0008152			
contig18304	GO:0006044	GO:0004342			
contig21827	GO:0005634	GO:0008270			
contig00694	GO:0016773	GO:0006139	GO:0005524		
contig01122	GO:0007264	GO:0005622	GO:0005525		
contig01972	GO:0005622	GO:0003676	GO:0000166		
contig04486	GO:0031072	GO:0006457	GO:0051082		
contig06200	GO:0004379	GO:0042967	GO:0006499		
contig10277	GO:0050662	GO:0044237	GO:0003824		
contig10311	GO:0051287	GO:0055114	GO:0016491		
contig10729	GO:0035091	GO:0007165	GO:0004871		
contig24459	GO:0000276	GO:0015078	GO:0015986		
contig05584	GO:0004872	GO:0005525	GO:0006614	GO:0007165	
contig07015	GO:0003735	GO:0006412	GO:0042254	GO:0005840	
contig18857	GO:0003887	GO:0006260	GO:0042575	GO:0003677	
contig23858	GO:0006270	GO:0017111	GO:0005524	GO:0003677	
contig01207	GO:0004177	GO:0006508	GO:0009987	GO:0008235	GO:0046872
contig07550	GO:0004003	GO:0005657	GO:0003677	GO:0005524	GO:0006289
contig13885	GO:0016192	GO:0006886	GO:0030126	GO:0005488	GO:0005198
contig17718	GO:0005634	GO:0006270	GO:0017111	GO:0005524	GO:0003677
contig11691	GO:0005856	GO:0006777	GO:0007529	GO:0008092	GO:0030054
	GO:0007165	GO:0060077	GO:0019897	GO:0005102	GO:0046872
	GO:0051260	GO:0005737	GO:0045184	GO:0045211	GO:0016740
	GO:0042803	GO:0005524	GO:0032947	GO:0003676	

Table E.2: **Genes Determined To Have Evolved Under A Decreased Evolutionary Rate.** Of the 169 *Enallagma* genes evolving at altered rates, 140 of these were shown to be evolving at a decreased rate. Of these 169, 33 were mapped to unique GO IDs.

minimumReadLength	20
overlapSeedStep	12
overlapSeedLength	16
overlapMinSeedCount	1
overlapSeedHitLimit	70
overlapHitPositionLimit	1000000
overlapMinMatchLength	40
overlapMinMatchIdentity	90
overlapMatchIdentScore	2
overlapMatchDiffScore	-3
overlapMatchUniqueThresh	12
isogroupThresh	500
isotigThresh	100
isotigContigCountThresh	100
isotigContigLengthThresh	3
aceMode	Auto
aceReadMode	Default
pairAlignMode	None
alignInfoMode	Auto
mapMinContigDepth	1
allContigThresh	100
largeContigThresh	500
expectedDepth	0
cDNAMode	TRUE
referenceMode	Auto
largeGenome	FALSE
ripMode	FALSE
heterozygoteMode	FALSE
assemblerBatchSize	0
numCPU	7
showSingleReadVariations	FALSE
nimblegenMappingMode	FALSE
backwardCompatibleContigging	FALSE
finishMode	FALSE
autoTrimming	TRUE

Table E.3: **Parameters Used With** `Newbler 2.3` **Software During *De Novo* Assembly Of The *Enallagma hageni* Transcriptome.**

Table E.4: **Orthologous Genes** Orthologous genes determined using the reciprocal best blast method (Section 2.2.3). We obtained 634 orthologs across the 11 species studies.

External Supplemental File "Table_S3.xlsx"

Table E.5: **Of The 634 Genes In The Orthologous, Protein-coding Set, 488 Were Mapped To At Least One GO ID** . These genes were mapped to 1669 GO IDs in total, with 691 of these GO IDs being unique.

External Supplemental File "Table_S4.xlsx"

E.2 Chapter 3

	LogLik	AIC	Nref	nuAf0	nuNA0	nuAf	nuNA	Tdiv	T_ad	p_ad	mAf_NA	mNA_Af	p_misid
IM_1mig	-28816.90718	57641.81436	217036.4427	325966.2846	43432.5798	1665793.697	242052.3236	6943.780607			0.67426271	1.03910373	0.02514024
IM_2mig	-28816.90926	57641.81853	216892.8026	324910.9708	43407.05157	1665642.725	241875.8539	6957.602421			0.67416857	1.04058087	0.0251479
IM_3mig	-28816.90971	57641.81943	216908.0683	324644.4556	43381.44664	1666656.713	241912.1133	6956.239437			0.67436007	1.04081619	0.02516009
IM_1mig_admix	-27809.55903	55631.11807	203388.4943	210168.5006	16918.62173	2302000.695	343726.4781	6518.642462	2913.941534	0.37075715	0.57251998	0.74355838	0.02589131
IM_2mig_admix	**-27809.49170**	**55630.98339**	**203102.4188**	**208524.1911**	**16687.83103**	**2290047.078**	**341745.4938**	**6552.92027**	**2952.754687**	**0.37609901**	**0.57226038**	**0.75106258**	**0.02590806**
IM_3mig_admix	-27809.55903	55631.11805	203387.0293	210170.8816	16918.61173	2301972.417	343727.0632	6518.661135	2913.964748	0.37075711	0.57250587	0.74354186	0.02589148
IM_1admix	-33034.23598	66076.47197	488035.5034	3656891.864	30911.96316	2236361.08	1610420.278	6989.458044	1752.289702	0.19344584			0.023957562
IM_2admix	-33033.10256	66074.20512	490514.8991	3978624.338	31508.68413	2150568.896	1684628.836	6945.168694	1653.207615	0.178604433			0.023849921
IM_3admix	-33034.49915	66076.9983	488571.0662	3736406.846	30896.85319	2208946.681	1647473.727	6979.680994	1730.08392	0.190122201			0.023956303
IM_1	-44515.21742	89034.43484	549946.0844	6803411.413	79318.76791	2258987.147	761908.7861	6291.05391					0.022796384
IM_2	-44514.23469	89032.46939	549570.1118	7257582.696	79112.33294	2215400.479	763694.8202	6284.559468					0.022814311
IM_3	-44513.68237	89031.36473	550007.5775	8463988.56	78584.6716	2113121.763	764950.469	6259.201599					0.022790432

Table E.6: **Models Optimized And Results From The $\partial a \partial i$ Inference Analysis**. We used $\partial a \partial i$ to learn parameters for the following four models: 1) a simple 2 population isolation model in which an ancestral population splits into two daughters (N. America and Africa) and each subsequent daughter population experiences growth, 2) an isolation-with-migration (IM) model as above but with asymmetric migration rates between N. America and Africa, 3) an isolation with admixture model that is the same as model 1 but with the addition of a single burst of admixture from Africa into N. America, and 4) an IM model (as in model 2) that adds admixture from Africa to N. America. Each of the four models were optimized for the SFS described above 3 separate times from different initial starting conditions.

	SVM_{auto}		SVM_{Ronen}		SVM_{SFS}	
Model	Acc (%)	AUC	Acc (%)	AUC	Acc (%)	AUC
Hard vs. Neutral	99.2	0.998	99.0	0.998	99.0	0.997
Hard vs. Soft & Neutral	96.1	0.990	96.1	0.990	95.9	0.988
Hard vs. Soft	94.4	0.983	94.3	0.983	94.0	0.981
Hard & Soft vs. Neutral	89.2	0.945	86.6	0.928	87.5	0.935
Soft vs. Neutral	83.2	0.901	78.6	0.863	80.3	0.876

Table E.7: **Results Of The Comparative SVM Analysis.** The SVM_{auto} utilizes feature vectors consisting of sfs and LD summary stats. The SVM_{Ronen} classifier utilizes full sfs data, but lacks LD information. The SVM_{SFS} is comprised of five sfs-based summary statistics (Section 3.2.4).

	Sweep		Mean # Permuted		
Interaction	Class	#Interactors	Interactors	Enrichment	p-value
Flybase genetic interactions	Both	53	54.5	0.97	p = 0.57
	Hard	2	10.07	0.20	p = 0.997
	Soft	30	20.18	1.49	p = 0.13
RNA-gene interactions	Both	0	0.0	N/A	p = 1.0
	Hard	0	0.0	N/A	p = 1.0
	Soft	0	0.0	N/A	p = 1.0
TF-gene interactions	Both	34	36.65	0.93	p = 0.53
	Hard	15	8.62	1.74	p = 0.133
	Soft	1	11.65	0.086	p = 0.89
Flybase other physical interactions	Both	16	22.28	0.72	p = 0.96
	Hard	4	4.09	0.98	p = 0.55
	Soft	6	8.78	0.68	p = 0.89
yeast two-hybrid	Both	6	6.13	0.98	p = 0.61
	Hard	0	0.82	0.0	p = 1.0
	Soft	4	2.81	1.42	p = 0.315
co-affinity purification	Both	30	38.20	0.785	p = 0.83
	Hard	5	6.04	0.83	p = 0.63
	Soft	10	14.69	0.68	p = 0.82

Table E.8: **Interactions In Selective Sweeps Called From The Filtered Set Of Fixed Differences.** Here we test for an excess of interacting pairs of genes both experiencing selective sweeps (Section 3.4.10)

Annotation	Hard	Soft	Both
synonymous	0.049	0.261	0.310
nonsynonymous	0.050	0.258	0.308
TF binding site	0.053	0.190	0.244
5′ UTR	0.062	0.220	0.325
3′ UTR	0.052	0.203	0.255

Table E.9: **Fraction Of Adaptive Fixations Across Annotation Classes In The Unfiltered Set.** We show hard and soft sweeps separately and combined (labeled as "Both").

Annotation	Hard	Soft	Both
synonymous	0.177	0.392	0.570
nonsynonymous	0.245	0.310	0.555
TF binding site	0.259	0.250	0.509
5′ UTR	0.284	0.304	0.588
3′ UTR	0.282	0.356	0.638

Table E.10: **Fraction Of Adaptive Fixations Across Annotation Classes In The Filtered Set.** We show hard and soft sweeps separately and combined (labeled as "Both")

Table E.11: **Unfiltered Hard Sweep GO Clusters**

"unfilt_hard_GO_cluster.xlsx"

Table E.12: **Unfiltered Soft Sweep GO Clusters**

"unfilt_soft_GO_cluster.xlsx"

Table E.13: **Filtered Hard Sweep GO Clusters**

"filt_hard_GO_cluster.xlsx"

Table E.14: **Filtered Soft Sweep GO Clusters**

"filt_soft_GO_cluster.xlsx"

Table E.15: **Hard Sweep Genes (Unfiltered)**

"hard_sweeps_unfiltered.xlsx"

Table E.16: **Soft Sweep Genes (Unfiltered)**

"soft_sweeps_unfiltered.xlsx"

Table E.17: **Hard Sweep Genes (Filtered)**

“hard_sweeps_filtered.xlsx”

Table E.18: **Soft Sweep Genes (Filtered)**

“soft_sweeps_filtered.xlsx”

E.3 Chapter 4

Equilibrium Demography	Bottleneck	Exponential Growth	Contraction-Then-Growth
$Ne \sim U(500, 25000)$	$NeA \sim U(100, 20000)$	$NeA \sim U(20, 20000)$	$NeA \sim U(5000, 25000)$
	$NeB \sim U(100, 5000)$	$T_G \sim U(10, 5000)$	$NeC \sim U(100, 4999)$
	$T_B = 3500 \star$	$Ne0 \sim U(500, 100000)$	$T_C = 2040 \star$
	$T_R \sim U(100, 3499)$		$T_G\ U(100, 2039)$
	$Ne0 \sim U(100, 40000)$		$Ne0 \sim U(5000, 50000)$

Table E.19: **Prior Distributions Used To Generate Coalescent Simulations For ABC Analysis**. For each demographic model examined, coalescent simulations were generated using parameters drawn from uniform distributions. Parameters with a $\star$ indicate that they were fixed for the ABC analysis.

Fraction of		Contraction			
Linked Loci	Bottleneck	& Growth	Growth	Ambiguous	Equilibrium
0.0	0	0	0	0	100
0.1	0	0	39	61	0
0.2	0	11	9	80	0
0.3	0	48	2	50	0
0.4	0	78	0	22	0
0.5	0	75	0	25	0
0.6	0	82	0	18	0
0.7	0	88	0	12	0
0.8	0	85	0	15	0
0.9	0	92	0	8	0
1.0	0	92	0	8	0

Table E.20: **Model Selection Using Appoximate Bayes Computation (Means And Variances)**. The column labeled "Ambiguous" indicates the number of observations for which no one model fit better than any other (Bayes Factor > 20).

Appendix F

Supplemental Figures

F.1 Chapter 2

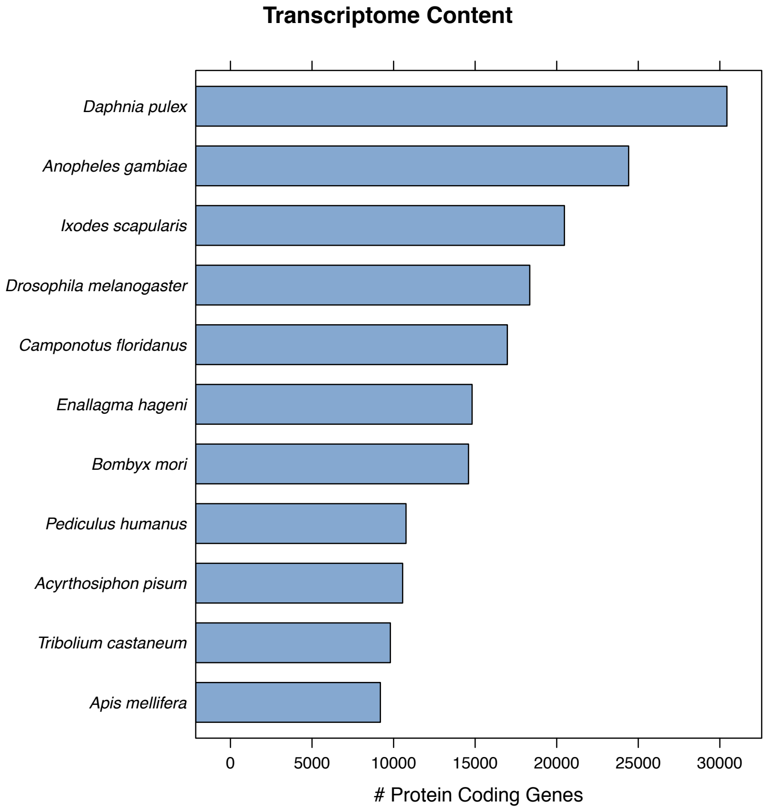


Figure F.1: **Arthropod Transcriptome Content.** The number of protein coding genes of the 11 species used in the analysis is shown. The *Enallagma hageni* transcriptome possesses 14,813 protein coding genes.

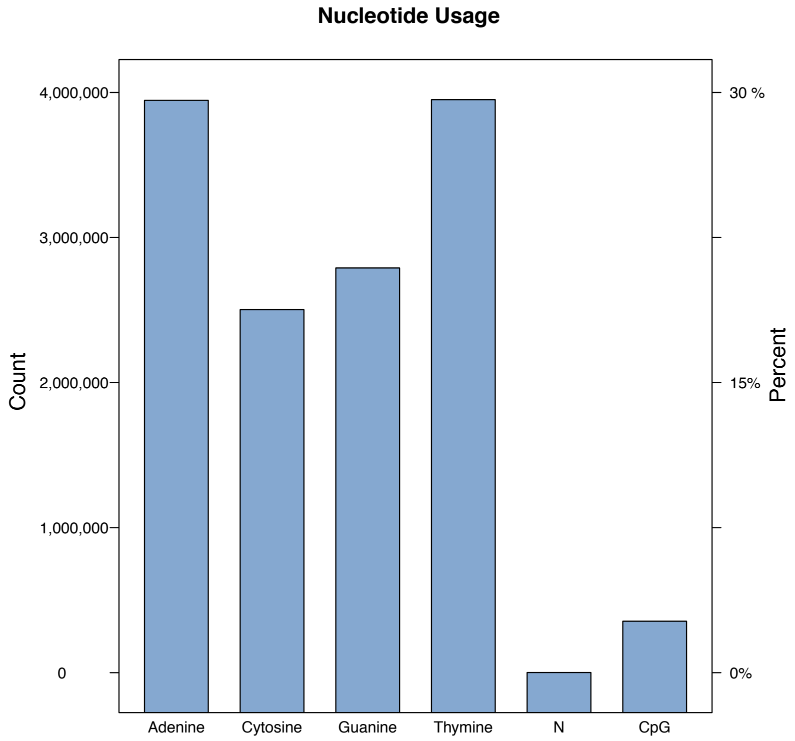


Figure F.2: **Nucleotide Profile Of Assembled Enallagma Contigs.** The assembled E. hageni transcriptome is comprised of 13,191,394 nucleotides. An AT bias is observed (59.86% AT, 40.13% GC, 0.01%N) and CpG sites occurred in 2.69% of the assembled transcriptome

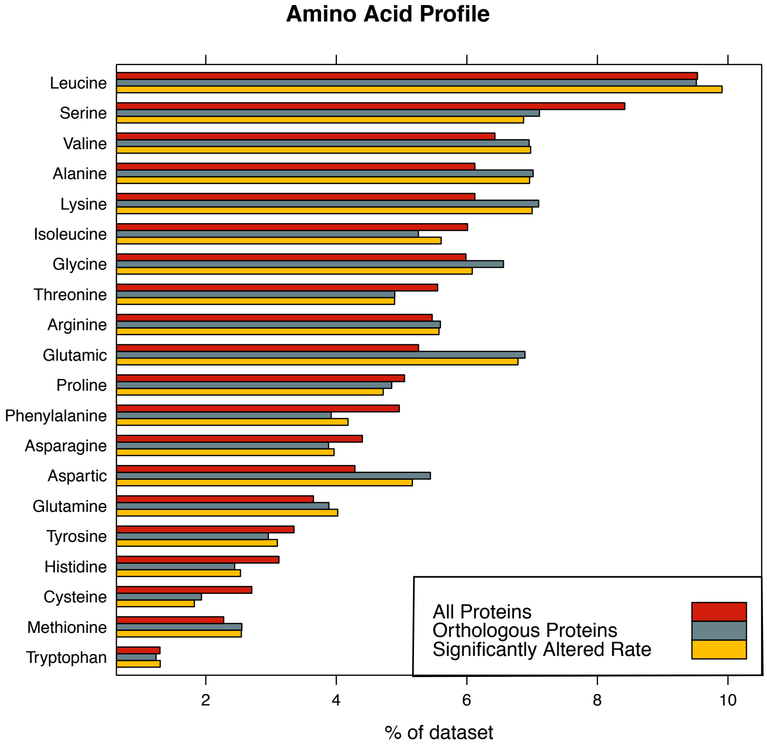
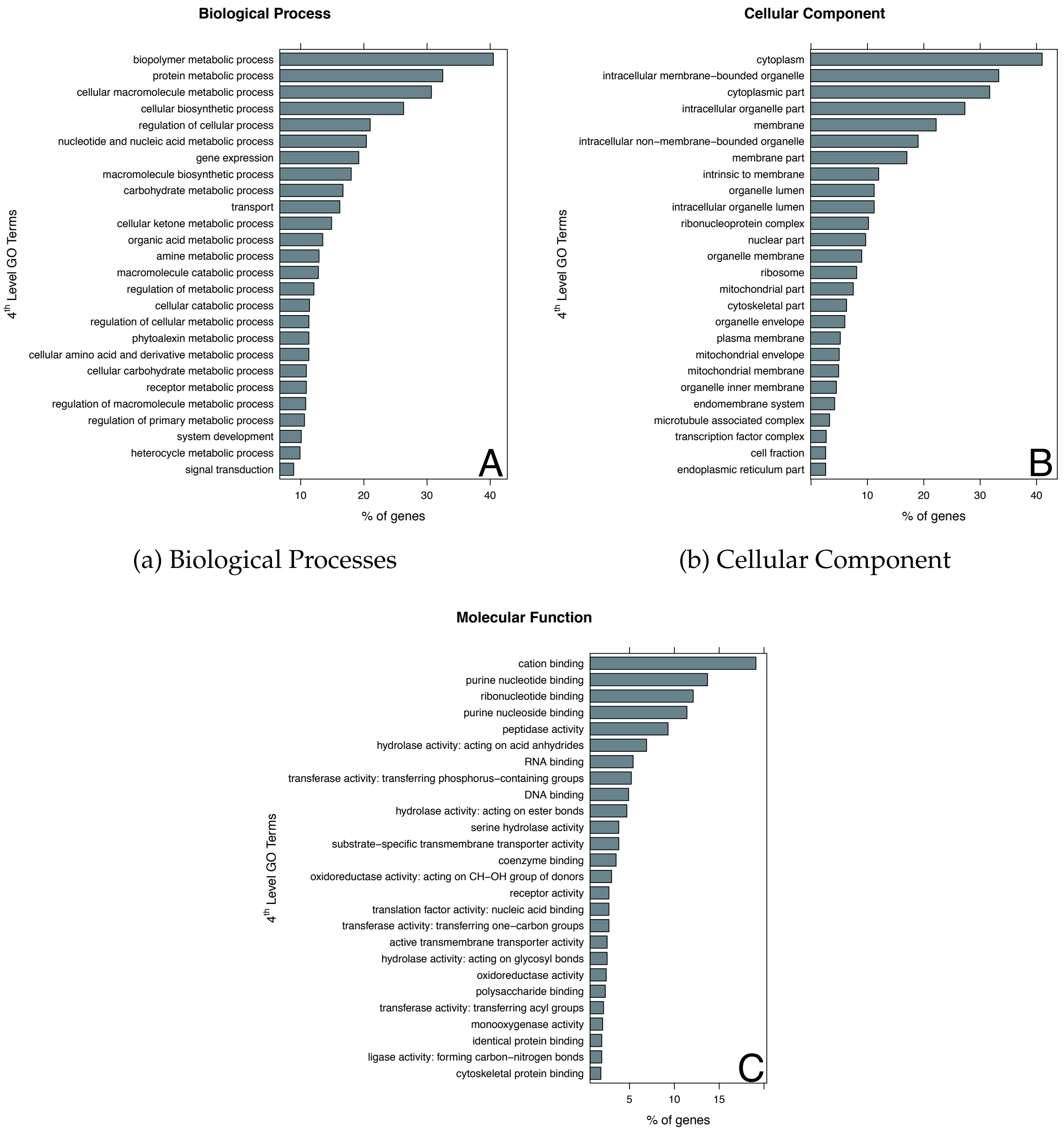


Figure F.3: ***Enallagma* Amino Acid Profile.** The amino acid profiles of three groups of translated *Enallagma* proteins are presented. The profile of all 1,621,208 amino acids comprising the 14,813 protein coding genes is shown in red. The 634 proteins orthologous across all 11 arthropod species in this study are indicated in grey and the 169 genes shown to have at an altered rate are shown in yellow.



(a) Biological Processes

(b) Cellular Component

(c) Molecular Function

Figure F.4: **4th Level GO Term Distributions For Annotated *Enallagma hageni* Genes.** At the 4th level of the GO term hierarchy, we mapped the dataset of genes to 1463 GO terms across the 3 ontologies. Shown are the top 25 most significant results in each of the 1st level categories: (a) Biological Process, (b) Cellular Component, and (c) Molecular Function.

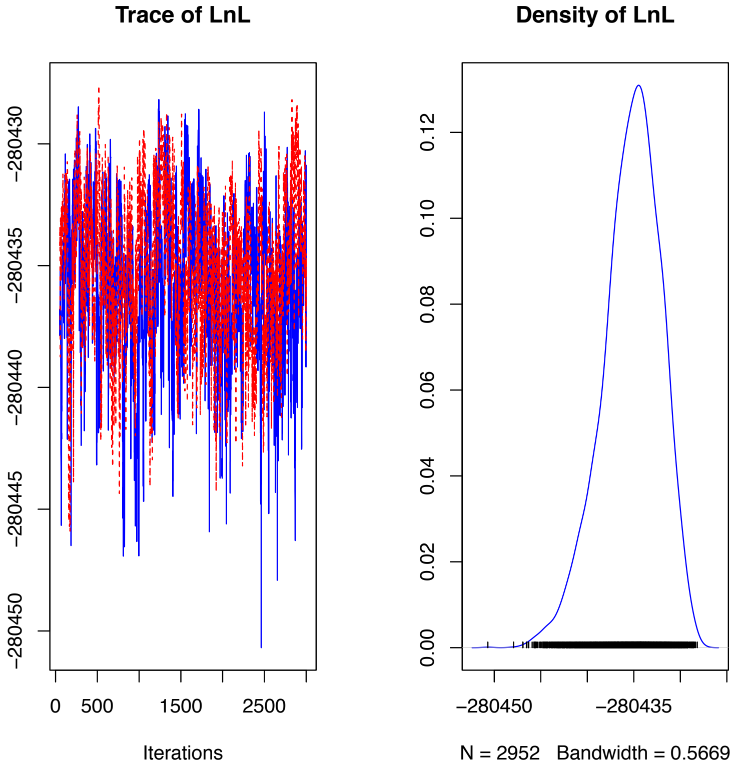


Figure F.5: **Trace And Density Plots Of The Posterior Probability Distribution Determined In Phylogenetic Analysis.** After thinning the samples of the posterior probability, we obtained 2952 draws from the posterior. Shown in (A) is the negative log-likelihood trace plot. In (B) I plot the density of the thinned posterior.

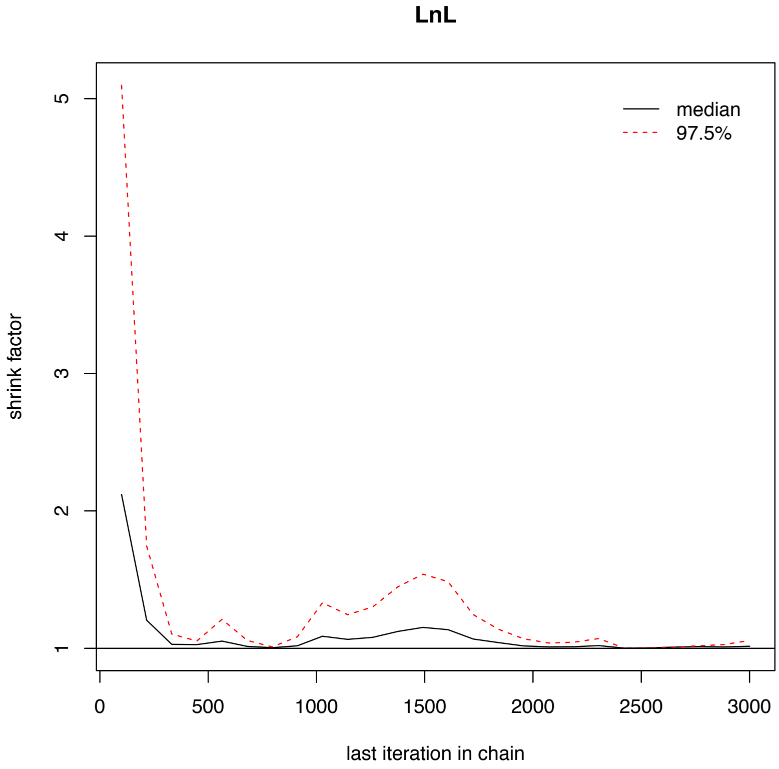
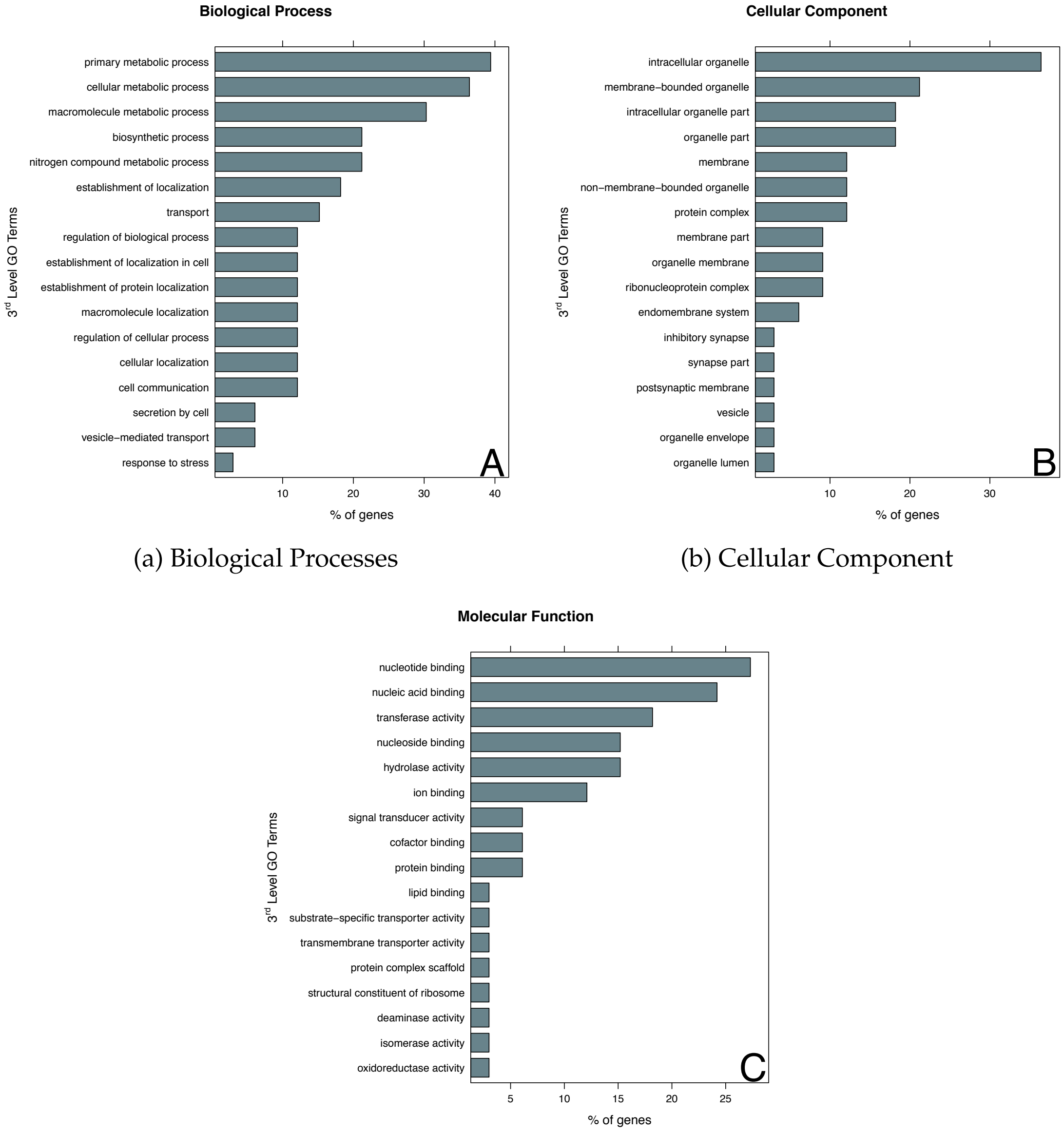


Figure F.6: **Gelman And Rubin Convergence Plot Of The MCMC Phylogenetic Analysis.** To test that our chains have converged to the stationary distribution, the thinned samples from both chains in the MCMC run are used to compute the Gelman-Rubin test for convergence. This test calculates the within-chain and between-chain variance and returns a potential scale reduction factor. If this factor is below $\approx$ 1.25, it is another assurance that the stationary distribution has been reached [2].

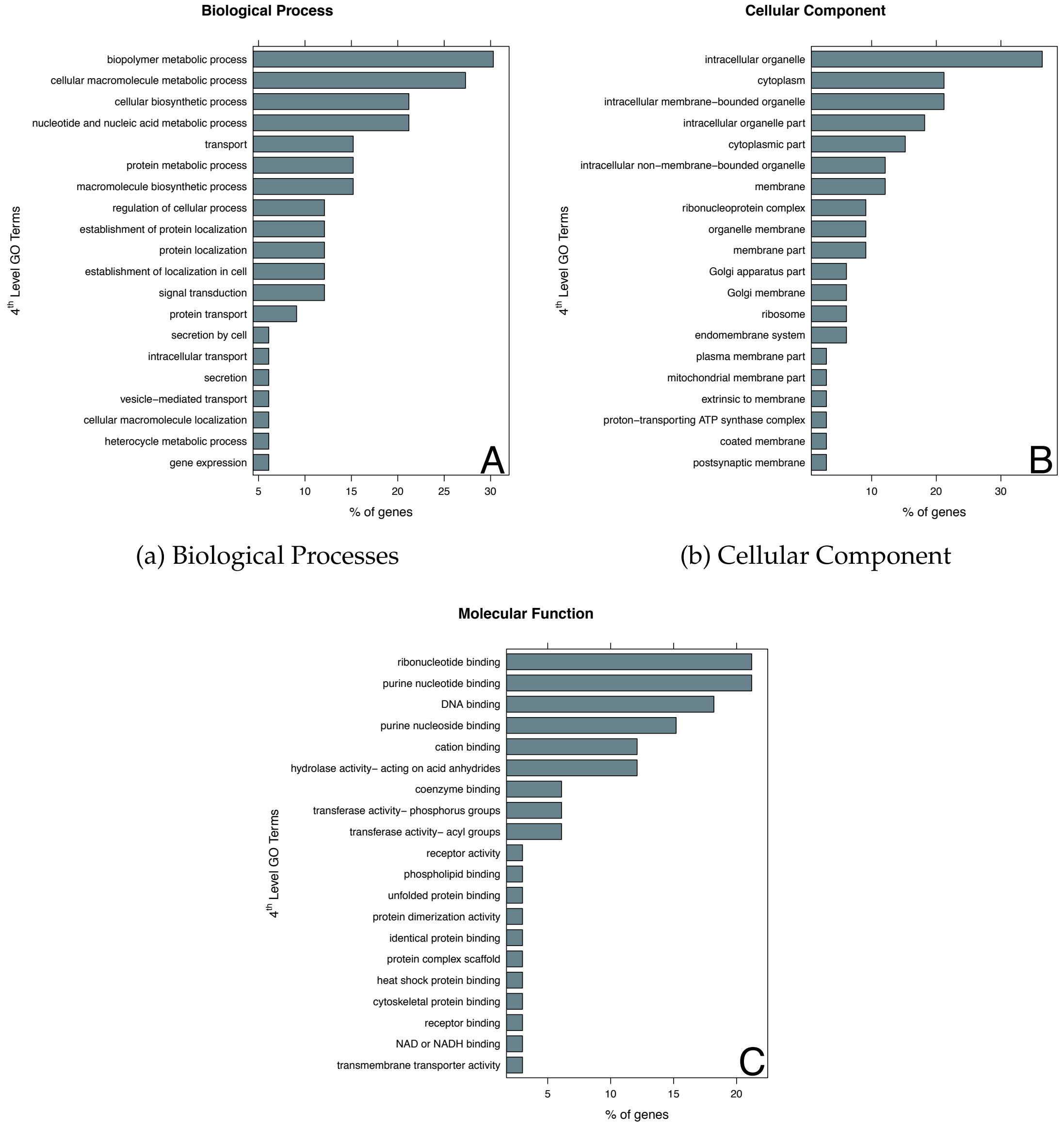


(a) Biological Processes

(b) Cellular Component

(c) Molecular Function

Figure F.7: **3rd Level GO Term Distribution For Decreased Rate Genes.** Of the 140 *Enallagma hageni* genes which were shown to be evolving at either a diminished rate, per the branch length tests, we were able to map 33 of these genes to 105 GO terms. Shown here are the top 17 most significant of these terms across the three orthologies: (a) Biological Process, (b) Cellular Component, and (c) Molecular Function.



(a) Biological Processes

(b) Cellular Component

(c) Molecular Function

Figure F.8: **4th Level GO Term Distribution For Decreased Rate Genes.** The top 20 most significant GO terms are shown for each of the three ontologies: (a) Biological Process, (b) Cellular Component, and (c) Molecular Function.

F.2 Chapter 3

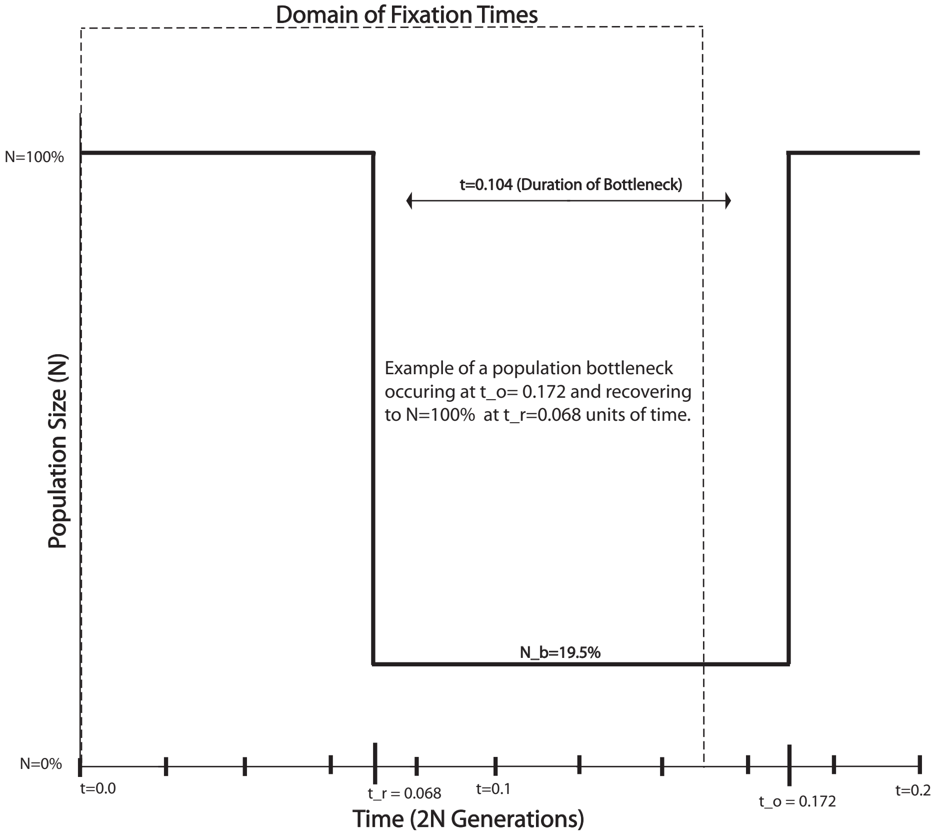

Domain of Fixation Times
N=100%
t=0.104 (Duration of Bottleneck)
Example of a population bottleneck
occuring at t_o= 0.172 and recovering
to N=100% at t_r=0.068 units of time.
Population Size (N)
N_b=19.5%
N=0%
t=0.0
t_r = 0.068
t=0.1
t_o = 0.172
t=0.2
Time (2N Generations)

Figure F.9: **Prior To Training And Testing Our SVM On The Full Admixture Model, We Examined How Robust Our Classifier Was To A Single Population Bottleneck.** Our model assumes following: a population bottleneck begins 0.172 units of $2N$ time in the past and completes at 0.068 units of $2N$ generations. The severity of the bottleneck during that time is 0.195, that is, the population is reduced to 19.5% of its original size and upon the completion of the bottleneck returns to its original size. We allow for fixations to happen anywhere between the immediate present (0.0 units of $2N$ time) and 0.15 units of $2N$ time into the past. Thus, the fixation can occur at any point starting shortly after the bottleneck begins until present, after which the population size has already recovered.

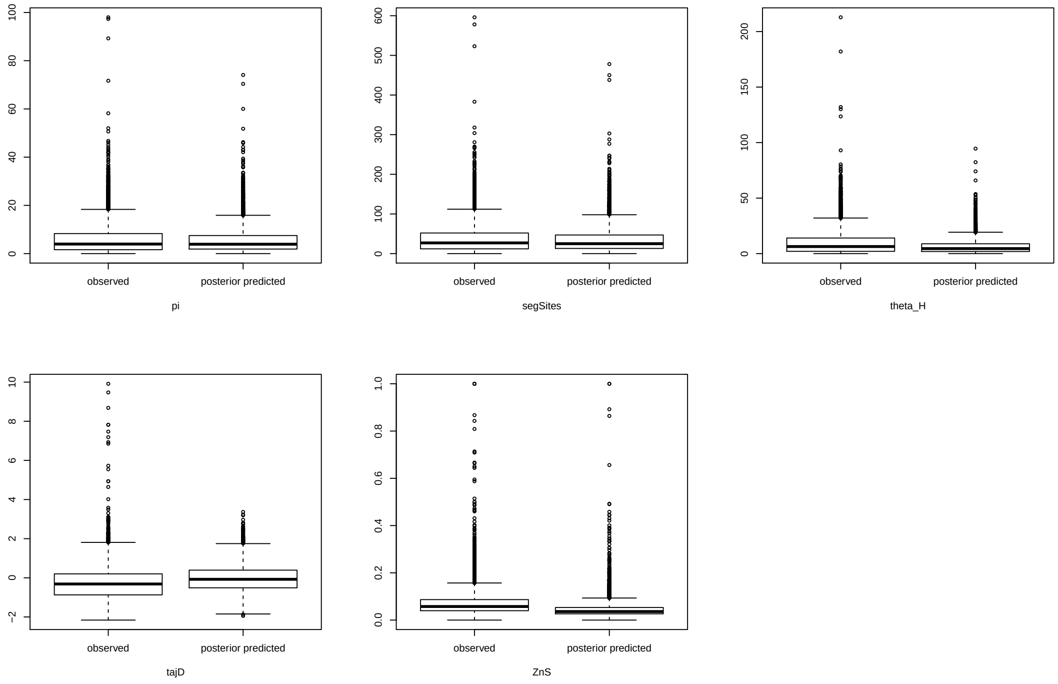


Figure F.10: **Posterior Predictive Simulations For The Admixture Model.** Parameters obtained using $\partial a \partial i$ were used to generate coalescent simulations. Common population genetic summary statistics show excellent agreement to the North American data.

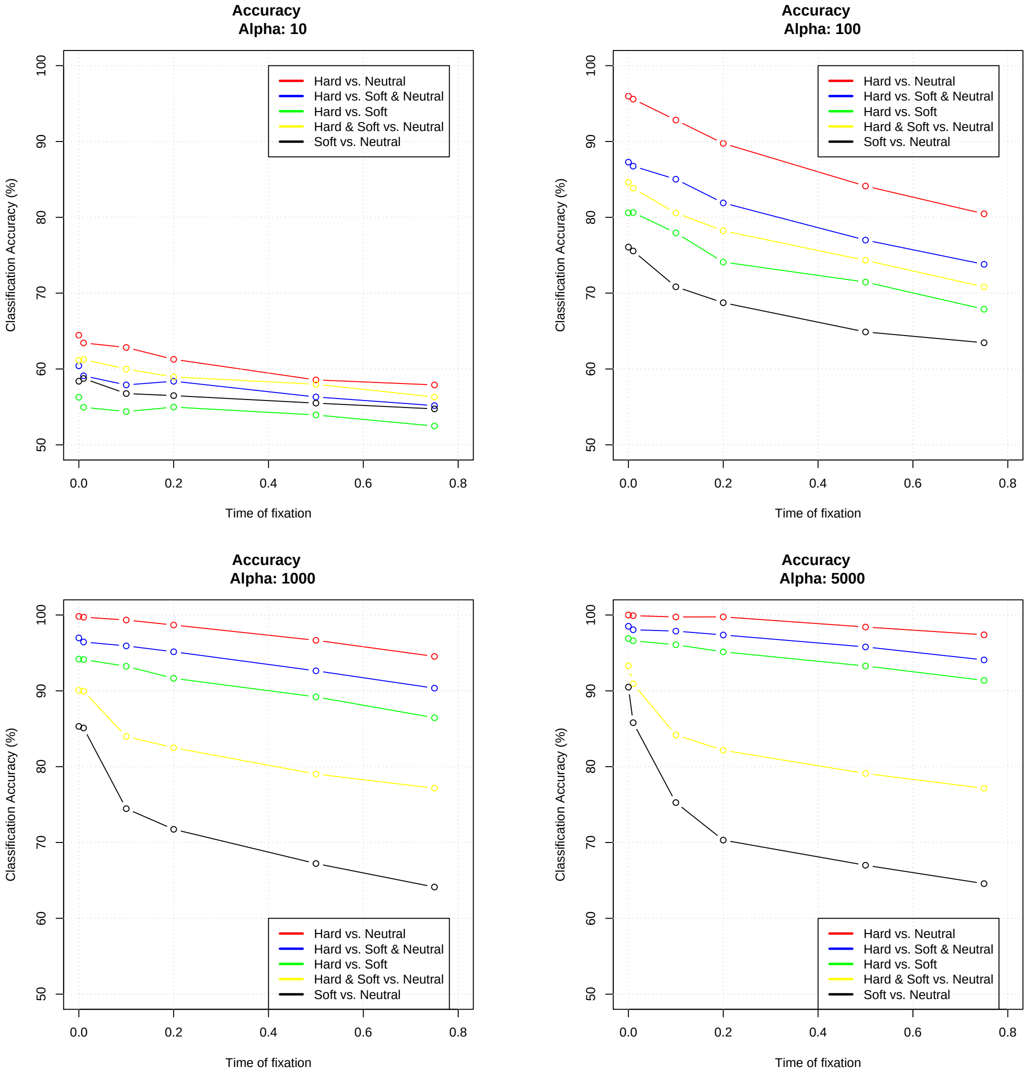


Figure F.11: **Support Vector Machine Performance Under Constant Population Size.** To test the power and performance of our classifier initially, we simulate data across a range of fixation times and selection strengths under a stationary demographic history and examine classification accuracy.

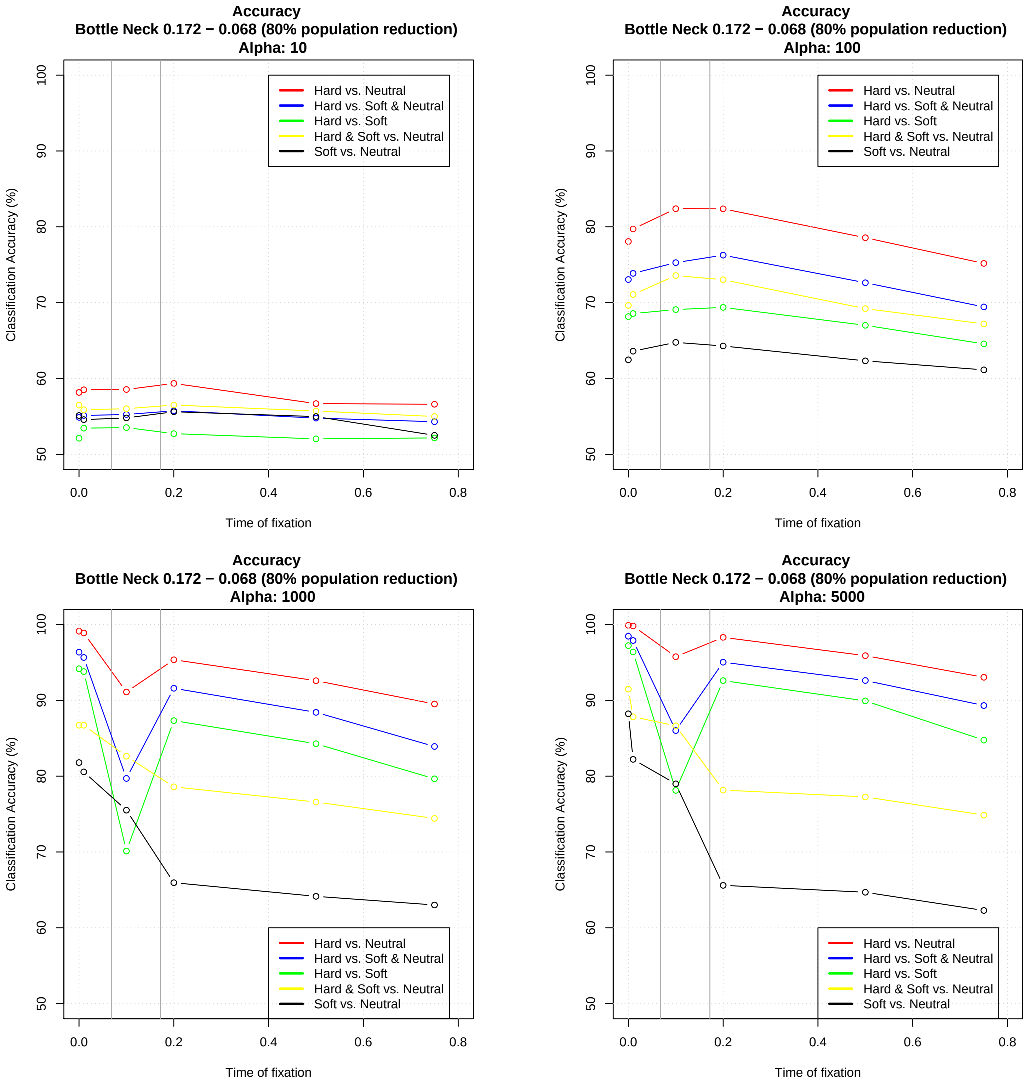


Figure F.12: **Support Vector Machine Performance Under A Population Bottleneck.** We tested the power of our classifier with varying fixation times and strengths of selection, however in this test we also introduce a population bottleneck. Note the duration of this bottleneck is the period delineated by the two vertical lines. See Figure F.9.

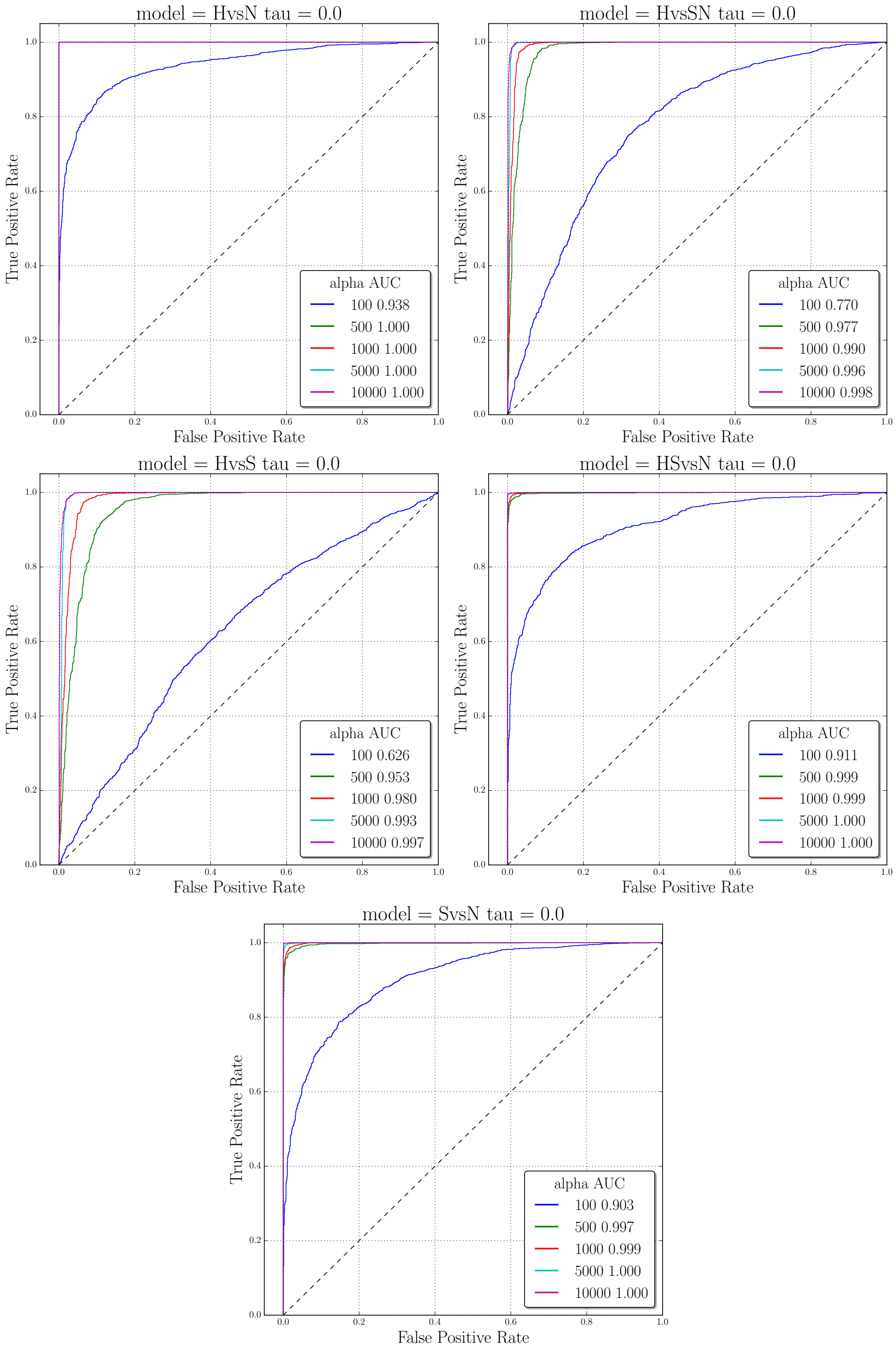


Figure F.13: **ROC Curves For A Fixed Value Of $\tau = 0.0$ And Various Strengths Of Selection (α).** All testing and training was done under the admixture demographic model.

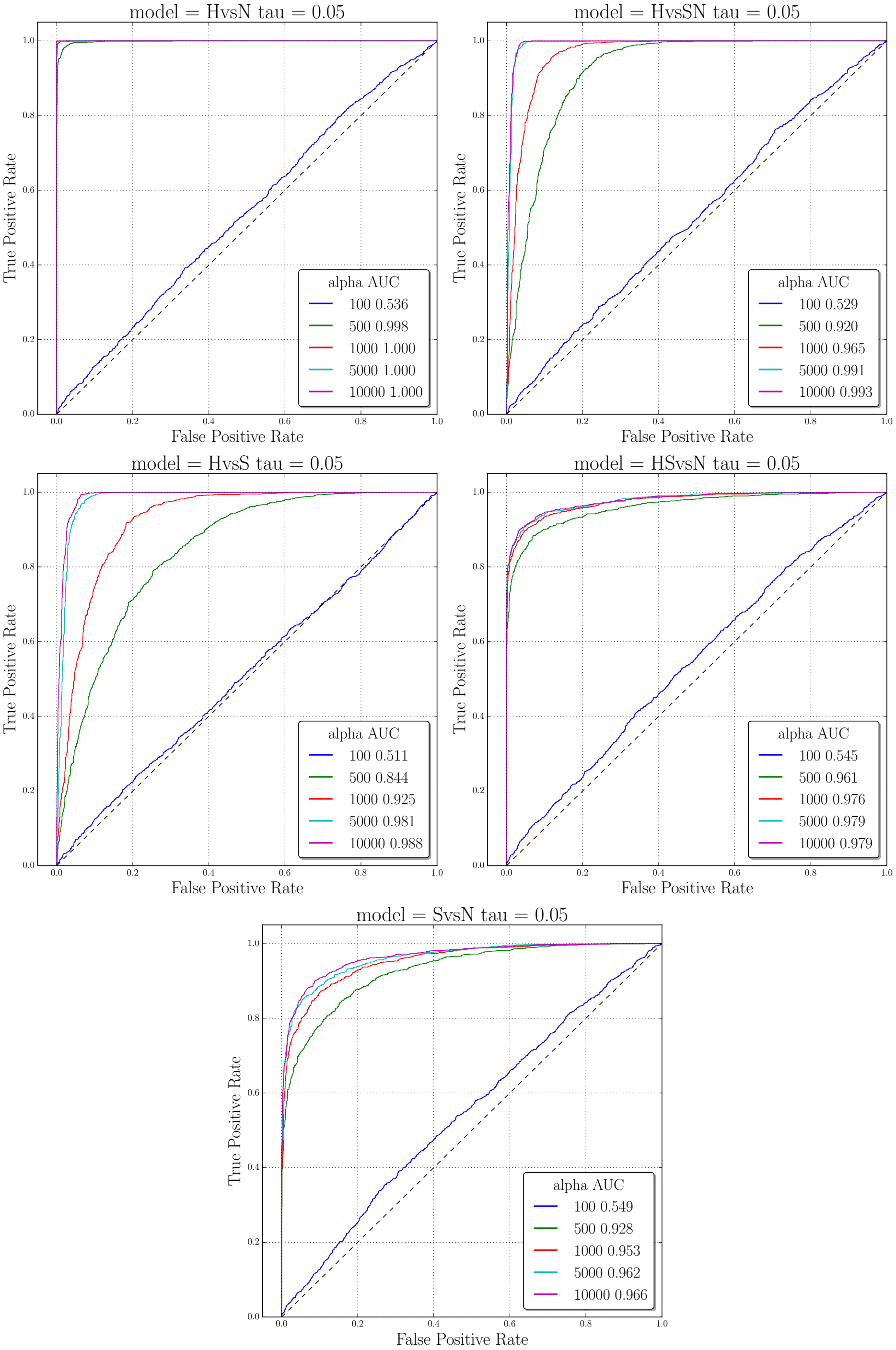


Figure F.14: **ROC Curves For A Fixed Value Of $\tau = 0.05$ And Various Strengths Of Selection (α).** All testing and training was done under the admixture demographic model.

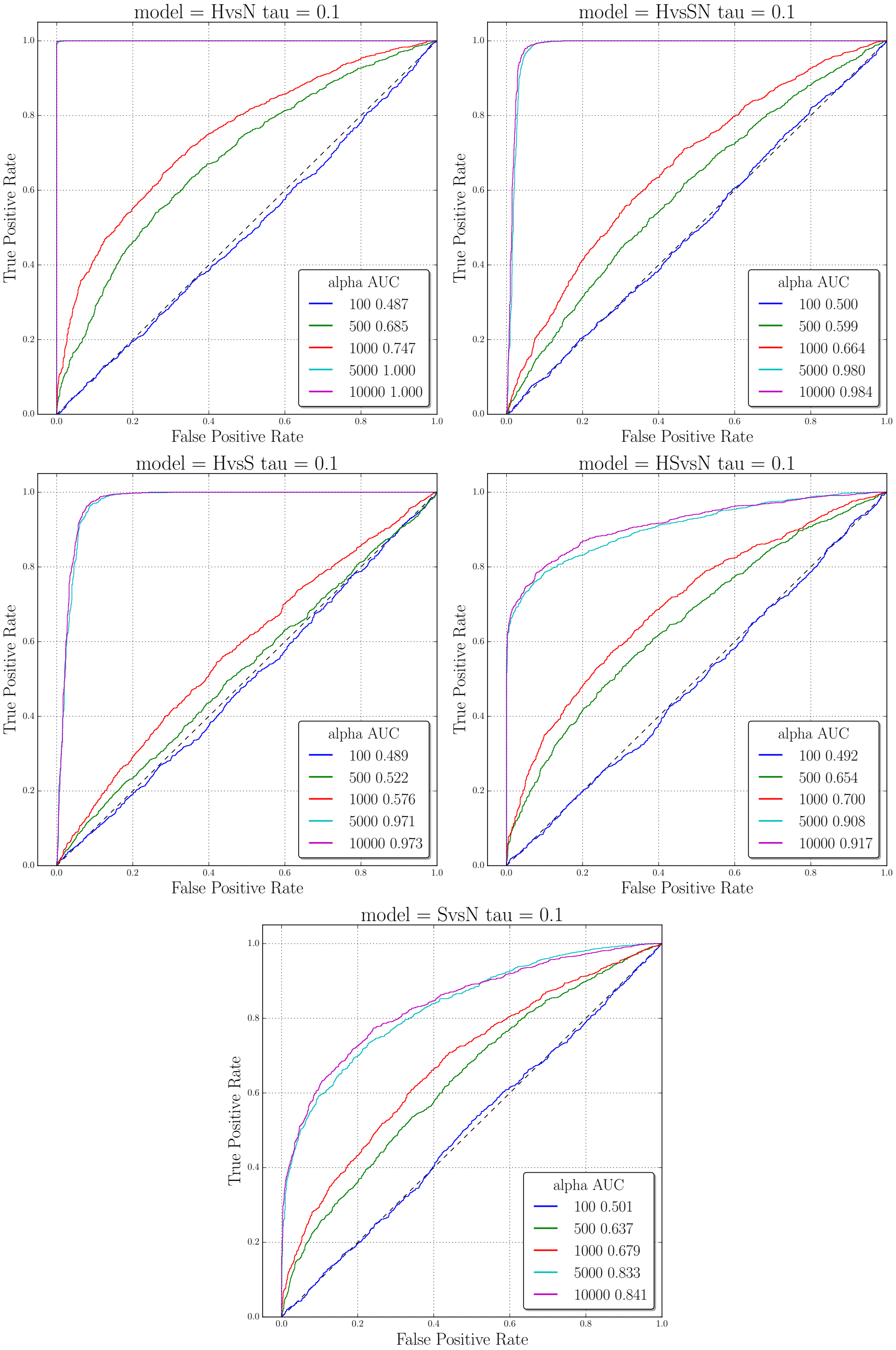


Figure F.15: **ROC Curves For A Fixed Value Of $\tau = 0.1$ And Various Strengths Of Selection (α).** All testing and training was done under the admixture demographic model.

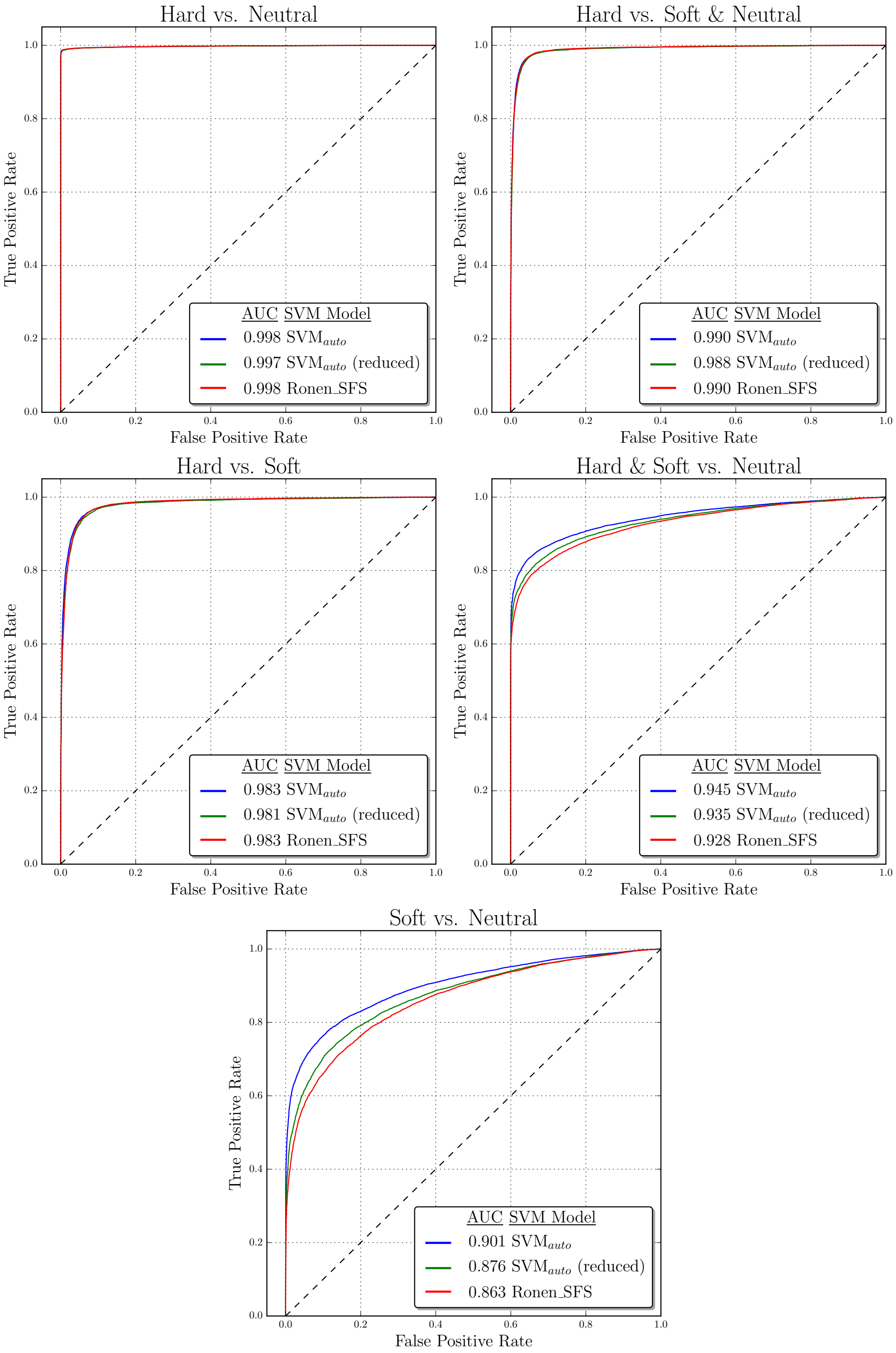


Figure F.16: **Support Vector Machine Comparisons Between Other Classifiers.** In each of the five one-vs-one comparisons we examine the ROC plots and AUC values. Our SVM$_{auto}$ classifier is shown as a blue line, Ronen et al.'s feature space is shown as a red line [3] , while a version of our SVM utilizing only SFS data is colored in green. In subplots A-C, there is little difference in the three classifiers, however, when we look at Hard & Soft vs. Neutral and Soft vs. Neutral, we observe that our SVM$_{auto}$ classifier outperforms both Ronen and SFS. Again, all testing and training was done under our admixture model Figure 3.2.

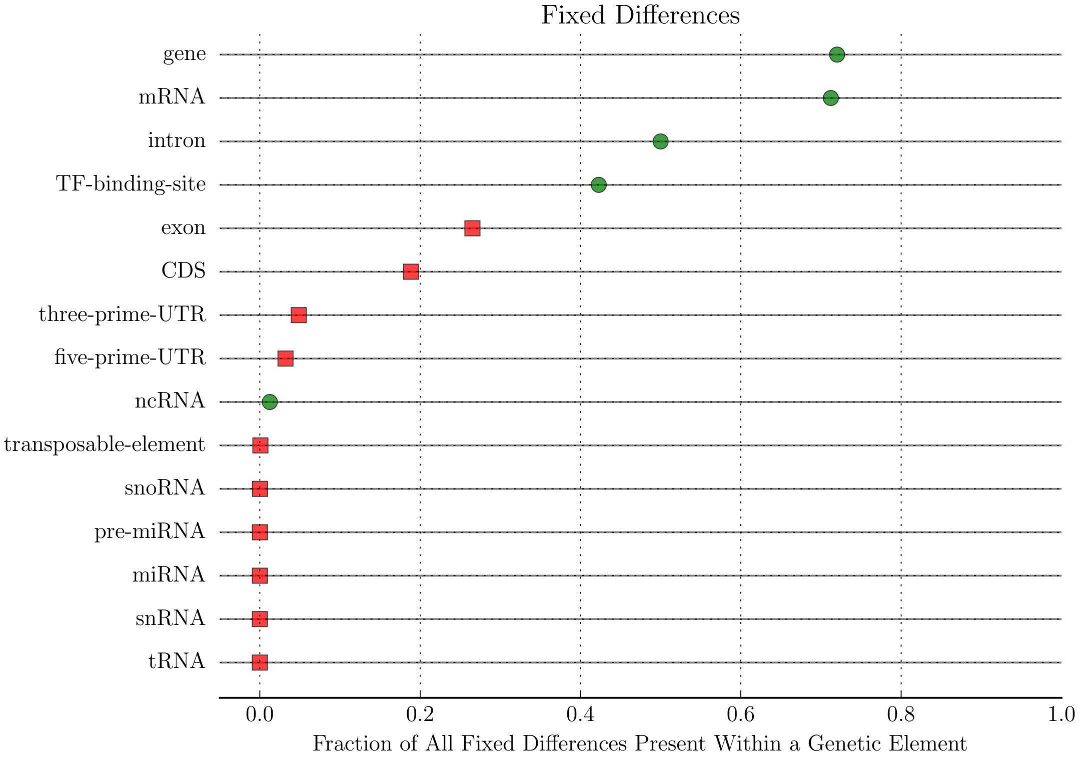


Figure F.17: **Fraction Of Fixed Positions Occurring In Various Genetic Elements For The Unfiltered Set Of Fixed Differences.** Green circles are those elements which are enriched for the presence of fixed positions.

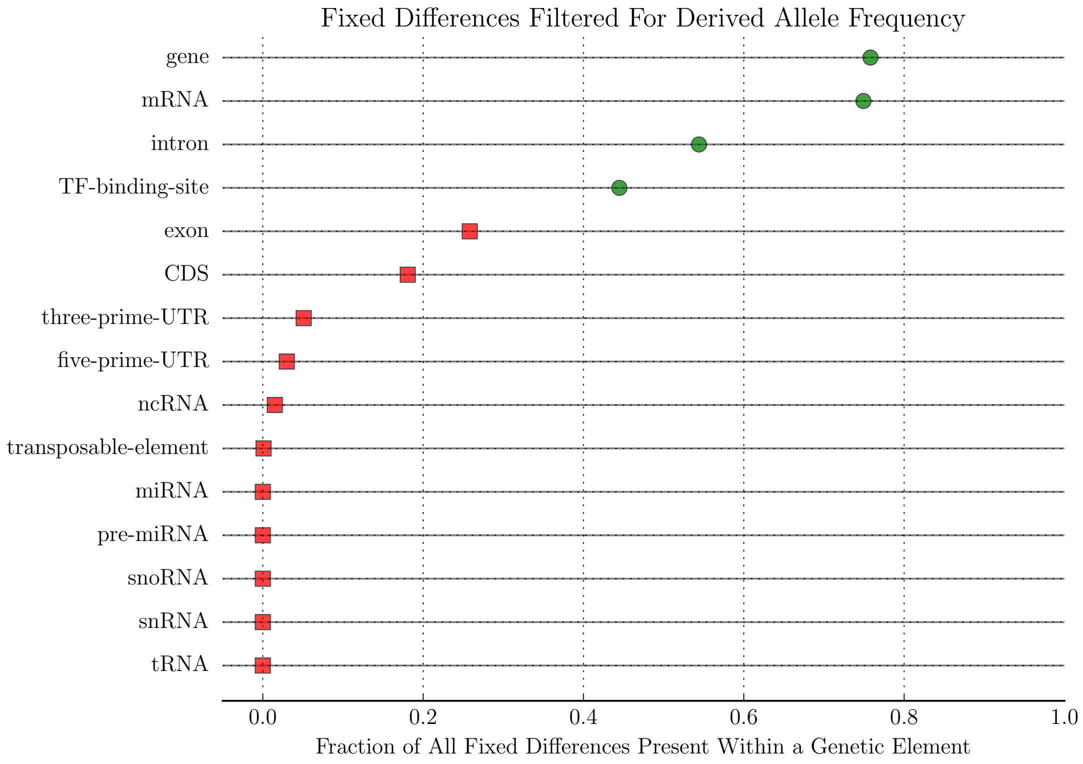


Figure F.18: **Fraction Of Fixed Positions Occurring In Various Genetic Elements In The Filtered Fixed Differences Dataset.** Green circles are those elements which are enriched for the presence of fixed positions.

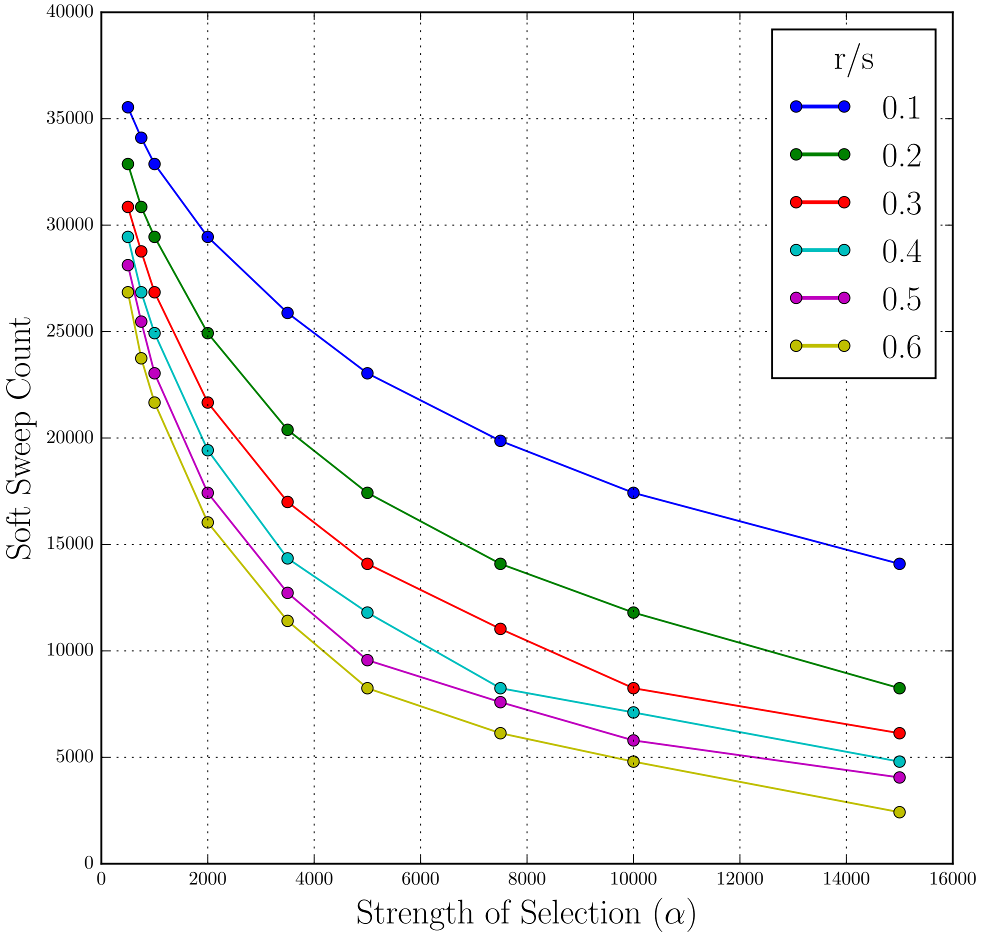


Figure F.19: **Visualizing The "Soft Shoulder Effect" [4].** Neutral fixations occurring near hard sweeps may be erroneously called as soft sweeps. While we determine a value of "r/s" that allows for a 5% misclassification rate in Section 3.4.9, here we show how the number of called soft sweeps can radically differ given varying assumptions of α and "r/s".

F.3 Chapter 4

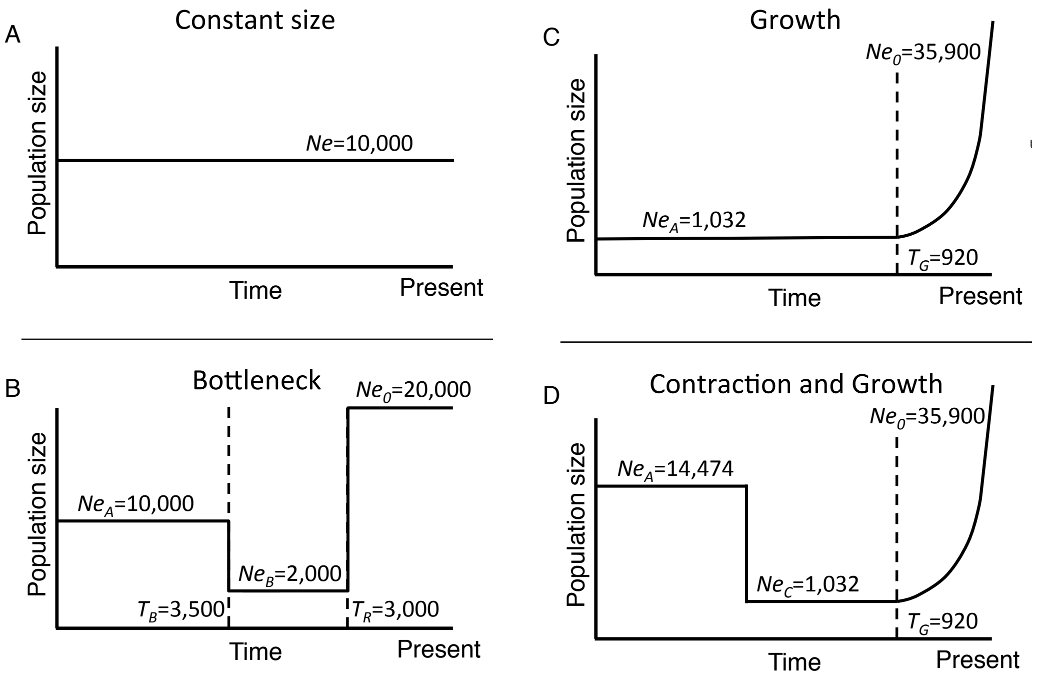


Figure F.20: **Demographic Models And Parameters Used In The $\partial a \partial i$ And ABC Analysis.** For each model, a diagram of the population size history is shown (not to scale) along with the values of each parameter. (A) A model with constant population size. (B) A population bottleneck (parameterization from Marth et al. [5]). (C) Recent exponential population growth. (D) A three-epoch model with a population contraction and recent exponential growth (a simplified version of the European model from Gravel et al. [6]).

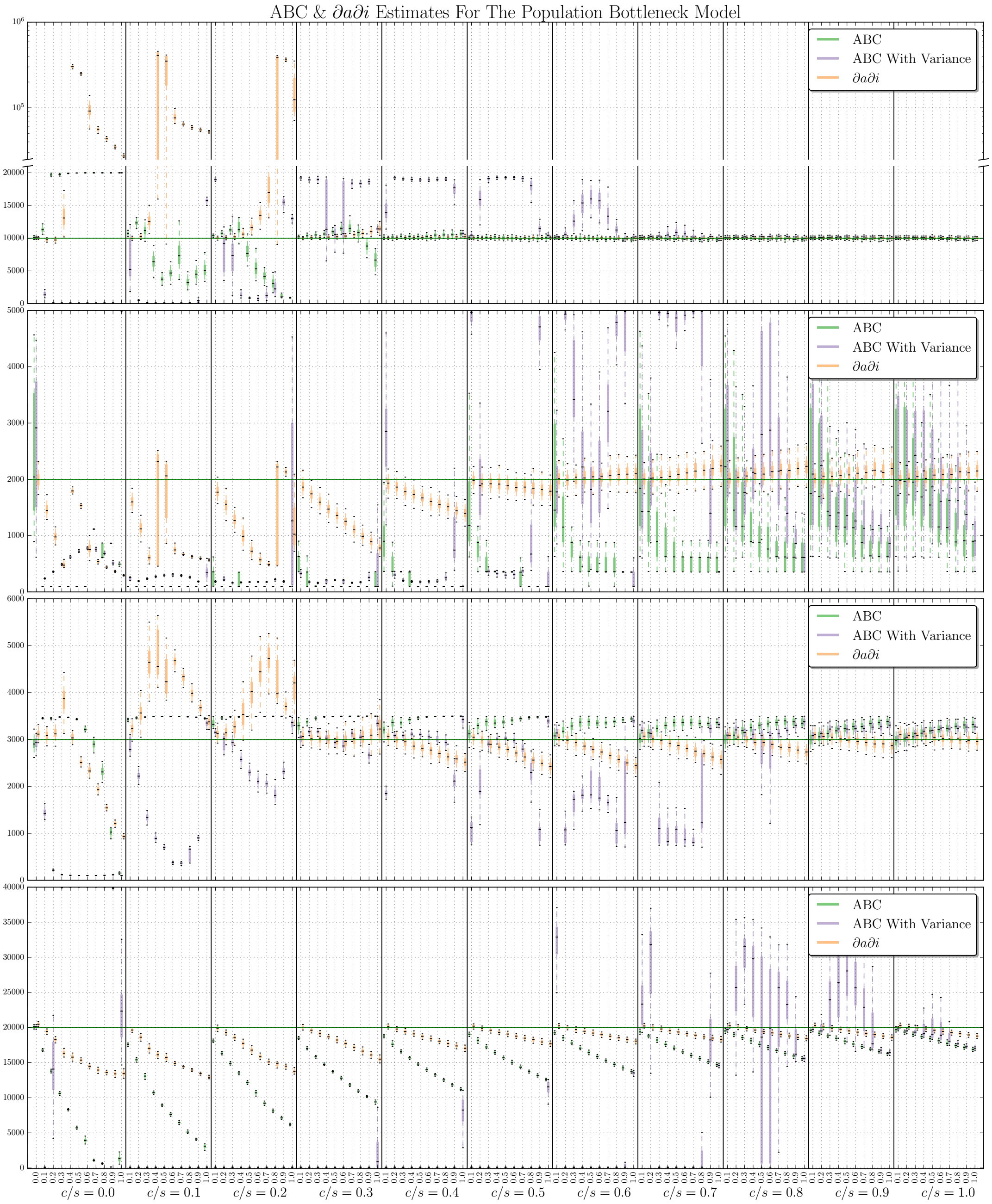


Figure F.21: **Bottleneck Model Parameter Estimates From ∂a∂i, ABC Using Summary Statistic Means, And ABC Using Both Means And Variances.** Parameter estimation was performed on simulated data sets either evolving neutrally, or with some fraction of loci used for inference linked to a selective sweep at some distance (measured by c/s). Each box plot summarizes estimates from 100 replicates for each scenario. Note that T_B, the bottleneck onset time, is absent from this figure because it was fixed it to the true value (Section 4.4).

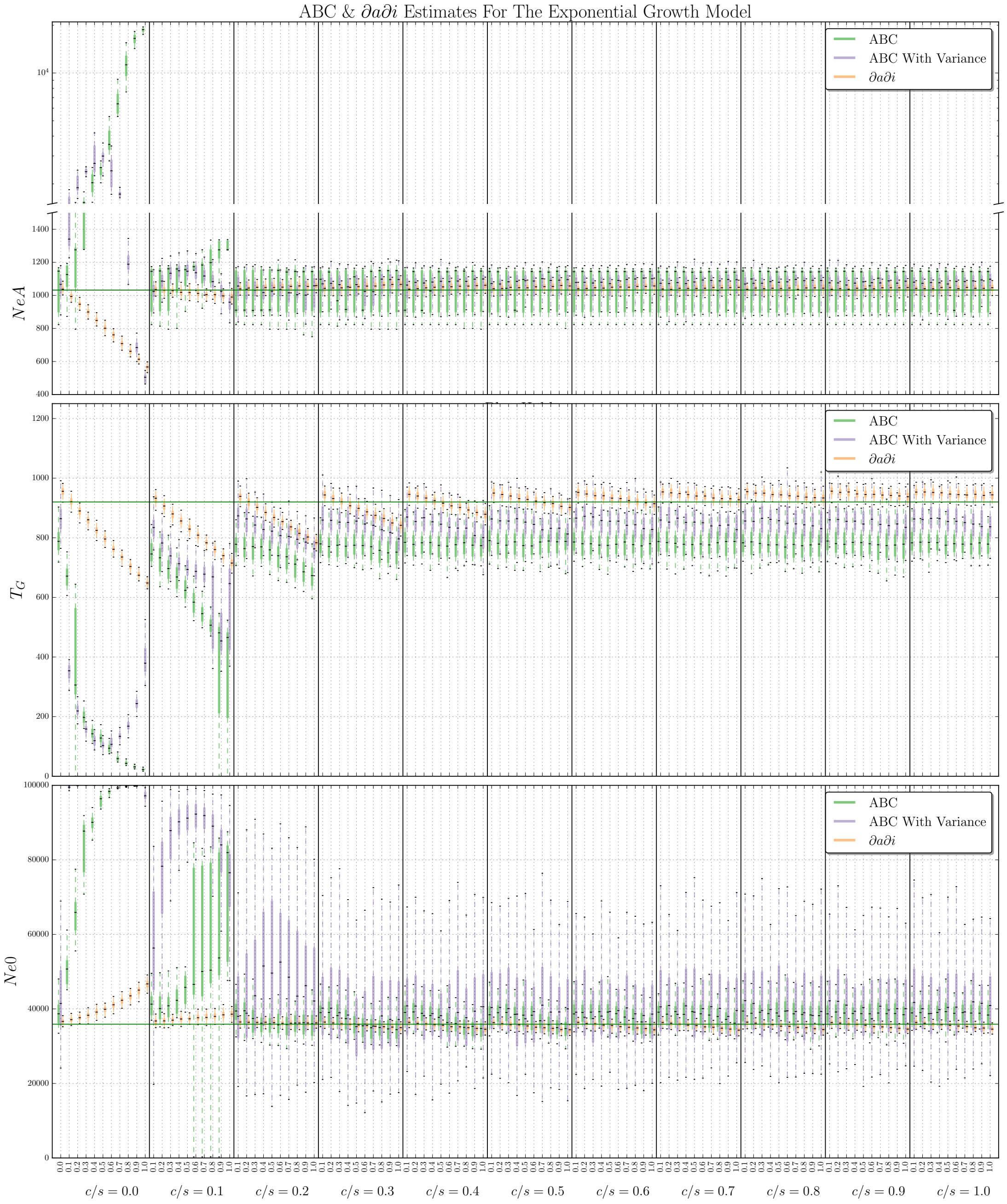


Figure F.22: **Growth Model Parameter Estimates From $\partial a \partial i$, ABC Using Summary Statistic Means, And ABC Using Both Means And Variances.** Parameter estimation was performed on simulated data sets either evolving neutrally, or with some fraction of loci used for inference linked to a selective sweep at some distance (measured by c/s). Each box plot summarizes estimates from 100 replicates for each scenario.

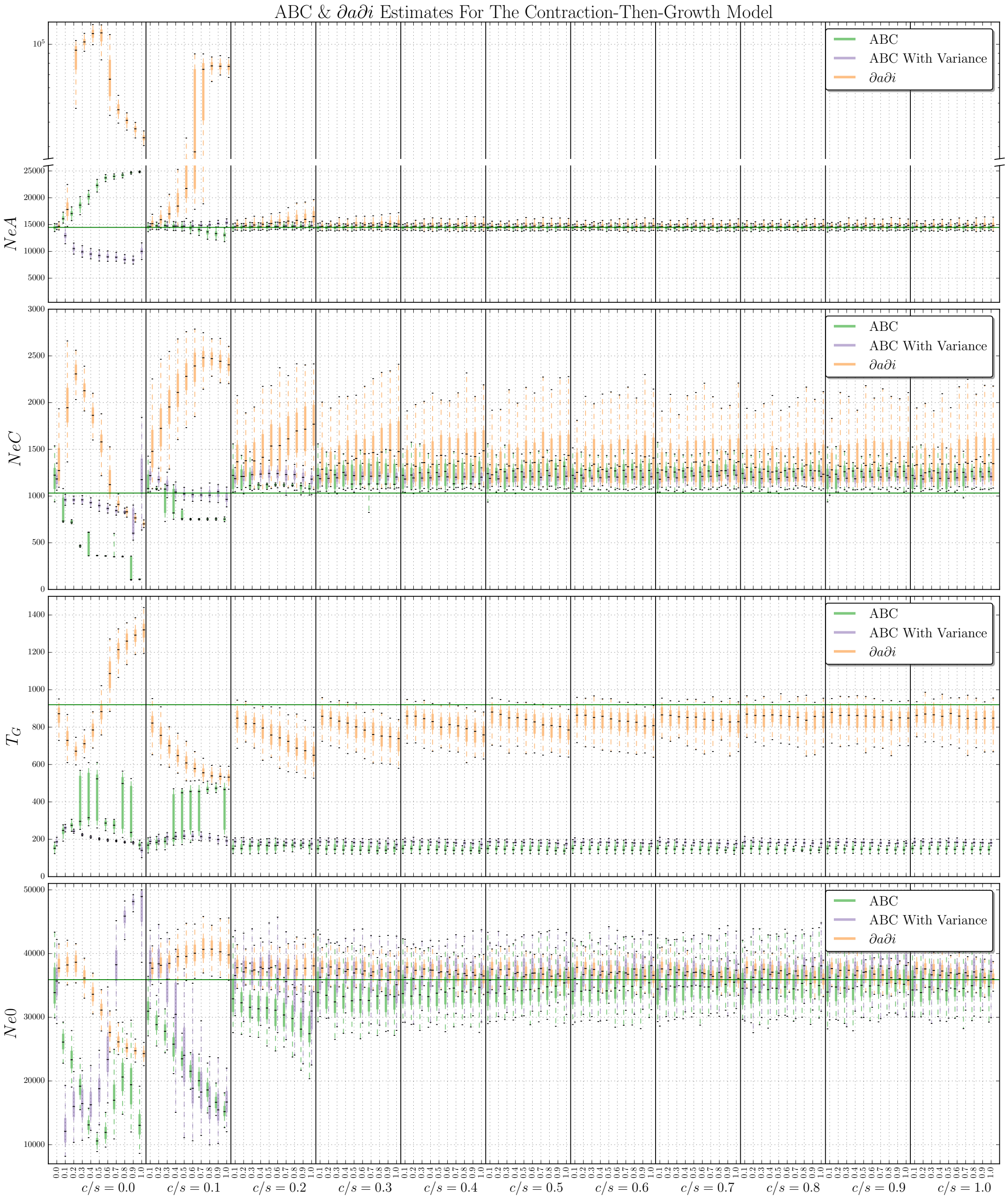


Figure F.23: **Contraction-Then-Growth Model Parameter Estimates From *∂a∂i*, ABC Using Summary Statistic Means, And ABC Using Both Means And Variances.** Parameter estimation was performed on simulated data sets either evolving neutrally, or with some fraction of loci used for inference linked to a selective sweep at some distance (measured by c/s). Each box plot summarizes estimates from 100 replicates for each scenario.

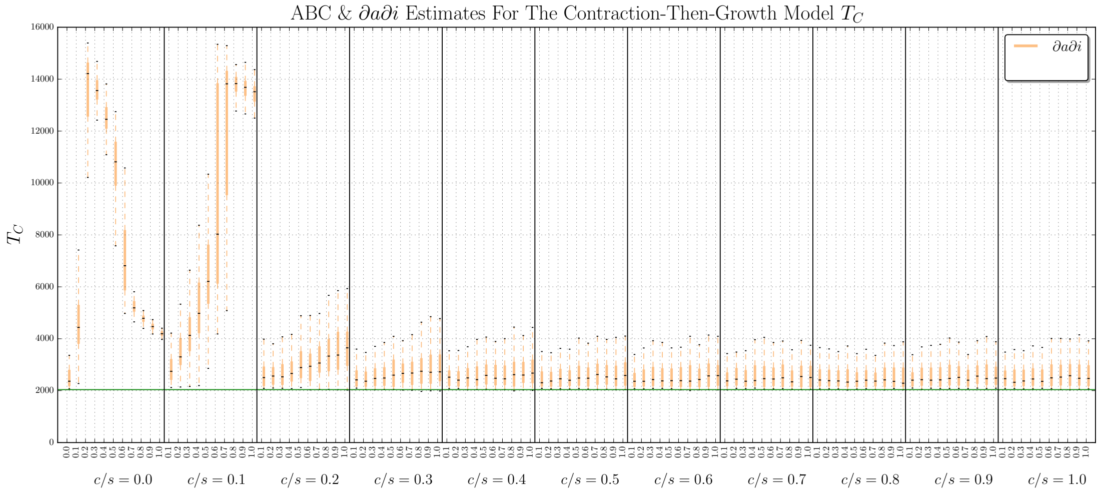


Figure F.24: **Contraction-Then-Growth Model T_C Parameter Estimates From $\partial a \partial i$.** Parameter estimation was performed on simulated data sets either evolving neutrally, or with some fraction of loci used for inference linked to a selective sweep at some distance (measured by c/s). Each box plot summarizes estimates from 100 replicates for each scenario.

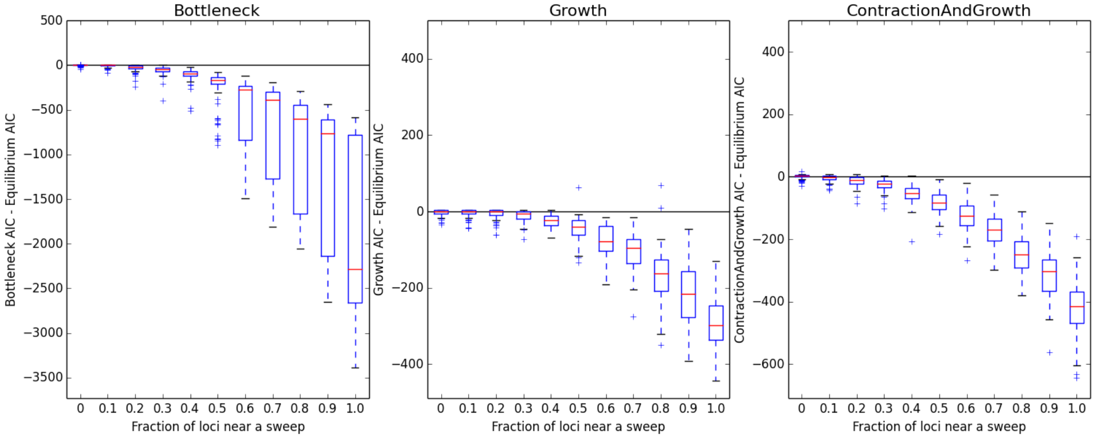


Figure F.25: **Differences In AIC Between Equilibrium And Non-equilibrium Models When Fitted By $\partial a \partial i$ To Simulated Constant-size Populations With Varying Degrees Of Positive Selection** . For the growth model, a small number of simulated optimized very poorly, leading to large AICs, and therefore large differences between the growth and equilibrium AIC. The upper limit of the y-axis of this plot was truncated to allow visualization of AIC differences for the bulk of the data for which optimization was more successful (though box and whisker lengths still reflect the presence of these outliers in the set).

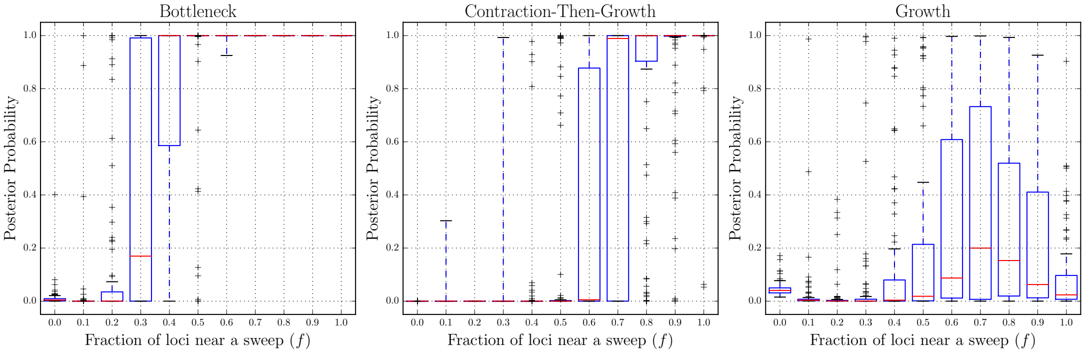


Figure F.26: **Estimated Posterior Probabilities Of The Non-Equilibrium Model.** Shown are the posterior probabilities of the alternative model when model selection is conducted in the following: Bottleneck vs. Equilibrium, Contraction & Growth vs. Equilibrium, and Growth vs. Equilibrium.

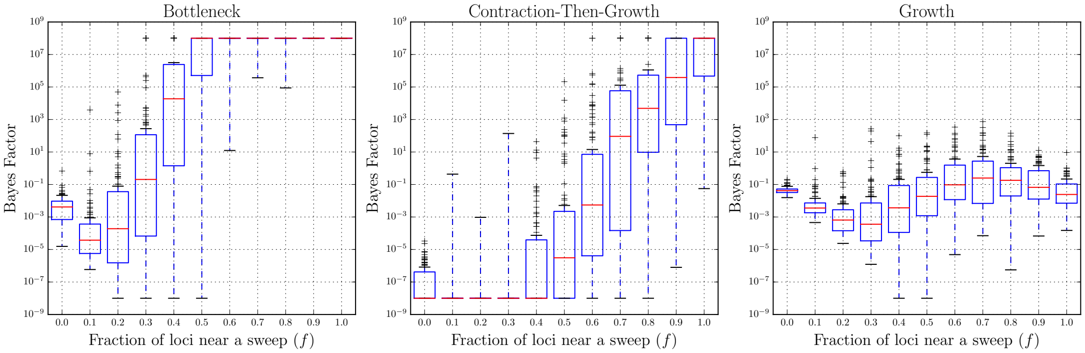


Figure F.27: **Estimated Bayes Factors Of The Non-Equilibrium Model.** Shown are the Bayes factors of the indicated model when model selection is conducted in the following: Bottleneck vs. Equilibrium, Contraction & Growth vs. Equilibrium, and Growth vs. Equilibrium.

Colophon

This thesis was created using LaTeX 2$_\varepsilon$ and BibTeX, the former originally developed by Leslie Lamport and based on Donald Knuth's TeX. The source code for this document was edited using Sublime Text 3 with the LaTeXTools package. The body text is set in 12 point URW Palladio, created originally by Design and Development GmbH in the style of Palintino. All captions were set in URW Palladio, while mathematical notation was set in Pazo Math and Myriad Pro. The bulk of the figures contained in this thesis were generated using matplotlib in python. Tables were created in LaTeX, except for those shown in Chapter 2 and Appendix E.1. Much of this document was formatted using the *RUThesis* template, which was found at `http://www.math.rutgers.edu/grad/phd_requirements/thesis.html`.

Bibliography

[1] Alexander G Shanku, Mark A. McPeek, and Andrew D. Kern. Functional Annotation and Comparative Analysis of a Zygopteran Transcriptome. *G3 (Bethesda)*, 3(4):763–770, January 2013. v

[2] Andrew Gelman and Donald B Rubin. Inference from Iterative Simulation Using Multiple Sequences. *Statistical Science*, 7(4):457–472, 1992. xxviii, 179

[3] R Ronen, N Udpa, E Halperin, and V Bafna. Learning Natural Selection from the Site Frequency Spectrum. *GENETICS*, June 2013. xxxi, 26, 28, 64, 78, 91, 189

[4] Daniel R Schrider, Fábio K Mendes, Matthew W Hahn, and Andrew D. Kern. Soft Shoulders Ahead: Spurious Signatures of Soft and Partial Selective Sweeps Result from Linked Hard Sweeps. *GENETICS*, 200(1):267–284, May 2015. xxxi, 79, 93, 134, 192

[5] Gabor T Marth, Eva Czabarka, Janos Murvai, and Stephen T Sherry. The allele frequency spectrum in genome-wide human variation data reveals signals of differential demographic history in three large world populations. *GENETICS*, 166(1):351–372, January 2004. xxxii, 31, 36, 136, 193

[6] Simon Gravel, Brenna M Henn, Ryan N Gutenkunst, Amit R Indap, Gabor T Marth, Andrew G Clark, Fuli Yu, Richard A Gibbs, Carlos D Bustamante, David L Altshuler, Richard M Durbin, Gonçalo R Abecasis, David R Bentley, Aravinda Chakravarti, Francis S Collins, Francisco M De La Vega, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B Gabriel, Bartha M Knoppers, Eric S Lander, Hans Lehrach, Elaine R Mardis, Gil A McVean, Debbie A Nickerson, Leena Peltonen, Alan J Schafer, Stephen T Sherry, Jun Wang, Richard K Wilson, David Deiros, Mike Metzker, Donna Muzny, Jeff Reid, David Wheeler, Jingxiang Li, Min Jian, Guoqing Li, Ruiqiang Li, Huiqing Liang, Geng Tian, Bo Wang, Jian Wang, Wei Wang, Huanming Yang, Xiuqing Zhang, Huisong Zheng, Lauren Ambrogio, Toby Bloom, Kristian Cibulskis, Tim J Fennell, David B Jaffe, Erica Shefler, Carrie L Sougnez, Niall Gormley, Sean Humphray, Zoya Kingsbury, Paula Koko-Gonzales, Jennifer

Stone, Kevin J McKernan, Gina L Costa, Jeffry K Ichikawa, Clarence C Lee, Ralf Sudbrak, Tatiana A Borodina, Andreas Dahl, Alexey N Davydov, Peter Marquardt, Florian Mertes, Wilfiried Nietfeld, Philip Rosenstiel, Stefan Schreiber, Aleksey V Soldatov, Bernd Timmermann, Marius Tolzmann, Jason Affourtit, Dana Ashworth, Said Attiya, Melissa Bachorski, Eli Buglione, Adam Burke, Amanda Caprio, Christopher Celone, Shauna Clark, David Conners, Brian Desany, Lisa Gu, Lorri Guccione, Kalvin Kao, Andrew Kebbel, Jennifer Knowlton, Matthew Labrecque, Louise McDade, Craig Mealmaker, Melissa Minderman, Anne Nawrocki, Faheem Niazi, Kristen Pareja, Ravi Ramenani, David Riches, Wanmin Song, Cynthia Turcotte, Shally Wang, David Dooling, Lucinda Fulton, Robert Fulton, George Weinstock, John Burton, David M Carter, Carol Churcher, Alison Coffey, Anthony Cox, Aarno Palotie, Michael Quail, Tom Skelly, James Stalker, Harold P Swerdlow, Daniel Turner, Anniek De Witte, Shane Giles, Matthew Bainbridge, Danny Challis, Aniko Sabo, Jin Yu, Xiaodong Fang, Xiaosen Guo, Yingrui Li, Ruibang Luo, Shuaishuai Tai, Honglong Wu, Hancheng Zheng, Xiaole Zheng, Yan Zhou, Erik P Garrison, Weichun Huang, Amit Indap, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Aaron R Quinlan, Chip Stewart, Michael P Stromberg, Alistair N Ward, Jiantao Wu, Charles Lee, Ryan E Mills, Xinghua Shi, Mark J Daly, Mark A DePristo, Aaron D Ball, Eric Banks, Brian L Browning, Kiran V Garimella, Sharon R Grossman, Robert E Handsaker, Matt Hanna, Chris Hartl, Andrew M Kernytsky, Joshua M Korn, Heng Li, Jared R Maguire, Steven A McCarroll, Aaron McKenna, James C Nemesh, Anthony A Philippakis, Ryan E Poplin, Alkes Price, Manuel A Rivas, Pardis C Sabeti, Stephen F Schaffner, Ilya A Shlyakhter, David N Cooper, Edward V Ball, Matthew Mort, Andrew D Phillips, Peter D Stenson, Jonathan Sebat, Vladimir Makarov, Kenny Ye, Seungtai C Yoon, Adam Boyko, Jeremiah Degenhardt, Mark Kaganovich, Alon Keinan, Phil Lacroute, Xin Ma, Andy Reynolds, Laura Clarke, Fiona Cunningham, Javier Herrero, Stephen Keenen, Eugene Kulesha, Rasko Leinonen, William M McLaren, Rajesh Radhakrishnan, Richard E Smith, Vadim Zalunin, Xiangqun Zheng-Bradley, Jan O Korbel, Adrian M Stütz, Markus Bauer, R Keira Cheetham, Tony Cox, Michael Eberle, Terena James, Scott Kahn, Lisa Murray, Kai Ye, Yutao Fu, Fiona C L Hyland, Jonathan M Manning, Stephen F McLaughlin, Heather E Peckham, Onur Sakarya, Yongming A Sun, Eric F Tsung, Mark A Batzer, Miriam K Konkel, Jerilyn A Walker, Marcus W Albrecht, Vyacheslav S Amstislavskiy, Ralf Herwig, Dimitri V Parkhomchuk, Richa Agarwala, Hoda M Khouri, Aleksandr O Morgulis, Justin E Paschall, Lon D Phan, Kirill E Rotmistrovsky, Robert D Sanders, Martin F Shumway, Chunlin Xiao, Adam Auton, Zamin Iqbal, Gerton Lunter, Jonathan L Marchini, Loukas Moutsianas, Simon Myers, Afidalina Tumian, and Kni... Demographic history and rare allele sharing among human populations. *Proceedings of the National Academy of Sciences*, 108(29):11983–11988, July 2011. xxxii, 36, 133, 136, 193

[7] W.W. Goodwin. *Plutarch's Morals*. Number v. 3 in Plutarch's Morals. Little, Brown,, 1874. URL `http://data.perseus.org/citations/urn:cts:greekLit:tlg0094.tlg003.perseus-eng1:5.19`. 1

[8] J. Huxley. *Evolution: The Modern Synthesis*. Science editions. John Wiley & Sons, 1942. URL `https://books.google.com/books?id=CHrgAAAAMAAJ`. 2

[9] J D Watson and F H C Crick. Molecular Structure of Nucleic Acids: A Structure for Deoxyribose Nucleic Acid. *NATURE*, 171(4356):737–738, April 1953. 2

[10] F Sanger and A R Coulson. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.*, 94(3):441–448, May 1975. 3

[11] F Sanger, G M Air, B G Barrell, N L Brown, A R Coulson, J C Fiddes, C A Hutchison, P M Slocombe, and M Smith. Nucleotide sequence of bacteriophage —[phi]—X174 DNA. *, Published online: 24 February 1977; — doi:10.1038/265687a0*, 265(5596):687–695, February 1977. 3

[12] F Sanger, S Nicklen, and A R Coulson. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U.S.A.*, 74(12):5463–5467, December 1977. 4

[13] Michael L Metzker. Emerging technologies in DNA sequencing. *GENOME RES*, 15(12):1767–1776, December 2005. 4

[14] S Anderson. Shotgun DNA sequencing using cloned DNase I-generated fragments. *NUCLEIC ACIDS RES*, 9(13):3015–3027, July 1981. 5

[15] R Staden. A strategy of DNA sequencing employing computer programs. *NUCLEIC ACIDS RES*, 6(7):2601–2610, June 1979. 5

[16] S Anderson, A T Bankier, B G Barrell, M H de Bruijn, A R Coulson, J Drouin, I C Eperon, D P Nierlich, B A Roe, F Sanger, P H Schreier, A J Smith, R Staden, and I G Young. Sequence and organization of the human mitochondrial genome. *NATURE*, 290(5806):457–465, April 1981. 5

[17] F Sanger, A R Coulson, G F Hong, D F Hill, and G B Petersen. Nucleotide sequence of bacteriophage lambda DNA. *Journal of molecular biology*, 162(4):729–773, December 1982. 5

[18] R D Fleischmann, M D Adams, O White, R A Clayton, E F Kirkness, A R Kerlavage, C J Bult, J F Tomb, B A Dougherty, J M Merrick, and et al. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *SCIENCE*, 269(5223):496–512, July 1995. 5

[19] M D Adams, S E Celniker, R A Holt, C A Evans, J D Gocayne, P G Amanatides, S E Scherer, P W Li, R A Hoskins, R F Galle, R A George, S E Lewis, S Richards, M Ashburner, S N Henderson, G G Sutton, J R Wortman, M D Yandell, Q Zhang, L X Chen, R C Brandon, Y H Rogers, R G Blazej, M Champe, B D Pfeiffer, K H Wan, C Doyle, E G Baxter, G Helt, C R Nelson, G L Gabor, J F Abril, A Agbayani, H J An, C Andrews-Pfannkoch, D Baldwin, R M Ballew, A Basu, J Baxendale, L Bayraktaroglu, E M Beasley, K Y Beeson, P V Benos, B P Berman, D Bhandari, S Bolshakov, D Borkova, M R Botchan, J Bouck, P Brokstein, P Brottier, K C Burtis, D A Busam, H Butler, E Cadieu, A Center, I Chandra, J M Cherry, S Cawley, C Dahlke, L B Davenport, P Davies, B de Pablos, A Delcher, Z Deng, A D Mays, I Dew, S M Dietz, K Dodson, L E Doup, M Downes, S Dugan-Rocha, B C Dunkov, P Dunn, K J Durbin, C C Evangelista, C Ferraz, S Ferriera, W Fleischmann, C Fosler, A E Gabrielian, N S Garg, W M Gelbart, K Glasser, A Glodek, F Gong, J H Gorrell, Z Gu, P Guan, M Harris, N L Harris, D Harvey, T J Heiman, J R Hernandez, J Houck, D Hostin, K A Houston, T J Howland, M H Wei, C Ibegwam, M Jalali, F Kalush, G H Karpen, Z Ke, J A Kennison, K A Ketchum, B E Kimmel, C D Kodira, C Kraft, S Kravitz, D Kulp, Z Lai, P Lasko, Y Lei, A A Levitsky, J Li, Z Li, Y Liang, X Lin, X Liu, B Mattei, T C McIntosh, M P McLeod, D McPherson, G Merkulov, N V Milshina, C Mobarry, J Morris, A Moshrefi, S M Mount, M Moy, B Murphy, L Murphy, D M Muzny, D L Nelson, D R Nelson, K A Nelson, K Nixon, D R Nusskern, J M Pacleb, M Palazzolo, G S Pittman, S Pan, J Pollard, V Puri, M G Reese, K Reinert, K Remington, R D Saunders, F Scheeler, H Shen, B C Shue, I Sidén-Kiamos, M Simpson, M P Skupski, T Smith, E Spier, A C Spradling, M Stapleton, R Strong, E Sun, R Svirskas, C Tector, R Turner, E Venter, A H Wang, X Wang, Z Y Wang, D A Wassarman, G M Weinstock, J Weissenbach, S M Williams, WoodageT, K C Worley, D Wu, S Yang, Q A Yao, J Ye, R F Yeh, J S Zaveri, M Zhan, G Zhang, Q Zhao, L Zheng, X H Zheng, F N Zhong, W Zhong, X Zhou, S Zhu, X Zhu, H O Smith, R A Gibbs, E W Myers, G M Rubin, and J C Venter. The genome sequence of Drosophila melanogaster. *SCIENCE*, 287(5461):2185–2195, March 2000. 6

[20] G M Rubin and E B Lewis. A brief history of Drosophila's contributions to genome research. *SCIENCE*, 287(5461):2216–2218, March 2000. 6

[21] J Palca. Human genome. Department of Energy on the map. *NATURE*, 321(6068):371, May 1986. 6

[22] Eric S Lander, Lauren M Linton, Bruce Birren, Chad Nusbaum, Michael C Zody, Jennifer Baldwin, Keri Devon, Ken Dewar, Michael Doyle, William FitzHugh, Roel Funke, Diane Gage, Katrina Harris, Andrew Heaford, John Howland, Lisa Kann, Jessica Lehoczky, Rosie LeVine, Paul McEwan, Kevin McKernan, James Meldrim, Jill P Mesirov, Cher Miranda, William Morris, Jerome Naylor, Christina Raymond, Mark Rosetti, Ralph Santos, Andrew Sheridan, Carrie Sougnez, Nicole Stange-Thomann, Nikola Stojanovic, Aravind Subramanian, Dudley Wyman, Jane Rogers, John Sulston, Rachael Ainscough, Stephan Beck, David Bentley, John Burton, Christopher Clee, Nigel Carter, Alan Coulson, Rebecca Deadman, Panos Deloukas, Andrew Dunham, Ian Dunham, Richard Durbin, Lisa French, Darren Grafham, Simon Gregory, Tim Hubbard, Sean Humphray, Adrienne Hunt, Matthew Jones, Christine Lloyd, Amanda McMurray, Lucy Matthews, Simon Mercer, Sarah Milne, James C Mullikin, Andrew Mungall, Robert Plumb, Mark Ross, Ratna Shownkeen, Sarah Sims, Robert H Waterston, Richard K Wilson, LaDeana W Hillier, John D McPherson, Marco A Marra, Elaine R Mardis, Lucinda A Fulton, Asif T Chinwalla, Kymberlie H Pepin, Warren R Gish, Stephanie L Chissoe, Michael C Wendl, Kim D Delehaunty, Tracie L Miner, Andrew Delehaunty, Jason B Kramer, Lisa L Cook, Robert S Fulton, Douglas L Johnson, Patrick J Minx, Sandra W Clifton, Trevor Hawkins, Elbert Branscomb, Paul Predki, Paul Richardson, Sarah Wenning, Tom Slezak, Norman Doggett, Jan-Fang Cheng, Anne Olsen, Susan Lucas, Christopher Elkin, Edward Uberbacher, Marvin Frazier, Richard A Gibbs, Donna M Muzny, Steven E Scherer, John B Bouck, Erica J Sodergren, Kim C Worley, Catherine M Rives, James H Gorrell, Michael L Metzker, Susan L Naylor, Raju S Kucherlapati, David L Nelson, George M Weinstock, Yoshiyuki Sakaki, Asao Fujiyama, Masahira Hattori, Tetsushi Yada, Atsushi Toyoda, Takehiko Itoh, Chiharu Kawagoe, Hidemi Watanabe, Yasushi Totoki, Todd Taylor, Jean Weissenbach, Roland Heilig, William Saurin, Francois Artiguenave, Philippe Brottier, Thomas Bruls, Eric Pelletier, Catherine Robert, Patrick Wincker, André Rosenthal, Matthias Platzer, Gerald Nyakatura, Stefan Taudien, Andreas Rump, Douglas R Smith, Lynn Doucette-Stamm, Marc Rubenfield, Keith Weinstock, Hong Mei Lee, JoAnn Dubois, Huanming Yang, Jun Yu, Jian Wang, Guyang Huang, Jun Gu, Leroy Hood, Lee Rowen, Anup Madan, Shizen Qin, Ronald W Davis, Nancy A Federspiel, A Pia Abola, Michael J Proctor, Bruce A Roe, Feng Chen, Huaqin Pan, Juliane Ramser, Hans Lehrach, Richard Reinhardt, W Richard McCombie, Melissa de la Bastide, Neilay Dedhia, Helmut Blöcker, Klaus Hornischer, Gabriele Nordsiek, Richa Agarwala, L Aravind, Jeffrey A Bailey, Alex Bateman,

Serafim Batzoglou, Ewan Birney, Peer Bork, Daniel G Brown, Christopher B Burge, Lorenzo Cerutti, Hsiu-Chuan Chen, Deanna Church, Michele Clamp, Richard R Copley, Tobias Doerks, Sean R Eddy, Evan E Eichler, Terrence S Furey, James Galagan, James G R Gilbert, Cyrus Harmon, Yoshihide Hayashizaki, David Haussler, Henning Hermjakob, Karsten Hokamp, Wonhee Jang, L Steven Johnson, Thomas A Jones, Simon Kasif, Arek Kaspryzk, Scot Kennedy, W James Kent, Paul Kitts, Eugene V Koonin, Ian Korf, David Kulp, Doron Lancet, Todd M Lowe, Aoife McLysaght, Tarjei Mikkelsen, John V Moran, Nicola Mulder, Victor J Pollara, Chris P Ponting, Greg Schuler, Jörg Schultz, Guy Slater, Arian F A Smit, Elia Stupka, Joseph Szustakowki, Danielle Thierry-Mieg, Jean Thierry-Mieg, Lukas Wagner, John Wallis, Raymond Wheeler, Alan Williams, Yuri I Wolf, Kenneth H Wolfe, Shiaw-Pyng Yang, Ru-Fang Yeh, Francis Collins, Mark S Guyer, Jane Peterson, Adam Felsenfeld, Kris A Wetterstrand, Richard M Myers, Jeremy Schmutz, Mark Dickson, Jane Grimwood, David R Cox, Maynard V Olson, and Rajin... Kaul. Initial sequencing and analysis of the human genome. *NATURE*, 409(6822):860–921, February 2001. 6, 9

[23] Jonathan Max Gitlin. Calculating the economic impact of the Human Genome Project. May 2011. URL `https://www.genome.gov/27544383`. 6

[24] Michael L Metzker. Sequencing technologies - the next generation. *NAT REV GENET*, 11(1):31–46, January 2010. 7

[25] Lin Liu, Yinhu Li, Siliang Li, Ni Hu, Yimin He, Ray Pong, Danni Lin, Lihua Lu, and Maggie Law. Comparison of Next-Generation Sequencing Systems. *Journal of Biomedicine and Biotechnology*, 2012(7):1–11, 2012. 7, 9

[26] Marcel Margulies, Michael Egholm, William E Altman, Said Attiya, Joel S Bader, Lisa A Bemben, Jan Berka, Michael S Braverman, Yi-Ju Chen, Zhoutao Chen, Scott B Dewell, Lei Du, Joseph M Fierro, Xavier V Gomes, Brian C Godwin, Wen He, Scott Helgesen, Chun He Ho, Gerard P Irzyk, Szilveszter C Jando, Maria L I Alenquer, Thomas P Jarvie, Kshama B Jirage, Jong-Bum Kim, James R Knight, Janna R Lanza, John H Leamon, Steven M Lefkowitz, Ming Lei, Jing Li, Kenton L Lohman, Hong Lu, Vinod B Makhijani, Keith E McDade, Michael P McKenna, Eugene W Myers, Elizabeth Nickerson, John R Nobile, Ramona Plant, Bernard P Puc, Michael T Ronan, George T Roth, Gary J Sarkis, Jan Fredrik Simons, John W Simpson, Maithreyan Srinivasan, Karrie R Tartaro, Alexander Tomasz, Kari A Vogt, Greg A Volkmer, Shally H Wang, Yong Wang, Michael P Weiner, Pengguang Yu, Richard F Begley, and Jonathan M Rothberg. Genome sequencing in microfabricated high-density picolitre reactors. *NATURE*, 437(7057):376–380, September 2005. 8

[27] GenomeWeb. Roche shutting down 454 sequencing business. 2013. URL `https://www.genomeweb.com/sequencing/roche-shutting-down-454-sequencing-business`. 8

[28] Illumina. History of illumina sequencing. 2016. URL `http://www.illumina.com/technology/next-generation-sequencing/solexa-technology.html`. 8

[29] David R Bentley, Shankar Balasubramanian, Harold P Swerdlow, Geoffrey P Smith, John Milton, Clive G Brown, Kevin P Hall, Dirk J Evers, Colin L Barnes, Helen R Bignell, Jonathan M Boutell, Jason Bryant, Richard J Carter, R Keira Cheetham, Anthony J Cox, Darren J Ellis, Michael R Flatbush, Niall A Gormley, Sean J Humphray, Leslie J Irving, Mirian S Karbelashvili, Scott M Kirk, Heng Li, Xiaohai Liu, Klaus S Maisinger, Lisa J Murray, Bojan Obradovic, Tobias Ost, Michael L Parkinson, Mark R Pratt, Isabelle M J Rasolonjatovo, Mark T Reed, Roberto Rigatti, Chiara Rodighiero, Mark T Ross, Andrea Sabot, Subramanian V Sankar, Aylwyn Scally, Gary P Schroth, Mark E Smith, Vincent P Smith, Anastassia Spiridou, Peta E Torrance, Svilen S Tzonev, Eric H Vermaas, Klaudia Walter, Xiaolin Wu, Lu Zhang, Mohammed D Alam, Carole Anastasi, Ify C Aniebo, David M D Bailey, Iain R Bancarz, Saibal Banerjee, Selena G Barbour, Primo A Baybayan, Vincent A Benoit, Kevin F Benson, Claire Bevis, Phillip J Black, Asha Boodhun, Joe S Brennan, John A Bridgham, Rob C Brown, Andrew A Brown, Dale H Buermann, Abass A Bundu, James C Burrows, Nigel P Carter, Nestor Castillo, Maria Chiara E Catenazzi, Simon Chang, R Neil Cooley, Natasha R Crake, Olubunmi O Dada, Konstantinos D Diakoumakos, Belen Dominguez-Fernandez, David J Earnshaw, Ugonna C Egbujor, David W Elmore, Sergey S Etchin, Mark R Ewan, Milan Fedurco, Louise J Fraser, Karin V Fuentes Fajardo, W Scott Furey, David George, Kimberley J Gietzen, Colin P Goddard, George S Golda, Philip A Granieri, David E Green, David L Gustafson, Nancy F Hansen, Kevin Harnish, Christian D Haudenschild, Narinder I Heyer, Matthew M Hims, Johnny T Ho, Adrian M Horgan, Katya Hoschler, Steve Hurwitz, Denis V Ivanov, Maria Q Johnson, Terena James, T A Huw Jones, Gyoung-Dong Kang, Tzvetana H Kerelska, Alan D Kersey, Irina Khrebtukova, Alex P Kindwall, Zoya Kingsbury, Paula I Kokko-Gonzales, Anil Kumar, Marc A Laurent, Cynthia T Lawley, Sarah E Lee, Xavier Lee, Arnold K Liao, Jennifer A Loch, Mitch Lok, Shujun Luo, Radhika M Mammen, John W Martin, Patrick G McCauley, Paul McNitt, Parul Mehta, Keith W Moon, Joe W Mullens, Taksina Newington, Zemin Ning, Bee Ling Ng, Sonia M Novo, Michael J O'Neill, Mark A Osborne, Andrew Osnowski, Omead Ostadan, Lambros L Paraschos, Lea Pickering, Andrew C Pike, Alger C Pike, D Chris Pinkard, Daniel P Pliskin, Joe Podhasky, Victor J Quijano,

Come Raczy, Vicki H Rae, Stephen R Rawlings, Ana Chiva Rodriguez, Phyllida M Roe, John Rogers, Maria C Rogert Bacigalupo, Nikolai Romanov, Anthony Romieu, Rithy K Roth, Natalie J Rourke, Silke T Ruediger, Eli Rusman, Raquel M Sanches-Kuiper, Martin R Schenker, Josefina M Seoane, Richard J Shaw, Mitch K Shiver, Steven W Short, Ning L Sizto, Johannes P Sluis, Melanie A Smith, Jean Ernest Sohna Sohna, Eric J Spence, Kim Stevens, Neil Sutton, Lukasz Szajkowski, Carolyn L Tregidgo, Gerardo Turcatti, Stephanie Vandevondele, Yuli Verhovsky, Selene M Virk, Suzanne Wakelin, Gregory C Walcott, Jingwen Wang, Graham J Worsley, Juying Yan, Ling Yau, Mike Zuerlein, Jane Rogers, James C Mullikin, Matthew E Hurles, Nick J McCooke, John S West, Frank L Oaks, Peter L Lundberg, David Klenerman, Richard Durbin, and Anthony J Smith. Accurate whole human genome sequencing using reversible terminator chemistry. *NATURE*, 456(7218):53–59, November 2008. 8

[30] Illumina. Sequencing platform comparison tool. 2016. URL `https://www.illumina.com/systems/sequencing-platform-comparison.html`. 9

[31] Elaine R Mardis. The impact of next-generation sequencing technology on genetics. *TRENDS GENET*, 24(3):133–141, March 2008. 9, 52

[32] Elaine R Mardis. A decade's perspective on DNA sequencing technology. *NATURE*, 470(7333):198–203, February 2011. 9

[33] Chen-Shan Chin, Jon Sorenson, Jason B Harris, William P Robins, Richelle C Charles, Roger R Jean-Charles, James Bullard, Dale R Webster, Andrew Kasarskis, Paul Peluso, Ellen E Paxinos, Yoshiharu Yamaichi, Stephen B Calderwood, John J Mekalanos, Eric E Schadt, and Matthew K Waldor. The Origin of the Haitian Cholera Outbreak Strain. *N Engl J Med*, 364(1):33–42, January 2011. 10

[34] Sergey Koren, Gregory P Harhay, Timothy PL Smith, James L Bono, Dayna M Harhay, Scott D Mcvey, Diana Radune, Nicholas H Bergman, and Adam M Phillippy. Reducing assembly complexity of microbial genomes with single-molecule sequencing. *GENOME BIOL*, 14(9):R101, September 2013. 10

[35] Tyson A Clark, Iain A Murray, Richard D Morgan, Andrey O Kislyuk, Kristi E Spittle, Matthew Boitano, Alexey Fomenkov, Richard J Roberts, and Jonas Korlach. Characterization of DNA methyltransferase specificities using single-molecule, real-time DNA sequencing. *NUCLEIC ACIDS RES*, 40(4):e29–e29, February 2012. 10

[36] Iain A Murray, Tyson A Clark, Richard D Morgan, Matthew Boitano, Brian P Anton, Khai Luong, Alexey Fomenkov, Stephen W Turner, Jonas Korlach, and Richard J Roberts. The methylomes of six bacteria. *NUCLEIC ACIDS RES*, 40(22):11450–11462, December 2012. 11

[37] Zhong Wang, Mark Gerstein, and Michael Snyder. RNA-Seq: a revolutionary tool for transcriptomics. *NAT REV GENET*, 10(1):57–63, January 2009. 11, 12

[38] A G Carig, D Nizetic, J D Hoheisel, G Zehetner, and H Lehrach. Ordering of cosmid clones covering the Herpes Simplex virus type I (HSV-I) genome: a test case for fingerprinting by hybridisation. *NUCLEIC ACIDS RES*, 18(9):2653–2660, 1990. 11

[39] William Bains and Geoff C Smith. A novel method for nucleic acid sequence determination. *Journal of Theoretical Biology*, 135(3):303–307, December 1988. 11

[40] Victor Trevino, Francesco Falciani, and Hugo A Barrera-Saldaña. DNA Microarrays: a Powerful Genomic Tool for Biomedical and Clinical Research. *Molecular Medicine*, 13(9-10):527–541, 2007. 12

[41] Laura J van 't Veer, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, George J Schreiber, Ron M Kerkhoven, Chris Roberts, Peter S Linsley, René Bernards, and Stephen H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *NATURE*, 415(6871):530–536, January 2002. 12

[42] Dinesh Singh, Phillip G Febbo, Kenneth Ross, Donald G Jackson, Judith Manola, Christine Ladd, Pablo Tamayo, Andrew A Renshaw, Anthony V D'Amico, Jerome P Richie, Eric S Lander, Massimo Loda, Philip W Kantoff, Todd R Golub, and William R Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, March 2002. 12

[43] Arndt Brachat, Benoit Pierrat, Alexandros Xynos, Karin Brecht, Marjo Simonen, Adrian Brüngger, and Jutta Heim. A microarray-based, integrated approach to identify novel regulators of cancer drug response and apoptosis. - PubMed - NCBI. *Oncogene*, 21(54):8361–8371, November 2002. 12

[44] Patrick O Brown, Jonathan R Pollack, Charles M Perou, Ash A Alizadeh, Michael B Eisen, Alexander Pergamenschikov, Cheryl F Williams, Stefanie S Jeffrey, and David Botstein. Genome-wide analysis of DNA copy-number changes using cDNA microarrays. - PubMed - NCBI. *NAT GENET*, 23(1):41–46, September 1999. 12

[45] D J Cutler, M E Zwick, M M Carrasquillo, C T Yohn, K P Tobin, C Kashuk, D J Mathews, N A Shah, E E Eichler, J A Warrington, and A Chakravarti. High-throughput variation detection and genotyping using microarrays. *GENOME RES*, 11(11):1913–1925, November 2001. 12

[46] Shanrong Zhao, Wai-Ping Fung-Leung, Anton Bittner, Karen Ngo, and Xuejun Liu. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PLOS ONE*, 9(1):e78644, 2014. 12

[47] Michał J Okoniewski and Crispin J Miller. Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations. *BMC Bioinformatics*, 7(1):1–14, 2006. 12

[48] Jason R Miller, Sergey Koren, and Granger Sutton. Assembly Algorithms for Next-Generation Sequencing Data. *Genomics*, 95(6):315–327, June 2010. 13, 14

[49] Pavel A Pevzner, Haixu Tang, and Michael S Waterman. An Eulerian path approach to DNA fragment assembly. *P NATL ACAD SCI USA*, 98(17):9748–9753, August 2001. 13

[50] N. G. de Bruijn. A Combinatorial Problem. *Koninklijke Nederlandse Akademie v. Wetenschappen*, 49:758–764, 1946. 13

[51] R M Idury and M S Waterman. A new algorithm for DNA sequence assembly. *J. Comput. Biol.*, 2(2):291–306, 1995. 13, 14

[52] Mihai Pop. Genome assembly reborn: recent computational challenges. *Briefings in bioinformatics*, 10(4):354–366, July 2009. 13

[53] Stephen F Altschul, Warren Gish, Webb Miller, Eugene W Myers, and David J Lipman. Basic local alignment search tool. *Journal of molecular biology*, 215(3):403–410, October 1990. 14

[54] M Sabrina Pankey, Vladimir N Minin, Greg C Imholte, Marc A Suchard, and Todd H Oakley. Predictable transcriptome evolution in the convergent and complex bioluminescent organs of squid. *Proceedings of the National Academy of Sciences*, 111(44):E4736–42, November 2014. 15

[55] Amy L Toth and Gene E Robinson. Evo-devo and the evolution of social behavior. *TRENDS GENET*, 23(7):334–341, January 2007. 15

[56] Amy L Toth, Kranthi Varala, Michael T Henshaw, Sandra L Rodriguez-Zas, Matthew E Hudson, and Gene E Robinson. Brain transcriptomic analysis in paper wasps identifies genes associated with behaviour across social insect lineages. *Proceedings. Biological sciences / The Royal Society*, 277(1691):2139–2148, July 2010. 15

[57] Ali J Berens, James H Hunt, and Amy L Toth. Comparative transcriptomics of convergent evolution: different genes but conserved pathways underlie caste phenotypes across lineages of eusocial insects. *MOL BIOL EVOL*, 32(3):690–703, March 2015. 15

[58] Liandong Yang, Ying Wang, Zhaolei Zhang, and Shunping He. Comprehensive transcriptome analysis reveals accelerated genic evolution in a Tibet fish, Gymnodiptychus pachycheilus. *Genome Biol Evol*, 7(1):251–261, January 2015. 16

[59] Raquel Assis, Qi Zhou, and Doris Bachtrog. Sex-biased transcriptome evolution in Drosophila. *Genome Biol Evol*, 4(11):1189–1200, 2012. 16

[60] Ana Riesgo, Sónia C S Andrade, Prashant P Sharma, Marta Novo, Alicia R Pérez-Porro, Varpu Vahtera, Vanessa L González, Gisele Y Kawauchi, and Gonzalo Giribet. Comparative description of ten transcriptomes of newly sequenced invertebrates and efficiency estimation of genomic sampling in non-model taxa. *Frontiers in Zoology*, 9(1):33, November 2012. 17

[61] Eva Jiménez-Guri, Jaime Huerta-Cepas, Luca Cozzuto, Karl R Wotton, Hui Kang, Heinz Himmelbauer, Guglielmo Roma, Toni Gabaldón, and Johannes Jaeger. Comparative transcriptomics of early dipteran development. *BMC Genomics 2009 10:219*, 14(1):1, February 2013. 17

[62] David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, Frank W Albert, Ulrich Zeller, Philipp Khaitovich, Frank Grützner, Sven Bergmann, Rasmus Nielsen, Svante Pääbo, and Henrik Kaessmann. The evolution of gene expression levels in mammalian organs. *NATURE*, 478(7369):343–348, October 2011. 17

[63] Charles Darwin. *On the Origin of Species by Means of Natural Selection*. Murray, London, 1859. Or the Preservation of Favored Races in the Struggle for Life. 18

[64] Joshua M Akey, Ge Zhang, Kun Zhang, Li Jin, and Mark D Shriver. Interrogating a High-Density SNP Map for Signatures of Natural Selection. *GENOME RES*, 12(12):1805–1814, December 2002. 18, 62

[65] Penelope R Haddrill, Kevin R Thornton, Brian Charlesworth, and Peter Andolfatto. Multilocus patterns of nucleotide variability and the demographic and selection history of Drosophila melanogaster populations. *GENOME RES*, 15(6):790–799, 2005. 18, 62

[66] Rasmus Nielsen, Carlos Bustamante, Andrew G Clark, Stephen Glanowski, Timothy B Sackton, Melissa J Hubisz, Adi Fledel-Alon, David M Tanenbaum, Daniel Civello, Thomas J White, John J Sninsky, Mark D Adams, and Michele Cargill. A scan for positively selected genes in the genomes of humans and chimpanzees. *PLOS BIOL*, 3(6):e170, 2005. 18, 62

[67] KS Pollard, SR Salama, B King, and AD Kern. Forces shaping the fastest evolving regions in the human genome. *PLOS GENETIC*, 2006. 18, 62

[68] BF Voight, S Kudaravalli, X Wen, and JK Pritchard. A map of recent positive selection in the human genome. *PLOS BIOL*, 4(3):e72, 2006. 18, 62

[69] W Stephan and H Li. The recent demographic and adaptive history of Drosophila melanogaster. *HEREDITY*, 98(2):65–68, February 2007. 18, 62

[70] David J. Begun, Alisha K Holloway, Kristian Stevens, LaDeana W Hillier, Yu-Ping Poh, Matthew W Hahn, Phillip M Nista, Corbin D. Jones, Andrew D. Kern, Colin N Dewey, Lior Pachter, Eugene Myers, and Charles H Langley. Population Genomics: Whole-Genome Analysis of Polymorphism and Divergence in Drosophila simulans. *PLOS BIOL*, 5(11):e310, November 2007. 18, 62, 95, 99

[71] Charles H Langley, Kristian Stevens, Charis Cardeno, Yuh Chwen G Lee, Daniel R Schrider, John E Pool, Sasha A Langley, Charlyn Suarez, Russell B Corbett-Detig, Bryan Kolaczkowski, et al. Genomic variation in natural populations of drosophila melanogaster. *GENETICS*, 192(2):533–598, 2012. 18, 62, 64

[72] JM. Maynard Smith and J. Haigh. The hitch-hiking effect of a favourable gene. *GENET RES*, (23):23–35, 1974. 19, 62, 63, 114

[73] J H Gillespie. *The causes of molecular evolution*. Oxford University Press, New York, 1991. 19, 63, 92

[74] J H Gillespie. Genetic drift in an infinite population. The pseudohitchhiking model. *GENETICS*, 155(2):909–919, 2000. 19

[75] Benjamin H Good and Michael M Desai. Deleterious Passengers in Adapting Populations. *GENETICS*, 198(3):1183–1208, November 2014. 19

[76] J C Fay and C I Wu. Hitchhiking under positive Darwinian selection. *GENETICS*, 155:1405–1413, 2000. 19, 23, 28, 71, 146

[77] N L Kaplan, R R Hudson, and C H Langley. The ”hitchhiking effect” revisited. *GENETICS*, 123(4):887–899, December 1989. 19, 63, 130

[78] J M Braverman, R R Hudson, N L Kaplan, C H Langley, and W Stephan. The hitchhiking effect on the site frequency spectrum of DNA polymorphisms. *GENETICS*, 140(2):783–796, June 1995. 19, 21, 101

[79] Joachim Hermisson and Pleuni S. Pennings. Soft sweeps: molecular population genetics of adaptation from standing genetic variation. *GENETICS*, 169(4):2335–2352, 2005. 19, 62, 63, 92, 134

[80] Sewall Wright. The Evolution of Dominance. *The American naturalist*, 63(689):556–561, November 1929. 20

[81] Motoo Kimura. *The neutral theory of molecular evolution*. Cambridge University Press, New York, 1983. 20

[82] Pardis C Sabeti, David E Reich, John M Higgins, Haninah Z P Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J Mcdonald, Hans C Ackerman, Sarah J Campbell, David Altshuler, Richard Cooper, Dominic Kwiatkowski, Ryk Ward, and Eric S Lander. Detecting recent positive selection in the human genome from haplotype structure. *NATURE*, 419(6909):832–837, October 2002. 20, 134

[83] Pardis C Sabeti, Patrick Varilly, Ben Fry, Jason Lohmueller, Elizabeth Hostetter, Chris Cotsapas, Xiaohui Xie, Elizabeth H Byrne, Steven A McCarroll, Rachelle Gaudet, Stephen F Schaffner, Eric S Lander, International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier,

Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, David Altshuler, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, Tatsuhiko Tsunoda, Todd A Johnson, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Changqing Zeng, Hui Zhao, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Richard A Gibbs, John W Belmont, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, David A Wheeler, Imtaz Yakub, Stacey B Gabriel, Robert C Onofrio, Daniel J Richter, Liuda Ziaugra, Bruce W Birren, Mark J Daly, David Altshuler, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, and John... Wallenburg. Genome-wide detection and characterization of positive selection in human populations. *NATURE*, 449(7164):913–918, October 2007. 20, 29

[84] S Wright. The distribution of gene frequencies under irreversible mutation. *P NATL ACAD SCI USA*, 24(7):253–259, 1938. 20

[85] M Kimura. Diffusion models in population genetics. *Journal of Applied*

Probability, 1(2):177–232, 1964. 20

[86] Fumio Tajima. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *GENETICS*, 123:585–595, 1989. 20, 28, 63, 71, 145

[87] Molly Przeworski. The signature of positive selection at randomly chosen loci. *GENETICS*, 160(3):1179–1189, 2002. 21, 114

[88] Yuseob Kim and Wolfgang Stephan. Detecting a local signature of genetic hitchhiking along a recombining chromosome. *GENETICS*, 160(2):765–777, February 2002. 21, 23, 27, 63, 101

[89] W H Press, S A Teukolsky, W T Vetterling, and B P Flannery. *Numerical Recipes in C*. Cambridge University Press, 1992. 22

[90] M Kimura. Theoretical foundation of population genetics at the molecular level. *THEOR POPUL BIOL*, 2:174–208, 1971. 22

[91] Jeffrey D Jensen, Yuseob Kim, Vanessa Bauer DuMont, Charles F Aquadro, and Carlos D Bustamante. Distinguishing between selective sweeps and demography using DNA polymorphism data. *GENETICS*, 170(3):1401–1410, July 2005. 23, 24, 31, 114

[92] Yuseob Kim and Rasmus Nielsen. Linkage disequilibrium as a signature of selective sweeps. *GENETICS*, 167(3):1513–1524, 2004. 23, 27, 71, 98, 147

[93] Rasmus Nielsen, Scott Williamson, Yuseob Kim, Melissa J Hubisz, Andrew G Clark, and Carlos Bustamante. Genomic scans for selective sweeps using SNP data. *GENOME RES*, 15(11):1566–1575, November 2005. 24, 63

[94] Lan Zhu and Carlos D Bustamante. A composite-likelihood approach for detecting directional selection from DNA sequence data. *GENETICS*, 170(3):1411–1421, 2005. 24

[95] P Pavlidis, S Hutter, and W Stephan. A population genomic approach to map recent positive selection in model species. *MOL ECOL*, 17(16):3585–3598, 2008. 24

[96] Guy Sella, Dmitri A Petrov, Molly Przeworski, and Peter Andolfatto. Pervasive natural selection in the Drosophila genome? *PLOS GENETIC*, 5(6):e1000495, June 2009. 24, 95, 114

[97] K R Thornton, J D Jensen, C Becquet, and P Andolfatto. Progress and prospects in mapping recent selection in the genome. *HEREDITY*, 98(6):340–348, June 2007. 24

[98] Jeffrey D Jensen, Kevin R Thornton, and Charles F Aquadro. Inferring Selection in Partially Sequenced Regions. *MOL BIOL EVOL*, 25(2):438–446, February 2008. 24

[99] J F C Kingman. On the Genealogy of Large Populations. *Journal of Applied Probability*, 19:27, 1982. 25, 150

[100] R R Hudson. Properties of a neutral allele model with intragenic recombination. *THEOR POPUL BIOL*, 23(2):183–201, April 1983. 25, 153

[101] Normal L Kaplan, Thomas Darden, and Richard R Hudson. The Coalescent Process in Models With Selection. *GENETICS*, 120:819–829, 1988. 25

[102] Richard R Hudson and Normal L Kaplan. The Coalescent Process in Models With Selection and Recombination. *GENETICS*, 120:831–840, 1988. 25

[103] R R Hudson and N L Kaplan. Deleterious Background Selection with recombination. *GENETICS*, 141:1605–1617, 1995. 25

[104] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian computation in population genetics. *GENETICS*, 162(4):2025–2035, December 2002. 25, 26

[105] J K Pritchard, M T Seielstad, A Perez-Lezaun, and M W Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *MOL BIOL EVOL*, 16(12):1791–1798, December 1999. 26, 114, 149

[106] Molly Przeworski. Estimating the time since the fixation of a beneficial allele. *GENETICS*, 164(4):1667–1676, 2003. 26

[107] Jeffrey D Jensen, Kevin R Thornton, and Peter Andolfatto. An approximate bayesian estimator suggests strong, recurrent selective sweeps in Drosophila. *PLOS GENETIC*, 4(9):e1000198, 2008. 26

[108] Nandita R. Garud, Philipp W. Messer, Erkan O. Buzbas, and Dmitri A. Petrov. Recent selective sweeps in north american drosophila melanogaster show signatures of soft sweeps. *PLoS GENET*, 11(2):e1005004, 02 2015. doi: 10.1371/journal.pgen.1005004. URL `http://dx.doi.org/10.1371%2Fjournal.pgen.1005004`. 26, 62, 134

[109] Pavlos Pavlidis, Jeffrey D Jensen, and Wolfgang Stephan. Searching for footprints of positive selection in whole-genome SNP data from nonequilibrium populations. *GENETICS*, 185(3):907–922, July 2010. 26, 27, 63, 64, 90, 98

[110] K Lin, H Li, C. Schlotterer, and A. Futschik. Distinguishing Positive Selection From Neutral Evolution: Boosting the Performance of Summary Statistics. *GENETICS*, 187(1):229–244, January 2011. 26

[111] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006. ISBN 0387310738. 27, 102

[112] M.S. Bartlett, Gwen Littlewort, M. Frank, C. Lainscsek, Ian Fasel, and J. Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. 2:568–573 vol. 2, June 2005. ISSN 1063-6919. doi: 10.1109/CVPR.2005.297. 27

[113] N. Benchettara, R. Kanawati, and C. Rouveirol. Supervised machine learning applied to link prediction in bipartite social networks. pages 326–330, Aug 2010. doi: 10.1109/ASONAM.2010.87. 27

[114] Francisco Pereira, Tom Mitchell, and Matthew Botvinick. Machine learning classifiers and fmri: A tutorial overview. *NeuroImage*, 45(1, Supplement 1):S199 – S209, 2009. ISSN 1053-8119. doi: http://dx.doi.org/10.1016/j.neuroimage.2008.11.007. URL `http://www.sciencedirect.com/science/article/pii/S1053811908012263`. Mathematics in Brain Imaging. 27

[115] N Alachiotis, A Stamatakis, and P Pavlidis. OmegaPlus: a scalable tool for rapid detection of selective sweeps in whole-genome datasets. *BIOINFORMATICS*, 28(17):2274–2275, September 2012. 28

[116] Pavlos Pavlidis, Daniel Zivkovic, Alexandros Stamatakis, and Nikolaos Alachiotis. SweeD: Likelihood-based detection of selective sweeps in thousands of genomes. *MOL BIOL EVOL*, January 2013. 28

[117] Richard M Durbin, David L Altshuler, Gonçalo R Abecasis, David R Bentley, Aravinda Chakravarti, Andrew G Clark, Francis S Collins, Francisco M De La Vega, Peter Donnelly, Michael Egholm, Paul Flicek, Stacey B Gabriel, Richard A Gibbs, Bartha M Knoppers, Eric S Lander, Hans Lehrach, Elaine R Mardis, Gil A McVean, Debbie A Nickerson, Leena Peltonen, Alan J Schafer, Stephen T Sherry, Jun Wang, Richard K Wilson, David Deiros, Mike Metzker, Donna Muzny, Jeff Reid, David Wheeler, Jingxiang Li, Min Jian, Guoqing Li, Ruiqiang Li, Huiqing Liang, Geng Tian, Bo Wang, Jian Wang, Wei Wang, Huanming Yang, Xiuqing Zhang, Huisong Zheng, Lauren Ambrogio, Toby Bloom, Kristian Cibulskis, Tim J Fennell, David B Jaffe, Erica Shefler, Carrie L Sougnez, Niall Gormley, Sean Humphray, Zoya Kingsbury, Paula Koko-Gonzales, Jennifer Stone, Kevin J McKernan, Gina L Costa, Jeffry K Ichikawa, Clarence C Lee, Ralf Sudbrak, Tatiana A Borodina, Andreas

Dahl, Alexey N Davydov, Peter Marquardt, Florian Mertes, Wilfiried Nietfeld, Philip Rosenstiel, Stefan Schreiber, Aleksey V Soldatov, Bernd Timmermann, Marius Tolzmann, Jason Affourtit, Dana Ashworth, Said Attiya, Melissa Bachorski, Eli Buglione, Adam Burke, Amanda Caprio, Christopher Celone, Shauna Clark, David Conners, Brian Desany, Lisa Gu, Lorri Guccione, Kalvin Kao, Andrew Kebbel, Jennifer Knowlton, Matthew Labrecque, Louise McDade, Craig Mealmaker, Melissa Minderman, Anne Nawrocki, Faheem Niazi, Kristen Pareja, Ravi Ramenani, David Riches, Wanmin Song, Cynthia Turcotte, Shally Wang, David Dooling, Lucinda Fulton, Robert Fulton, George Weinstock, John Burton, David M Carter, Carol Churcher, Alison Coffey, Anthony Cox, Aarno Palotie, Michael Quail, Tom Skelly, James Stalker, Harold P Swerdlow, Daniel Turner, Anniek De Witte, Shane Giles, Matthew Bainbridge, Danny Challis, Aniko Sabo, Fuli Yu, Jin Yu, Xiaodong Fang, Xiaosen Guo, Yingrui Li, Ruibang Luo, Shuaishuai Tai, Honglong Wu, Hancheng Zheng, Xiaole Zheng, Yan Zhou, Gabor T Marth, Erik P Garrison, Weichun Huang, Amit Indap, Deniz Kural, Wan-Ping Lee, Wen Fung Leong, Aaron R Quinlan, Chip Stewart, Michael P Stromberg, Alistair N Ward, Jiantao Wu, Charles Lee, Ryan E Mills, Xinghua Shi, Mark J Daly, Mark A DePristo, Aaron D Ball, Eric Banks, Brian L Browning, Kiran V Garimella, Sharon R Grossman, Robert E Handsaker, Matt Hanna, Chris Hartl, Andrew M Kernytsky, Joshua M Korn, Heng Li, Jared R Maguire, Steven A McCarroll, Aaron McKenna, James C Nemesh, Anthony A Philippakis, Ryan E Poplin, Alkes Price, Manuel A Rivas, Pardis C Sabeti, Stephen F Schaffner, Ilya A Shlyakhter, David N Cooper, Edward V Ball, Matthew Mort, Andrew D Phillips, Peter D Stenson, Jonathan Sebat, Vladimir Makarov, Kenny Ye, Seungtai C Yoon, Carlos D Bustamante, Adam Boyko, Jeremiah Degenhardt, Simon Gravel, Ryan N Gutenkunst, Mark Kaganovich, Alon Keinan, Phil Lacroute, Xin Ma, Andy Reynolds, Laura Clarke, Fiona Cunningham, Javier Herrero, Stephen Keenen, Eugene Kulesha, Rasko Leinonen, William M McLaren, Rajesh Radhakrishnan, Richard E Smith, Vadim Zalunin, Xiangqun Zheng-Bradley, Jan O Korbel, Adrian M Stütz, Markus Bauer, R Keira Cheetham, Tony Cox, Michael Eberle, Terena James, Scott Kahn, Lisa Murray, Kai Ye, Yutao Fu, Fiona C L Hyland, Jonathan M Manning, Stephen F McLaughlin, Heather E Peckham, Onur Sakarya, Yongming A Sun, Eric F Tsung, Mark A Batzer, Miriam K Konkel, Jerilyn A Walker, Marcus W Albrecht, Vyacheslav S Amstislavskiy, Ralf Herwig, Dimitri V Parkhomchuk, Richa Agarwala, Hoda M Khouri, Aleksandr O Morgulis, Justin E Paschall, Lon D Phan, Kirill E Rotmistrovsky, Robert D Sanders, Martin F Shumway, Chunlin Xiao, Adam Auton, Zamin Iqbal, Gerton Lunter, Jonathan L Marchini, Loukas Moutsianas, Simon Myers, Afidalina Tumian, James Knight, Roger Winer, and Craig... A map of human genome variation from population-scale sequencing. *NATURE*, 467(7319):1061–1073, 2010. 29

[118] Hua Chen, Nick Patterson, and David Reich. Population differentiation as a test for selective sweeps. *Genome Res.*, 20(3):393–402, March 2010. 29

[119] Joseph K Pickrell, Graham Coop, John Novembre, Sridhar Kudaravalli, Jun Z Li, Devin Absher, Balaji S Srinivasan, Gregory S Barsh, Richard M Myers, Marcus W Feldman, and Jonathan K Pritchard. Signals of recent positive selection in a worldwide sample of human populations. *Genome Res.*, 19(5):826–837, May 2009. 29

[120] International HapMap Consortium, Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, Shiran Pasternak, David A Wheeler, Thomas D Willis, Fuli Yu, Huanming Yang, Changqing Zeng, Yang Gao, Haoran Hu, Weitao Hu, Chaohua Li, Wei Lin, Siqi Liu, Hao Pan, Xiaoli Tang, Jian Wang, Wei Wang, Jun Yu, Bo Zhang, Qingrun Zhang, Hongbin Zhao, Hui Zhao, Jun Zhou, Stacey B Gabriel, Rachel Barry, Brendan Blumenstiel, Amy Camargo, Matthew Defelice, Maura Faggart, Mary Goyette, Supriya Gupta, Jamie Moore, Huy Nguyen, Robert C Onofrio, Melissa Parkin, Jessica Roy, Erich Stahl, Ellen Winchester, Liuda Ziaugra, David Altshuler, Yan Shen, Zhijian Yao, Wei Huang, Xun Chu, Yungang He, Li Jin, Yangfan Liu, Yayun Shen, Weiwei Sun, Haifeng Wang, Yi Wang, Ying Wang, Xiaoyan Xiong, Liang Xu, Mary M Y Waye, Stephen K W Tsui, Hong Xue, J Tze-Fei Wong, Luana M Galver, Jian-Bing Fan, Kevin Gunderson, Sarah S Murray, Arnold R Oliphant, Mark S Chee, Alexandre Montpetit, Fanny Chagnon, Vincent Ferretti, Martin Leboeuf, Jean-François Olivier, Michael S Phillips, Stéphanie Roumy, Clémentine Sallée, Andrei Verner, Thomas J Hudson, Pui-Yan Kwok, Dongmei Cai, Daniel C Koboldt, Raymond D Miller, Ludmila Pawlikowska, Patricia Taillon-Miller, Ming Xiao, Lap-Chee Tsui, William Mak, You Qiang Song, Paul K H Tam, Yusuke Nakamura, Takahisa Kawaguchi, Takuya Kitamoto, Takashi Morizono, Atsushi Nagashima, Yozo Ohnishi, Akihiro Sekine, Toshihiro Tanaka, Tatsuhiko Tsunoda, Panos Deloukas, Christine P Bird, Marcos Delgado, Emmanouil T Dermitzakis, Rhian Gwilliam, Sarah Hunt, Jonathan Morrison, Don Powell, Barbara E Stranger, Pamela Whittaker, David R Bentley, Mark J Daly, Paul I W de Bakker, Jeff Barrett, Yves R Chretien, Julian Maller, Steve McCarroll, Nick Patterson, Itsik Pe'er, Alkes Price, Shaun Purcell, Daniel J Richter, Pardis Sabeti, Richa Saxena, Stephen F Schaffner, Pak C Sham, Patrick Varilly, Lincoln D Stein, Lalitha Krishnan, Albert Vernon Smith, Marcela K Tello-Ruiz, Gudmundur A Thorisson, Aravinda Chakravarti, Peter E Chen, David J

Cutler, Carl S Kashuk, Shin Lin, Gonçalo R Abecasis, Weihua Guan, Yun Li, Heather M Munro, Zhaohui Steve Qin, Daryl J Thomas, Gilean McVean, Adam Auton, Leonardo Bottolo, Niall Cardin, Susana Eyheramendy, Colin Freeman, Jonathan Marchini, Simon Myers, Chris Spencer, Matthew Stephens, Peter Donnelly, Lon R Cardon, Geraldine Clarke, David M Evans, Andrew P Morris, Bruce S Weir, James C Mullikin, Stephen T Sherry, Michael Feolo, Andrew Skol, Houcan Zhang, Ichiro Matsuda, Yoshimitsu Fukushima, Darryl R Macer, Eiko Suda, Charles N Rotimi, Clement A Adebamowo, Ike Ajayi, Toyin Aniagwu, Patricia A Marshall, Chibuzor Nkwodimmah, Charmaine D M Royal, Mark F Leppert, Missy Dixon, Andy Peiffer, Renzong Qiu, Alastair Kent, Kazuto Kato, Norio Niikawa, Isaac F Adewole, Bartha M Knoppers, Morris W Foster, Ellen Wright Clayton, Jessica Watkin, Donna Muzny, Lynne Nazareth, Erica Sodergren, George M Weinstock, Imtaz Yakub, Bruce W Birren, Richard K Wilson, Lucinda L Fulton, Jane Rogers, John Burton, Nigel P Carter, Christopher M Clee, Mark Griffiths, Matthew C Jones, Kirsten McLay, Robert W Plumb, Mark T Ross, Sarah K Sims, David L Willey, Zhu Chen, Hua Han, Le Kang, Martin Godbout, John C Wallenburg, Paul L'Archevêque, Guy Bellemare, Koji Saeki, Hongguang Wang, Daochang An, Hongbo Fu, Qing Li, Zhen Wang, Renwu Wang, Arthur L Holden, Lisa D Brooks, Jean E McEwen, Mark S Guyer, Vivian Ota Wang, Jane L Peterson, Michael Shi, Jack Spiegel, Lawrence M Sung, Lynn F Zacharia, Francis S Collins, Karen Kennedy, Ruth Jamieson, and John Stewart. A second generation human haplotype map of over 3.1 million SNPs. *NATURE*, 449(7164):851–861, October 2007. 29

[121] Pleuni S. Pennings and Joachim Hermisson. Soft sweeps III: the signature of positive selection from recurrent mutation. *PLOS GENETIC*, 2(12):e186, December 2006. 30, 94

[122] Pleuni S. Pennings and Joachim Hermisson. Soft sweeps II–molecular population genetics of adaptation from recurrent mutation or migration. *MOL BIOL EVOL*, 23(5):1076–1084, May 2006. 30, 134

[123] Benjamin M Peter, Emilia Huerta-Sanchez, and Rasmus Nielsen. Distinguishing between selective sweeps from standing variation and from a de novo mutation. *PLoS GENETIC*, 8(10):e1003011, 2012. 30, 62, 63

[124] Katy L Simonsen, Gary A Churchill, and Charles F Aquadro. Properties of Statistical Tests of Neutrality for DNA Polymorphism Data. *GENETICS*, 141:413–429, 1995. 31, 63, 72, 90, 114

[125] Jeffrey D Jensen, Yuseob Kim, Vanessa Bauer DuMont, Charles F Aquadro, and Carlos D Bustamante. Distinguishing between selective

sweeps and demography using dna polymorphism data. *GENETICS*, 170(3):1401–10, Jul 2005. doi: 10.1534/genetics.104.038224. 31, 63, 72, 90

[126] N H Barton. The effect of hitch-hiking on neutral genealogies. *Genetical Reseach Cambridge*, 72:123–133, 1998. 31

[127] Joseph H Greenberg, Christy G Turner, Stephen L Zegura, Lyle Campbell, James A Fox, W S Laughlin, Em xf6 ke J E Szathmary, Kenneth M Weiss, and Ellen Woolford. The Settlement of the Americas: A Comparison of the Linguistic, Dental, and Genetic Evidence [and Comments and Reply]. *Current Anthropology*, 27(5):477–497, 1986. 31

[128] Jeffrey D Wall, Peter Andolfatto, and Molly Przeworski. Testing models of selection and demography in Drosophila simulans. *GENETICS*, 162(1):203–216, 2002. 31

[129] Alison M Adams and Richard R Hudson. Maximum-Likelihood Estimation of Demographic Parameters Using the Frequency Spectrum of Unlinked Single-Nucleotide Polymorphisms. *GENETICS*, 168(3):1699–1712, November 2004. 31, 33

[130] J Hey. On the number of New World founders: a population genetic portrait of the peopling of the Americas. *PLOS BIOL*, 2005. 31

[131] A J Drummond, A Rambaut, B Shapiro, and O G Pybus. Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences. *MOL BIOL EVOL*, 22(5):1185–1192, January 2005. 31

[132] JE Stajich and MW Hahn. Disentangling the effects of demography and selection in human history. *MOL BIOL EVOL*, 22(1):63–73, January 2005. 31

[133] Ryan N Gutenkunst, Ryan D Hernandez, Scott H Williamson, and Carlos D Bustamante. Inferring the Joint Demographic History of Multiple Populations from Multidimensional SNP Frequency Data. *PLOS GENETIC*, 5(10):e1000695 EP –, October 2009. 31, 33, 65, 67, 99, 113, 114, 115, 137

[134] Pablo Duchen, Daniel Zivkovic, Stephan Hutter, Wolfgang Stephan, and Stefan Laurent. Demographic Inference Reveals African and European Admixture in the North American Drosophila melanogaster Population. *GENETICS*, 193(1):291–301, January 2013. 31, 35, 65, 67, 68, 99, 125, 133

[135] A R Rogers and H Harpending. Population growth makes waves in the distribution of pairwise genetic differences. *MOL BIOL EVOL*, 9(3):552–569, May 1992. 31

[136] John Wakeley and Jody Hey. Estimating Ancestral Population Parameters. *GENETICS*, 145(3):847–855, January 1997. 31

[137] Naoyuki Takahata and Masatoshi Nei. GENE GENEALOGY AND VARIANCE OF INTERPOPULATIONAL NUCLEOTIDE DIFFERENCES. *GENETICS*, 110(2):325–344, June 1985. 32

[138] Joseph Felsenstein. Estimating effective population size from samples of sequences: a bootstrap Monte Carlo integration method. *GENET RES*, 60(03):209–220, December 1992. 32

[139] Mary K Kuhner, Jon Yamato, and Joseph Felsenstein. Maximum Likelihood Estimation of Population Growth Rates Based on the Coalescent. *GENETICS*, 149(1):429–434, May 1998. 32

[140] R Nielsen. Estimation of Population Parameters and Recombination Rates From Single Nucleotide Polymorphisms. *GENETICS*, 2000. 32

[141] Anna Pluzhnikov, Anna Di Rienzo, and Richard R Hudson. Inferences About Human Demography Based on Multilocus Analyses of Noncoding Sequences. *GENETICS*, 161(3):1209–1218, July 2002. 32

[142] Y X Fu and W H Li. Statistical tests of neutrality of mutations. *GENETICS*, 133(3):693–709, March 1993. 32

[143] Gunter Weiss and Arndt von Haeseler. Inference of Population History Using a Likelihood Approach. *GENETICS*, 149(3):1539–1546, July 1998. 32, 33

[144] L Frisse, R R Hudson, A Bartoszewicz, J D Wall, J Donfack, and A Di Rienzo. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet*, 69(4):831–843, October 2001. 33

[145] Benjamin F Voight, Alison M Adams, Linda A Frisse, Yudong Qian, Richard R Hudson, and Anna Di Rienzo. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *P NATL ACAD SCI USA*, 102(51):18508–18513, December 2005. 33

[146] Vincent Plagnol and Jeffrey D Wall. Possible Ancestral Structure in Human Populations. *PLOS GENETIC*, 2(7):e105, July 2006. 33

[147] Joshua G Schraiber and Joshua M Akey. Methods and models for unravelling human evolutionary history. *NAT REV GENET*, 16(12):727–740, December 2015. 33

[148] A. Gelman. *Bayesian data analysis*. CRC press, 2004. 34

[149] Mark A Beaumont, Wenyang Zhang, and David J Balding. Approximate Bayesian Computation in Population Genetics. *GENETICS*, 162(4):2025–2035, December 2002. 34, 114, 140, 149

[150] Paul Marjoram, John Molitor, Vincent Plagnol, and Simon Tavare. Markov chain Monte Carlo without likelihoods. *P NATL ACAD SCI USA*, 100(26):15324–15328, December 2003. 34

[151] Daniel Wegmann, Christoph Leuenberger, and Laurent Excoffier. Efficient approximate Bayesian computation coupled with Markov chain Monte Carlo without likelihood. *GENETICS*, 182(4):1207–1218, August 2009. 34

[152] Mattias Jakobsson Sen Li. Estimating demographic parameters from large-scale population genomic data using Approximate Bayesian Computation. *BMC Genet*, 13(1):22, 2012. 34

[153] Kevin Thornton and Peter Andolfatto. Approximate Bayesian Inference Reveals Evidence for a Recent, Severe Bottleneck in a Netherlands Population of Drosophila melanogaster. *GENETICS*, 172(3):1607–1619, January 2006. 35, 133, 149

[154] Aaron B A Shafer, Lucie M Gattepaille, Robert E A Stewart, and Jochen B W Wolf. Demographic inferences using short-read genomic data in an approximate Bayesian computation framework: in silico evaluation of power, biases and proof of concept in Atlantic walrus. *MOL ECOL*, 24(2):328–345, January 2015. 35

[155] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *NATURE*, 475(7357):493–496, July 2011. 35, 39, 40, 114, 123, 132, 133

[156] Nelson J. R. Fagundes, Nicolas Ray, Mark Beaumont, Samuel Neuenschwander, Francisco M. Salzano, Sandro L. Bonatto, and Laurent Excoffier. Statistical evaluation of alternative models of human evolution. *P NATL ACAD SCI USA*, 104(45):17614–17619, January 2007. 36, 113, 133

[157] Alon Keinan, James C Mullikin, Nick Patterson, and David Reich. Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. *NAT GENET*, 39(10):1251–1255, September 2007. 36

[158] S F Schaffner. Calibrating a coalescent simulation of human genome sequence variation. *GENOME RES*, 15(11):1576–1583, November 2005. 36

[159] D Garrigan, S B Kingan, M M Pilkington, J A Wilder, M P Cox, H Soodyall, B Strassmann, G Destro-Bisol, P de Knijff, A Novelletto, J Friedlaender, and M F Hammer. Inferring Human Population Sizes, Divergence Times and Rates of Gene Flow From Mitochondrial, X and Y Chromosome Resequencing Data. *GENETICS*, 177(4):2195–2207, December 2007. 36

[160] Vincent Plagnol and Jeffrey D Wall. Possible Ancestral Structure in Human Populations. *PLOS GENETIC*, 2(7):e105, July 2006. 36

[161] Gilean A T McVean and Niall J Cardin. Approximating the coalescent with recombination. *The Royal Society of London. Biological sciences*, 360(1459):1387–1393, July 2005. 36, 114, 157

[162] Jeffrey D Wall and Michael F Hammer. Archaic admixture in the human genome. *Genomes and evolution*, 16(6):606–610, December 2006. 39

[163] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *NAT GENET*, 46(8):919–925, June 2014. 40, 132

[164] Sabrina Simon, Sascha Strauss, Arndt von Haeseler, and Heike Hadrys. A Phylogenomic Approach to Resolve the Basal Pterygote Divergence. *MOL BIOL EVOL*, 26(12):2719–2730, January 2009. 43

[165] A. Bourret, M A McPeek, and J Turgeon. Regional divergence and mosaic spatial distribution of two closely related damselfly species (Enallagma hageni and Enallagma ebrium). *Journal of Evolutionary Biology*, 25(1):196–209, November 2011. 43

[166] Julie Turgeon, Robby Stoks, Ryan A Thum, Jonathan M Brown, and Mark A. McPeek. Simultaneous Quaternary Radiations of Three Damselfly Clades across the Holarctic. *The AM NAT*, 165(4):E78–E107, April 2005. 43, 53

[167] David Outomuro, Folmer Bokma, and Frank Johansson. Hind Wing Shape Evolves Faster than Front Wing Shape in Calopteryx Damselflies. *Evol Biol*, 39(1):116–125, 2012. 43

[168] J K Abbott, S Bensch, T P Gosden, and E I Svensson. Patterns of differentiation in a colour polymorphism and in neutral markers reveal rapid genetic changes in natural damselfly populations. *MOL ECOL*, 17(6):1597–1604, March 2008. 43

[169] A. Iserbyt, J. Bots, H. Van Gossum, and K. Jordaens. Did historical events shape current geographic variation in morph frequencies of a polymorphic damselfly? *Journal of Zoology*, 282(4):256–265, December 2010. 43

[170] Gary G. Mittelbach, Douglas W. Schemske, Howard V. Cornell, Andrew P. Allen, Jonathan M Brown, Mark B. Bush, Susan P. Harrison, Allen H. Hurlbert, Nancy Knowlton, Harilaos A. Lessios, Christy M. McCain, Amy R. McCune, Lucinda A. McDade, Mark A. McPeek, Thomas J. Near, Trevor D. Price, Robert E. Ricklefs, Kaustuv Roy, Dov F. Sax, Dolph Schluter, James M. Sobel, and Michael Turelli. Evolution and the latitudinal diversity gradient: speciation, extinction and biogeography. *Ecol Letters*, 10(4):315–331, April 2007. 43

[171] Stefanie Slos, Luc D. Meester, and Robby Stoks. Behavioural activity levels and expression of stress proteins under predation risk in two damselfly species. *Ecological Entomology*, 34(3):297–303, June 2009. 43

[172] Francis Strobbe, Mark A. McPeek, Marjan Block, and Robby Stoks. Fish predation selects for reduced foraging activity. *Behav Ecol Sociobiol*, 65(2):241–247, August 2010. 43

[173] Diana Bellin, Alberto Ferrarini, Antonio Chimento, Olaf Kaiser, Natasha Levenkova, Pascal Bouffard, and Massimo Delledonne. Combining next-generation pyrosequencing with microarray for large scale expression analysis in non-model species. *BMC Genomics 2009 10:219*, 10(1):555, 2009. 43

[174] Y Surget-Groba and J I Montoya-Burgos. Optimization of de novo transcriptome assembly from next-generation sequencing data. *GENOME RES*, 20(10):1432–1440, October 2010. 43

[175] Tal Nawy. Non–model organisms. *Nat Meth*, 9(1):37–37, December 2011. 43

[176] J Turgeon and M A McPeek. Phylogeographic analysis of a recent radiation of Enallagma damselflies (Odonata: Coenagrionidae). *MOL ECOL*, 11(10):1989–2001, October 2002. 43

[177] Corrie Saux, Chris Simon, and Greg S Spicer. Phylogeny of the Dragonfly and Damselfly Order Odonata as Inferred by Mitochondrial 12S Ribosomal RNA Sequences. *Annals of the Entomological Society of America*, 96(6):693–699, November 2003. 43

[178] Henri J Dumont, Andy Vierstraete, and Jacques R Vanfleteren. A molecular phylogeny of the Odonata (Insecta). *Systematic Entomology*, 35(1):6–18, August 2009. 43

[179] Karen Meusemann, Björn M von Reumont, Sabrina Simon, Falko Roeding, Sascha Strauss, Patrick Kück, Ingo Ebersberger, Manfred Walzl, Günther Pass, Sebastian Breuers, Viktor Achter, Arndt von Haeseler,

Thorsten Burmester, Heike Hadrys, J Wolfgang Wägele, and Bernhard Misof. A phylogenomic approach to resolve the arthropod tree of life. *MOL BIOL EVOL*, 27(11):2451–2464, November 2010. 51

[180] Keisuke Ishiwata, Go Sasaki, Jiro Ogawa, Takashi Miyata, and Zhi-Hui Su. Phylogenetic relationships among insect orders based on three nuclear protein-coding gene sequences. *Molecular Phylogenetics and Evolution*, 58(2):169–180, February 2011. 51

[181] Michelle D Trautwein, Brian M Wiegmann, Rolf Beutel, Karl M Kjer, and David K Yeates. Advances in Insect Phylogeny at the Dawn of the Postgenomic Era. *Annu Rev Entomol*, 57(1):449–468, January 2012. 51

[182] Mark A. McPeek, Ann K. Schrot, and Jonathan M. T3 Brown. Adaptation to Predators in a New Community: Swimming Performance and Predator Avoidance in Damselflies. *Ecology*, 77(2):617–629, January 1996. 52

[183] Mark A. McPeek. Biochemical Evolution Associated with Antipredator Adaptation in Damselflies. *EVOLUTION*, 53(6):1835–1845, January 1999. 52

[184] M A McPeek. Predisposed to adapt? Clade-level differences in characters affecting swimming performance in damselflies. *EVOLUTION*, 54(6):2072–2080, December 2000. 52

[185] Olivier Harismendy, Pauline C Ng, Robert L Strausberg, Xiaoyun Wang, Timothy B Stockwell, Karen Y Beeson, Nicholas J Schork, Sarah S Murray, Eric J Topol, Samuel Levy, and Kelly A Frazer. Evaluation of next generation sequencing platforms for population targeted sequencing studies. *GENOME BIOL*, 10(3):R32, 2009. 52

[186] Evandro Novaes, Derek R Drost, William G Farmerie, Georgios J Pappas, Dario Grattapaglia, Ronald R Sederoff, and Matias Kirst. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. *BMC Genomics 2009 10:219*, 9(1):312, 2008. 52

[187] Eli Meyer, Galina Aglyamova, Shi Wang, Jade Buchanan-Carter, David Abrego, John Colbourne, Bette Willis, and Mikhail Matz. Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics 2009 10:219*, 10(1):219, May 2009. 52

[188] Rasmus Wernersson. Virtual Ribosome–a comprehensive DNA translation tool with support for integration of sequence feature annotation. *NUCLEIC ACIDS RES*, 34(Web Server issue):W385–8, July 2006. 54

[189] Gabriel Moreno-Hagelsieb and Kristen Latimer. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *BIOINFORMATICS*, 24(3):319–324, February 2008. 55

[190] Toni Gabaldón. Large-scale assignment of orthology: back to phylogenetics? *GENOME BIOL*, 9(10):235, 2008. 55

[191] Li Li, Christian J Stoeckert, and David S Roos. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *GENOME RES*, 13(9):2178–2189, September 2003. 55

[192] A Darling and L Carey. The design, implementation, and evaluation of mpiBLAST. In *Proceedings of ClusterWorld*. 2003. 56

[193] M A Larkin, G Blackshields, N P Brown, R Chenna, P A McGettigan, H McWilliam, F Valentin, I M Wallace, A Wilm, R Lopez, J D Thompson, T J Gibson, and D G Higgins. Clustal W and Clustal X version 2.0. *BIOINFORMATICS*, 23(21):2947–2948, November 2007. 56

[194] Sudhindra R. Gadagkar, Michael S. Rosenberg, and Sudhir Kumar. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. *Journal of experimental zoology. Part B, Molecular and developmental evolution*, 304B(1):64–74, 2005. 56

[195] Gerard Talavera and Jose Castresana. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. *SYST BIOL*, 56(4):564–577, August 2007. 56

[196] Diego Darriba, Guillermo L Taboada, Ramón Doallo, and David Posada. ProtTest 3: fast selection of best-fit models of protein evolution. *BIOINFORMATICS*, 27(8):1164–1165, April 2011. 56

[197] Federico Abascal, Rafael Zardoya, and David Posada. ProtTest: selection of best-fit models of protein evolution. *BIOINFORMATICS*, 21(9):2104–2105, May 2005. 56

[198] M.O. Dayhoff and R.M. Schwartz. Chapter 22: A model of evolutionary change in proteins. 1978. 57

[199] David T Jones, William R Taylor, and Janet M Thornton. The rapid generation of mutation data matrices from protein sequences. *BIOINFORMATICS*, 8(3):275–282, 1992. 57

[200] Simon Whelan and Nick Goldman. A General Empirical Model of Protein Evolution Derived from Multiple Protein Families Using a Maximum-Likelihood Approach. *MOL BIOL EVOL*, 18(5):691–699, January 2001. 57

[201] M Hasegawa, H Kishino, and T Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J MOL EVOL*, 22(2):160–174, 1985. 57

[202] Y Cao, J Adachi, A Janke, S Paabo, and M Hasegawa. Phylogenetic-Relationships Among Eutherian Orders Estimated From Inferred Sequences of Mitochondrial Proteins - Instability of a Tree-Based on a Single-Gene. *J MOL EVOL*, 39(5):519–527, 1994. 57

[203] Tobias Müller and Martin Vingron. Modeling Amino Acid Replacement. *Journal of Computational Biology*, 7(6):761–776, December 2000. 57

[204] J ADACHI, PJ Waddell, W Martin, and M Hasegawa. Plastid genome phylogeny and a model of amino acid substitution for proteins encoded by chloroplast DNA. *J MOL EVOL*, 50(4):348–358, 2000. 57

[205] MW Dimmic, JS Rest, DP Mindell, and RA Goldstein. rtREV: An amino acid substitution matrix for inference of retrovirus and reverse transcriptase phylogeny. *J MOL EVOL*, 55(1):65–73, 2002. 57

[206] Federico Abascal, David Posada, and Rafael Zardoya. MtArt: A new model of amino acid replacement for arthropoda. *MOL BIOL EVOL*, 24(1):1–5, 2007. 57

[207] David C. Nickle, Laura Heath, Mark A. Jensen, Peter B. Gilbert, James I. Mullins, and Sergei L Kosakovsky Pond. HIV-Specific Probabilistic Models of Protein Evolution. *PLOS ONE*, 2(6):e503 EP —-, January 2007. 57

[208] Si Quang Le and Olivier Gascuel. An improved general amino acid replacement matrix. *MOL BIOL EVOL*, 25(7):1307–1320, 2008. 57

[209] S Henikoff. Amino acid substitution matrices from protein blocks. In *Proceedings of the National ...*. 1992. 57

[210] D Posada and KA Crandall. Selecting the best-fit model of nucleotide substitution. *SYST BIOL*, 50(4):580–601, 2001. 57

[211] Fredrik Ronquist and John P Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *BIOINFORMATICS*, 19(12):1572–1574, August 2003. 57

[212] John P Huelsenbeck, Fredrik Ronquist, Rasmus Nielsen, and Jonathan P Bollback. Bayesian Inference of Phylogeny and Its Impact on Evolutionary Biology. *SCIENCE*, 294(5550):2310–2314, January 2001. 57

[213] G Altekar, S Dwarkadas, JP Huelsenbeck, and F Ronquist. Parallel metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *BIOINFORMATICS*, 20(3):407–415, 2004. 57

[214] Ziheng Yang. PAML 4: phylogenetic analysis by maximum likelihood. *MOL BIOL EVOL*, 24(8):1586–1591, August 2007. 58

[215] Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological), Vol. 57, No. 1. (1995), pp. 289-300, doi:10.2307/2346101*, 57(1):289–300, 1995. 58, 109

[216] Ana Conesa, Stefan Götz, Juan Miguel García-Gómez, Javier Terol, Manuel Talón, and Montserrat Robles. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *BIOINFORMATICS*, 21(18):3674–3676, 2005. 59

[217] Seung Yon Rhee, Valerie Wood, Kara Dolinski, and Sorin Draghici. Use and misuse of the gene ontology annotations. *NAT REV GENET*, 9(7):509–515, May 2008. 59

[218] David Groppe. Benjamini & Hochberg/Yekutieli Procedure for Controlling False Discovery Rate - File Exchange - MATLAB Central. 59

[219] Yoav Benjamini and Daniel Yekutieli. The Control of the False Discovery Rate in Multiple Testing under Dependency. *The Annals of Statistics*, 29(4):1165–1188, January 2001. 59

[220] Jia Ye, Lin Fang, Hongkun Zheng, Yong Zhang, Jie Chen, Zengjin Zhang, Jing Wang, Shengting Li, Ruiqiang Li, Lars Bolund, and Jun Wang. WEGO: a web tool for plotting GO annotations. *NUCLEIC ACIDS RES*, 34(suppl 2):W293–W297, July 2006. 60

[221] Douglas S Falconer, Trudy FC Mackay, and Richard Frankham. Introduction to quantitative genetics (4th edn). *TRENDS GENET*, 12(7):280, 1996. 62

[222] H A Orr and A J Betancourt. Haldane's sieve and adaptation from the standing genetic variation. *GENETICS*, 157(2):875–884, February 2001. 62, 95, 134

[223] Molly Przeworski, Graham Coop, and Jeffrey D Wall. The signature of positive selection on standing genetic variation. *EVOLUTION*, 59(11):2312–2323, November 2005. 62, 101, 102, 134

[224] Ryan D Hernandez, Joanna L Kelley, Eyal Elyashiv, S Cord Melton, Adam Auton, Gilean McVean, Guy Sella, Molly Przeworski, et al. Classic selective sweeps were rare in recent human evolution. *SCIENCE*, 331(6019):920–924, 2011. 62

[225] Rowan DH Barrett and Dolph Schluter. Adaptation from standing genetic variation. *TRENDS ECOL EVOL*, 23(1):38–44, 2008. 62

[226] Trudy F C Mackay, Stephen Richards, Eric A Stone, Antonio Barbadilla, Julien F Ayroles, Dianhui Zhu, Sonia Casillas, Yi Han, Michael M Magwire, Julie M Cridland, Mark F Richardson, Robert R H Anholt, Maite Barrón, Crystal Bess, Kerstin Petra Blankenburg, Mary Anna Carbone, David Castellano, Lesley Chaboub, Laura Duncan, Zeke Harris, Mehwish Javaid, Joy Christina Jayaseelan, Shalini N Jhangiani, Katherine W Jordan, Fremiet Lara, Faye Lawrence, Sandra L Lee, Pablo Librado, Raquel S Linheiro, Richard F Lyman, Aaron J Mackey, Mala Munidasa, Donna Marie Muzny, Lynne Nazareth, Irene Newsham, Lora Perales, Ling-Ling Pu, Carson Qu, Miquel Ràmia, Jeffrey G Reid, Stephanie M Rollmann, Julio Rozas, Nehad Saada, Lavanya Turlapati, Kim C Worley, Yuan-Qing Wu, Akihiko Yamamoto, Yiming Zhu, Casey M Bergman, Kevin R Thornton, David Mittelman, and Richard A Gibbs. The Drosophila melanogaster Genetic Reference Panel. *NATURE*, 482(7384):173–178, February 2012. 64, 67, 81, 96, 132

[227] John E Pool, Russell B Corbett-Detig, Ryuichi P Sugino, Kristian A Stevens, Charis M Cardeno, Marc W Crepeau, Pablo Duchen, J J Emerson, Perot Saelao, David J. Begun, and Charles H Langley. Population Genomics of Sub-Saharan Drosophila melanogaster: African Diversity and Non-African Admixture. *PLOS GENETIC*, 8(12):e1003080, December 2012. 64, 68, 83, 96

[228] Justin B Lack, Charis M Cardeno, Marc W Crepeau, William Taylor, Russell B Corbett-Detig, Kristian A Stevens, Charles H Langley, and John E Pool. The Drosophila Genome Nexus: A Population Genomic Resource of 623 Drosophila melanogaster Genomes, Including 197 from a Single Ancestral Range Population. *GENETICS*, January 2015. 64, 67, 96, 97

[229] J R David and P Capy. Genetic variation of Drosophila melanogaster natural populations. *TRENDS GENET*, 4(4):106–111, April 1988. 64, 92

[230] Joyce Y Kao, Asif Zubair, Matthew P Salomon, Sergey V Nuzhdin, and Daniel Campo. Population genomic analysis uncovers african and european admixture in drosophila melanogaster populations from the southeastern united states and caribbean islands. *Molecular ecology*, 24(7):1499–1509, 2015. 67

[231] H Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, December 1974. 67, 99, 138

[232] D J Begun and C F Aquadro. African and North American populations of Drosophila melanogaster are very different at the DNA level. *NATURE*, 365(6446):548–550, 1993. 68

[233] John C. Platt, Nello Cristianini, and John Shawe-Taylor. Large margin dags for multiclass classification. pages 547–553, 2000. URL `http://papers.nips.cc/paper/1773-large-margin-dags-for-multiclass-classification.pdf`. 68

[234] J K Kelly. A test of neutrality based on interlocus associations. *GENETICS*, 146(3):1197–1206, July 1997. 71, 146

[235] Josep M Comeron, Ramesh Ratnappan, and Samuel Bailin. The Many Landscapes of Recombination in Drosophila melanogaster. *PLOS GENETIC*, 8(10):e1002905, October 2012. 81, 97, 107, 131

[236] P J Daborn, J L Yen, M R Bogwitz, G Le Goff, E Feil, S Jeffers, N Tijet, T Perry, D Heckel, P Batterham, R Feyereisen, T G Wilson, and R H ffrench Constant. A single p450 allele associated with insecticide resistance in Drosophila. *SCIENCE*, 297(5590):2253–2256, September 2002. 83

[237] Joshua M Schmidt, Robert T Good, Belinda Appleton, Jayne Sherrard, Greta C Raymant, Michael R Bogwitz, Jon Martin, Phillip J Daborn, Mike E Goddard, Philip Batterham, and Charles Robin. Copy Number Variation and Transposable Elements Feature in Recent, Ongoing Adaptation at the Cyp6g1 Locus. *PLOS GENETIC*, 6(6):e1000998, June 2010. 83

[238] Bryan Kolaczkowski, Andrew D. Kern, Alisha K Holloway, and David J. Begun. Genomic differentiation between temperate and tropical Australian populations of Drosophila melanogaster. *GENETICS*, 187(1):245–260, January 2011. 83, 97

[239] C H Langley, K Stevens, C Cardeno, Y C G Lee, D R Schrider, J E Pool, S A Langley, C Suarez, R B Corbett-Detig, B Kolaczkowski, S Fang, P M Nista, A K Holloway, A D Kern, C N Dewey, Y S Song, M W Hahn, and D J Begun. Genomic Variation in Natural Populations of Drosophila melanogaster. *GENETICS*, 192(2):533–598, October 2012. 83, 95, 132

[240] Sascha Glinka, David De Lorenzo, and Wolfgang Stephan. Evidence of Gene Conversion Associated with a Selective Sweep in Drosophila melanogaster. *MOL BIOL EVOL*, 2006. 83

[241] J E Pool. A Scan of Molecular Variation Leads to the Narrow Localization of a Selective Sweep Affecting Both Afrotropical and Cosmopolitan

Populations of Drosophila melanogaster. *GENETICS*, 172(2):1093–1105, November 2005. 83

[242] R Gasperini and G Gibson. Absence of protein polymorphism in the Ras genes of Drosophila melanogaster. *J MOL EVOL*, 49(5):583–590, November 1999. 83

[243] Michael E Katz and Frank McCormick. Signal transduction from multiple Ras effectors. *CURR OPIN GENET DEV*, 7(1):75–79, February 1997. 83

[244] M J Cann, E Chung, and L R Levin. A new family of adenylyl cyclase genes in the male germline of Drosophila melanogaster. *DEV GENES EVOL*, 210(4):200–206, March 2000. 83

[245] Mary F Durham, Michael M Magwire, Eric A Stone, and Jeff Leips. Genome-wide analysis in Drosophila reveals age-specific effects of SNPs on fitness traits. *NAT COMMUN*, 5 SP -, July 2014. 83

[246] Kevin Thornton and Manyuan Long. Rapid divergence of gene duplicates on the Drosophila melanogaster X chromosome. *MOL BIOL EVOL*, 19(6):918–925, June 2002. 85, 95

[247] V B DuMont. Multiple Signatures of Positive Selection Downstream of Notch on the X Chromosome in Drosophila melanogaster. *GENETICS*, 171(2):639–653, July 2005. 85

[248] Simon Boitard, Christian Schlötterer, Viola Nolte, Ram Vinay Pandey, and Andreas Futschik. Detecting Selective Sweeps from Pooled Next-Generation Sequencing Samples. *MOL BIOL EVOL*, 29(9):2177–2186, January 2012. 85

[249] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *NATURE*, 4(1):44–57, December 2008. 85, 109

[250] Da Wei Huang, Brad T Sherman, and Richard A Lempicki. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *NUCLEIC ACIDS RES*, 37(1):1–13, January 2009. 85, 109

[251] A L Hughes. Coevolution of the vertebrate integrin alpha- and beta-chain genes. *MOL BIOL EVOL*, 1992. 87

[252] Jianzhi Zhang. Rates of conservative and radical nonsynonymous nucleotide substitutions in mammalian nuclear genes. *J MOL EVOL*, 50(1):56–68, 2000. 87

[253] Thilakam Murali, Svetlana Pacifico, Jingkai Yu, Stephen Guest, George G 3rd Roberts, and Russell L Jr Finley. DroID 2011: a comprehensive, integrated resource for protein, transcription factor, RNA and gene interactions for Drosophila. *NUCLEIC ACIDS RES*, 39(Database issue):D736–43, January 2011. 88

[254] Jingkai Yu, Svetlana Pacifico, Guozhen Liu, and Russell L Jr Finley. DroID: the Drosophila Interactions Database, a comprehensive resource for annotated gene and protein interactions. *BMC GENOM*, 9:461, 2008. 88

[255] Susan Tweedie, Michael Ashburner, Kathleen Falls, Paul Leyland, Peter McQuilton, Steven Marygold, Gillian Millburn, David Osumi-Sutherland, Andrew Schroeder, Ruth Seal, and Haiyan Zhang. FlyBase: enhancing Drosophila Gene Ontology annotations. *NUCLEIC ACIDS RES*, 37(Database issue):D555–9, January 2009. 88

[256] Marc S Halfon, Steven M Gallo, and Casey M Bergman. REDfly 2.0: an integrated database of cis-regulatory modules and transcription factor binding sites in Drosophila. *NUCLEIC ACIDS RES*, 36(Database issue):D594–8, January 2008. 88

[257] Sushmita Roy, Jason Ernst, Peter V Kharchenko, Pouya Kheradpour, Nicolas Negre, Matthew L Eaton, Jane M Landolin, Christopher A Bristow, Lijia Ma, Michael F Lin, Stefan Washietl, Bradley I Arshinoff, Ferhat Ay, Patrick E Meyer, Nicolas Robine, Nicole L Washington, Luisa Di Stefano, Eugene Berezikov, Christopher D Brown, Rogerio Candeias, Joseph W Carlson, Adrian Carr, Irwin Jungreis, Daniel Marbach, Rachel Sealfon, Michael Y Tolstorukov, Sebastian Will, Artyom A Alekseyenko, Carlo Artieri, Benjamin W Booth, Angela N Brooks, Qi Dai, Carrie A Davis, Michael O Duff, Xin Feng, Andrey A Gorchakov, Tingting Gu, Jorja G Henikoff, Philipp Kapranov, Renhua Li, Heather K MacAlpine, John Malone, Aki Minoda, Jared Nordman, Katsutomo Okamura, Marc Perry, Sara K Powell, Nicole C Riddle, Akiko Sakai, Anastasia Samsonova, Jeremy E Sandler, Yuri B Schwartz, Noa Sher, Rebecca Spokony, David Sturgill, Marijke van Baren, Kenneth H Wan, Li Yang, Charles Yu, Elise Feingold, Peter Good, Mark Guyer, Rebecca Lowdon, Kami Ahmad, Justen Andrews, Bonnie Berger, Steven E Brenner, Michael R Brent, Lucy Cherbas, Sarah C R Elgin, Thomas R Gingeras, Robert Grossman, Roger A Hoskins, Thomas C Kaufman, William Kent, Mitzi I Kuroda, Terry Orr-Weaver, Norbert Perrimon, Vincenzo Pirrotta, James W Posakony, Bing Ren, Steven Russell, Peter Cherbas, Brenton R Graveley, Suzanna Lewis, Gos Micklem, Brian Oliver, Peter J Park, Susan E Celniker, Steven Henikoff, Gary H Karpen, Eric C Lai, David M

MacAlpine, Lincoln D Stein, Kevin P White, and Manolis Kellis. Identification of functional elements and regulatory circuits by Drosophila modENCODE. *SCIENCE*, 330(6012):1787–1797, December 2010. 88

[258] Benjamin P Lewis, Christopher B Burge, and David P Bartel. Conserved seed pairing, often flanked by adenosines, indicates that thousands of human genes are microRNA targets. *CELL*, 120(1):15–20, January 2005. 88

[259] J Graham Ruby, Alexander Stark, Wendy K Johnston, Manolis Kellis, David P Bartel, and Eric C Lai. Evolution, biogenesis, expression, and target predictions of a substantially expanded set of Drosophila microRNAs. *GENOME RES*, 17(12):1850–1864, December 2007. 88

[260] Michael Schnall-Levin, Yong Zhao, Norbert Perrimon, and Bonnie Berger. Conserved microRNA targeting in Drosophila is as widespread in coding regions as in 3'UTRs. *P NATL ACAD SCI USA*, 107(36):15751–15756, September 2010. 88

[261] L Giot, J S Bader, C Brouwer, A Chaudhuri, B Kuang, Y Li, Y L Hao, C E Ooi, B Godwin, E Vitols, G Vijayadamodar, P Pochart, H Machineni, M Welsh, Y Kong, B Zerhusen, R Malcolm, Z Varrone, A Collis, M Minto, S Burgess, L McDaniel, E Stimpson, F Spriggs, J Williams, K Neurath, N Ioime, M Agee, E Voss, K Furtak, R Renzulli, N Aanensen, S Carrolla, E Bickelhaupt, Y Lazovatsky, A DaSilva, J Zhong, C A Stanyon, R L Jr Finley, K P White, M Braverman, T Jarvie, S Gold, M Leach, J Knight, R A Shimkets, M P McKenna, J Chant, and J M Rothberg. A Protein Interaction Map Of Drosophila Melanogaster. *SCIENCE*, 302(5651):1727–1736, December 2003. 88

[262] K G Guruharsha, Jean-Francois Rual, Bo Zhai, Julian Mintseris, Pujita Vaidya, Namita Vaidya, Chapman Beekman, Christina Wong, David Y Rhee, Odise Cenaj, Emily McKillip, Saumini Shah, Mark Stapleton, Kenneth H Wan, Charles Yu, Bayan Parsa, Joseph W Carlson, Xiao Chen, Bhaveen Kapadia, K VijayRaghavan, Steven P Gygi, Susan E Celniker, Robert A Obar, and Spyros Artavanis-Tsakonas. A protein complex network of Drosophila melanogaster. *CELL*, 147(3):690–703, October 2011. 89

[263] Tamas Lukacsovich, Kazuya Yuge, Wakae Awano, Zoltan Asztalos, Shunzo Kondo, Naoto Juni, and Daisuke Yamamoto. The ken and barbie gene encoding a putative transcription factor with a BTB domain and three zinc finger motifs functions in terminalia development of Drosophila. *ARCH INSECT BIOCHEM PHYSIOL*, 54(2):77–94, October 2003. 89

[264] Benjamin Houot, Stéphane Fraichard, Ralph J Greenspan, and Jean-François Ferveur. Genes Involved in Sex Pheromone Discrimination in Drosophila melanogaster and Their Background-Dependent Effect. *PLOS ONE*, 7(1):e30799, January 2012. 89

[265] Dariel Ashton-Beaucage, Christian M Udell, Patrick Gendron, Malha Sahmi, Martin Lefrançois, Caroline Baril, Anne-Sophie Guenier, Jean Duchaine, Daniel Lamarre, Sébastien Lemieux, and Marc Therrien. A functional screen reveals an extensive layer of transcriptional and splicing control underlying RAS/MAPK signaling in Drosophila. *PLOS BIOL*, 12(3):e1001809, March 2014. 89

[266] Young-Jun Kim, Hong Bao, Liana Bonanno, Bing Zhang, and Mihaela Serpe. Drosophila Neto is essential for clustering glutamate receptors at the neuromuscular junction. *Genes Dev*, 26(9):974–987, May 2012. 89

[267] Kenneth Weber, Nancy Johnson, David Champlin, and April Patty. Many P-element insertions affect wing shape in Drosophila melanogaster. *GENETICS*, 169(3):1461–1475, March 2005. 89

[268] Bridget C Lear, Eric J Darrah, Benjamin T Aldrich, Senetibeb Gebre, Robert L Scott, Howard A Nash, and Ravi Allada. UNC79 and UNC80, putative auxiliary subunits of the NARROW ABDOMEN ion channel, are indispensable for robust circadian locomotor rhythms in Drosophila. *PLoS ONE*, 8(11):e78147, 2013. 89

[269] C Schütt and R Nöthiger. Structure, function and evolution of sex-determining systems in Dipteran insects. *Development*, 127(4):667–677, February 2000. 89

[270] X Zheng, R K Mann, N Sever, and P A Beachy. Genetic and biochemical definition of the Hedgehog receptor. - PubMed - NCBI. *Genes Dev*, 24(1):57–71, January 2010. 89

[271] Darius Camp, Ko Currie, Alain Labbé, Donald J van Meyel, and Frédéric Charron. Ihog and Boi are essential for Hedgehog signaling in Drosophila. *Neural Dev*, 5(1):28, 2010. 90

[272] Daniel Lachaise, Marie-Louise Cariou, Jean R David, Françoise Lemeunier, Léonidas Tsacas, and Michael Ashburner. Historical biogeography of the drosophila melanogaster species subgroup. In *Evolutionary biology*, pages 159–225. Springer, 1988. 92

[273] Stanley A Sawyer, John Parsch, Zhi Zhang, and Daniel L Hartl. Prevalence of positive selection among nearly neutral amino acid replacements in Drosophila. *P NATL ACAD SCI USA*, 104(16):6504–6510, April 2007. 92

[274] Peter D Keightley and Adam Eyre-Walker. Estimating the rate of adaptive molecular evolution when the evolutionary divergence between species is small. *J MOL EVOL*, 74(1-2):61–68, 2012. 92

[275] Peter Andolfatto. Adaptive evolution of non-coding DNA in Drosophila. *NATURE*, 437(7062):1149–1152, 2005. 92, 95

[276] Talia Karasov, Philipp W Messer, and Dmitri A Petrov. Evidence that adaptation in drosophila is not limited by mutation at single sites. *PLoS GENET*, 6(6):e1000924, 2010. 94

[277] Andrew D. Kern, Corbin D. Jones, and David J. Begun. Genomic Effects of Nucleotide Substitutions in Drosophila simulans. *GENETICS*, 162(4):1753, December 2002. 95

[278] Shmuel Sattath, Eyal Elyashiv, Oren Kolodny, Yosef Rinott, and Guy Sella. Pervasive adaptive protein evolution apparent in diversity patterns around amino acid substitutions in drosophila simulans. *PLoS Genet*, 7(2):e1001302, 2011. doi: 10.1371/journal.pgen.1001302. 95

[279] Yuh Chwen G Lee, Charles H Langley, and David J Begun. Differential strengths of positive selection revealed by hitchhiking effects at small physical scales in drosophila melanogaster. *MOL BIOL EVOL*, 31(4):804–16, Apr 2014. doi: 10.1093/molbev/mst270. 95

[280] Li Zhao, Janneke Wit, Nicolas Svetec, and David J. Begun. Parallel gene expression differences between low and high latitude populations of drosophila melanogaster and d. simulans. *PLoS Genet*, 11(5):e1005184, 05 2015. doi: 10.1371/journal.pgen.1005184. URL `http://dx.doi.org/10.1371%2Fjournal.pgen.1005184`. 95

[281] B Charlesworth, J A Coyne, and N H Barton. The Relative Rates of Evolution of Sex Chromosomes and Autosomes. *AM NAT*, 130(1):113–146, July 1987. 95

[282] K. Thornton. Recombination and the properties of Tajima's D in the context of approximate likelihood calculation. *GENETICS*, 2005. 95

[283] Brian A Counterman, Daniel ORTíZ-Barrientos, and Mohamed A F Noor. Using Comparative Genomic Data to Test for Fast-X Evolution. *EVOLUTION*, 58(3):656–660, March 2004. 95

[284] Heidi Musters, Melanie A Huntley, and Rama S Singh. A Genomic Comparison of Faster-Sex, Faster-X, and Faster-Male Evolution Between Drosophila melanogaster and Drosophila pseudoobscura. *J MOL EVOL*, 62(6):693–700, April 2006. 95

[285] Andrew G Clark, Michael B Eisen, Douglas R Smith, Casey M Bergman, Brian Oliver, Therese A Markow, Thomas C Kaufman, Manolis Kellis, William Gelbart, Venky N Iyer, Daniel A Pollard, Timothy B Sackton, Amanda M Larracuente, Nadia D Singh, Jose P Abad, Dawn N Abt, Boris Adryan, Montserrat Aguade, Hiroshi Akashi, Wyatt W Anderson, Charles F Aquadro, David H Ardell, Roman Arguello, Carlo G Artieri, Daniel A Barbash, Daniel Barker, Paolo Barsanti, Phil Batterham, Serafim Batzoglou, Dave Begun, Arjun Bhutkar, Enrico Blanco, Stephanie A Bosak, Robert K Bradley, Adrianne D Brand, Michael R Brent, Angela N Brooks, Randall H Brown, Roger K Butlin, Corrado Caggese, Brian R Calvi, A Bernardo de Carvalho, Anat Caspi, Sergio Castrezana, Susan E Celniker, Jean L Chang, Charles Chapple, Sourav Chatterji, Asif Chinwalla, Alberto Civetta, Sandra W Clifton, Josep M Comeron, James C Costello, Jerry A Coyne, Jennifer Daub, Robert G David, Arthur L Delcher, Kim Delehaunty, Chuong B Do, Heather Ebling, Kevin Edwards, Thomas Eickbush, Jay D Evans, Alan Filipski, Sven Findeiss, Eva Freyhult, Lucinda Fulton, Robert Fulton, Ana C L Garcia, Anastasia Gardiner, David A Garfield, Barry E Garvin, Greg Gibson, Don Gilbert, Sante Gnerre, Jennifer Godfrey, Robert Good, Valer Gotea, Brenton Gravely, Anthony J Greenberg, Sam Griffiths-Jones, Samuel Gross, Roderic Guigo, Erik A Gustafson, Wilfried Haerty, Matthew W Hahn, Daniel L Halligan, Aaron L Halpern, Gillian M Halter, Mira V Han, Andreas Heger, LaDeana Hillier, Angie S Hinrichs, Ian Holmes, Roger A Hoskins, Melissa J Hubisz, Dan Hultmark, Melanie A Huntley, David B Jaffe, Santosh Jagadeeshan, William R Jeck, Justin Johnson, Corbin D. Jones, William C Jordan, Gary H Karpen, Eiko Kataoka, Peter D Keightley, Pouya Kheradpour, Ewen F Kirkness, Leonardo B Koerich, Karsten Kristiansen, Dave Kudrna, Rob J Kulathinal, Sudhir Kumar, Roberta Kwok, Eric Lander, Charles H Langley, Richard Lapoint, Brian P Lazzaro, So-Jeong Lee, Lisa Levesque, Ruiqiang Li, Chiao-Feng Lin, Michael F Lin, Kerstin Lindblad-Toh, Ana Llopart, Manyuan Long, Lloyd Low, Elena Lozovsky, Jian Lu, Meizhong Luo, Carlos A Machado, Wojciech Makalowski, Mar Marzo, Muneo Matsuda, Luciano Matzkin, Bryant McAllister, Carolyn S McBride, Brendan McKernan, Kevin McKernan, Maria Mendez-Lago, Patrick Minx, Michael U Mollenhauer, Kristi Montooth, Stephen M Mount, Xu Mu, Eugene Myers, Barbara Negre, Stuart Newfeld, Rasmus Nielsen, Mohamed A F Noor, Patrick O'Grady, Lior Pachter, Montserrat Papaceit, Matthew J Parisi, Michael Parisi, Leopold Parts, Jakob S Pedersen, Graziano Pesole, Adam M Phillippy, Chris P Ponting, Mihai Pop, Damiano Porcelli, Jeffrey R Powell, Sonja Prohaska, Kim Pruitt, Marta Puig, Hadi Quesneville, Kristipati Ravi Ram, David Rand, Matthew D Rasmussen, Laura K Reed, Robert Reenan, Amy Reily, Karin A Remington, Tania T Rieger,

Michael G Ritchie, Charles Robin, Yu-Hui Rogers, Claudia Rohde, Julio Rozas, Marc J Rubenfield, Alfredo Ruiz, Susan Russo, Steven L Salzberg, Alejandro Sanchez-Gracia, David J Saranga, Hajime Sato, Stephen W Schaeffer, Michael C Schatz, Todd Schlenke, Russell Schwartz, Carmen Segarra, Rama S Singh, Laura Sirot, Marina Sirota, Nicholas B Sisneros, Chris D Smith, Temple F Smith, John Spieth, Deborah E Stage, Alexander Stark, Wolfgang Stephan, Robert L Strausberg, Sebastian Strempel, David Sturgill, Granger Sutton, Granger G Sutton, Wei Tao, Sarah Teichmann, Yoshiko N Tobari, Yoshihiko Tomimura, Jason M Tsolas, Vera L S Valente, Eli Venter, J Craig Venter, Saverio Vicario, Filipe G Vieira, Albert J Vilella, Alfredo Villasante, Brian Walenz, Jun Wang, Marvin Wasserman, Thomas Watts, Derek Wilson, Richard K Wilson, Rod A Wing, Mariana F Wolfner, Alex Wong, Gane Ka-Shu Wong, Chung-I Wu, Gabriel Wu, Daisuke Yamamoto, Hsiao-Pei Yang, and Shiaw... Yang. Evolution of genes and genomes on the Drosophila phylogeny. *NATURE*, 450(7167):203–218, 2007. 95

[286] J F Baines, S A Sawyer, D L Hartl, and J Parsch. Effects of X-Linkage and Sex-Biased Gene Expression on the Rate of Adaptive Protein Evolution in Drosophila. *MOL BIOL EVOL*, 25(8):1639–1650, April 2008. 95

[287] Andrea J Betancourt, Daven C Presgraves, and Willie J Swanson. A Test for Faster X Evolution in Drosophila. *MOL BIOL EVOL*, 19(10):1816–1819, October 2002. 95

[288] K. Thornton. X chromosomes and autosomes evolve at similar rates in drosophila: No evidence for faster-x protein evolution. *GENOME RES*, 16(4):498–504, March 2006. 95

[289] Tim Connallon. Adaptive protein evolution of X-linked and autosomal genes in Drosophila: implications for faster-X hypotheses. *MOL BIOL EVOL*, 24(11):2566–2572, 2007. 95

[290] Jonathan K Pritchard, Joseph K Pickrell, and Graham Coop. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *CURR BIOL*, 20(4):R208–15, Feb 2010. doi: 10.1016/j.cub.2009.11.055. 96

[291] Obi L Griffith, Stephen B Montgomery, Bridget Bernier, Bryan Chu, Katayoon Kasaian, Stein Aerts, Shaun Mahony, Monica C Sleumer, Mikhail Bilenky, Maximilian Haeussler, Malachi Griffith, Steven M Gallo, Belinda Giardine, Bart Hooghe, Peter Van Loo, Enrique Blanco, Amy Ticoll, Stuart Lithwick, Elodie Portales-Casamar, Ian J Donaldson, Gordon Robertson, Claes Wadelius, Pieter De Bleser, Dominique Vlieghe, Marc S Halfon, Wyeth Wasserman, Ross Hardison, Casey M

Bergman, Steven J M Jones, and Open Regulatory Annotation Consortium. ORegAnno: an open-access community-driven resource for regulatory annotation. *NUCLEIC ACIDS RES*, 36(Database issue):D107–13, January 2008. 99

[292] Helen Attrill, Kathleen Falls, Joshua L Goodman, Gillian H Millburn, Giulia Antonazzo, Alix J Rey, Steven J Marygold, and the FlyBase consortium. FlyBase: establishing a Gene Group resource for Drosophila melanogaster. *NUCLEIC ACIDS RES*, 44(D1):D786–D792, January 2016. 99

[293] P W Jansen and R E Perez. Constrained structural design optimization via a parallel augmented Lagrangian particle swarm optimization approach. *Computers and Structures*, 89(13-14):1352–1366, July 2011. 100, 138

[294] Dieter Kraft et al. *A software package for sequential quadratic programming*. DFVLR Obersfaffeuhofen, Germany, 1988. 100, 138

[295] Ruben E Perez, Peter W Jansen, and Joaquim R R A Martins. pyOpt: a Python-based object-oriented framework for nonlinear constrained optimization. *Struct Multidisc Optim*, 45(1):101–118, May 2011. 100, 138

[296] Daniel R Schrider, David Houle, Michael Lynch, and Matthew W Hahn. Rates and Genomic Consequences of Spontaneous Mutational Events in Drosophila melanogaster. *GENETICS*, 194(4):937–954, January 2013. 100

[297] W. J Ewens. *Mathematical population genetics*. Springer, New York, 2nd ed edition, 2004. 101

[298] F Tajima. Relationship Between DNA Polymorphism and Fixation Time. *GENETICS*, 125(2):447–454, January 1990. 101

[299] Andrew D. Kern and David Haussler. A population genetic hidden Markov model for detecting genomic regions under selection. *MOL BIOL EVOL*, 27(7):1673–1685, 2010. 101

[300] Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. 102, 103, 160

[301] Vladimir N. Vapnik. The Nature of Statistical Learning Theory. Springer-Verlag New York, Inc., New York, NY, USA, 2000. 102

[302] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *J MACH LEARN RES*, 12:2825–2830, 2011. 103

[303] Tom Fawcett. An Introduction To Roc Analysis. *PATTERN RECOGN LETT*, 27(8):861–874, June 2006. 105, 106

[304] JBS Haldane. The Combination Of Linkage Values, And The Calculation Of Distance Between The Loci Of Linked Factors. *J GENET*, 8:299–309, 1919. 107

[305] John E Pool, Ines Hellmann, Jeffrey D Jensen, and Rasmus Nielsen. Population genetic inference from genomic sequence variation. *GENOME RES*, 20(3):291–300, January 2010. 113

[306] Joseph K. Pickrell and Jonathan K Pritchard. Inference of Population Splits and Mixtures from Genome-Wide Allele Frequency Data. *PLOS GENETIC*, 8(11):e1002967 EP –, January 2012. 113

[307] Vitor Sousa and Jody Hey. Understanding the origin of species with genome-scale data: modelling gene flow. *NAT REV GENET*, pages –, May 2013. 113

[308] P Hájková, C Pertoldi, B Zemanová, K Roche, B Hájek, J Bryja, and J Zima. Genetic structure and evidence for recent population decline in Eurasian otter populations in the Czech and Slovak Republics: implications for conservation. *Journal of Zoology*, 272(1):1–9, May 2007. 113

[309] Ryan C Garrick, Brittney Kajdacsi, Michael A Russello, Edgar Benavides, Chaz Hyseni, James P Gibbs, Washington Tapia, and Adalgisa Caccone. Naturally rare versus newly rare: demographic inferences on two timescales inform conservation of Galápagos giant tortoises. *Ecology and Evolution*, 5(3):676–694, February 2015. 113

[310] Yun-Xin Fu. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. *GENETICS*, 147(2):915–925, October 1997. 113

[311] Takeo Maruyama and Paul A Fuerst. Population Bottlenecks and Nonequilibrium Models in Population Genetics. II. Number of Alleles in a Small Population That Was Formed by a Recent Bottleneck. *GENETICS*, 111(3):675–689, November 1985. 113

[312] Paul Marjoram and Jeff D Wall. Fast ”Coalescent” Simulation. *BMC Genomics 2009 10:219*, 7(1):16, 2006. 114

[313] Russell B Corbett-Detig, Daniel L Hartl, and Timothy B Sackton. Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS BIOL*, 13(4):e1002112, April 2015. 114

[314] Joshua M Akey, Michael A Eberle, Mark J Rieder, Christopher S Carlson, Mark D Shriver, Deborah A Nickerson, and Leonid Kruglyak. Population history and natural selection shape patterns of genetic variation in 132 genes. *PLOS BIOL*, 2(10):e286, 2004. 114

[315] Elodie Gazave, Li Ma, Diana Chang, Alex Coventry, Feng Gao, Donna Muzny, Eric Boerwinkle, Richard A Gibbs, Charles F Sing, Andrew G Clark, and Alon Keinan. Neutral genomic regions refine models of recent rapid human population growth. *P NATL ACAD SCI USA*, 111(2):757–762, January 2014. 114, 135

[316] Gregory B Ewing and Jeffrey D Jensen. The consequences of not accounting for background selection in demographic inference. *MOL ECOL*, 25(1):135–141, January 2016. 114, 130, 134

[317] Kevin R Thornton. Automating approximate Bayesian computation by local linear regression. *BMC Genetics*, 10(1):35, July 2009. 115, 140

[318] Katalin Csillery, Olivier Francois, and Michael G. B. Blum. abc: an r package for approximate bayesian computation (abc). *Methods in Ecology and Evolution*, 2012. doi: http://dx.doi.org/10.1111/j.2041-210X.2011.00179.x. 115, 141

[319] Robert E Kass and Adrian E Raftery. Bayes Factors. *Journal of the American Statistical Association*, 90(430):773–795, February 2012. 129

[320] Richard R Hudson and Norman L Kaplan. Gene Trees with Background Selection. In *Non-Neutral Evolution*, pages 140–153. Springer US, Boston, MA, 1994. 130

[321] Augustine Kong, Gudmar Thorleifsson, Daniel F Gudbjartsson, Gisli Masson, Asgeir Sigurdsson, Aslaug Jonasdottir, G Bragi Walters, Adalbjorg Jonasdottir, Arnaldur Gylfason, Kari Th Kristinsson, Sigurjon A Gudjonsson, Michael L Frigge, Agnar Helgason, Unnur Thorsteinsdottir, and Kari Stefansson. Fine-scale recombination rate differences between sexes, populations and individuals. *NATURE*, 467(7319):1099–1103, October 2010. 131, 137

[322] Benjamin M Peter, Emilia Huerta-Sanchez, and Rasmus Nielsen. Distinguishing between Selective Sweeps from Standing Variation and from a De Novo Mutation. *PLOS GENETIC*, 8(10):e1003011, October 2012. 131

[323] Nick G C Smith and Adam Eyre-Walker. Adaptive protein evolution in Drosophila. *NATURE*, 415(6875):1022–1024, February 2002. 132

[324] Nicolas Bierne and Adam Eyre-Walker. The genomic rate of adaptive amino acid substitution in Drosophila. *MOL BIOL EVOL*, 21(7):1350–1360, July 2004. 132

[325] Ryan D Hernandez, Joanna L Kelley, Eyal Elyashiv, S Cord Melton, Adam Auton, Gilean McVean, 1000 Genomes Project, Guy Sella, and Molly Przeworski. Classic selective sweeps were rare in recent human evolution. *SCIENCE*, 331(6019):920–924, February 2011. 132

[326] Adam R Boyko, Scott H Williamson, Amit R Indap, Jeremiah D Degenhardt, Ryan D Hernandez, Kirk E Lohmueller, Mark D Adams, Steffen Schmidt, John J Sninsky, Shamil R Sunyaev, Thomas J White, Rasmus Nielsen, Andrew G Clark, and Carlos D Bustamante. Assessing the evolutionary impact of amino acid mutations in the human genome. *PLOS GENETIC*, 4(5):e1000083, May 2008. 132

[327] D Enard, P W Messer, and D A Petrov. Genome-wide signals of positive selection in human evolution. *GENOME RES*, March 2014. 132

[328] Hui Li, Namita Mukherjee, Usha Soundararajan, Zsanett Tarnok, Csaba Barta, Shagufta Khaliq, Aisha Mohyuddin, Sylvester L B Kajuna, S Qasim Mehdi, Judith R Kidd, and Kenneth K Kidd. Geographically separate increases in the frequency of the derived ADH1B*47His allele in eastern and western Asia. *Am J Hum Genet*, 81(4):842–846, October 2007. 132

[329] George H Perry, Nathaniel J Dominy, Katrina G Claw, Arthur S Lee, Heike Fiegler, Richard Redon, John Werner, Fernando A Villanea, Joanna L Mountain, Rajeev Misra, Nigel P Carter, Charles Lee, and Anne C Stone. Diet and the evolution of human amylase gene copy number variation. *NAT GENET*, 39(10):1256–1260, October 2007. 132

[330] Sarah A Tishkoff, Floyd A Reed, Alessia Ranciaro, Benjamin F Voight, Courtney C Babbitt, Jesse S Silverman, Kweli Powell, Holly M Mortensen, Jibril B Hirbo, Maha Osman, Muntaser Ibrahim, Sabah A Omar, Godfrey Lema, Thomas B Nyambo, Jilur Ghori, Suzannah Bumpstead, Jonathan K Pritchard, Gregory A Wray, and Panos Deloukas. Convergent adaptation of human lactase persistence in Africa and Europe. - PubMed - NCBI. *NAT GENET*, 39(1):31–40, December 2006. 132

[331] Luis B Barreiro, Guillaume Laval, Hélène Quach, Etienne Patin, and Lluís Quintana-Murci. Natural selection has driven population differentiation in modern humans. - PubMed - NCBI. *NAT GENET*, 40(3):340–345, February 2008. 132

[332] Jarosław Bryk, Emilie Hardouin, Irina Pugach, David Hughes, Rainer Strotmann, Mark Stoneking, and Sean Myles. Positive Selection in East Asians for an EDAR Allele that Enhances NF-κB Activation. *PLOS ONE*, 3(5):e2209, May 2008. 132

[333] W Stephan, T H E Wiehe, and M W Lenz. The effect of strongly selected substitutions on neutral polymorphism: analytical results based on diffusion theory. *THEOR POPUL BIOL*, 41:237–254, 1992. 133

[334] Philipp W Messer and Dmitri A Petrov. Frequent adaptation and the McDonald-Kreitman test. *Proceedings of the National Academy of Sciences*, 110(21):8615–8620, May 2013. 133, 134

[335] J H McDonald and M Kreitman. Adaptive protein evolution at the Adh locus in Drosophila. *NATURE*, 351(6328):652–654, June 1991. 133

[336] Adam Eyre-Walker and Peter D Keightley. Estimating the rate of adaptive molecular evolution in the presence of slightly deleterious mutations and population size change. *MOL BIOL EVOL*, 26(9):2097–2108, September 2009. 133, 135

[337] Jacob A Tennessen, Abigail W Bigham, Timothy D O'Connor, Wenqing Fu, Eimear E Kenny, Simon Gravel, Sean McGee, Ron Do, Xiaoming Liu, Goo Jun, Hyun Min Kang, Daniel Jordan, Suzanne M Leal, Stacey Gabriel, Mark J Rieder, Goncalo Abecasis, David Altshuler, Deborah A Nickerson, Eric Boerwinkle, Shamil Sunyaev, Carlos D Bustamante, Michael J Bamshad, Joshua M Akey, Broad GO, Seattle GO, and on behalf of the NHLBI Exome Sequencing Project. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. *SCIENCE*, 337(6090):64–69, January 2012. 133

[338] Hideki Innan and Yuseob Kim. Pattern of polymorphism after strong artificial selection in a domestication event. *P NATL ACAD SCI USA*, 101(29):10667–10672, 2004. 134

[339] R R Hudson. How can low levels of DNA sequence variation in regions of the Drosophila genome with low recombination rates be explained? *P NATL ACAD SCI USA*, 91:6815–6818, 1994. 134

[340] Benjamin F Voight, Sridhar Kudaravalli, Xiaoquan Wen, and Jonathan K Pritchard. A map of recent positive selection in the human genome. *PLOS BIOL*, 4(3):e72, March 2006. 134

[341] Kosuke M Teshima and Molly Przeworski. Directional positive selection on an allele of arbitrary dominance. *GENETICS*, 172(1):713–718, January 2006. 134

[342] Jonathan K Pritchard, Joseph K. Pickrell, and Graham Coop. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. - PubMed - NCBI. *Current Biology*, 20(4):R208–R215, February 2010. 134

[343] Jeremy J Berg and Graham Coop. A population genetic signal of polygenic adaptation. - PubMed - NCBI. *PLOS GENETIC*, 10(8):e1004412, August 2014. 134

[344] Lisha A Mathew and Jeffrey D Jensen. Evaluating the ability of the pairwise joint site frequency spectrum to co-estimate selection and demography. *Front Genet*, 6(235):268, 2015. 134

[345] D J Begun and C F Aquadro. Levels of naturally occurring DNA polymorphism correlate with recombination rates in D. melanogaster. *NATURE*, 356(6369):519–520, April 1992. 135

[346] Sara Sheehan, Kelley Harris, and Yun S. Song. Estimating Variable Effective Population Sizes from Multiple Genomes: A Sequentially Markov Conditional Sampling Distribution Approach. *GENETICS*, 194(3):647–662, January 2013. 135

[347] Augustine Kong, Michael L Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, Wendy S W Wong, Gunnar Sigurdsson, G Bragi Walters, Stacy Steinberg, Hannes Helgason, Gudmar Thorleifsson, Daniel F Gudbjartsson, Agnar Helgason, Olafur Th Magnusson, Unnur Thorsteinsdottir, and Kari Stefansson. Rate of de novo mutations and the importance of father/'s age to disease risk. *NATURE*, 488(7412):471–475, August 2012. 137

[348] Grant Hamilton, Mark Stoneking, and Laurent Excoffier. Molecular analysis reveals tighter social regulation of immigration in patrilocal populations than in matrilocal populations. *P NATL ACAD SCI USA*, 102(21):7476–7480, May 2005. 140

[349] W H Li M Nei. Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, 76(10):5269, October 1979. 144

[350] Sewall Wright. Statistical Theory of Evolution. *Journal of the American Statistical Association*, 26(173):201–208, January 1931. 150

[351] R A Fisher. *The Genetical Theory of Natural Selection*. Clarendon Press, Oxford, 1930. 150

[352] Jotun Hein, Mikkel H Schierup, and Carsten Wiuf. *Gene Genealogies, Variation and Evolution: A Primer in Coalescent Theory*. Oxford University Press, 2005. 152, 155, 156, 158

[353] R C Griffiths and P MARJORAM. Ancestral Inference from Samples of DNA Sequences with Recombination. *Journal of Computational Biology*, 3(4):479–502, January 1996. 154

[354] Carsten Wiuf and Jotun Hein. Recombination as a Point Process along Sequences. *THEOR POPUL BIOL*, 55(3):248–259, June 1999. 155, 156

[355] A Eriksson, B Mahjani, and B Mehlig. Sequential Markov coalescent algorithms for population models with demographic structure. *THEOR POPUL BIOL*, 76(2):84–91, September 2009. 157, 158, 159

[356] S Sathiya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vector machines with Gaussian kernel. *Neural Comput*, 15(7):1667–1689, July 2003. 161

[357] CW Hsu, CC Chang, and CJ Lin. A practical guide to support vector classification. 2010. 161